A Survey on Influence and Information Diffusion in Twitter Using Big Data Analytics

Radia El Bacha(✉) and Thi Thi Zin(✉)

Graduate School of Engineering, University of Miyazaki, Miyazaki, Japan
radia.elbacha@gmail.com, thithi@cc.miyazaki-u.ac.jp

Abstract. By now, even if we are still geographically situated, we're able to reach, connect and know about each other through social networks like never before. Among all popular Social Networks, Twitter is considered as the most open social media platform used by celebrities, politicians, journalists and recently attracted a lot of attention among researcher mainly because of its unique potential to reach this large number of diverse people and for its interesting fast-moving timeline where lots of latent information can be mined such as finding influencers or understanding influence diffusion process. This studies have a significant value to various applications, e.g., understanding customer behavior, predicting flu trends, event detection and more. The purpose of this paper is to investigate the most recent research methods related to this topic and to compare them to each other. Finally, we hope that this summarized literature gives directions to other researchers for future studies on this topic.

Keywords: Influence · Social networks · Information cascade · Twitter
Big data analytics

1 Introduction

With the advent of Web 2.0, user experience in the Web shifted from a monologue-oriented towards a dialogue-oriented environment where the user become able to interact and share information through services that are now inevitable in our daily lives such as Social Network Services (SNS). As such, the growth of generated data has been phenomenal. In 2013, the size of the 'digital universe' reached 4.4 zettabytes and it is estimated to grow up to 44 zettabytes by 2020 [1]. All in a model that is often not structured. But most of all is the velocity of data generation. For instance, every second there are more than 9,100 tweets sent on Twitter [2]. This is what bring as to a new digital era, the era of Big Data.

Currently, one of the hottest topic related to Big Data is typically SNS. The large amount of data exchanged in SNS attracted many researcher, companies and data scientist to deep dive into those data in order to understand information diffusion models and get the hidden value from it.

In this paper, we investigate research results related to information diffusion and influence in SNS, particularly on Twitter since it is a popular and fast growing platform with 330 million active users [2], with public profiles rather than protected ones which make it a good source of data targeted by researchers on this topic.

T. T. Zin and J. C.-W. Lin (Eds.): ICBDL 2018, AISC 744, pp. 39–47, 2019.
https://doi.org/10.1007/978-981-13-0869-7_5

By investigating a Twitter dataset, lots of hidden information can be mined. Many research effort has been put into this topic but overall the related literature can be classified into three approaches: Network Topology based approach, Twitter metrics based approach which investigate on user actions on Twitter or another approach that combine both of them.

This paper is organized as follows. Section 2 describes Twitter Network properties and APIs (Application Programing Interface). In Sect. 3, we describe and compare influence models used in literature. Some applications to other field are presented in Sect. 4, and finally, we finish this work with some conclusions in Sect. 5.

2 Twitter Network

Nowadays, Twitter is one of the main SNS in the world. Launched in 2006 but rapidly become one of the most popular micro-blogging sites where people can exchange small elements of content such as short sentences of 140 characters [3], individual images, URL or video links. Twitter users follow others or are followed but reciprocity is not necessary. It is important to know that being a follower on Twitter means that the user receives all the messages (called tweets) from those the user follows.

2.1 Twitter Metrics

In Twitter they are four actions that a user can make to interact within the network: Tweet, Retweet, Mention and Reply. An interesting thing about twitter is its Markup culture such as the *symbol* '@' followed by a Twitter username which is used to mention a user address in a tweet, it's a link to that Twitter profile and the *symbol* '**#**' followed by a keyword which is called a *hashtag*. It can be placed anywhere in the tweet. Searching for a hashtag on Twitter shows all other tweets marked with that hashtag. Often trending topics are hashtags that become very popular. The notation **RT** or sometimes **R/T** in a tweet message stands for retweet. This means that this is a retweet of someone else tweet.

Twitter user's actions are widely used as metrics or measure to define influential users or influence propagation. The details of these different metrics are summarized in Table 1.

