TwiScraper: A Collaborative Project to Enhance Twitter Data Collection

Didier Henry

TwiScraper: A Collaborative Project to Enhance Twitter Data Collection.

Didier Henry
didier.henry@univ-antilles.fr
LAMIA, Université des Antilles
Guadeloupe

ABSTRACT

In past years several works have noted that Twitter data are essential in diverse fields and may have a lot of applications. Nevertheless, the API proposed by Twitter sternly restricts access to public data generated by users. These restrictions have the consequences of greatly slowing down the contributions of researchers and of limiting their scope. In this paper we introduce TwiScraper, a collaborative project to enhance Twitter data collection by scraping methods. We present a module allowing user-centered data collection: Twi-FFN.

KEYWORDS

Twitter data collection, Twitter datasets, Twitter network, Twitter users, Web scraping

ACM Reference Format:
Didier Henry. 2021. TwiScraper: A Collaborative Project to Enhance Twitter Data Collection.. In *Proceedings of the Fourteenth ACM International Conference on Web Search and Data Mining (WSDM '21), March 8–12, 2021, Virtual Event, Israel.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3437963.3441716

1 INTRODUCTION

In recent years Twitter has been the subject of numerous studies. For instance, Bollen et al. [1] have analyzed messages shared on Twitter to predict changes in the stock markets. In their approach, Tumasjan et al. [24] have used messages diffused on Twitter to predict the results of a political election. Some studies have focused on detecting real-world events [12, 22]. Twitter has also been widely used in emergency situations such as forest fires [9], floods [25, 26] and hurricanes [14]. Therefore, researchers have observed that messages posted on Twitter may be useful for predicting or detecting events and saving lives.

Nevertheless, anyone can be the source of information and messages posted on Twitter are not systematically controlled. Thus, Twitter frequently sees the appearance of rumors and misinformation. Due to dramatic consequences that fake information may lead on individuals or society, several researchers [5, 8, 20, 30] have focused on fake news and rumor detection. For instance, Henry and Stattner [13] have used messages and users data to quickly and precisely detect a death hoax diffusion.

WSDM '21, March 8–12, 2021, Virtual Event, Israel

ACM ISBN 978-1-4503-8297-7/21/03.
https://doi.org/10.1145/3437963.3441716

Other researchers [19, 28] have worked on information dissemination in Twitter. For example, Ferrara et al.[11] have observed that negative messages do not spread extremely faster than neutral ones. Son et al.[23] have remarked that messages containing hashtags or URL spread faster than other ones. Some researchers[17] have observed that user information may encourage the dissemination of information. Yang et al.[29] have noticed that messages posted by greatly active users spread faster than other users. Several studies [6, 7, 10, 27] have demonstrated that the user network may have an impact on information diffusion.

Therefore, to perform studies on Twitter researchers need to have access to posted messages, user network and user information. However, the Twitter API sternly restricts access to those data, particularly regarding data concerning the user and his network. Effectively, Twitter API allows a limit number of requests by using an authentication token consequently it is impractical to gather large datasets. In addition, Twitter data collection in parallel is not possible. Several researchers [3, 4, 15] have mentioned the limit of their recent work because of Twitter API limitations. Some researchers [2, 18] have proposed tools to improve collect of messages posted on Twitter. Recently, Pratikakis [21] has developed *twAwler* aims to optimize the time to collect messages, users information and users network. However, this tool is based on the Twitter API, therefore the data collection time is still very important.

This is the reason why we found innovative and interesting to introduce in this paper *TwiScraper* a collaborative project to enhance Twitter data collection by scraping methods. Nowadays, scraping is widely used, especially by start-ups, to aggregate a large amount of content in a short time. From a practical point of view, the scraping of public data consists in moving on a site without creating an account and without having registered and consequently without having accepted the terms of use of the site in question. Similarly to tools or methods [2, 18] proposed in the literature, we collect public data contained in HTML pages of Twitter.

The rest of the article is organized as follows: Section 2 introduces some Twitter API limitations. Section 3 presents a module of TwiScraper allowing users network extraction: Twi-FFN. Section 4 discusses the legal aspect of the use of scraping methods for Twitter data collection. Finally, section 5 concludes this paper.

