2013

Impact of Multimedia in Sina Weibo

Xun ZHAO
Singapore Management University, xun.zhao.2011@smu.edu.sg

IMPACT OF MULTIMEDIA IN SINA WEIBO

XUN ZHAO

SINGAPORE MANAGEMENT UNIVERSITY

2013

Impact of Multimedia in Sina Weibo

by

Xun Zhao

Submitted to School of Information Systems in partial fulfillment of the requirements for the Degree of

Master of Science in Information Systems

Thesis Committee:

Feida Zhu(Supervisor/Chair)
Assistant Professor of Information Systems
Singapore Management University

Jialie Shen
Assistant Professor of Information Systems
Singapore Management University

Baihua Zheng
Associate Professor of Information Systems
Singapore Management University

Singapore Management University

2013

Impact of Multimedia in Sina Weibo

Xun Zhao

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 thesis 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. In addition to consider the multimedia impact on popularity, we also compare the user influence in multimedia and text setting. 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. We prove that tweets with multimedia contents survive longer than text tweets. Finally, multimedia contents tend to attract more attention while text maintains discussion. Our results 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. Our result is useful for web developers and microblogging marketers.

Table of Contents

ACKNOWLEDGEMENT

This dissertation is partially supported by the Singapore National Research Foundation under its International Research Centre at Singapore Funding Initiative and administered by the IDM Program Office. This dissertation 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.

1 Introduction

1.1 Background

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[14].

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. Regarding the nature of those social networks, [16] points out that twitter is more of a news media than a social network, [5] points out Twitter follower count alone could not reflect the popularity of users. Researchers in Data Mining field have applied Pattern Mining and Graph Mining algorithms to study the structure of the social networks. Machine Learning people tries to group similar users, explore post preferences and build recommendation systems out of those giant social networks. NLP people are attracted by the rich resource of the textual information, they applied bag-of-words to facilitate the analysis and designed a series of topic models capturing the short and rich content nature of social network messages.

Sina Weibo is a popular, Twitter-like microblogging service platform originated from China. It features more than 500 million users, of whom 49 million are active

users in February 2013. 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. Considering the fact that a Chinese character conveys more information than an English character, with Sina Weibo incorporating the feature of multimedia content sharing at the beginning of its foundation, the data in Sina Weibo is more diverse and intriguing for scientific research.

1.2 Motivation

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 like "Gangnam Style" command viral popularity everywhere they go ranging from personal blogs, video sharing sites, to social network services. For example, according to our findings, over half of tweets published in Sina Weibo are linked with multimedia contents. 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 retweeted by users for a much longer period of time, which we call they *survive* longer.

1.3 Research Objectives and Contributions

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?
   (c) How much textual information do multimedia tweets contain?
   (d) Is there any significant difference in using mentions("@") and hashtags?
2. **Tweet Consumption.**
   (a) Would multimedia content influence the popularity of tweets?
   (b) Is multimedia content related to the life span of tweets?

The results in this thesis shows that multimedia tweets dominant pure text tweets 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, text length of tweets and the popularity of user. The findings are are useful for web developers, researchers and microblogging marketers.

2 Preliminaries

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.

We use the number of direct retweets and the sum of all retweets in a tweets retweet network to measure the popularity of a tweet.

$$p_1 = \sum_{i=0}^{n} r_i \tag{1}$$

$$p_2 = \sum_{i=0}^{n} \sum_{j=0}^{m} r_{ij} \tag{2}$$

Where n is the number of retweets within each layer and m is the height of the retweet tree.

We use the number of followers of the user as an index of user popularity.

According to [20], 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

$$N(t) = N_0 e^{-bt} \tag{3}$$

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

$\tau = \frac{1}{b}$ is defined as the average life span of tweets.

The number of mentions("@") captures how many connections the tweet has. The number of hashtags("#") captures how easy the tweet could be found. We use them in the text body analysis to reflect certain characters of the tweet.

3 Detailed Literature Review

A lot of research effort has been dedicated to provide a rough tour guide of popular microblogging services. [16] uses a huge data to illustrate the user composition, trending topics et. of Twitter. [14] studies the underlying motivation of certain user activity. A lot of attention has been drawn to study Chinese social networks. [35] examines key topics that trend on Sina Weibo and contrast them with Twitter. The trends in Sina Weibo almost entirely created through retweets of media content such as jokes, while the trends in Twitter relate more to global event and news stories. [26] studied how Chinese Internet users use microblogging service in disaster response.

