Geo Twitter Data Collection and Visualization System

Hideyuki Fujita

Graduate School of Information Systems,
University of Electro-Communications
(Tokyo, Japan)
Backgrounds

Mobile social media

- generating valuable data for analyzing human behavior and events in the real world
- (mobile use of) Facebook, Twitter, Instagram, Flickr, Foursquare, etc.

Twitter

- 500 million users
- 400 million tweets per day
- 0.77% geotagged (with the location coordinates)
- 64% posted from mobile devices

Report in July 2012 by Semiocast, inc

- becoming mobile media
  - sharing realtime information including information related to current location
Geotagged tweets
(2009.1-2013.5)

Twitter official blog, https://blog.twitter.com/2013/geography-tweets-3
Geotagged tweets
(2009.1-2013.5)

Twitter official blog, https://blog.twitter.com/2013/geography-tweets-3
Geotagged tweets (2009.1-2013.5)

Twitter official blog: https://blog.twitter.com/2013/geography-tweets-3
Geo-Twitter Application: Related works

- **Interactive map application** for situational awareness  
  MacEachren et al., 2011
- **Realtime mapping** of local news  
  Sankaranarayanan et al., 2009
- **Realtime event detection and location / trajectory prediction** of earthquakes and typhoons  
  Sakaki et al., 2010

**Key technologies**
- Event extraction from text  
  - Natural language processing, machine learning
- Spatial analysis
- Location based data collection
Twitter Data collection: Problem and Objective

**Twitter API (Application Programing Interface)**
- Twitter's official service for providing *sampling data* through HTTP communication.
- easy to get *small amount of data*

**Problem** in collecting large amount of data
- The amount of sampling data is small in straightforward use of Twitter API.
- Continuous collection of data costs much effort.
- Having many researchers collecting the same data is not efficient.

**Objective**
- Efficient data collection system for geo-tweet data
- Data visualization system for geo-tweet data

**Future plan**
- *Data sharing system* for researchers using geo-tweet data
Data collection method

Limitation of Twitter Search API
• returns maximum 1,500 tweets under one search filter with location and date-period

Method
• divide area into small areas (grid)
• divide date-period into tweetID-periods
  tweet ID: integer ID attached to all tweets in ascending sequence
• collect data within each divided area and period
• aggregate collected data
Evaluation

Area: about $2 \times 2$ km around Tokyo Station
Period: 1 day

Num. of collected tweets:
- Common method using Streaming API: 31,711
- Common method using Search API: 1,500
- Proposed method: 97,787
Practical issues for collecting large area and long period

Access rate limitation to Search API per IP address
  • Connection is refused when the limit is exceeded.

Unstability of the API (best effort service)
  • Without explicit error message, the number of tweets in Search API response often becomes much smaller than usual.
Solutions for practical issues (1 of 2)

Data collection by distributed system

• access the API from multiple servers with multiple different IP addresses

Pilot data collection for monitoring Twitter API status

• continuously monitor the number of tweets collected in a certain small grid cell to determine the status of the API

• halt the data collection of the whole area when the number of collected data in the pilot data collection is much smaller than usual (smaller than 10% of the average), restart the data collection when the API returns stable.
Solutions for practical issues (2 of 2)

*Re-collection* of data that the system failed to collect

- check posted date time of collected tweets of each grid cell
- If there are certain periods when tweets were not collected, try to collect the data for those periods again in the grid cell

Repeat request when receiving an explicit API error
Distributed system for practical data collection

Master server (1 machine)
- *Pilot data collection* for monitoring Twitter API status
- Getting and caching *Date Boundary Tweet ID*
- Assigning collection areas and periods to data collection servers

Data collection servers (multiple machines)
- Data collection within assigned area and period
- *Data re-collection*
Experiment and Result

Area    about \(20 \times 20\) km around central Tokyo
Grid size  about \(2 \times 2\) km
Period    2 weeks (from 25 July 2011 0:00 JST)
Num. of tweets  3,476,059
Num. of users    216,430
Daily variation (central Tokyo 20x20km)

- Weekday > Weekend
Daily variation (Odaiba area 2x2km, 1 day)

- A popular shopping and amusement area
- Weekend > Weekday
Hourly variation (central Tokyo 20x20km, 2 days)

- A small spike at around noon (lunch break?)
- A large spike at midnight

Largest (Thursday)
Smallest (Sunday)
Hourly variation (around Tokyo station 2x2km)

- Small spike around 4 AM corresponds to a small earthquake.
Most users posted fewer than 4 geo-tweets in 2 weeks.
Number of grid cells user posted geo-tweets
(central Tokyo 20x20km, 2 weeks)

- More than half of the users posted geo-tweets in at least two different cells.
- One user posted geo-tweets in 56 different cells.
Conclusion

• Distributed data collection system for geo-tweet
  • collected several times more data than commonly used methods

• Spatio-temporal visualization system for geo-tweet

Future plan

• Scaling up the system
  • enlarge the area for collecting geo-tweet data
• Integrating realtime data collection system
• Data sharing system for researchers using geo-tweet data
Response of Twitter API (abstract)

Tweet text
Tweet ID
User ID
Destination user ID (optional)
  • only for tweets posted as replies to others (with “@user”)
User profile (optional)
  • including location name input by the user
Location coordinates (optional)
  • only for tweets tagged with the location coordinates (0.77%)
Types of Twitter API

**Streaming API**
- sends tweets continuously in realtime while connected by an API client

**Search API**
- returns a set of tweets that match a specified query when accessed by an API client

To collect tweets within a specified area
- **Streaming API** with location filter (geographic coordinates of an area)
- **Search API** with location and period (from and to date) search filter
Location information of Twitter

• Not all the tweets have location information

Location coordinates (latitude, longitude)
  • attached only when the user opt in geotagging with the location coordinates
  • mostly from devices with GPS / Wi-Fi positioning systems
  • 0.77% of all tweets

Location name in user profile
  • input by the users. Fake, joke, wrong name
    • Search API extract only tweets with “correct” location names

Location name in tweet text
  • extracted by Natural Language Processing technique
    • not high accuracy at this moment (less than 50%)
Common method for collecting geo-tweet data continuously (1 of 2)
Caching data in realtime by connecting Streaming API with location filter

Advantage
- collecting realtime data

Disadvantage
- The number of target tweets is relatively small.
- cannot collect past data
Common method for collecting geo-tweet data continuously (2 of 2)
Collecting data by accessing Search API at certain intervals with location and period search filter

**Advantage**
- The number of target tweets is relatively large.

**Disadvantage**
- The search period is limited to the 5 days before the current date.
- Impossible to collect all the tweets in areas where the number of tweets per day is over 1,500
- Search API Limitations:
  - The maximum number of tweets under one search condition: 1,500
  - The minimum search area: $1 \times 1$ km
  - The minimum search period: 1 day
Diffusion of Retweet

2011 the sum of 11 consecutive prime numbers.

Heavy rain warning issued for Tokyo.