DOI: 10.15439/2016F149
ACSIS, Vol. 8. ISSN 2300-5963

Knowledge Gained from Twitter Data

Wiesław Wolny
University of Economics in Katowice
ul. 1 Maja 50, 40-287 Katowice, Poland
Email: wieslaw.wolny@ue.katowice.pl

***Abstract*—Social media constitute a challenging new source of information for intelligence gathering and decision making. Twitter is one of the most popular social media sites and often becomes the primary source of information. Twitter messages are short and well suited for knowledge discovery. Twitter provides both researchers and practitioners a free Application Programming Interface (API) which allows them to gather and analyse large data sets of tweets. Twitter data are not only tweet texts, as Twitter's API provides more information to perform interesting research studies. The paper briefly describes process of data gathering and the main areas of data mining, knowledge discovery and data visualisation from Twitter data.**

I. Introduction

TWITTER is a social networking site directed towards short-form, fast communication. Launched in 2006, Twitter rapidly gained global popularity and has become one of the ten most visited websites in the world. As of March 2016, Twitter boasts 302 million active users who collectively produce 500 million tweets per day, and these numbers are continually growing. These characteristics have made it a primary source of real-time information.

Given this enormous volume of data, analysts have come to recognise Twitter as a virtual treasure trove of information for data mining, social network analysis and information for sensing public opinion trends. For companies, Twitter can be used to build business intelligence tools focused on Twitter data acquisition and processing. It can be used in many ways to collect raw Twitter data and to transform it into valuable business intelligence data.

Twitter is an exceptional tool for knowledge discovery due to six key features:

- Twitter's API is clean and well-documented, with rich developer tools.
- Twitter data are rich in information and have a data format that is convenient for analysis
- Twitter's terms of use for data are relatively liberal. Tweets are public and reachable to anyone. Twitter is based on the asymmetric following model that allows access to any account without request for approval.

However, analysis of Twitter messages (tweets) is regarded as a challenging problem due to some difficulties:

- large amounds of data that cannot be easily handled.
- tweets are short,
- over 40% [1] of tweets are of an informal type of discourse that does not cover any functional topics,

Despite these difficulties discovering knowledge can be simple and bring significant value.

II. Twitter data terminology

The Twitter messages are called tweets. Twitter users post messages that show up in the streams of all of the people who have subscribed to them. Unlike traditional blogs, microblogs are typically limited in the amount of text that can be posted. Twitter's limit is 140 characters.

Tweets often contain links to on-line resources, such as web pages, images, or videos, and more often than not, they refer to other users (called mentions). As is the case with most microblogging, when a message is posted, updates are seen by all users who have chosen to "follow" the author who posted the message (submitter).

In Twitter, all the posts are public. Most people may not receive them if they do not follow the submitter, but messages can be searched for with a keyword or topic and found by someone who is talking about a specific subject.

Twitter has its own conventions that renders it distinct from other textual data. Understanding the language and terminology that is used is important for effective knowledge acquisition.

There are some particular features used in Twitter:

1) Tweet — is a message posted on Twitter, consisting of 140 characters or fewer. It can contain text, photos, links and videos.
2) Twitter name — Twitter usernames appear with an at sign "@" before the name.
3) Hashtags — The # symbol, called a hashtag, is used to mark keywords or topics in a Tweet. It was created organically by Twitter users as a way to categorize messages.
4) Mention — Users of Twitter use the "@" symbol to refer to other users. Referring in this manner automatically alerts them.
5) "Reply" — is used to respond to a tweet. Replying to a tweet is a way of building relationships with followers and friends and joining in conversations.
6) A Retweet — is where one chooses to take a tweet from someone else and tweet it to one's own followers. It can either be done directly with the retweet button or by adding one's own message and including the letters "RT" ahead of the content that is being retweeted.

Twitter names, hashtags and mentions provide an easy way of identifying people and topics, and thus allow to search for and filter information on any subject of interest. Twitter messages have also many unique attributes connected with a tweet which are available with the Twitter API or other tools.

978-83-60810-90-3/$25.00©2016, IEEE

III. Gathering Twitter data

Any social media investigation is only as good as the data used for its analysis. The process of social media analysis involves essentially four steps: data identification, data analysis, data interpretation and, finally, information presentation.

The main problem is how to extract the information that is available on Twitter and how it can be used to draw meaningful insight. To achieve this, first there is a need to build a data analyser for tweets. Tweets are available to researchers and practitioners through public Twitter APIs. Twitter allows developers to collect data via Twitter REST API (https://dev.twitter.com/rest/public/) and the Streaming API (https://dev.twitter.com/streaming/overview).