2 TWITTER API LIMITATIONS

Twitter launched in 2006 in the United States is today one of the main social media with around 300 million active users and nearly 500 million messages posted every day in different languages around the world. Twitter users post messages called "tweets" limited to 280 characters (since November 2017) freely and are connected to each other by a one-way link. Twitter (unlike most social

media) allows access to metadata concerning public posted messages and concerning the user and his network if the user account is public. Indeed, the social media Twitter offers an API[1] allowing access to certain data after an authentication token. This section presents some services of the API and their limitations.

2.1 Streaming tweets

This service allows to collect tweets in real time by asking a query (with parameters such as language, location, etc.) based on keywords. Access to this service is limited to 180 requests per 15-minute window. Thus, as soon as the number of tweets related keywords reaches a certain threshold, this service does not allow continuous access to the new tweets published during a moment. In addition, due to the limitation of the number of requests it seems impossible to collect tweets on several subjects at the same time.

2.2 Search tweets

This service makes it possible to extract by asking a query (with parameters such as language, location, date of tweet, etc.) of tweets present in the Twitter database (going back a maximum of 7 days) depending keywords. Nevertheless, access to this service is limited to 180 requests per 15-minute window, with a maximum of 100 tweets returned per request. In other words, this service enables to perform only 540 requests in the same hour because it is necessary to consider the time it takes for requests to return results. Thus, this service permits to collect until 54 000 tweets in one hour, it is not much for popular topics. In addition, this method is impractical to collect tweets related several topics at the same time. Furthermore, the limitation of tweets published in the past 7 days is very handicapping for the retrospective analysis of the evolution of information dissemination on Twitter.

2.3 Tweet timelines

This service allows to collect user past tweets. However, access to this service is limited to 900 requests per 15-minute window up to a maximum of 200 tweets per distinct request. Furthermore, this method can only return up to 3 200 of a user most recent tweets. Thus, it is not possible to retrieve all of the tweets from a user who posted more than 3 200 tweets.

2.4 User information

This service provides access to a variety of information related to users profile such as description, number of followers / followings, number of tweets posted, date of creation of the profile, etc. Access to this service is limited to 900 requests per 15-minute window. In other words, this service allows to collect only 2 700 users information in one hour because requests take a few minutes. This is insufficient for large datasets. Indeed, considering that we collected 100 000 tweets in two hours with the search tweets service and that each user posted a single tweet, it would take about 37 hours to collect information on these users.

[1] https://developer.twitter.com/en/docs

2.5 Followings and followers

This service is used to retrieve the list of identifiers of a user followers/followings, up to a limit of 200 identifiers per request. In addition, this service is limited to 15 requests per 15-minute window. This service allowing to recover the network of the users is extremely restricted. Indeed, if the targeted users have about 12 000 followers, this service makes it possible to retrieve the list of followers of only one user in one hour. Same things for user followings. So in this case, to extract the complete network of one user would take more than one hour. Therefore, using this method, the time required to recover the network of hundreds of thousands of users would be counted in weeks or even months.

2.6 Consequences on researchers activities

The Twitter API allows free but very limited access to different types of data linked to public messages broadcast or to users of the platform. The time restrictions of the various API services make it impractical to gather large datasets in a short time. In addition, certain limitations prevent the collection of certain data as a whole. We argue that these restrictions have the consequences of greatly slowing down the contributions of researchers and of limiting their scope. Indeed, on account of time limits of Twitter API sometimes researchers focus on small data samples [8] or on biased data. For instance, most recent studies [16] on rumors detection focus only on tweets published in English and do not consider those disseminated in other languages which should be useful to improve the accuracy of detection. Thus, despite the important time required to collect the data and then analyze it certain works are not presented in conferences because the results of studies are too specific and can not be generalized.

3 TWISCRAPER

In order to conduct studies on Twitter less costly in terms of data collection time and also more faithful to reality, we suggest that the community work together on the TwiScraper project. Thus, through various methods of scraping Twitter pages and by offering regular updates of the various modules set up, researchers will be able to carry out their work on Twitter more efficiently and could find new research partners.

The final objective of the TwiScraper project is to offer the scientific community a set of modules allowing with less time constraint:

- the collection of messages posted on Twitter according to several parameters such as keywords, location, accounts cited, etc.,
- the collection of information concerning users such as their number of followers, tweets, description of their profile, their location, etc.,
- the collection of their network, that is to say their followers list, on the one hand, and their following list on the other.