3.1 Data Mining Perspective

Researchers in Data Mining field have developed Pattern Mining and Graph Mining algorithms in the context of social network. [9] proposes to model the customer network as a Markov random field. It shows the advantages of this approach using a social network. [1] utilizes pattern mining algorithm to discover user activity behavior patterns in event log information. [25] presents a large-scale measurement study and analysis of the structure of multiple online social networks. The results confirm the power-law, small-world, and scale-free properties of online social networks. [10] introduce a system for sensing complex social systems. The author demonstrate the ability to use standard Bluetooth-enabled telephones to measure information acess and recognize social patterns in daily user activity.

3.2 Machine Learning Perspective

Machine Learning people have been developing models for social networks ever since the foundation of social network services. [27] propose an approach to combine first-order logic and probabilistic graphical models in a single representation. Weights

are efficiently learned from relational databases by iteratively optimizing a pseudo-likelihood measure. [23] develops approaches to link prediction based on measures for analyzing "proximity" of nodes in a network. [18] propose and evaluate a honeypot-based approach for uncovering social spammers in MySpace and Twitter. The author develops machine learning base classifiers for identifying previously unknown spammers.

3.3 Natural Language Processing Perspective

Natural Language Processing researchers are attracted to social network research because of the enormous textual data generated every day. Ever since David Blei proposed the generative probabilistic model [3] to deal with discrete data such as text corpora. A lot of following models have been proposed to incorporate different features of social networks. [30] captures not only the low-dimensional structure of data, but also how the structure changes over time. [21] introduce the pachinko allocation model, which captures arbitrary, nested, and possibly sparse correlations between topics using a directed acyclic graph. [33] propose a collaborative web recommendation framework, which employs LDA to model underlying topic-simplex space and discover the associations between user sessions and multiple topics via probability inference.

3.4 Information Securtiy Perspective

On the other hand, since data security is the opposite of data mining, researchers in Information Security have also investigated a lot of energy concerning the private information in social networks. [24] explore the effectiveness of possible sanitization techniques that can be used to combat inference attacks under different scenarios. [4] examine the difficulty of collecting profile and graph information from Facebook and describe several novel ways in which data can be extracted by third parties. [2] describes Haystack, an object storage system optimized for Facebook's Photos applica-

tion. The fact is that Facebook actually stores user information on third party servers, which means data security in facebook is not guaranteed.

The above research effort mainly focus on utilizing the textual information and social link information in social network sites. Multimedia information in social networks are mostly researched by Computer Vision scientist. [7] combines content analysis based on text tags and image data with structural analysis based on geospatial data to organize a large collection of geotagged photos. [22] study image classification on a dataset of 30 million images and learn models for these landmarks with a multiclass support vector machine, using vector-quantized interest point descriptors as features. [6] propose and evaluate a probabilistic framework for estimating a Twitter user's city-level location based purely on the content of the user's tweets. [15] presents a method for estimating geographic location for sequences of time-stamped photographs. A prior distribution over travel describes the likelihood of traveling from one location to another during a given time interval.

To the best of the author's knowledge, there is no published work which looks at the statistics of multimedia content in microblogging service from a data analytic perspective. This thesis is the first work which combines and compares the spastical difference of textual and multimedia information in microblogging service.

4 Methodology and Solution Approach

The data we use contains 230 million tweets published by 1812701 users from Jan. 2011 to Jul. 2011. For each tweet, we categorize the tweet into Text Tweet, Image Tweet and URL Tweet as previous discussed. We also have the user profiles and retweet chains at our disposal.

To evaluate the impact of multimedia content in Sina Weibo, we use a binary comparison method to show the difference of multmedia content and text only content. From a data analytic perspective, we conduct our comparison in the following dimensions:

1. Composition. We compare the composition of Multimedia Tweet and Text Tweet in a general case and in a popular subset of tweets.

2. Tweet Popularity. We use the direct retweet number and the overall sum of retweet to measure the popularity of a tweet. Direct Retweet Number indicate how broad a tweet could influence. Sum of retweet tells us the over popularity of the tweet.