APIs to access Twitter data can be classified into two types based on their design and access method [4]:

- REST APIs are based on the REST architecture that is now popularly used for designing web APIs. These APIs use the pull strategy for data retrieval, i.e. a user must explicitly request information in order to collect it.
- Streaming APIs provides a continuous stream of public information from Twitter. These APIs use the push strategy for data retrieval. Once a request for information is made, the Streaming APIs provide a continuous stream of updates with no further input from the user.

The response from Twitter APIs is in JavaScript Object Notation (JSON) format. JSON data can be converted to CSV and imported to databases. NoSQL databases, such as MongoDB, allows to store and access JSON data directly, without conversion.

Since gathering of data have particular target, further preprocessing and filtering of collected data can be done using @twitter_names and #hashtags as a search arguments for Twitter API. This method is more precise and provides better results in narrowing data than other text mining approaches.

IV. Storage of data

Explosion in the size of data generated on social media calls for a new data storage paradigm. Commonly used relational databases are insufficiently effective in storing very large datasets. Also the JSON-based data format, used in social media, requires additional conversion to a relational form. At the forefront of this movement is NoSQL, which promises to store large amounts of data in a more accessible way than the traditional, relational model.

There are several NoSQL implementations. One of the most popular ones is MongoDB (https://www.mongodb.org). MongoDB is a proper solution for storing Twitter data due to its adherence to the following principles:

- MongoDB is an open–source document database that provides high performance, high availability and automatic scaling.
- A record in MongoDB is a document, which is a data structure composed of field and value pairs.
- MongoDB documents are similar to JSON objects. This makes it very easy to store raw documents from Twitter's APIs.

In addition to these abilities, it also works well in a single-instance environment.

V. Data structure

Twitter APIs, besides basic information such as the tweet text and the author of the tweet, returns data structure contains additional information which can be used to provide further analysis. For each maximum 140 character tweet, API returns a JSON document containing over 160 items of metadata presented as key and value pairs.

Some the most useful keys for knowledge acquisition are:

- 'coordinates', 'geo' and 'place' — determine the location of the tweet's author;
- 'lang' — allows to easily specify the language of the tweet without text analysis;
- 'created_at' — the date and time of the tweet
- 'entities' — "symbols", "hashtags", "user_mentions" and "urls" included in the tweet;
- 'retweeted_status' — information about retweets;
- 'source' — application or website uses to create a tweet;
- 'text" — text of the tweet;
- "id" — unique identification of the tweet;
- "user" — contains 38 fields about the user. The most useful may be:
  - 'name' — user name;
  - 'time_zone' — time zone of computer or mobile;
  - 'created_at' — date and time of account creation;
  - 'description' — additional information about the user;
  - 'friends_count' — number of friends;
  - 'followers_count' — number of followers;
  - 'location' — actual location.

The above-mentioned keys can make it easier to analyse the text data, but also to perform various specific analyses of users, users nets, time and geolocation information.

VI. Methods of Twitter Data Analysis

Analysing Twitter data is searching through massive amounts of unstructured data. Filtering the data by Twitter names, topic or hashtags may reduce the data size, but it can still be enormous. Also, most Tweets contain no useful information.

Many different types of analysis can be performed with obtained Twitter data. The first can be simple text mining of posts, yet information provided within a tweet's text allows to conduct more Twitter-specific analysis, e.g. user information, connections between users, and localisation at the level of country even any place on the map of the world.

A. Text Mining

Text Mining refers to the process of deriving information from a text. A typical approach in Tweeter analysis is a document-level approach to scan a set of tweets written in a natural language and either to model the document set for predictive classification purposes or to populate a database or search index with the information that is extracted.

Most NLP-based text mining methods perform without particular success in social media. The informal and specialised language that is used in tweets as well as the nature of the microblogging domain make Twitter text mining analysis a very different task. Almost all forms of social media are very noisy and full of all kinds of spelling, grammatical, and punctuation errors. Text mining of tweets can be easier because Twitter API provides information about the language that is used, hashtags and usernames, hence there is no need for its detection.

B. Collecting a User's information

On Twitter, users create profiles to describe themselves to other users on Twitter. When a user creates or reconfigures an account, he/she provides some personal information, such as his/her name, username, password, email address, or phone number. The user may also provide with profile information, such as a short biography, location, website, date of birth, or a picture. On the Twitter Services, the name and username are listed publicly, including on the user profile page and in the search results. A user's profile is a rich source of information about him or her.

The Twitter REST API function users show (https://dev.twitter.com/rest/reference/get/users/show) is an easy way to obtain valuable information about a user, including:

- Real name,
- Description - which typically contains additional information about user,
- Entities such as hashtags, links and media, which can point to further sources of data,
- Followers_count,
- Friends_count,
- Location,
- Language.

