Data Mining in Social Network

Presenter: Keren Ye

References

Kwak, Haewoon, et al. "What is Twitter, a social network or a news media?." Proceedings of the 19th international conference on World wide web. ACM, 2010.

Pak, Alexander, and Patrick Paroubek. "Twitter as a Corpus for Sentiment Analysis and Opinion Mining." LREc. Vol. 10. 2010.

Data Mining in Social Network

What is Twitter, a social network or a news media?

Twitter

Basic Features

- Tweet about any topic within 140-character limit
- Follow others to receive their tweets

Twitter Space Crawl

Twitter Space Crawl

- Application Programming Interface (API)
- Data collection
  - Profiles of all users: June 6th - June 31st, 2009
  - Profiles of users who mentioned trending topics: June 6th - September 24th, 2009

Twitter Space Crawl

User Profile

41.7 million (4,170,000) user profiles.

1.47 billion (1,470,000,000) directed relations of following and being followed

Trending Topics + Associated Tweets

4,262 unique trending topics and their tweets

Query API every five minutes for trending topic title (Top-10)

Grab all the tweets that mention the trending topic

Twitter Space Crawl

Removing Spam Tweets

Why

- Undermine the accuracy of PageRank
- Spam keywords hinder relevant web page extraction
- Add noise and bias in analysis

How

- Filters tweets from users who have been on Twitter for less than a day
- Removes tweets that contain three or more trending topics

Basic Analysis

Followings and Followers (CCDF)