3. User Popularity. We use follower count to measure user popularity.

4. User Activeness. We first rank user activeness based on the number of tweet and multimedia tweet he posted. Then we use spearman correlation coefficient to analyze the similarity of the two ranks.

5. Tweet Life Span. We first get the publish time distribution of a set of popular tweet. Following literature, we use the reciprocal of decay rate to indicate the life span of a tweet.

6. Retweet Length. We define the longest path of the retweet chain as Retweet Length. Retweet length indicates the deepness of tweet content.

7. Text Body Analysis. We use the length of text, number of mentions, and numbers of hashtags to reflects certain characteristics of the text body.

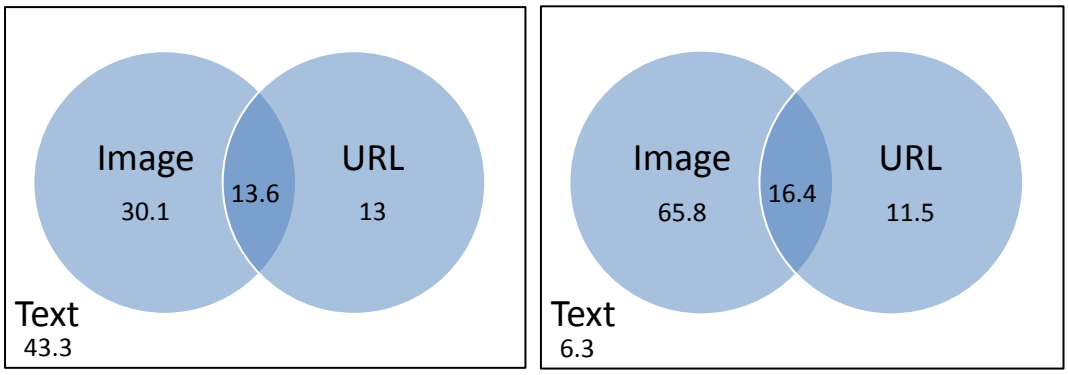
5 General Analysis

5.1 Multimedia Content 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 [29], [16]. 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. [5] has reported that popular tweets are more likely to be posted by celebrities and news medias. [37] has reported the topics of popular tweets are different from ordinary tweet. [34] 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.

5.2 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[5], 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



(a) General Tweet Composition (b) Popular Tweet Composition

Figure 1: Venn Diagram for Composition of Multimedia Tweet

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.

5.3 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 [5] 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[11] 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

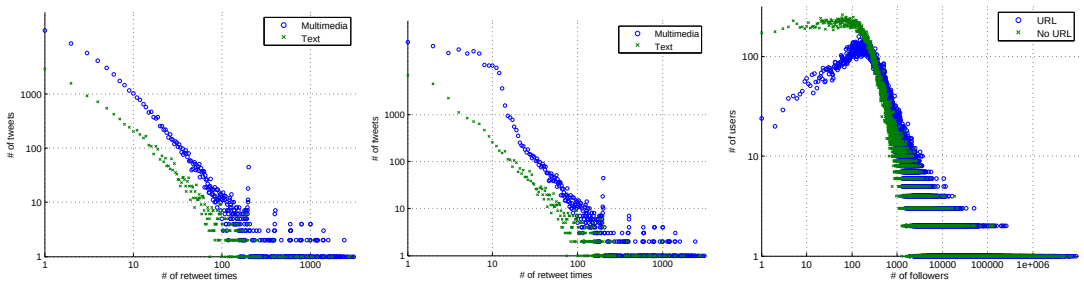
magnitude. Interestingly, we also observe a long tail in both multimedia setting and text setting when $retweettimes > 100$. This abnormal pattern indicates the number of very popular tweet is larger than power law distribution suggests, reflecting that very popular tweets do exist in a considerable amount. This finding has important implications for microblog based marketing. Marketers would get a great pay off by aiming at those top popular tweets.

We also use Directed Acyclic Graph(DAG) node sum as another metric to indicate tweet popularity. Instead of counting direct retweet times, DAG node sum captures all the direct and indirect retweets from an original tweet's retweet graph. Figure 2(b) shows that for not so popular tweets, the dominance of popularity is even more obvious for multimedia than that in direct retweet times.