An even more valuable function is the followers list (https://dev.twitter.com/rest/reference/get/followers/list). As the name suggest, it allows to access a whole list of followers in one query. The returned data may contain the same information for all followers as the function users/show. Using this function for Twitter's most followed users allows to collect information about millions of users without exceeding twitter API limits.

Information about the friends list is provided by function friends list (https://dev.twitter.com/rest/reference/get/friends/list).

Crawling using the above functions can be used to recognise networks of users.

C. Network Information

A Twitter user network refers to connected user accounts based on various types of relatedness. Structured content, in the form of friends and followers, @replies and @mentions, #hashtags and retweets, makes the Twitter user population networked in multiple ways. Each of these features can be considered as a kind of connection that can exist between two Twitter users. Kumar, Morstatter and Liu [4] categorised two main types of networks:

- Information Flow Networks.
- Friend-Follower networks,

A first kind of network shows who was mentioned or replied-to in the users' Tweets. Second kind of network is based on list of friends and followers of user.

Another type of network is one associated with time-bounded events, such as conferences. Many events like conferences now communicate a common "hashtag" or keyword to identify messages related to the event. Hansen, Smith and Shneiderman [5] [6] created EventGraps. EventGraphs help make sense of the collections of connections that form when people follow, reply or mention one another and a keyword.

To analyse and visualise social media network data from Twitter, the most popular software to use is NodeXL, a free and open add-in for Excel 2007/2010/2013. NodeXL is a project from the Social Media Research Foundation (http://www.smrfoundation.org/). NodeXL is a general purpose network analysis application that supports network overview, discovery and exploration.

D. Sentiment Analysis

The nature of microblogs is that people post real-time messages about their opinions on a variety of topics, discuss current issues, and complain and express either positive or negative sentiment for products they use in daily life. Data from these sources can be used in opinion mining and sentiment analysis tasks, e.g. manufacturing companies may be interested in the following questions:

- What do people think about their product (service, company, etc.)?
- How positive (or negative) are people about our product?
- What would people prefer our product to be like?

All of this information can be obtained from social networks. Opinions and related concepts such as sentiments and emotions are the subjects of study of sentiment analysis and opinion mining.

Sentiment analysis is a growing area of the Natural Language Processing. With the growing population of blogs and social networks, sentiment analysis have become a field of interest for many researches. A very broad overview of the existing studies was presented in [7].

E. Geolocated information

Twitter is also characterised by the diversity of its users, in terms of location. Harvesting this geospatial information provides a unique opportunity to gain valuable insight into information flow and social networking within society.

An important aspect of Twitter data is that some data are geotagged, which means that the posting user has attached a GPS coordinates to the tweet when uploading the information. Such information can be particularly important in order to understand where the user is and what he/she is referring to.

Fischer [8] tracked geo-tagged tweets from Twitter's public API for the last three and a half years. He claimed that there are about 10 million public geotagged tweets every day, which is about 120 per second. This is still a small portion of all tweets, and it is often necessary to use all sources of location information to determine the Tweet's location.

VII. Visualisation of Twitter data

Textual information is generated when users publish on Twitter. Analysing Twitter users and conversations is more than tabulating counts and trends — it is about connections and interactions between people. Twitter enables the collective creation and sharing of digital artifacts. The use of these tools inherently creates network data. These networks represent the connections between content creators as they view, reply, annotate or explicitly link to one another's content.

Twitter provides other embedded information, such as location data. All kinds of gathered data can also be analysed in the time dimension. Visualisation techniques can help to efficiently analyse and understand how and why users interact on Twitter. The display data task requires a remarkable collection of tools and skills.

A. Text visualisation

Text visualisations provide visual representations of documents or small corpora with the primary aim of supporting language analysis. The most popular method of text visualisation are tag or word clouds. Tag clouds are a simple but effective way of representing the distribution of words in a document or corpus, such as a tweet. They are widely employed for both casual use and serious analysis [9]. Clouds give greater prominence to words that appear more frequently in the source text. Clouds can be tweaked with different fonts, layouts, and color schemes.

Another method of text visualisation is the word tree, which is a technique that transforms text into a hierarchical representation based on a selected word or phrase [10].

B. Network visualization

By using network analysis, one can visualise complex sets of relationships such as graphs or sociograms of connected symbols and calculate precise measures of the size, shape, and density of the network as a whole and the positions of each element within it [11]. Structured content, in the form of friends and followers, @replies and @mentions, #hashtags and retweets, makes the Twitter user population networked in multiple ways. Each of these features can be considered as a kind of connection that can exist between two Twitter users.