In this paper, we present the module *Twi-FFN*[2] (**F**ollowers **F**ollowing **N**etworks) which allows the collection of the list of followers and the collection of the list of following of a user by providing his username. *Twi-FFN* was developed in Java, it is based on the Jsoup library and is simply used on the command line.

[2] http://didier-henry.com/home/tools/

Table 1: Comparison of the collection time between the Twitter API and Twi-FFN module.

Number of followers	API Twitter time	Twi-FFN average time
10^1	3 sec	1 sec
10^2	3 sec	5 sec
10^3	20 sec	30 sec
10^4	≈ 45 min	5 min 20 sec
10^5	≈ 8 h 30 min	45 min
10^6	≈ 3 days + 12 h	2 h 45 min

This module uses the HTML pages of Twitter site to collect the network of a user (see Figure 1).

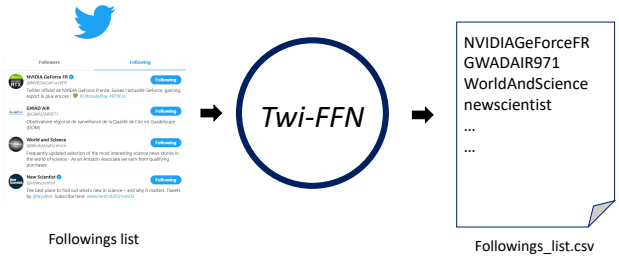

Figure 1: Twi-FFN module data collection process

According to the official Twitter API documentation[3] for collecting the list of followers or following, it is possible to make 15 requests per 15 minutes, each request returns a list of up to 200 users. Therefore, in one hour, the API can collect up to 12,000 followers or 12,000 following. Thus, in one day, for example, it is possible to collect 288,000 followers. This is insufficient because most of the datasets collected by researchers contain hundreds of thousands of messages posted by hundreds of thousands of users. In addition, some users have more than 10,000 or even 100,000 followers, recovering the network of all these users could take several months.

In order to show the relevance of *Twi-FFN* module, we carried out several collections for users with different numbers of followers. For each category of a number of followers, we carried out collection measurements on 100 different users. The results of these experiments are presented in the table 1. Thus, we observe that to collect the list of followers of a user with approximately 1 million followers, it would take almost 3 days and 12 hours with the Twitter API against only approximately 2 hours and 45 minutes with the Twi-FFN module.

In addition, we argue that the Twi-FFN module can be used in a multiprocessing architecture in order to recover the network of several users simultaneously, which is impossible with the Twitter API. This feature of Twi-FFN will improve research on a wide variety of subjects such as the evolution of user networks, link or content recommendation systems or even more realistic models of information dissemination.

[3]https://developer.twitter.com/en/docs/twitter-api/v1/accounts-and-users/follow-search-get-users/api-reference/get-followers-list

4 DISCUSSION

This section introduces recent legal cases or examples which demonstrate that it is completely legal to use Twi-FFN module and participate in the development of TwiScraper.

4.1 Opodo VS Ryanair

The airline Ryanair is scraped by an online flight comparator Opodo but only part of the ticket price information is returned to end customers, making the offer unattractive. In 2010 Ryanair took Opodo to court over this matter. In France, the airline was ultimately rejected on appeal[4]. Among the reasons that led to the court of appeal to this decision, the transformation carried out on the data was considered sufficient to justify a substantial investment on the part of Opodo. In addition, the general conditions invoked as violated by Ryanair are applicable only when purchasing a plane ticket. This scraping activity therefore does not violate these conditions and was found to be legal by the court.

4.2 Resultly VS QVC

QVC is a known online TV retailer. Resultly is a start-up shopping app which proposes a set of items for sale by scraping online retailers websites among which there is QVC. In May 2014, the excessive number of requests of Resultly's bot scraper overburden QVC's servers producing breakdowns. QVC blocked access to Resultly's scraper and sought a preliminary injunction[5] based on the Computer Fraud and Abuse Act (CFAA) which prohibits intentionally causing damage. In March 2015, the court denied the motion for a preliminary injunction[6], finding Resultly, a non-competitor, lacked intent to cause damage to QVC's servers. In addition, the court noted that QVC failed to demonstrate a irreparable prejudice because evidence indicated QVC's capability to defend itself against any future breakdowns provoked by bots.