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 [5]. Figure 2 (c) 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,



(a) Tweet Popularity using RT Times (b) Tweet Popularity Using DAG node sum (c) User Popularity Distribution

Figure 2: Tweet and User Popularity Distribution

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.

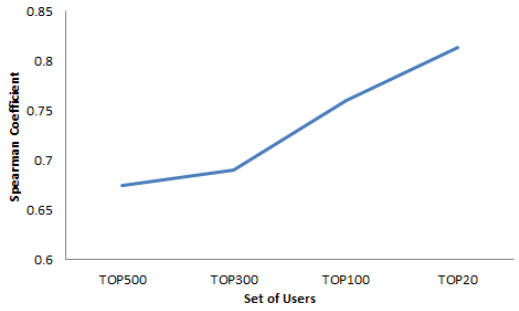
5.4 Comparing User Activeness

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[5]. 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[28]

$$\rho = 1 - \frac{6 \sum (x_i - y_i)^2}{N^3 - N} \tag{4}$$

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

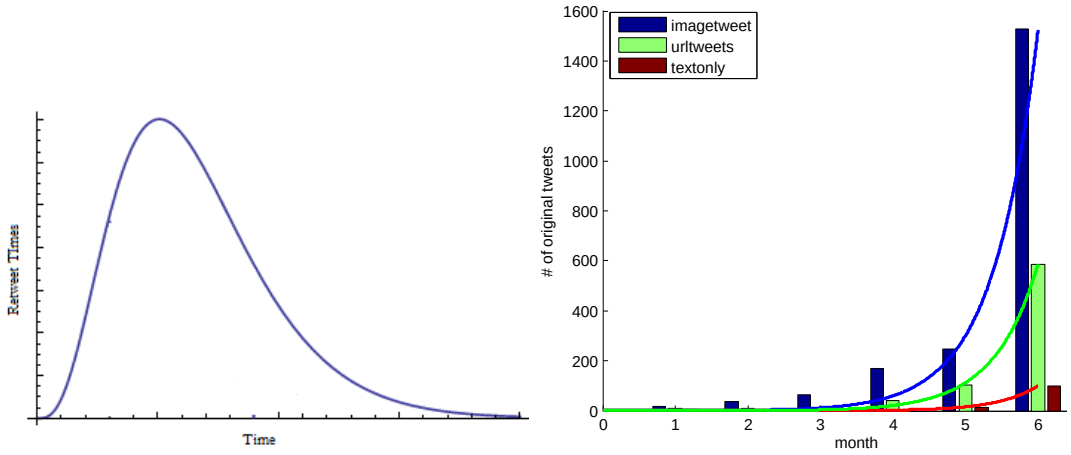
Figure 3: Spearman Correlation of User Post Activeness

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.

5.5 Life Span Analysis

Many factors, such as user popularity and topic of tweet[19] could affect the life span of a tweet. Previous studies [14][16] have reported that messages in microblogging services such as Twitter spread and disappear rather fast. [17] 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 [17]. It quickly receives a lot of retweets



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

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

Figure 4: Tweet and User Popularity Distribution

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.

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

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

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.

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.

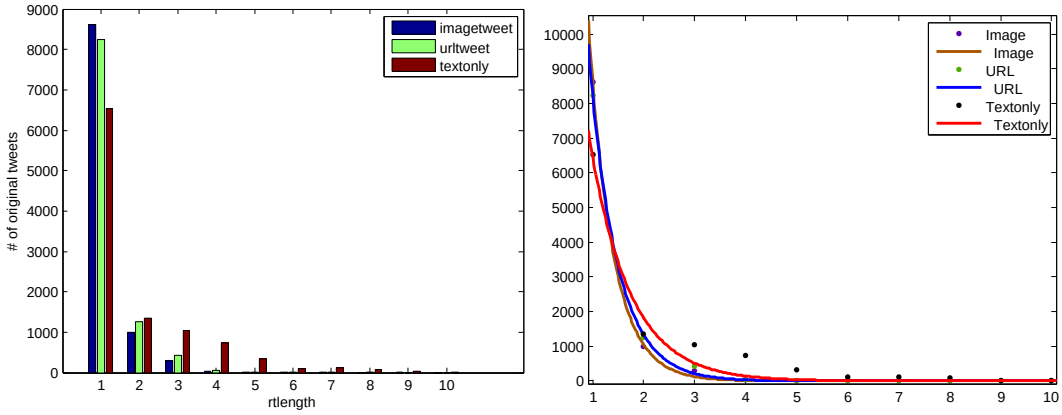
In order to get a larger sample size, we also set different retweet popularity threshold(100, 200, 300, etc.) in this experiment. In all of these attempts, the program would not finish running because of large sample size. Our findings point out that multimedia content have a longer life span than traditional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