Complementary cumulative distribution function

$$\bar{F}(x) = \mathrm{P}(X > x) = 1 - F(x).$$

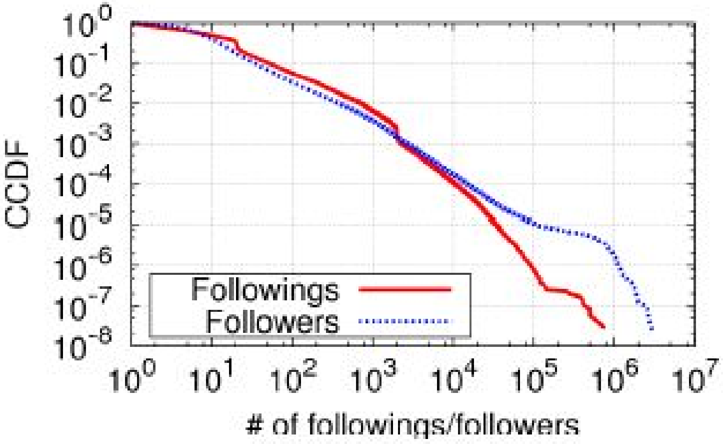


Figure 1: Number of followings and followers

Basic Analysis

Followers vs. Tweets

y: number of followers a user has

y: number of tweets the user tweets

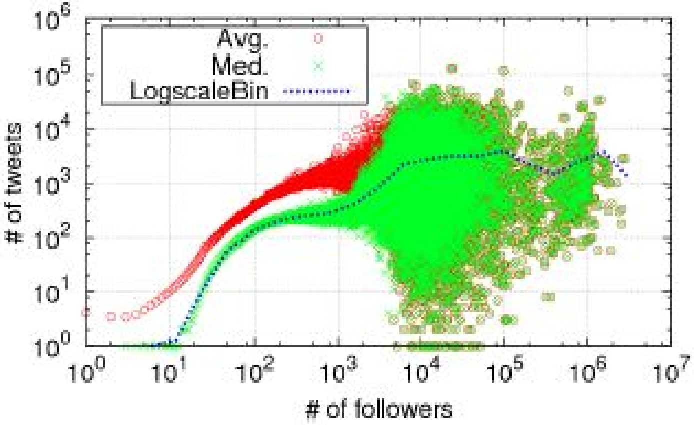


Figure 2: The number of followers and that of tweets per user

Basic Analysis

Followings vs. Tweets

y: number of followings a user has

y: number of tweets the user tweets

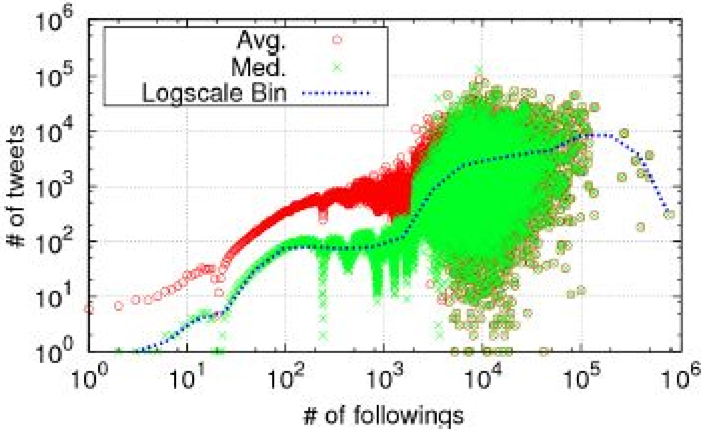


Figure 3: The number of followings and that of tweets per user

Basic Analysis

Reciprocity

Top users by the number of followers in Twitter are mostly celebrities and mass media

77:9% of user pairs with any link between them are connected one-way

only 22:1% have reciprocal relationship between them - r-friends

67:6% of users are not followed by any of their followings in Twitter

A source of information? A social networking site?

Basic Analysis

Degree of seperation

- Small world phenomenon - Stanley Milgram’s
    - “Any two people could be connected on average within six hops from each other”
- Main difference
    - The directed nature of Twitter relationship - only 22:1% of user pairs are reciprocal
    - Can we expect that two users in Twitter to be longer than other known networks
        - MSN - 180 million users, 6.0, 7.8 for medium and 90% degree of separation respectively

Basic Analysis

Degree of separation

Choose a seed randomly

Compute the shortest paths between the seed and the rest of the network - 4.12

Social network? Source of information?

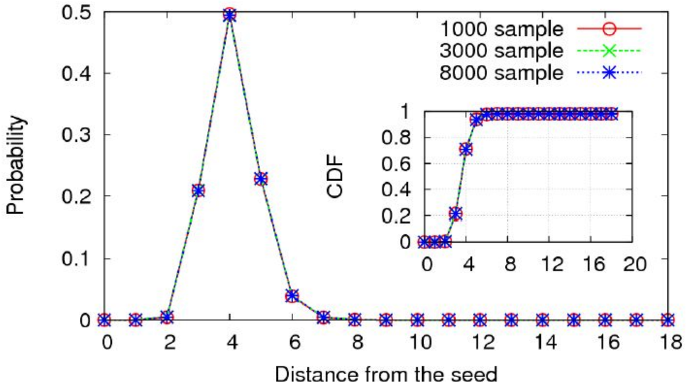


Figure 4: Degree of separation

Basic Analysis

Homophily

- A contact between similar people occurs at a higher rate than among dissimilar people
- Investigate homophily in two context
  - Geographic location
  - Popularity

Basic Analysis

Homophily

Geographic Location

Popularity

Social network? Source of information?

