CS224W Project Proposal: Information Flows on Twitter

Huang-Wei Chang, Te-Yuan Huang

1 Introduction

Twitter has become a very popular microblog website and had attracted millions of users up to 2009. It is generally considered as a social networking website but gradually also used as a media for people or companies to spread out news or marketing information. In (Kwak, 2010) the authors made the previous claim and showed some statistics of the Twitter data to support it. For example, on Twitter a user u_1 can *follow* another user u_2 without u_2's permission and u_2 may not follow u_1's post (the posts are called *tweets*). In other words, the following relationship can be asymmetric. The authors in (Kwak, 2010) found only 22.1% of the following relationships are reciprocal, which implies the information flow maybe more like broadcasting instead of interaction. This is very different from the online messaging services such as MSN or Google Chat. However, because Twitter provides a very convenient way to *re-post* others' posts on a user's own page (this action is called *retweet*) and to reply (comment) others' posts, there can be much more interactions among users of Twitter than the usual blog services or personal webpages.

As an online news media, Twitter have the potential of spreading information beyond the physical geographical distances and many other restrictions. However, on the other hand, as a social networking platform, the information flowing within a group should be what the people in that group are concerned about. The multifaceted nature of Twitter makes it an intriguing resource to study how information flows over the online environments. Besides, there is another reason why Twitter is a valuable resource to study online information flows. In stead of only being exposed to a post by whom a user follows, a user can retweet, reply or mark that post favorite; these actions enable us to know the user actually had read the post. Therefore, it is possible to do more reliable analysis of the information flows over Twitter users.

Motivated by above observations, we plan to investigate the information flows over Twitter as our final project topic. Besides the pattern of information flow by looking at retweets and replies, we are also interested in knowing how location and content information affects the information flow. To be specific, we want to build a model for the information cascades on Twitter. We plan to take the following three steps in our final project: (1) an epidemic model, of which the design is motivated by the design of Twitter, (2) explore the behaviors of information flow on the Twitter dataset, especially their relationship with location and content (3) design a model for the cascades from the observations we obtain in (2), and a machine learn algorithm to fit the model to the data.

2 The Epidemic Model

There are three actions users can take after they read a post by the ones they follow: reply, retweet, and mark as favorite. Therefore, if the followers take any of the three actions, we know they have *read* the post for sure. Among those three actions, only through retweeting, the reader's followers can see the post automatically. In other words, only when a reader retweets, the reader start to propagate the information received from others. Based on the design, we use the following epidemic model (see Figure 1): In the beginning the status of a user u is *susceptible*. If one of the users whom u follows become *contagious*, u becomes *exposed* immediately. Then if u takes any of the three aforementioned actions it becomes *infected*. Moreover, if the action taken is retweet, then u not only becomes infected but also contagious. We call the transfer rate from the exposed state to the infected state as the *infection rate* and the transfer rate from being contagious as the *contagion rate* [1].

One reason of taking this model is we want to distinguish the case of having a post on the page (exposed) with the case of actually reading the post (infected). Of course, the way we define the status will underestimate

[1] The name *contagion rate* could be misleading; maybe we could coin a better one later.

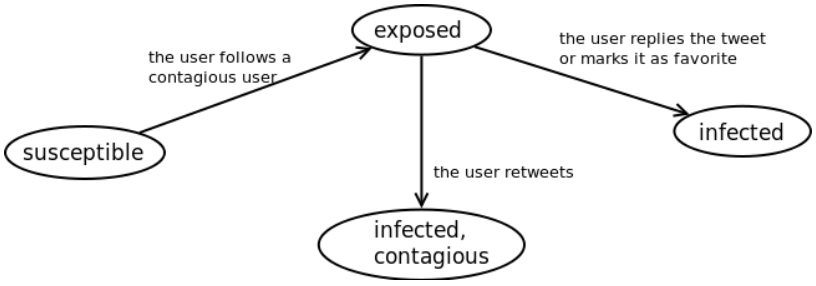

Figure 1: The Epidemic Model

the number of the second case because a user may read the post but doesn't take any of the three actions. One way of detecting more of that kind of cases is to analyze u's posts after the information appears on the user's page, and mark u as infected if the content is closely related to the original post. However, this will need much more work and we are not sure how effective it is [2]. Thus, in the first stage we plan to use the simplified model and if time permits we will do text analysis.

3 Dataset Description

The datasets we plan to use include the data provided with (Kwak, 2010) and the data we have retrieved, which is tagged with high resolution location information. We describe the data we retrieved in this section.

Earlier this year, Twitter starts allowing users to append their geo-location information when they are tweeting with their GPS-enabled mobile devices. Also, in this June, they released a set of API and through which we can retrieve tweets that is published within a pre-defined range. For the moment, we are collecting data from top 12 cities (NYC, LA, Chicago, Seattle, BayArea, Dallas, Huston, Philadelphia, San Antonio, San Diego, Phoenix). The data we collected contains the following information:

- Text content of the tweet.
- Is the tweet a reply? If yes, which tweet is this one replied to.
- Is the tweet a retweet? If yes, which tweet is this one retweet from.
- The source of the tweet (whether it is through web, API or twitter client on the mobile devices)?
- The timestamp of the tweet.
- Where the tweet is generated (GPS location, i.e., longitude and latitude)?
- The user's profile.