5.6 Retweet Length

While the number of retweet times measures how broad a tweet could influence, the length of the retweet chain implies how deep a tweet could reach. Number of retweet times tells how many tweets are intrigued by this tweet. Retweet length tells how long could this sort of interest last through conversation and interaction. shows that sentiment in hyperlinked blogs tend to first heat up then cool down in a repost chain.

In each category of Image Tweet, URL Tweet and Text Tweet, we select 100,000 original tweets which received retweets. For each of these tweets, we use a depth first search method to dig out how far each retweet reach. Although above analysis show tweets with multimedia contents tend to prevail in Sina weibo and attract more retweets, our findings suggests the opposite trend in retweet length. According to Figure 5 (a), Image Tweet has the most 1 hop retweets with URL Tweet and Text Tweet slightly fall behind. However, retweet length of Image Tweet and URL Tweet decline very fast when we further zoom in. Multimedia Tweet could hardly be retweeted 8 hops away. On the contrary, decline of Text Tweet is not significant compared with Multimedia Tweet. A considerable amount of Text Tweet are still active after several rounds of retweet. Figure 5 (b) shows the exponential fitting of the trend of decline of the three groups. The decay rate of Multimedia Tweet is significantly larger than that of Text Tweet.

Theses findings together with the Retweet Popularity Analysis suggest that although Multimedia Tweet are more kind of eye catching thus attracting a larger number of retweets at first, it is not able to continuously maintain the conversation. On the other hand, Text Tweet may fail to attract the attention at first glance, the intricate power of language often makes it attract a long chain of discussion. The influence of Multimedia Tweet is broad and radioactive, while the influence of Text Tweet is deep and penetrating.



(a) Retweet Length Comparison (b) Exponential Fitting

Figure 5: Retweet Length and Fitting Result

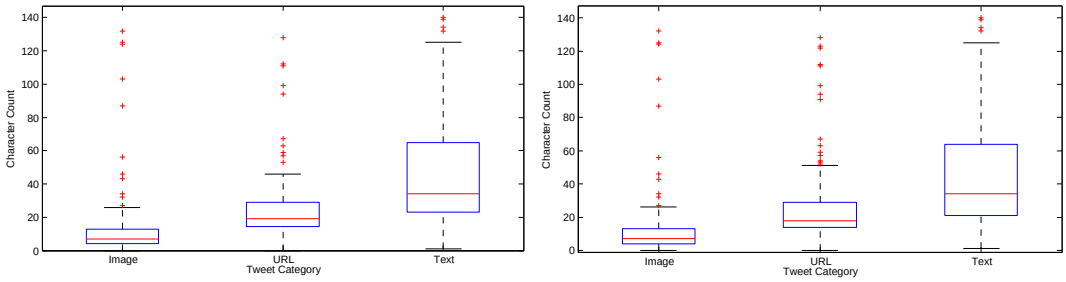
5.7 Text Body Analysis

As discussed in previous sections, Text Tweets in our dataset are tweets only contain text information. We've also noticed that most Multimedia Tweets are also associated with textual information, which usually serves to support or highlight the multimedia content. Sina Weibo has also set up a restraint to allow no more than 140 characters. A character could be a Chinese character, an English character, a punctuation or even a emoticon. Since most Chinese words are made up of two or three characters, 140 characters in Sina Weibo could convey a lot more information than 140 English characters in Twitter. Moreover, the inclusion of multimedia content makes tweets in Sina Weibo even more resourseful. However, compared with Text Tweet, would the presence of multimedia content influence certain characteristics of the text body in Multimedia Tweet? Specifically, we use three indexes to reflect certain aspect of the text body.

