COMS 6998: Social Networks Final Project

Ying Tan
Columbia University
New York, New York
yt2443@columbia.edu

Andy Enkeboll
Columbia University
New York, New York
ale2124@columbia.edu

ABSTRACT

Retweeting behavior on twitter is the behavior that user reposts comments from their friends. Few studies have investigated in combining a user's subjectivity motivation with his conformity to environment to predict a user's retweeting behavior. Based on the sentiment analysis, we combine a user's subjectivity motivation with a designed adoption model which measures a user's neighbors' influences, and then establish a mixture model with these two factors to do prediction. Then, we evaluate our model in twitter dataset to verify its prediction performance.

Keywords

Twitter, Retweet behavior, Information diffusion, Sentiment analysis

1. PROBLEM STATEMENT

1.1 Motivation

Currently, social networks plays an important role in spreading information in large scale. Social networks provide a platform for people to express their opinions and communicate with others through a series of interactive behaviors, such as retweeting, posting and clicking likes. This provides a vast source of information on a unprecedented scale. How to utilize such information from user's comments is a core problem. In this paper, we will focus on using short comments in social networks to predict users' forwarding behaviors. For business merchants, this will help them to adjust their marketing strategy promptly to continuously meet customers' new demands. For a government, this will help it to do online public opinion analysis. For politicians, this will help them to gather more votes from public during election (Etter et al. 2014). Thus, people's behaviors contain so much information which provide a great chance for us to understand human behaviors.

1.2 Basic Idea

In sociological psychology, a person will firstly make an individual free choice according to his subjective conscious, and then choose to change in belief or behavior in order to fit in with a group. This is called conformity. So, when predicting a user's behavior in social networks, we have such assumption:
Assumption 1: Users' behaviors in social network are influenced by their individual choices and choices of their neighbors.

More specifically, we decide to predict user's retweeting behavior in Twitter. Given a tweet w from a user u_1, we will predict the probability of another user u_2 to retweet w.

To measure how much u_2's individual choices will influence his retweeting behavior, we build our subjectivity model which has following assumption:

Assumption 2: Users' individual retweeting choices are influenced by their interests in the tweet, as well as the opinion similarity between their opinions and opinions expressed by the tweet.
For the factor of interest in assumption 2, we will measure how much interest u_2 has in tweet's content because a user will retweet a tweet only if he has interests in this tweet's content. Then, for the second factor, we will measure how much u_2 agrees or disagrees with this tweet, which is opinion similarity between u_2 and w. So, combining these two factors, we can measure how much u_2's individual choices will influence his retweeting behavior.

To measure how much u_2 will be influenced by his neighbors' choices, we build our adoption model based on the following assumption:

Assumption 3: The probability of a user to decide to retweet w is a monotone function of probability of his neighbors who will retweet it.
So, if u_2's neighbors have high probability to retweet w, even u_2 decides not to tweet it firstly, he may retweet it after a certain time.

1.3 Modeling retweeting behaviors

Our contribution in this paper can be summarized as follows:

- Given a tweet w posted by user u_1, we create a subjectivity model to calculate the probability of another user u_2 to retweet it.
- Based on the result of the first phase, we build an adoption model to simulate how u_2's behavior is affected by his neighbors, and finally predict how likely u_2 will retweet a tweet.

The main steps of modeling in our approach are shown below:

First Phase: We introduce $p_w(u)$ to denote how likely user u will retweet tweet w, and we calculate it in this way:

$$p_w(u) = \alpha DI_u(w) + (1-\alpha)sim(u,w) \quad (1)$$

where $DI_u(w)$ denotes u's interests in tweet w; $sim(u,w)$ denotes opinion similarity between u and w; $0 < \alpha < 1$, as a weight distributed on two different factors according to assumption 2. We measure $DI_u(w)$ by average weighted sum of u's overall sentiment strength towards all topics related to w, $sim(u,w)$ is measured by sum of similarities between u and v's sentiment strength distributions on all topics related to w.

Second Phase: Based on the first phase, we calculate final $p_w(u)$ according to our adoption model. The adoption model for updating $p_w(u)$ is:

$$p_w(u)^{t+1} = \beta * N_w(u)^t + (1-\beta) * p_w(u)^t \quad (2)$$

where $N_w(u)^t$ denotes the influence of all u's neighbors on u 's decisions of whether tweet w or not. $0 < \beta < 1$, as the weight of neighbor's influenece. $p_w(u)^t$ is the probability of u to retweet w at time t. We set $p_w(u)^0$ to the result based on equation (1). $N_w(u)^t$ is measured by weighted combination of probability of all u's neighbors to reweet w, and the weight factors are determined by opinion similarity between u and his neighbors.