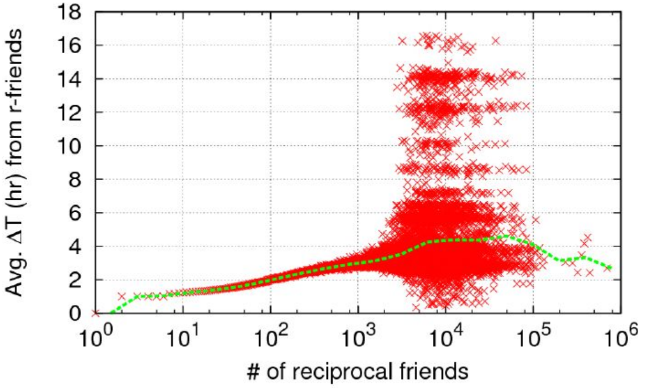


Figure 5: The average time differences between a user and r-friends

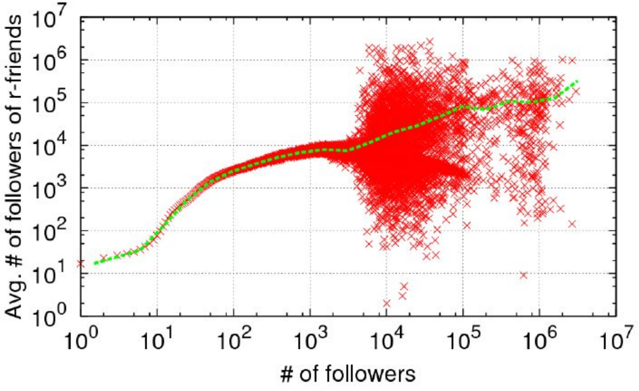


Figure 6: The average number of followers of r-friends per user

Trending the trends

Motivation

Interpret the act of following as subscribing to tweets

How trending topics rise in popularity, spread through the followers' network, and eventually die

Review

4,266 unique trending topics from June 3rd to September 25th, 2009

Apple's Worldwide Developers Conference, the E3 Expo, NBA Finals, and the Miss Universe Pageant

Trending the trends

Compare to Google Trend

Similarity

Only 126 (3.6%) out of 3,479 unique trending topics from Twitter exist in 4,597 unique hot keywords from Google

Freshness

On average 95% of topics each day are new in Google while only 72% of topics are new in Twitter

Interactions might be a factor to keep trending topics persist

Social Network?

Trending the trends

Compare to CNN Headline News

- Preliminary Results
    - More than half the time CNN was ahead in reporting
    - However, some news broke out on Twitter before CNN

Source of information?

Trending the trends

Singleton, Reply, Mention, and Retweet

Singleton: tweet with no reply or a retweet

Reply

Mention: tweet addressing a specific user, both replies and mentions include “@” followed by the addressed user’s Twitter id

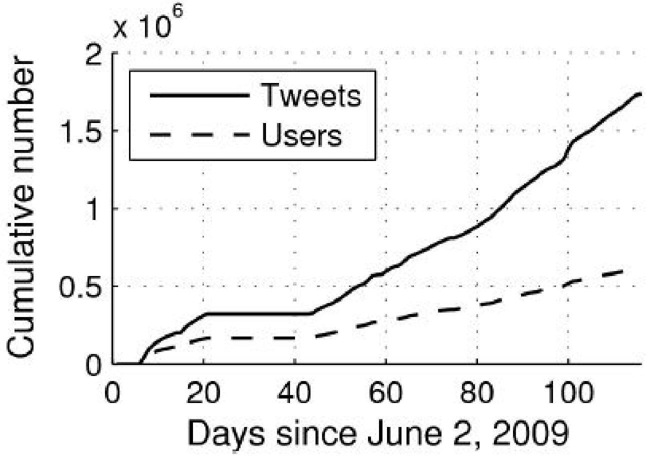
Retweet: marked with either “RT” followed by “@user id” or “via @user id”

Among all tweets mentioning 4,266 unique trending topics, singletons

are most common, followed by replies and retweets.

Trending the trends

Out of 41 million Twitter users, a large number of users (8; 262; 545) participated in trending topics and about 15% of those users participated in more than 10 topics during four months.