Table 1. Twitter metrics used in research literature

Twitter metrics	Explanation
Tweets (T)	Number of messages posted by a user
Retweets (RT)	Number of times a user's tweet had been republished by other users
Mentions (M)	Number of times a user was mentioned by other users
Replies (R)	Number of replies to another user tweets
Followers (F)	Number of users who follow a user (Indegree)
Time (TI)	Activeness of the user or time difference between his tweets and the first time they get retweeted by another user to evaluate the speed of getting reaction to his tweets in the network

2.2 Twitter API

To compute the metrics above, Twitter data must be collected. For this purpose, Twitter API can be used. Twitter provide two public APIs for developers. First one is called Rest API, it is suitable for searching historic of tweets or reading user profile information. The other one is the streaming API, it gives access to a sample of tweets as they are published in twitter. On average, about 9,100 tweets per second are posted on twitter and developers can crawl a small portion of it (<=1%).

Usually the output rendered by either of the APIs is in JSON (JavaScript Object Notation) format which consists of key-value pairs. A tweet metadata can have over 150 attributes in addition to the text content itself. The complete list of attributes for a tweet can be found on [4].

2.3 Data Collection

Due to privacy concerns, it is quite difficult to find a public Twitter dataset therefore more often researchers collect their own dataset from Twitter APIs.

Considering the limitations of the Twitter Rest API, more often researcher choose to use Twitter streaming API to collect the datasets to test their method. Twitter provides a list of possible request and their associated tokens [5]. For instance, to get the user tweets we need to request 'GET lists/statuses' which has 900 tokens.

Some researchers mentioned using a software system called TweetScope [6–19] in order to collect a dataset. This system can collect tweets streams by connecting to Twitter API.

After testing the proposed method on the collected Twitter dataset, usually Experimental results are compared to a survey conducted by the researchers to verify the accuracy of the results [7, 8].

3 Influence Models in Social Networks

One of the key aspects of information diffusion in social networks is Influence. Therefore, it is considered as one of the most important points that cannot be ignored when reviewing information diffusion research.

3.1 Definition of Influence

Many research was conducted on Twitter network data to investigate influence propagation. But, even if we focus only in one specific SNS, namely Twitter, still we can't find a global definition for influence, because it depends on the used influence metrics which usually vary from one researcher to another.

Some researchers consider a more influential user as someone who has the potential to lead others who are connected with him/her to act in a certain way considering Indegree and 2 other user activities on Twitter: retweets and mentions [8]. In a slightly different way, other researcher [9] suggest that influence is measured by the 'excitation'

a user causes in the network by receiving attention from other users. 'Excitation' was defined as how much interest or attention a user receives on the network and it was estimated considering three mechanisms of interaction: retweets, replies and mentions. In some other research methods, categories of influence were introduced [10] usually determined based on the specific role a user is playing in conversations instead of his static followership network. Most of all, retweet network is the most common method used to determine influence.

3.2 Information Diffusion in Twitter

Besides the influence of a node, it is important to know how information can propagate from a user to another distant user on the network. Offline, this is known as Word of mouth diffusion model, a traditional way in which information propagate from person to person in our daily life. Similarly, the same phenomenon can be seen on SNS, Jansen et al. [11] define this as the electronic word of mouth while Bakshy et al. [12] call this a cascade and suggest that the largest cascades tend to be generated by users who have been influential in the past and who have a large number of followers. For instance, in Fig. 1, let's consider node ***a*** as a Twitter user, suppose that he made a Tweet, all his followers can immediately see his tweet in their feeds among them user ***b***. Then ***b*** also decides to retweet this tweet to let his followers know about it. At t_i, ***c*** a follower of ***b*** will decide to retweet ***b*** then at moment T, ***d*** a follower of ***c*** will also make a retweet. This is how Tweets can propagate from ***a*** to ***d*** even if they are not directly connected in a cascade model. It is a very popular way of information diffusion on SNS to spread ideas, campaign, trends, fashions, etc.

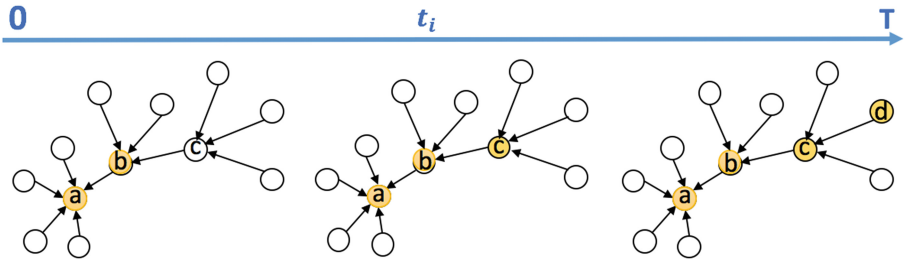

Fig. 1. A graph representing cascade diffusion model in a Twitter network sample

Galuba et al. [13] use this mechanism of information diffusion to predict the spreading of URLs on Twitter. In their study they insist on the importance of Retweets as a strong indication of the direction of information flow. The proposed model takes into consideration the influence of users on each other and time dimension as they introduce a new metric called diffusion delay which is the duration between a user tweet and the first retweet of his tweet.

