ISSN 1870-4069

Measuring Influence on Twitter Using Text and User Relationships

Carlos Rodríguez, Gabriela Ramírez

Universidad Autónoma Metropolitana Unidad Cuajimalpa
Information Technologies Department, Cuajimalpa, Mexico

crodriguez@correo.cua.uam.mx, gramirez@correo.cua.uam.mx

Abstract. Graph theory concepts as centrality measure can be used to identify users, modelled as nodes of a graph, that have more influence or popularity in a social network. That can be used to classify users. Centrality is one of the most studied concepts in the analysis of Social Networks and there are a great variety of ways to measure it in order to identify the most relevant users in such networks. One of the main issues is how these measures can be calculated in a computationally tractable way and to allow users to be classified as closely as possible to reality. In the literature it can be found many interesting articles that study the application of the aforementioned measures in social networks with millions of users and an enormous amount of messages that flow in those networks. In the present article we are going to combine the information given by the mentioned graph theory measures with text analysis tools to improve the detection of influential users in the **Twitter** Social Network.

Keywords. Centrality measures, text analysis, user influence, social networks.

1 Introduction

Knowing the influence of users and being able to predict it can help to detect viral markets, improve searches, obtain recommendations from experts, more efficiently disseminate information or better manage social relationships with customers of a given company. In this paper we want to study how the influential on **Twitter** users can be detected. Given that the social networks can be modelled as graphs where the nodes represent the users and the edges represent the communication links among them, many graph theory tools become very useful for detecting who are the users that have the biggest audience, who are the users whose message are more cited, who issued the messages that are forwarded the most, etc. Many measures of influence have been presented in the literature ranging from those based on simple methods to those that appeal to complex mathematical models. Measures that record and differentiate between activity, such as popularity, are mentioned in such research works. The first article that we consulted about centrality measures in a network was [4]. In this article the authors studied the algorithmic aspect of calculating the betweenness

centrality measure. Before the publication of [4] the algorithmic complexity of the best algorithm for calculating the betweenness centrality measure known at that time was $\Theta(n^2)$ in time and $\Theta(n^3)$ in space where n represented the number of actors in the network. Motivated by the fast growing of the social networks and the increased time for calculating the centrality measures on such networks, they were interested in calculating them efficiently. So their contribution [4] was to propose an algorithmic complexity improvement in time $O(nm)$ and in space $O(n + m)$ and for the case of weighted and unweighted their time complexity improvement was $O(mn + n^2 \log n)$ where m represented the number of links. With their algorithmic improvement they enlarged the range of networks for which the centrality analysis can be performed in an computationally tractable way. One of the articles about centrality measures applied to the subject of network efficiency that we consulted was [7]. The authors of [7] mention that the idea of structural centrality was applied with the end of characterise human communication in small groups of people and related this concept with the concept of influence in group processes. The authors of [7] introduced the information centrality measure, denoted as C^I in their paper. This measure is applicable in the case to groups and classes as well as in the case of individuals. The authors of [7] make the distinction between the *individual centrality* measure and the centrality based in the number of paths that pass through a node for reaching another node. Because of that they the notion of *information centrality* and related it with the notions of *degree centrality*, *closeness centrality* and *betweenness centrality* of the nodes, denoted as C^D, C^C and C^B respectively. In [2] the authors pointed out that Twitter is not so much a social network where a big number of participants are inactive accounts with low motivation to having dialogues. The authors of [2] say that the majority of the audience consumes and spreads the content published by small set influencer users, called alpha users, in a number of micro-networks. The authors of [2] say that the concept of the strength of weak ties is also applicable to Twitter, what means that the following users who are not part of a personal, strongly social network results in a greater amount of novel information. For this reason it is proposed in [2] a new and simple approach to measuring social networking potential (SNP) that combine content oriented ratio with a dialogue oriented ratio. The research purpose of [2] is to determine a grounded approach for measuring social networking potential of individual Twitter users. In the paper [3] studied the attributes and relative influence of 1.6M Twitter users and tracked 74 million diffusion events that took place on the Twitter follower graph during two month in 2009 and have found that the largest spreading of content tend to be generated by users who have been influential in the past and who have a large number of followers. The authors of [3] conclude that word-of-mouth diffusion can only be harnessed reliably by targeting large numbers of potential influencer. The authors of [3] obtain influencer information on Twitter by crawling the follower graph. In [5] the authors used a large amount of data collected from Twitter, we present an in-depth comparison of three measures of influence: indegree, retweets, and mentions and investigated the