2. RELATED WORK

Retweeting Analysis: There are so many works on predicting users' behaviors. Most of them use machine learning approachs to do classification and prediction. However, few of them use semantic features hidden in users' tweets, which are important factors to represent users' interests and opinions. For example, Xu and Yang only used content-based features like TF-IDF feature, latent topic feature etc in classification framework to do prediction (Xu and Yang 2012). There are some works using sentiment analysis: Naveed introduced more content-based features like sentiment, positive and negative terms etc to learn which tweet was retweeted in regression model (Naveed et al. 2011). Xie introduced a subjective model to predict retweeting behavior based on tweet's sentiment strength (Xie, Tang, and Wang 2014). But, they didn't consider the situation that a user's decision of retweeting may also be influenced by other people's decisions. Wu introduced a information diffusion model to predict forwarding behavior. But, he has not considered the sentiment factor of text feature when training logistic regression model (Wu, Ji, and Liu 2013). To our best knowledge, there were few efforts on combiningsentiment analysis and social network influence spreading together to predict users' retweeting actions. So, in this paper, we make some improvements by combining these two factors together.

Sentiment Analysis: Sentiment Analysis is crucial for prediction of user's behavior. Through the sentiment analysis of user's tweets, we find users' interests in tweets and their opinions. Thelwall (Thelwall, Buckley, and Paltoglou 2012) created SentiStrength, which exploits the de facto grammars and spells styles of cyberspace to extract sentiment strength from informal English text. In our paper, we use SentiStrength tool to calculate emotion tendency showed by tweets, and thus calculating user's interests and opinions on tweets.

3. RESULTS

To test our new model, we used a Twitter dataset which contains more than 50,000 tweets related to about 1, 000 users between Sept. 10th and Sep. 20th 2012. Then we built three test cases by selecting randomly 100, 200, 400 tweets posted by these users after Sep. 21th, 2012. We also excluded from the test cases the tweets which are too short, or contain too few meaningful words. So, the final test cases consists of 100, 183 and 333 tweets respectively. We implement the adoption model and mixture model which combine the subjectivity model with the adoption model on these datasets and compare their prediction precisions. Our test results are shown in Table 1.

Table 1: Prediction Results

	Adoption Model	Mixture Model
Dataset of 100 tweets	66.0%	82.0%
Dataset of 183 tweets	73.22%	91.26%
Dataset of 333 tweets	75.37%	89.49%

Claim 1: Combining subjectivity model based on sentiment analysis with the adoption model can get better prediction accuracy in our dataset than the prediction based on the adoption model only

We also simulated the process of adoption to retweet a given tweet in the mixture model .The process is shown in following figure:

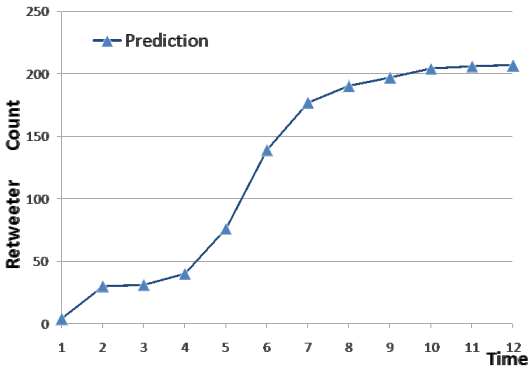

Figure 1: Trend of count of possible retweeters for a given tweet as time advance

Claim 2: As time goes on, the size of subset of users U_t which is infectious for the mixtured model will increase. Then, at some time slot t, the infectious subset U_t will remain relatively stable.

We proposed a mixture model combining the subjectivity model with adoption model to predict user's retweeting behavior in social networks. We compare our model with the adoption model using twitter dataset. We found out that combining the semantic analysis in adoption model can enhance the prediction accuracy.

4. REFERENCES

Etter, Vincent et al. (2014). "Mining Democracy". In: *2014 Conference on Online Social Network*, pp. 1–12.

Khosrovian, Keyvan, Dietmar Pfahl, and Vahid Garousi (2008). "GENSIM 2.0: a customizable process simulation model for software process evaluation". In: *ICSP'08 Proceedings of the Software process, 2008 international conference on Making globally distributed software development a success story*, pp. 294–306.