4.3 HiQ VS LinkedIn

HiQ exclusively collects information from public LinkedIn profiles of employees in a company using web scraping practices. HiQ analyzes this information to determine the profiles of employees for the benefit of their employer, in particular those who could be recruited by competitors. LinkedIn asked HiQ to stop collecting this data, under the pretext of certain laws such as the Computer Fraud and Abuse Act integrated into 18 US §1030[7] and a violation of the conditions of use of the site, which prohibits this type of practice. However, in 2017 the Court[8] did not follow this argument and considered that the data appearing on user profiles is public by being disseminated via the Internet and can therefore be freely collected via web scraping.

[4]https://www.legalis.net/jurisprudences/tribunal-de-grande-instance-de-paris-3eme-chambre-2eme-section-jugement-du-09-avril-2010/
[5]https://fr.scribd.com/doc/249068700/LinkedIn-v-Resultly-LLC-Complaint?secret_password=pEVKDbnvhQL52oKfdrmT
[6]https://www.leagle.com/decision/infdco20150317d82
[7]https://www.law.cornell.edu/uscode/text/18/1030
[8]https://regmedia.co.uk/2017/08/14/hiqlinkedintro.pdf

5 CONCLUSION

In this paper, we have introduced the collaborative project TwiScraper in order to enhance Twitter data collection by scraping methods. We have presented the module Twi-FFN for collecting the network of a user more efficiently than using the official Twitter API. Indeed, this new community tool allows the collection of the network of Twitter users in parallel. In addition, we have shown through examples that the scraping method used by this module is not illegal and therefore the data recovered using this tool can be used legally.

In perspectives, we plan to collaboratively develop other data extraction modules concerning user messages and information. With the help of the scientific community, regular updates of all of TwiScraper modules will help to energize future work on Twitter.

REFERENCES

[1] Johan Bollen, Huina Mao, and Xiaojun Zeng. 2011. Twitter mood predicts the stock market. *Journal of Computational Science* 2, 1 (2011), 1–8.

[2] Matko Bošnjak, Eduardo Oliveira, José Martins, Eduarda Mendes Rodrigues, and Luís Sarmento. 2012. Twitterecho: a distributed focused crawler to support open research with twitter data. In *Proceedings of the 21st International Conference on World Wide Web*. 1233–1240.

[3] Alina Campan, Tobel Atnafu, Traian Marius Truta, and Joseph Nolan. 2018. Is Data Collection through Twitter Streaming API Useful for Academic Research?. In *2018 IEEE International Conference on Big Data (Big Data)*. IEEE, 3638–3643.

[4] Emmanouil Chaniotakis and Constantinos Antoniou. 2015. Use of geotagged social media in urban settings: Empirical evidence on its potential from twitter. In *2015 IEEE 18th International Conference on Intelligent Transportation Systems*. IEEE, 214–219.

[5] Yi-Chin Chen, Zhao-Yang Liu, and Hung-Yu Kao. 2017. IKM at SemEval-2017 Task 8: Convolutional neural networks for stance detection and rumor verification. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*. 465–469.

[6] Justin Cheng, Lada Adamic, P Alex Dow, Jon Michael Kleinberg, and Jure Leskovec. 2014. Can cascades be predicted?. In *Proceedings of the 23rd international conference on World wide web*. ACM, 925–936.

[7] Justin Cheng, Lada A Adamic, Jon M Kleinberg, and Jure Leskovec. 2016. Do Cascades Recur?. In *Proceedings of the 25th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 671–681.

[8] Raveena Dayani, Nikita Chhabra, Taruna Kadian, and Rishabh Kaushal. 2015. Rumor detection in twitter: An analysis in retrospect. In *2015 IEEE International Conference on Advanced Networks and Telecommuncations Systems (ANTS)*. IEEE, 1–3.

[9] Bertrand De Longueville, Robin S Smith, and Gianluca Luraschi. 2009. Omg, from here, i can see the flames!: a use case of mining location based social networks to acquire spatio-temporal data on forest fires. In *Proceedings of the 2009 international workshop on location based social networks*. ACM, 73–80.

[10] P Alex Dow, Lada A Adamic, and Adrien Friggeri. 2013. The Anatomy of Large Facebook Cascades.. In *ICWSM*.