To the best of our knowledge, most of the research works about Information Diffusion focus on Cascade Models perceived from a producer perspective rather than a consumer perspective.

In fact, a Twitter user can have a dual role, content producer or a consumer of content. A user is producer if he tweet/retweet and a consumer when he sees the content produced by user he is following.

Rotabi et al. [14] proposed a method to measure the effect of cascade on the audience by considering views and engagement with the tweets in a user timeline. Surprisingly, they found that users show a larger engagement for cascades than for non-retweeted content on their timeline. The proposed method was also based on a new Twitter metric which is tweet view. It is determined by calculating if the tweet stayed enough time in the mobile screen of the user.

Table 2 gives a summary of reviewed literature about cascade Models. In the second column, we mark as yes (✓) or no (✗) if the cascades on Twitter are studied from a Producer Perspective (PP) or a Consumer Perspective (CP).

Table 2. Comparison of cascade models in literature

Reference	Perspective		Applications
	PP	CP	
Jansen et al. [11]	✓	✗	Investigate consumer opinions concerning brands through word of mouth diffusion model
Bakshy et al. [12]	✓	✗	Investigate events diffusion on Twitter by tracking Tweets which include URL
Galuba et al. [13]	✓	✗	Predictions of URL mention by users on Twitter
Rotabi et al. [14]	✗	✓	Modeling cascades for audience (Theoretical model)

3.3 User Influence Models in Literature

According to [15] almost half of the existing influence measures are based on the PageRank algorithm [16] or Social Networks Analysis methods [17]. PageRank was created by google founders [18], it is one of the algorithm used by Google to order search engine results. It gives a numerical weight to each web page to represent its importance among all other pages on the web. This weight is based on a link analysis. The weight of a page is defined recursively and depends on the number and incoming link to it. A page that is linked to many pages receives a high rank. Alp and Öğüdücü [7] proposed a Personalized version of PageRank algorithm based on a spread score and other baseline methodologies. Their analysis considers user actions and specific topics to identify topical influencers.

Cha et al. [8] proposed a method for measuring user influence in Twitter. This method investigates the network topology by considering Indegree measure (the number of followers of a user) and 2 user actions: retweets and mentions. They ranked users respectively by Indegree count, retweet count and mention count. Then calculate Spearman's rank correlation coefficient to quantify how a user's rank varies across different measures. The finding of this research suggest that Indegree measure doesn't reveal much about influence of a user and proves that retweet and mentions are better for

indicating influential users. Authors also has investigated the validity of this results over time and across a variety of topics. They found that the most influential users hold significant influence across different topics and need concerned efforts and involvement in a topic to maintain their influence on the network all the time.

In Twitter everything happens so fast, therefore immediacy of results has become very important. Cappelletti and Sastry [9] proposed a method to rank influential Twitter users in real-time for large scale events. Their rank was designed based on the concept of information amplification, which takes as influential those users who have a potential to reach a high audience. Two influence measures were presented, Cumulative influence which is achieved by receiving regular attention from other users and Instantaneous influence which can occur if an influential user is interested in an ordinary user, temporary the ordinary user is considered as influencer. The two measures were weighted by an amplification potential that combine two factors: Buzz and Structural advantage.

$$Buzz = \frac{\#Mentions}{\#Event\ Activity} \tag{1}$$

$$Structural\ Advantage = \frac{\#Followers}{\#followers + \#following} \tag{2}$$

#Event Activity is the number of tweets, retweets and mentions or replies of a user, all related to one event.

Structural Advantage is similar to popularity, it measures whether the surrounding network of the user is capable to provide him information or is seeking from him information, this is represented by Eq. (2).

This method shows a good performance for near real-time ranking of the top 4 or 5 influential users. Moreover, results show that it is much faster compared to PageRank algorithm which proves its quality for near real-time ranking. But, considering the accuracy of the results PageRank still win with a much higher accuracy of the ranking.