dynamics of user influence across topics and time. In [5] the authors observed that popular users who have high indegree are not necessarily influential in terms of spawning retweets or mentions. They also observed that most influential users can hold significant influence over a variety of topics. They concluded that topological measures such as indegree alone reveals very little about the influence of a user. Recently it was published in [9] a very complete surveys about centrality measures applied directly to the **Twitter** social network. The purpose of [9] is to collect and classify the different measures of influence of **Twitter** mentioning the ones based on the **PageRank** algorithm, those that use the content of the messages, others based on specific topics and others that try to make predictions. Additionally the mention some measures of activity and popularity, some mechanisms for correlating measures and some computational complexity aspects related to this context. The following are frequently used measures based on network topology: degree, closeness,betweenness, eigenvectors and eigenvalues of the adjacency matrix. The user and tweet relations are: user-to-user, user-to-tweet, tweet-to-user and tweet-to-tweet.

Metrics are simple mathematical expressions that provide basic information of a social network in numerical form. In bibliographical reference [9] the metrics involve: number of original tweets, number of *replies*, number of *mentions* and topological features of the network.

1.1 What does it Mean to be an Influential User?

This is a controversial topic because many criteria have been proposed as the ones that are innovative, prestigious, opinion leaders and authoritarian actors. Others associate them to being experts in a topic, opinion leaders, discussers or influencers about the opinions of others, inventors, disseminators, initiators of ideas and connectors. Thus they can be classified by the impact of their activity, diffusion capacity or by the content and authority of their messages. Other relevant users are em celebrities. They are classified by popularity in *broadcasters* and *in passives* (many *followers* and few *in followees*), contacts *acquaintances* # *in followers* $\approx$ # *in followees*) and *evangelists* (few *followers* and many *followees*) as is the case of *spammers* and *bots*. Some authors distinguish between being *in popular*, being *in influential*, *star* or *very read* by taking numerical metrics as the content of tweets. The author of [9] proposes to split the measures into three different types: activity measures, popularity measures (F1 better than F3) and influence measures (RT2 better than F2).

1.2 Activity Measures

The author of [9] consider that the active users are those who participate by sending: tweets, retweets, mentions and replies. For the calculation of the general activity it is proposed in [9] the following formula:

$$General\ Activity(i) = OT1 + RP1 + RT1 + FT1. \quad (1)$$

Table 1. Some important metrics on *Twitter*.

ID	Metric description
OT1	Number of tweets posted by the author
OT2	Number of shared URLs by his OTs
OT3	Number of hashtags included in their OT's
RP1	Number of RPs posted by the author
RP2	Number of tweets answered which conversation has been started by the author
RP3	Number of users that participated in RPs with the author
RT1	Number of RTs made by the author
RT2	Number of OT's posted by the author and retweeted by other users
RT3	Number of users that retweeted tweets of this author
FT1	Number of tweets marqued by others as favorites by the author
FT2	Number of tweets of the author marqued as favorites by other
FT3	Number of users that marcqued the tweets of the author as favorite
M1	Number of mentions by other user from the author
M2	Number of users mentioned by the author
M3	Number of mentions of the author by other users
M4	Number of users mentioning the author
F1	Number of followers
F2	Number of active followers in one subject
F3	Number of *followees*
F4	Number of active *followees* in one subject
F5	Number of followers sending tweets about a subject after the author
F6	Number of followers sending tweets about a subject before the author

1.3 Popularity Measures

A user is considered popular if it is recognized by many other authors on the network. A measure for this purpose is:

$$FollowerRank(i) = \frac{F1}{F1 + F3}, \tag{2}$$

There are variants of this measure such as em Tweeter Follower-Followee which is calculated as:

$$TTF(i) = \frac{F1}{F3}. \tag{3}$$

1.4 Influence Measures

According to the author of the article [9], an influential user is one whose actions in the network are able to affect the actions of other users in the network. Influential users tend to be active but few active users are influential. We can then think of some paradigms of social influence like: massive influence of a very

persuasive little group or connected and accidental influence due to unpredictable factors.

1.5 Influential Users on a Topic

Some authors have been interested in studying influential users in a specific subject. Some traditional centrality measures used to measure influence on Twitter are α centrality and based on the *eigenvectors* using time t as an additional parameter and considering retweets. Another measure of influence proposed is that of *Information Diffusion* which estimates the possible influence of user tweets between *followers* (followed by *followers*). This measure is calculated as follows:

$$ID(i) = \log(F5 + 1) - \log(F6 + 1). \tag{4}$$

Many other user influence measures are mentioned in [9].