1. Text Length. Text length generally reflect how much information the text body contains.

2. The number of "@". "@" is used to address other users. The number of "@" reflect how many connections does a tweet has.

3. The number of "#". Sina Weibo users put 2 "#" simultaneously to enclose the



(a) Text Length Box Plot Comparison (b) Remove mentions and hashtags

Figure 6: Retweet Length and Fitting Result

text which the author wish to highlight for easy search. Generally, the more "#" a tweet have, the easier it could be searched.

Text Length Comparison is shown in Fig.6. Generally, multimedia tweet has a shorter and more focused distribution of text length compared to text tweet. We conjecture this phenomena is due to the fact that publishers of text tweet tend to increase the information in their tweet in order to compete with multimedia tweet, at the risk of making it even more boring.

6 Topic Level Analysis

Tweets with different topic may have different popularity and life span indexes. In this section, we break our analysis further to compare the difference of popularity and life span in different topics.

We first divide the tweets into three subcategory containing the keywords **Travel**, **Food** and **News**. We use these three keywords because they are the top buzz words in Sina Weibo's trending list.

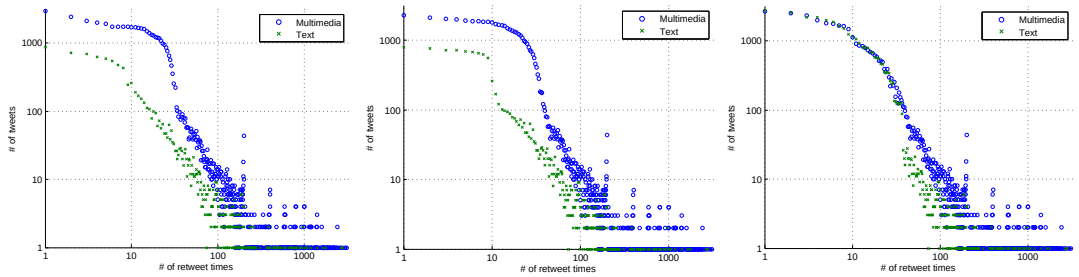
6.1 Topic Popularity Comparison

Following previous research, we use retweet times as a measure for a tweet's popularity. Figure 7 shows the difference of retweet times for tweets including **Travel**, **Food** and ***News*** for both multimedia tweets and text tweets. The results shows that multimedia tweets still prevail in both Travel and Food category, which is reflected as the blue circles always above the green cross in (a) and (b). However, in the case of "News", the curves are intertwined for less popular tweets, showing that multimedia tweets is no more popular than text tweet. If we further zoom in the figure, we find that the number of text tweets with rt below 10 are slightly bigger than the number of multimedia tweets.

The explanation may be that in traveling and dining industry, multimedia contents are more often used to lure customers, and multimedia dominance in these settings show customers do react (by retweeting more) to this idea. On the other hand, news media contain less multimedia information in Sina Weibo, and multimedia tweets do not necessarily attract more attention.

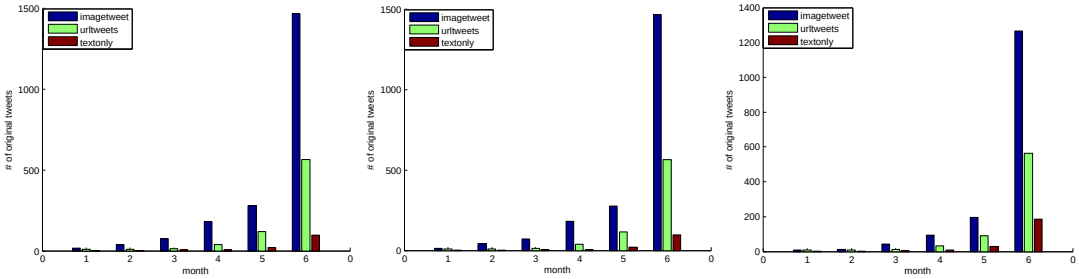
6.2 Topic Life Span Comparison

We also conduct life span comparison amongst hot topics. Figure 8 show all three groups in each topic drop exponentially from June to January. The amount of image



(a) Tweets with "Travel" keyword (b) Tweets with "Food" keyword (c) Tweets with "News" keyword

Figure 7: Retweet Length and Fitting Result



(a) Tweets with "Travel" keyword (b) Tweets with "Food" keyword (c) Tweets with "News" keyword

Figure 8: Retweet Length and Fitting Result

tweets are always dominant in each month followed by URL tweets, further suggesting multimedia content's power of attracting retweets over text.