Another relevant method but slightly different than the previous ones was proposed by Tinati et al. [10]. It's based only on the Twitter user's dynamic communication behaviors and doesn't include any network topology measure. In this method, a classification model based on the work of Edelmam's topology of influence was applied to a Twitter dataset in order to create a network where the interactions between users determine their role and influence against each other. This model has five categories of users based on their communicator roles – amplifier, curator, idea starter and commentator. All those roles were determined using the retweeting feature.

Table 3 gives a summary of methods used for finding influential users on Twitter. We compare them from many aspects. In the fourth column we mark as yes (✓) or no (✗) if the measures are based on Twitter metrics: tweet (T), retweets (RT), mentions (M), replies (RP), Followers (F), Time (TI). All these metrics are explained in Sect. 2.1.

Table 3. Comparison of influence methods

Reference	Research method	Network topology	User actions						Applications
			T	RT	M	R	F	TI	
Alp et al. [7]	– PageRank – Spread score – Baseline methodologies	✓	✓	✓	✗	✗	✗	✓	Identify topical social influencers in a network
Cha et al. [8]	– Correlation between measures of influence: – Spearman's rank	✓	✗	✓	✓	✗	✓	✗	Investigate if influence hold across different topics
Cappelletti and Sastry [9]	– Amplification Rank or IARank	✓	✗	✓	✓	✓	✓	✗	Influence rank in large Scale events: case study of London Fashion Week and London Olympics
Tinati et al. [10]	– Classification Model based on Edelman's Topology of influence	✗	✗	✓	✗	✗	✗	✓	Evaluation performed using different datasets ranging in Topic, size and geographic context

4 Applications

A number of studies has been conducted for identification of Influential users or spread of influence on Twitter in the context of social network analysis, computer science, and sociology. But recently we can find in literature creative applications of this research for different purposes, such as political sciences, business, health, among many others.

Some studies suggest that analyzing emergence of tweets including content related to flu or epidemics can be a productive way to evaluate discourse surrounding health and promote health awareness. Achrekar et al. [19] simulate twitter users as sensors and their tweets that mention key words related to flu as early indicators to track and predict the emergence and spread of flu or influenza-like illness in a population.

Another research conducted by Piccialli and Jung [20] introduces a case study of customer experience diffusion in twitter. This research demonstrate how information shared by companies is distributed and investigate about the main factors that stimulate the diffusion process more quickly and widely for advertisement or sale promotion.

In the same way but for different field, Chung et al. [21] investigate the use of twitter for health promotion, the goal of this study was to evaluate the content of tweets related to Breast Cancer Awareness Month. Understanding the way twitter is used to fight health issues is very important to improve this process.

Twitter can also have implications in political communication such as promotion of presidential election [22], in economics as well to predict stock market fluctuations [23] or even for detection of real-time events, such as earthquakes [25].

Sakaki et al. [25] took advantage from the popularity of Twitter in Japan as Statistics shows [24], Japan is ranked 2nd country on number of Twitter accounts outside U.S with 25.9 million accounts, to use it for detecting earthquakes in real-time.

5 Conclusion

There is a large literature about the various measures and methods used to find influential users and understand information propagation on Twitter. In this survey, we tried to focus more on recent literature and we were able to conclude that most of the research methods refer to either of user actions in Twitter or the network topology. Mostly the oldest research works on this topic (around 2009) focus on network topology using PageRank algorithm [18] whereas more recent ones focus on metrics based on user actions on the network.

Considering influence measures, we would like to emphasize the use of retweets. From all the metrics presented in Table 1, the metrics that is used in all of the methods was Retweets. It is very useful indeed, it enables us to find the source and the direction of propagation of tweets. Despite of this, research results have proven that time dimension is very important, therefore we recommend to consider time dimension in influence metrics.

Finally, we hope that this survey will provide researchers an overview of the variety of influence measures used recently to identify influential entities on Twitter Network.