Alongside the twitter trace with location tags, we also plan to use twitter's API to gather who a user is following and is followed by whom. Besides being used as a guide to build the network, the pattern can be used as one means to represent a user's interest (please see section 4.3).

4 Some Tasks We Plan to Do

4.1 Connectivity of The Twitter Network

In (Kwak, 2010) the authors found the average shortest path length of the directed following network is 4.12, which is surprisingly short by considering the size of the network and the directed nature of the edges. However, the way they computed the path length assumes every exposed node will become infected and contagious in our epidemic model (both infection rate and contagion rate are 1). It is interesting to remove this assumption and see how the average path length changes. For example, we could examine the relationship between the infection rate for an exposed user (or the contagion rate of an infected user) with the length of shortest path.

Besides, since Twitter can be considered as a news broadcast media, we are interested in how many cascades are generated or propagated by the main publishers (for the *main publishers* we mean the one who has a lot of followers) and how many are among r-friends (a pair of users who follow each other). We guess the cascading behavior will be a pyramid; that is, the initial source propagates information to its followers and then the information flows among the followers' r-friends.

[2] Note u may read the post but doesn't do anything at all. In this situation, we have no means to know whether it reads the post or not.

4.2 Information and Geo-Location

Intuitively, some information will be propagated widely but some will be read only locally; for example, consider the promotion information of a local restaurant. We would like to compute the average area the information can flow, and to check how many information flows can cross different geographical regions. Because the geographical property of a information flow is very possible to be related to the information content, another interesting analysis is to explore their relationship [3].

Next we want to consider the question: if a cascade can cross the geographical regions, how fast can it be? Note we can expect there will be delay in the cascade due to location difference because the users might be in different time zones even though they check their Twitter pages all at the same local time.

In (Kwak, 2010) the authors reported a temporal analysis of retweets; they mentioned most retweets occurs within a day but there are still about 10% retweets take place after one month. An interesting analysis is when and where the quick responses happen. Thanks for the timestamps and high resolution of GPS data attached with the posts and retweets, it is possible for us to predict where the user made the responses such as in the office or at home.

4.3 How Much Does The Content Matter?

We expect the behavior of information flow to be related to its content. For instance, if I'm not interested in ballet at all, I may ignore the information even though it is retweeted by my best friend. As a result, we would like to check what kind of information can produce longer cascade or can cover a larger region in short time. On the other hand, we also want to know whether there is a difference in what kind of information is more popular in different cities. One thing we plan to do is to compute the most popular topics of the information flows for each of the twelve cities in our dataset.

Here comes a problem: how can we know the content conveyed by a cascade? Generally speaking, it can be known only by doing text analysis, but if we are examining information flows generated by a main publisher we can use the profile of the publisher to guess the content. To be specific, most of the users with many followers are celebrities or organizations and we can use their profiles to tag their followers' interest. For example, if a user follows many singers, we can guess the user is interested in music.

4.4 Model Fitting

After designing the epidemic model, the next step is to infer the model parameters such as the infection and contagion rates (by our setting the exposure rate is one as long as the link exists) so that we can run simulations or algorithms to do predictions using the model. In stead of learning the parameters for each edge or node, we prefer to learn two functions $f_{infection}$ and $f_{contagion}$ that take features as input and output the predicted infection and contagion rates respectively. For an edge (u, v) the features we are considering to use include the relationship of u and v (r-friends), their distance, the content conveyed by the flow, some properties of u and v (such as how many followers, how many people v follows, and how often v retweet). We haven't figured out the exact forms of those prediction functions, but we intend to use a learning algorithm to learn the functions from data.

As to the evaluation methods, we will compute the likelihood of the model and compare it with some naive models such as using the same rates for all edges and nodes. With those models, we could also check if the location and content can provider better explanation of the cascading behaviors. Furthermore, we want to simulate cascades using the learned model and check if the simulated behaviors are consistent with the real data.

5 Other Possible Directions

There are two more possible directions we can do. The first one is text analysis for the tweets. As mentioned previously, in many situations if we can analyze the post content we can have better information of what is going on. For example, we can better understand if a post evokes discussions in the followers' pages. We can possibly also obtain more information of a user's interest or what topics the user cares by analyzing the user's posts. The second direction is to run cascade related algorithms over the learned model to see what will happen. For the cascade related algorithms we mean the ones we learned from the lectures like *influence maximization* and *breakout detection*.

References

H. Kwak, C. Lee, H. Park, S. Moon (2010). *What is Twitter, a Social Network or a News Media?*. WWW 2010.

[3] Please see section 4.3