After we further fit the data into the exponential decay fucntion,

$$N(t) = N_0 e^{-bt} \tag{5}$$

Table 2 suggests that as a general trend, the life span of multimedia contents are slightly longer in each of the three topic. Compared with the other two topics, tweets containing **News** have a significant shorter life span. This finding agrees with common sense that interestingness of news topics are more bursty and lacks continuous attention. Moreover, within the **News** group, text only news have the shortest life span, further suggesting that multimedia content plays a vital role in enhancing tweet life span.

Table 2: Decay rate coefficient fitting across topics and multimedia contents

Category/Coefficient	N_0	b	τ
Travel image	0.2252(-0.473,0.923)	1.463(0.948,1.984)	0.684
Travel url	0.0693(-0.010,0.149)	1.455(1.309,1.693)	0.687
Travel text	0.0154(-0.012,0.043)	1.501(1.150,1.760)	0.666
Food image	0.209(-0.443,0.861)	1.475(0.951,1.999)	0.678
Food url	0.058(-0.015,0.131)	1.504(1.319,1.741)	0.665
Food text	0.012(-0.011,0.034)	1.530(1.171,1.837)	0.654
News image	0.034(-0.054,0.122)	1.754(1.320,2.188)	0.570
News url	0.015(-0.014,0.043)	1.762(1.434,2.089)	0.568
News text	0.003(-0.003,0.009)	1.838(1.516,2.160)	0.544

7 Limitations and Future Work

In our previous work, we explore the impact of multimedia content in Sina Weibo from 3 aspects: (I)Tweet and User Popularity, (II) User Post Preference and (III) Tweet Life Span. Using retweet number as a measure for tweet popularity and follower count for user popularity, our findings shows that (I) Multimedia content promote tweet and user popularity, (II) Users exhibit similar preference in posting multimedia tweets and text tweets, and (III) Multimedia tweets have a longer life span.

The assumption in thesis is whether multimedia content has influence or not, it treats multimedia content as the sole factor to exert influence. The limit of this assumption is that it does not take into account that other factors such as user popularity, buzz words may also influence the popularity and life span of tweets.

Instead of studying whether multimedia contents influences or not, we ask ourselves: How much influence multimedia contents would have, compared with other factors, on the popularity and life span of tweets?

One solution is to build models which take different factors into account. For example, we may assume the popularity of tweets is linearly related to user follower count and whether it contains multimedia content, thus we may have

$$p = Au + Bm + C \tag{6}$$

where p denotes the popularity of the tweet, u denotes the user follower count of that tweet, m is whether this tweet contains multimedia, A, B and C are coefficient to be determined by data. We can also use other models to denote the relationship of these factors. Plugging the data into these models, we can find the best model with least fitting error. For the best model, we could compare the coefficient of those factors, then we could figure out how much influence multimedia will have compared with other factors.

8 Summary of Conclusions

In this thesis, we study the composition of multimedia content and analyze its impact in a popular microblogging service, Sina weibo. We use a binary comparison method to show the difference of popularity, life span, activeness etc. between multimedia contents and traditional textual information.

Our findings suggests 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. Retweet Length results reflect that the influence of Multimedia Tweet is broad and radioactive, while the influence of Text Tweet is deep and penetrating. Text body analysis shows multimedia tweet has a shorter and more focused distribution of text length compared to text tweet.

Our topical level analysis further suggest that in **News** tweets, multimedia contents do not necessarily promote the popularity of the tweet as it does in **Travel** and **Food**. News topics have the shortest life span in the three groups and our study further suggest that multimedia do prolong the life span of tweet.

Our findings is beneficial for web developers and social network marketers.

References

[1] WilM.P. Aalst and Minseok Song. Mining social networks: Uncovering interaction patterns in business processes. In *Business Process Management*, Lecture Notes in Computer Science, pages 244–260. Springer Berlin Heidelberg, 2004.