There are at least as many kind of networks as there are features listed here. All networks can be categorised into network of friends, network of followers and information flow networks — retweet propagation. These and many other kinds of networks are identified and described in depth in [12] and [4].

C. Geolocation visualization

Location information is typically used to gain insight into the prominent locations that are discussing an event. Maps are an obvious choice to visualise location information. A basic method of creating map identifying tweet locations is to simply highlight the individual tweet locations. Each tweet is identified by a dot on the map, and such dots are referred to as markers. Another way is drawing circles of size representing an number of tweets aggregated.

The second kind of map is a trends map. A trends map allows to make real-time mapping of Twitter trends on map. These are displayed as hashtags, @mentions or keywords superimposed over a world map. The map of course can be zoomed in for more detail, and trends from various cities can be selected.

Another kind of map is a heat map of tweets. It allows to quickly identify regions of interest or regions with a high density of Twitter users. A heat map of twitter data can be generated using the https://worldmap.harvard.edu/tweetmap/ website.

References

[1] Pearanalitycs, "Twitter study — august 2009," 2009. [Online]. Available: http://pearanalytics.com/wp-content/uploads/2009/08/Twitter-Study-August-2009.pdf

[2] A. Go, R. Bhayani, and L. Huang, "Twitter sentiment classification using distant supervision," *Processing*, pp. 1–6, 2009. [Online]. Available: http://www.stanford.edu/~alecmgo/papers/TwitterDistantSupervision09.pdf

[3] F. Morstatter, J. Pfeffer, H. Liu, and K. Carley, "Is the sample good enough? comparing data from twitter's streaming api with twitter's firehose," in *International AAAI Conference on Weblogs and Social Media*, 2013. [Online]. Available: http://www.aaai.org/ocs/index.php/ICWSM/ICWSM13/paper/view/6071

[4] S. Kumar, F. Morstatter, and H. Liu, *Twitter Data Analytics*. Springer, Aug. 2013.

[5] *44th Hawaii International International Conference on Systems Science (HICSS-44 2011), Proceedings, 4-7 January 2011, Koloa, Kauai, HI, USA*, IEEE Computer Society. IEEE Computer Society, January 5-8 2011. [Online]. Available: http://ieeexplore.ieee.org/xpl/mostRecentIssue.jsp?punumber=5716643

[6] D. L. Hansen, M. A. Smith, and B. Shneiderman, "Eventgraphs: Charting collections of conference connections," in *44th Hawaii International International Conference on Systems Science (HICSS-44 2011), Proceedings, 4-7 January 2011, Koloa, Kauai, HI, USA*, IEEE Computer Society. IEEE Computer Society, January 5-8 2011. doi: 10.1109/HICSS.2011.196. ISBN 978-0-7695-4282-9 pp. 1–10. [Online]. Available: http://dx.doi.org/10.1109/HICSS.2011.196

[7] B. Pang and L. Lee, "Opinion mining and sentiment analysis," *Found. Trends Inf. Retr.*, vol. 2, no. 1-2, pp. 1–135, Jan. 2008. doi: 10.1561/1500000011. [Online]. Available: http://dx.doi.org/10.1561/1500000011

[8] E. Fisher, "Making the most detailed tweet map ever," 03 2014. [Online]. Available: https://www.mapbox.com/blog/twitter-map-every-tweet/

[9] F. B. Viegas, M. Wattenberg, and J. Feinberg, "Participatory visualization with wordle," *Visualization and Computer Graphics, IEEE Transactions on*, vol. 15, no. 6, pp. 1137–1144, 2009.

[10] M. Wattenberg and F. B. Viégas, "The word tree, an interactive visual concordance," *IEEE Transactions on Visualization and Computer Graphics*, vol. 14, no. 6, pp. 1221–1228, Nov. 2008. doi: 10.1109/TVCG.2008.172. [Online]. Available: http://dx.doi.org/10.1109/TVCG.2008.172

[11] M. A. Smith, B. Shneiderman, N. Milic-Frayling, E. M. Rodrigues, V. Barash, C. Dunne, T. Capone, A. Perer, and E. Gleave, "Analyzing (social media) networks with nodexl." in *Proceedings of the 7th Conference on Creativity & Cognition, Berkeley, California, USA, October 26-30, 2009*, J. M. Carroll, Ed. ACM, 2009. ISBN 978-1-60558-713-4 pp. 255–264. [Online]. Available: http://dblp.uni-trier.de/db/conf/candt/candt2009.html#SmithSMRBDCPG09

[12] D. Hansen, B. Shneiderman, and M. A. Smith, *Analyzing Social Media Networks with NodeXL: Insights from a Connected World*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2010. ISBN 0123822297, 9780123822291