References

1. The digital universe of opportunities: rich data and the increasing value of the Internet of Things. https://www.emc.com/leadership/digital-universe/2014iview/executive-summary.htm. Accessed 2 Dec 2018
2. Twitter statistics. https://www.statisticbrain.com/twitter-statistics/. Accessed 2 Dec 2018
3. Makice, K.: Twitter API: up and running learn how to build applications with the Twitter API, 1st edn. O'Reilly Media, Sebastopol (2009)
4. Tweet data dictionary. https://developer.twitter.com/en/docs/tweets/data-dictionary/overview/tweet-object. Accessed 2 Dec 2018
5. Rate limits. https://developer.twitter.com/en/docs/basics/rate-limits. Accessed 2 Dec 2018
6. Trung, D.N., Jung, J.: Sentiment analysis based on fuzzy propagation in online social networks: a case study on TweetScope. Comput. Sci. Inf. Syst. **11**(1), 215–228 (2014)
7. Alp, Z.Z., Öğüdücü, S.G.: Topical influencers on twitter based on user behavior and network topology. Knowl. Based Syst. **141**, 211–221 (2018)
8. Cha, M., Haddadi, H., Benevenuto, F., Gummadi, P.K.: Measuring user influence in twitter: the million follower fallacy. In: ICWSM 2010, pp. 10–17 (2010)
9. Cappelletti, R., Sastry, N.: IARank: ranking users on twitter in near real-time, based on their information amplification potential. In: International Conference on Social Informatics 2012, Lausanne, pp. 70–77 (2012)
10. Tinati, R., Carr, L., Hall, W., Bentwood, J.: Identifying communicator roles in Twitter. In: Proceedings of the 21st International Conference on World Wide Web, pp. 1161–1168. ACM, New York (2012)

11. Jansen, B.J., Zhang, M., Sobel, K., Chowdury, A.: Twitter power: tweets as electronic word of mouth. JASIST **60**, 2169–2188 (2009)
12. Bakshy, E., Hofman, J.M., Mason, W., Watts, D.J.: Everyone's an influencer: quantifying influence on Twitter. In: Proceedings of the 4th ACM International Conference on Web Search and Data Mining (WSDM 2011), pp. 65–74 (2011)
13. Galuba, W., Aberer, K., Chakraborty, D., Despotovic, Z., Kellerer, W.: Outtweeting the twitterers - predicting information cascades in microblogs. In: Proceedings of the 3rd Conference on Online Social Networks (WOSN 2010) (2010)
14. Rotabi, R., Kamath, K., Kleinberg, J., Sharma, A.: Cascades: a view from audience. In: Proceedings of the 26th International Conference on World Wide Web, pp. 587–596 (2017)
15. Riquelme, F., González-Cantergiani, P.: Measuring user influence on Twitter: a survey. Inf. Process. Manage. **52**(5), 949–975 (2016)
16. Li, M., Wang, X., Gao, K., Zhang, S.: A survey on information diffusion in online social networks: models and methods. Information **8**, 118 (2017)
17. Wu, X., Zhang, H., Zhao, X., Li, B., Yang, C.: Mining algorithm of microblogging opinion leaders based on user-behavior network. Appl. Res. Comput. **32**, 2678–2683 (2015)
18. Brin, S., Page, L.: The anatomy of a large-scale hypertextual web search engine. In: Proceedings of the Seventh International Conference on World Wide Web 7 (WWW7), Amsterdam, The Netherlands, pp. 107–117 (1998)
19. Achrekar, H., Gandhe, A., Lazarus, R., Yu, S.-H., Liu, B.: Predicting flu trends using twitter data. In: IEEE Conference on Computer Communications Workshops 2011 (INFOCOM WKSHPS), Shanghai, pp. 702–707 (2011)
20. Piccialli, F., Jung, J.E.: Understanding customer experience diffusion on social networking services by big data analytics. Mobile Netw. Appl. **22**, 605–612 (2017)
21. Chung, J.E.: Retweeting in health promotion: analysis of tweets about breast cancer awareness month. Comput. Hum. Behav. **74**, 112–119 (2017)
22. Kreiss, D.: Seizing the moment: the presidential campaigns' use of Twitter during the 2012 electoral cycle. New Media Soc. **18**, 1473–1490 (2014)
23. Bollen, J., Mao, H., Zeng, X.: Twitter mood predicts the stock market. J. Comput. Sci. **2**(1), 1–8 (2011)
24. Twitter by the numbers: stats, demographics & fun facts. https://www.omnicoreagency.com/twitter-statistics/. Accessed 2 Dec 2018
25. Sakaki, T., Okazaki, M., Matsuo, Y.: Earthquake shakes Twitter users: realtime event detection by social sensors. In: Proceedings of the 19th International Conference on World Wide Web, WWW 2010, pp. 851–860. ACM, New York (2010)