[11] Emilio Ferrara and Zeyao Yang. 2015. Quantifying the effect of sentiment on information diffusion in social media. *PeerJ Computer Science* 1 (2015), e26.

[12] Janaína Gomide, Adriano Veloso, Wagner Meira Jr, Virgílio Almeida, Fabrício Benevenuto, Fernanda Ferraz, and Mauro Teixeira. 2011. Dengue surveillance based on a computational model of spatio-temporal locality of Twitter. In *Proceedings of the 3rd International Web Science Conference*. ACM, 3.

[13] Didier Henry and Erick Stattner. 2019. Predictive Models for Early Detection of Hoax Spread in Twitter. In *2019 International Conference on Data Mining Workshops (ICDMW)*. IEEE, 61–64.

[14] Amanda Lee Hughes and Leysia Palen. 2009. Twitter adoption and use in mass convergence and emergency events. *International Journal of Emergency Management* 6, 3-4 (2009), 248–260.

[15] Mücahit Kantepe and Murat Can Ganiz. 2017. Preprocessing framework for twitter bot detection. In *2017 International Conference on Computer Science and Engineering (UBMK)*. IEEE, 630–634.

[16] Quanzhi Li, Qiong Zhang, Luo Si, and Yingchi Liu. 2019. Rumor Detection on Social Media: Datasets, Methods and Opportunities. *arXiv preprint arXiv:1911.07199* (2019).

[17] Zhunchen Luo, Miles Osborne, Jintao Tang, and Ting Wang. 2013. Who will retweet me?: finding retweeters in twitter. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*. ACM, 869–872.

[18] Richard McCreadie, Ian Soboroff, Jimmy Lin, Craig Macdonald, Iadh Ounis, and Dean McCullough. 2012. On building a reusable Twitter corpus. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*. 1113–1114.

[19] Nasir Naveed, Thomas Gottron, Jérôme Kunegis, and Arifah Che Alhadi. 2011. Bad news travel fast: A content-based analysis of interestingness on twitter. In *Proceedings of the 3rd International Web Science Conference*. ACM, 8.

[20] Lahari Poddar, Wynne Hsu, Mong Li Lee, and Shruti Subramaniyam. 2018. Predicting Stances in Twitter Conversations for Detecting Veracity of Rumors: A Neural Approach. In *2018 IEEE 30th International Conference on Tools with Artificial Intelligence (ICTAI)*. IEEE, 65–72.

[21] Polyvios Pratikakis. 2018. twAwler: A lightweight twitter crawler. *arXiv preprint arXiv:1804.07748* (2018).

[22] Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. 2010. Earthquake shakes Twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web*. ACM, 851–860.

[23] Insoo Son, Dongwon Lee, and Youngkyu Kim. 2013. Understanding the Effect of Message Content and User Identity on Information Diffusion in Online Social Networks.. In *PACIS*. 8.

[24] Andranik Tumasjan, Timm Oliver Sprenger, Philipp G Sandner, and Isabell M Welpe. 2010. Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment. *ICWSM* 10 (2010), 178–185.

[25] Sarah Vieweg. 2010. Microblogged contributions to the emergency arena: Discovery, interpretation and implications. *Computer Supported Collaborative Work* (2010), 515–516.

[26] Sarah Vieweg, Amanda L Hughes, Kate Starbird, and Leysia Palen. 2010. Microblogging during two natural hazards events: what twitter may contribute to situational awareness. In *Proceedings of the SIGCHI conference on human factors in computing systems*. ACM, 1079–1088.

[27] Lilian Weng, Filippo Menczer, and Yong-Yeol Ahn. 2014. Predicting successful memes using network and community structure. In *Eighth international AAAI conference on weblogs and social media*.

[28] Shaomei Wu, Chenhao Tan, Jon M Kleinberg, and Michael W Macy. 2011. Does Bad News Go Away Faster?. In *ICWSM*. Citeseer.

[29] Jiang Yang and Scott Counts. 2010. Predicting the Speed, Scale, and Range of Information Diffusion in Twitter. *ICWSM* 10 (2010), 355–358.

[30] Zhe Zhao, Paul Resnick, and Qiaozhu Mei. 2015. Enquiring minds: Early detection of rumors in social media from enquiry posts. In *Proceedings of the 24th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 1395–1405.