Naveed, Nasir et al. (2011). "Bad news travel fast: a content-based analysis of interestingness on Twitter". In: *Proceedings of the 3rd International Web Science Conference* 8.

Thelwall, Mike, Kevan Buckley, and Georgios Paltoglou (2012). "Sentiment strength detection for the social web". In: *Journal of the American Society for Information Science and Technology* 63.1, pp. 163–173.

Wu, Kai, Xinsheng Ji, and Caixia Liu (2013). "Information Diffusion Model for Microblog". In: *Software Engineering and Service Science (ICSESS), 2013 4th IEEE International Conference on*, pp. 212–215.

Xie, Songxian, Jintao Tang, and Ting Wang (2014). "Resonance Elicits Diffusion: Modeling Subjectivity for Retweeting Behavior Analysis". In: *Cognitive Computation*, pp. 1–13.

Xu, Zhiheng and Qing Yang (2012). "Analyzing User Retweet Behavior on Twitter". In: *Advances in Social Networks Analysis and Mining (ASONAM), 2012 IEEE/ACM International Conference on*. Inst. of Autom., Beijing, China, pp. 46–50.

5. ADDITIONAL INFORMATION

There are two phases in modeling retweeting behavior, the details are outlined below.

5.1 Modeling: The First Phase

During data pre-processing, for each user u, we accumulate his 50 tweets together to form a single document, and use Gensim tool (Khosrovian, Pfahl, and Garousi 2008) to extract a set of topics T_u. To predict whether u will retweet a test tweet w, we extract a new set $T_w = t_1, t_2, ...t_k$ from T_u which contains k topics whose degrees of relevance with w are higher according to equation(4).

To calculate u's interests in a specific tweet w, we have the equation:

$$DI_u(w) = \frac{\sum_{i=1}^{k} DI_u(t_i) DR_w(t_i)}{\sum_{j=1}^{k} DR_w(t_j)} \tag{3}$$

where $DR_w(t_i)$ is w's degree of relevance with each topic $t_i \in T_w$, and $DI_u(t_i)$ is u's interests in a topic $t_i \in T_w$.

To calculate w's degree of relevance with each topic t_i , we define

$$DR_w(t_i) = \sum_{v_w} WP_v(t_i) num(v_w) \tag{4}$$

where v_w is the word from tweet w, $num(v_w)$ is the number of times v_w appears in tweet w. $WP_v(t_i)$ is the weight of word v_w in topic $t_i \in T_w$.

To calculate u's interests in a topic t_i, we have

$$DI_u(t_i) = \frac{\sum_{w \in W_u(t_i)} S_w}{\sum_{w \in W_u} S_w} \tag{5}$$

where $W_u(t_i)$ is the set of tweets forwarded by u which are related to t_i. Each w in this set satisfies $DR_w(t_i) > 0$, W_u is the set of all tweets forwarded by u, $S_w = s^p(w) + |s^n(w)|$ is tweet w's sentiment strength. $s^p(w)$ represents positive emotion strength of w and $s^n(w)$ is its negative emotion strength. We use *SentiStrength* tool (Thelwall, Buckley, and Paltoglou 2012) to calculate $s^p(w)$ and $s^n(w)$.

To calculate opinion similarity between u_2 and w, we have

$$sim(u, w) = \frac{maxDist - dist(u, w)}{maxDist} \tag{6}$$

where $maxDist$ equals to 9, which is maximum difference between any two opinions, and $dist(u, w)$ is the difference between user's opinions on w and opinions expressed in w's content. Then, to calculate u's opinions on w, we represent u's opinions on a tweet w as a vector $VP_u(w)$of length 10, which contains 10 different emotional weights for w, and each entry in this vector represents a sentiment strength value of 10 different emotion kinds (positive sentiments from 1 to 5, negative ones from -1 to -5) according to SentiStrength. In the same case, we represent opinions revealed in tweet w as a vector of length 10: VP_w.