(a) Topic ’apple’

Trending the trends

Impact of retweet

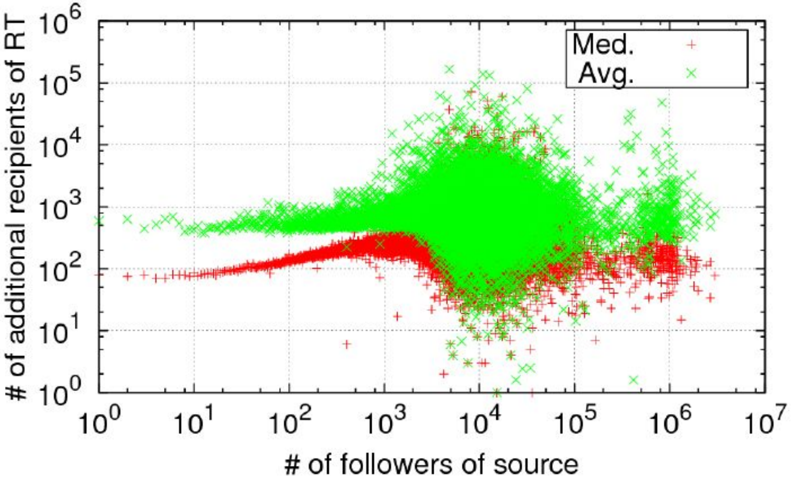


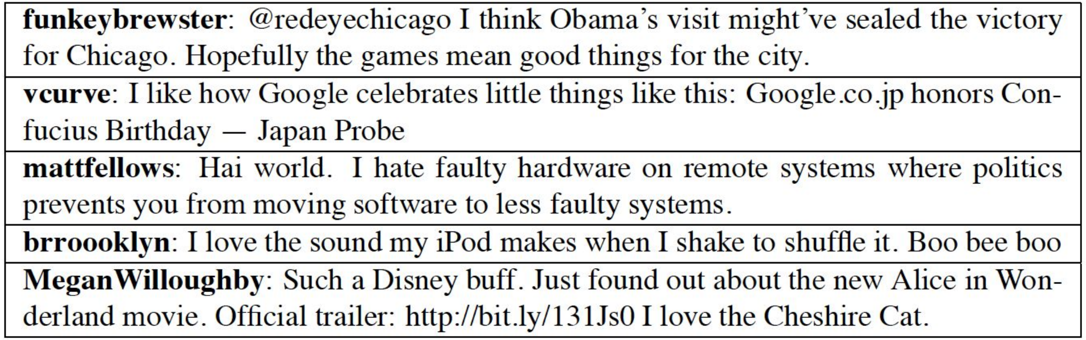
Figure 14: Average and median numbers of additional recipients of the tweet via retweeting

Data Mining in Social Network

Twitter as a Corpus for Sentiment Analysis and Opinion Mining

Motivation

Recognize positive / negative / objective sentiment



funkeybrewster: @redeyechicago I think Obama's visit might've sealed the victory for Chicago. Hopefully the games mean good things for the city.
vcurve: I like how Google celebrates little things like this: Google.co.jp honors Confucius Birthday — Japan Probe
mattfellows: Hai world. I hate faulty hardware on remote systems where politics prevents you from moving software to less faulty systems.
brroooklyn: I love the sound my iPod makes when I shake to shuffle it. Boo bee boo
MeganWilloughby: Such a Disney buff. Just found out about the new Alice in Wonderland movie. Official trailer: http://bit.ly/131Js0 I love the Cheshire Cat.

Table 1: Examples of Twitter posts with expressed users' opinions

Corpus collection

Use the Twitter API

The whole data set is huge, a subset is enough for training purpose

Using sentiment related emoji to get the positive / negative training corpus

Happy emoticons: “:-)”, “:)”, “=)”, “:D” etc.

Sad emoticons: “:-(”, “:(”, “=(”, “;(” etc.

For objective training corpus

Retrieve text messages from Twitter accounts of popular newspapers and magazines

Training the classifier

Feature

Feature Extraction

Model

Model Evaluation

Training the classifier

Feature

- Presence of a n-gram as a binary feature
- E.g., “I love the sound my iPodmakeswhen I shake to shuffle it. Boo bee boo”
  - Unigram (1-gram): presence of “I”, “love”, “the”, ...
  - Bigram (2-gram): presence of “I love”, “love the”, “the sound”, ...