[2] Doug Beaver, Sanjeev Kumar, Harry C. Li, Jason Sobel, and Peter Vajgel. Finding a needle in haystack: facebook's photo storage. In *Proceedings of the 9th USENIX conference on Operating systems design and implementation*, OSDI'10, pages 1–8, 2010.

[3] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, pages 993–1022, 2003.

[4] Joseph Bonneau, Jonathan Anderson, and George Danezis. Prying data out of a social network. In *Proceedings of the 2009 International Conference on Advances in Social Network Analysis and Mining*, ASONAM '09, pages 249–254, 2009.

[5] 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.

[6] Zhiyuan Cheng, James Caverlee, and Kyumin Lee. You are where you tweet: a content-based approach to geo-locating twitter users. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, CIKM '10, pages 759–768, 2010.

[7] 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.

[8] 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.

[9] Pedro Domingos and Matt Richardson. Mining the network value of customers. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '01, pages 57–66, 2001.

[10] Nathan Eagle and Alex (Sandy) Pentland. Reality mining: sensing complex social systems. *Personal Ubiquitous Comput.*, pages 255–268, 2006.

[11] D. Easley and J. Kleinberg. *Networks, crowds, and markets*. Cambridge Univ Press, 2010.

[12] 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.

[13] 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.

[14] 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.

[15] E. Kalogerakis, O. Vesselova, J. Hays, A.A. Efros, and A. Hertzmann. Image sequence geolocation with human travel priors. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 253–260, 29 2009-Oct. 2.

[16] 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.

[17] F. Lardinois. The short lifespan of a tweet: Retweets only happen within the first hour. In *http://www.readwriteweb.com*, 2010.

[18] Kyumin Lee, James Caverlee, and Steve Webb. Uncovering social spammers: social honeypots + machine learning. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '10, pages 435–442, 2010.

[19] 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.

[20] 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.

[21] Wei Li and Andrew McCallum. Pachinko allocation: Dag-structured mixture models of topic correlations. In *Proceedings of the 23rd international conference on Machine learning*, ICML '06, pages 577–584, 2006.

[22] Yunpeng Li, D.J. Crandall, and D.P. Huttenlocher. Landmark classification in large-scale image collections. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 1957–1964, 29 2009-Oct. 2.

[23] David Liben-Nowell and Jon Kleinberg. The link-prediction problem for social networks. *Journal of the American Society for Information Science and Technology*, pages 1019–1031, 2007.

[24] Jack Lindamood, Raymond Heatherly, Murat Kantarcioglu, and Bhavani Thuraisingham. Inferring private information using social network data. In *Proceedings of the 18th international conference on World wide web*, WWW '09, pages 1145–1146, 2009.

[25] Alan Mislove, Massimiliano Marcon, Krishna P. Gummadi, Peter Druschel, and Bobby Bhattacharjee. Measurement and analysis of online social networks. In *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*, pages 29–42, 2007.

[26] Yan Qu, Chen Huang, Pengyi Zhang, and Jun Zhang. Microblogging after a major disaster in china: a case study of the 2010 yushu earthquake. In *Proceedings of the ACM 2011 conference on Computer supported cooperative work*, CSCW '11, 2011.

[27] Matthew Richardson and Pedro Domingos. Markov logic networks. *Machine Learning*, pages 107–136, 2006.

[28] C. Spearman. The proof and measurement of association between two things. *The American journal of psychology*, 15(1):72–101, 1904.

[29] 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.

[30] Xuerui Wang and Andrew McCallum. Topics over time: a non-markov continuous-time model of topical trends. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '06, pages 424–433, 2006.

[31] Sina Weibo. http://www.weibo.com.

[32] J. Weng and B.S. Lee. Event detection in twitter. In *Proceedings of 5th International Conference on Weblogs and Social Media (ICWSM)*, 2011.

[33] Guandong Xu, Yanchun Zhang, and Xun Yi. Modelling user behaviour for web recommendation using lda model. In *Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology - Volume 03*, WI-IAT '08, pages 529–532, 2008.

[34] L. Yu, S. Asur, and B.A. Huberman. What trends in chinese social media. *arXiv preprint arXiv:1107.3522*, 2011.

[35] Louis Lei Yu, Sitaram Asur, and Bernardo A. Huberman. What trends in chinese social media. *CoRR*, abs/1107.3522, 2011.

[36] 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.

[37] 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.