Under this definition, we have:

$$dist(u, w) = |\sum_{i=1}^{10} i * VP_u(w)[i] - \sum_{j=1}^{10} j * VP_w[j]| \tag{7}$$

where $VP_u(w)[i]$ means the i th component of u's opinions on w, and $VP_w[j]$ means the j th component of opinions revealed by w. We calculate VP_w in such way

$$\begin{cases} VP_w[i+5] = 0.5 \ when \ s^p(w) = i, 1 \le i \le 5 \\ VP_w[j+6] = 0.5 \ when \ s^n(w) = j, -5 \le j \le -1 \\ VP_w[q] = 0 \ when \ q \ne i, j \end{cases} \tag{8}$$

To calculate $VP_u(w)$, we have

$$VP_u(w)[i] = \frac{\sum_{i=1}^{k} DI_u(t_i) VP_u(t_i)[i]}{\sum_{j=1}^{k} DR_w(t_i)} \tag{9}$$

where $VP_u(t_i)$ is u's opinions on a topic $t_i \in T_w$ as a vector of length 10, which can be calculated in such way:

$$VP_u(t_i) = \frac{num_u(t_i, s)}{num_u(t_i)} \tag{10}$$

where $num_u(t_i)$ is the number of tweets w forwarded by user u which is related to topic t_i, which means $DR_w(t_i) > 0$; $num_u(t_i, s)$ is the number of tweets w which satisfy: $DR_w(t_i) > 0$ and $S_w = s$, where

$$\begin{cases} s = i - 6 \ when \ 1 \le i \le 5 \\ s = i - 5 \ when \ 5 < i \le 10 \end{cases} \tag{11}$$

5.2 Modeling: The Second Phase

Based on the first phase, we calculate $p_w(u)$. The adoption model for updating $p_w(u)$ is:

$$p_w(u)^{t+1} = \beta * N_w(u)^t + (1 - \beta) * p_w(u)^t \tag{12}$$

where $N_w(u)^t$ is the influence of all u's neighbors on u at time t, it is calculated in such way:

$$N_w(u)^t = \sum_v \frac{N_w(u, v)^t sim(u, v, w)}{C} \tag{13}$$

where v is neighbors of u, $N_w(u, v)^t$ denotes v's influence on u 's decisions of whether retweet w or not at time t.

$sim(u, v, w)$ is the similarity between u and v's opinions on twitter w. C is the maximum numbers of neighbors of all users. To measure the influence of v on u, we have

$$N_u(v)^t = \frac{p_w(v)^t - \bar{p_w}^{t-1}}{pmax_w^{t-1} - \bar{p_w}^{t-1}} when\ p_w(v)^t > \bar{p_w}^{t-1} \quad (14)$$

where $p_w(v)^t$ shows the probability of v to retweet w at time t; $\bar{p_w}^{t-1}$ denotes the average value of probability of all users to retweet w at time $t-1$. $pmax_w^{t-1}$ shows the maximum value of probability to retweet w among all users.

5.3 Data Preprocessing

Because so many tweets contain useless information, so we design a way to collect training tweets. We randomly choose 1000 tweets which have been posted between 2012.09.10-2012.09.20 over twice, whose content contain at least 20 characters. Then, we create a set which contains the users who retweet them, and these users' followers. From this set, we filter the users who have not retweeted over 20 times during 2012.09.10-2012.09.20, and then randomly select 1000 users from it to do training. For each user in 1000, we collect 50 tweets posted by him which also satisify the fromer tweet constraint, and finally form the training dataset which contains about 51500 tweets. For each tweet, we use CMU ARK Twitter Part-of-Speech Tagger to remove noise and stop words. For each user's 50 tweets, we use Gensim 0.10.3 to divide them into 5 topics.

We prepare three testing tweet dataset with different sizes: 200, 400, 800. We only kept the tweets which have been posted after 2012.09.21 at least once, and contain over 20 characters. So, after filtering, three test datasets contain 100, 183, 333 tweets. For each testing tweet dataset, we also create a set which contains the users who retweet these tweets, and these users' followers. Then, we select 400 users from this set randomly. This forms the testing user dataset for each testing tweet dataset.

5.4 Test Cases

We implement two different models on the three test datasets. Because of time limit, we have not prepared large testing dataset. The first model just consider the influence of a user's neighbors on his retweet behavior. The second model consider the impact of the user's subjectivity on his retweet behavior, as well as the adoption process.

Then, we will explain our measure of prediction accuracy. For each test tweet w, we will calculate the probability of each user in testing user set to tweet w. Then, we rank these users according to their probability. According to this ranked list, we choose the former $x\%$ users to be the candidate set to tweet w. We assume our prediction to be correct if the real tweeter appears in our candidate set. In the result of table 1, we select $x\%$ to be 2%. The result with $x = 1$ can be shown in following table:

Table 2: Prediction Results when x%=1%

	Adoption Model	Mixture Model
Dataset of 100 tweets	44.0%	60.0%
Dataset of 183 tweets	54.64%	69.39%
Dataset of 333 tweets	52.25%	69.36%

We have implemented this mixture model in java, and more results could be seen when running our code.