Training the classifier

Feature extraction

- Filtering
  - Remove URL links, Twitter user names and emoticons
- Tokenization
  - Segment text by splitting it by spaces and punctuation marks
- Remove stopwords
- Construct n-gram
  - Negation is attached to a word which precedes it or follows it. E.g., “I do+not”, “do+not like”.

Training the classifier

Naive Bayes Model

s - sentiment

M - Twitter Message

$$P(s|M) = \frac{P(s) \cdot P(M|s)}{P(M)}$$

$$P(s|M) \sim P(M|s)$$

Training the classifier

Naive Bayes Model - An example

“I love the sound my iPodmakeswhen I shake to shuffle it. Boo bee boo”

P(s=+|M) ~ P(+) P(I|+) P(love|+) P(the|+) P(sound|+) ...

P(s=-|M) ~ P(-) P(I|-) P(love|-) P(the|-) P(sound|-) ...

By counting the number in training set, we can get:

P(+), P(-)

P(I|+), P(I|-), P(love|+), P(love|-), ...

$$P(s|M) = \frac{P(s) \cdot P(M|s)}{P(M)}$$

$$P(s|M) \sim P(M|s)$$

Training the classifier

Other details of the model

POS-tags as extra information

Discriminate common n-grams since they do not strongly indicate sentiment

Training the classifier

Model Evaluation

Precision: measures the proportion of correctly tagged tokens within the set of all the tokens that were non ambiguously tagged by the evaluated system. It is therefore a measure of the accuracy of the tagging effectively performed by the system.

Decision: measures the proportion of tokens non ambiguously tagged within the set of all token processed by the evaluated system. It therefore quantifies to which extent the evaluated system effectively tags the input data.

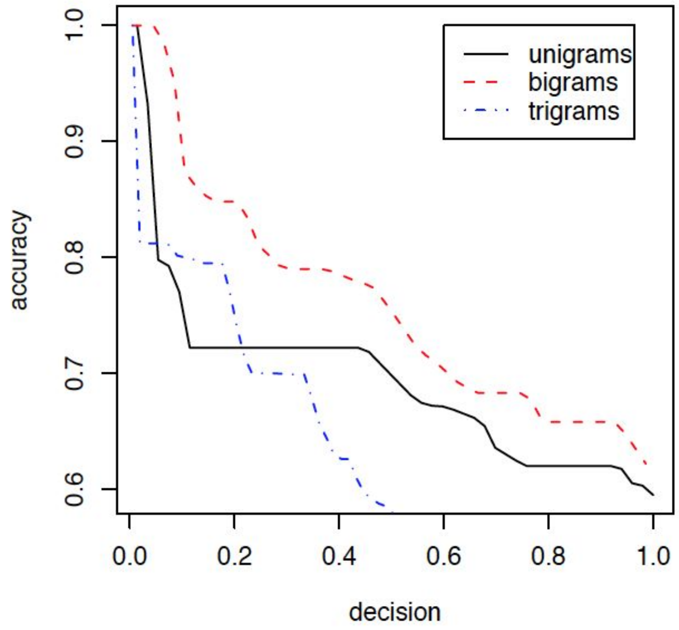
$$accuracy = \frac{N(\text{correct classifications})}{N(\text{all classifications})}$$

$$decision = \frac{N(\text{retrieved documents})}{N(\text{all documents})}$$

Training the classifier

$$accuracy = \frac{N(\text{correct classifications})}{N(\text{all classifications})}$$

$$decision = \frac{N(\text{retrieved documents})}{N(\text{all documents})}$$



Conclusion

Essence of data mining

- Find interesting patterns

General idea of the two papers

- Subjective way - propose problem, explain the reason.
- Objective way - propose problem, solve it.

Domain knowledge of the two

- Statistics and data visualization
- Machine learning technology

Thanks