1.6 Aplications on the Web

Some application that runs in real time for the study of presidential elections and that has been applied to the detection of influential users in other social networks use : Data-mining, Text-mining, Graph theory based algorithms and Sentimient analysis. There are Web sites like *Klout, PeerIndex, InfluenceTracker, Twitter Grader, Favstar, BehaviorMatrix, Kred* or *Twitalyzer*, among others, to rank the most relevant Twitter users according to their activity, popularity or influence. Most applications measure global influences.

2 Text Analysis on Twitter

User influence on social media as Twitter, among other electronic social medias, has been object of study in sociology, communication, marketing, and political science. This notion is the basis for understanding how businesses operate. This same notion helps to understand how a small group of agents in a social network can change the opinion of the rest of the participants in a social network. If we are able to detect who are members of these small group of agents in a social network, we will be able to detect the opinion leaders, that is to say, those who can polarize the opinion on some topic in a discussion that takes place in a social network for the benefit of an advertising campaign. In this [8] the authors present an empirical analysis on opinion leaders identification problem in social networking medium as Twitter. The proposed approach for opinion leaders identification in [8] is based on the idea that the leadership/influential level of an author can be detected by considering its writing style, and its behavior within the Twitter community. According to this approach the authors of [8] propose several stylistics attributes (lexical richness, language complexity, etc), as well as different behavioral features (post's frequency, directed tweets, etc.), that are

computed directly from users twitter accounts. When they have calculated all these features, they trained a classification model for identifying opinion leaders through machine learning algorithms and automatically identified influential users in a social network. The approach of [8] introduce the use text analysis techniques and behavioral features in order to detect the opinion leaders in a social network as Twitter. This work inspired us to propose a method for detection of opinion leaders that takes into account for this end, elements given by the text analysis in combination with the centrality measures mentioned in the introduction section of the present paper. In the next we will describe our proposed method.

3 Description of our Method

In this paper we propose that the identification of influential users on Twitter should not only be based on the analysis of metrics obtained of the user profile. From our point of view it must also be taken into account for this classification other features related to the style of reflected writing in their tweets and by the way the user interacts within the network with other users. The analysis of the metrics generated by a graph of relation of mentions of the user will give us greater elements to classify a user as influential or non-influential. For this en we will develop a web application, which allows users to be identified, on Twitter using textual attributes and attributes extracted from the graph of relationships between users. First we will implement a graph-based representation for a set related users on Twitter. After that we will Obtain and combine two types of attributes: attributes of the generated graph and shared text see [8] for more details. Then we will evaluate the identification of influential users with the use of a tagged collection. Finally we will build a web application that given a user name help us to determine if it is influential or not influential. For the purpose of this work we are going to consider a node (@userA) as a Twitter user and an edge as the relationship that is generated with another user (@ userB, @ userC, @ userD, etc) at the time @userA or @userB by means of a message or tweet, mention the other. No matter if whether @userA mentions @userB or @userB mentions @userA. In this sense, it is an unguided graph. Let's see the following figure where the graph is illustrated.

The literature identifies the following three types of influence for a user

1. Degree of influence, it refers the total number of followers, that is, the size of the audience.
2. Influence of retuits, that is the number of retuits that the user receives and indicates the amount of content a user generates which is transmitted through Twitter.
3. Mention influence, that measures the number of times a user is mentioned by others, indicating how many times this user initiates an interaction with other users.

In the article [8] it is proposed that the identification of influential users on Twitter should not only determined by these three parameters, but also by

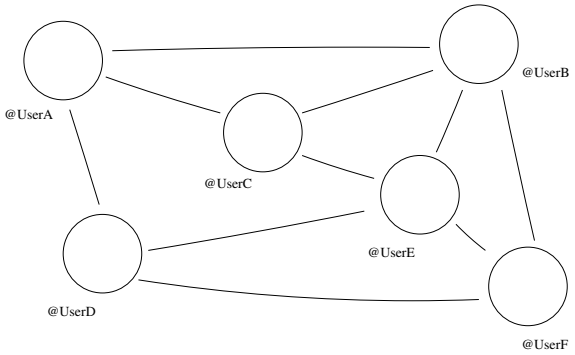

Fig. 1. Example of a non-directed graph on Twitter.

the style of writing and user behavior within the Twitter community as this is relevant to identify influential users. Thus, the influence of a user should not only be based on the analysis of the metrics of the user profile, in addition, the style of and how it interacts with other users makes it possible to identify more objectively those users capable of generating actions in others. We have called all the metrics we can get from a user's analysis of Twitter attributes. These attributes are numbered by the authors of [8] and are divided into two groups.

1. Style attibutes:
   (a) Words by tweet,
   (b) Size of words,
   (c) Length of the username,
   (d) Vocabulary Wealth,
   (e) Hapax,
   (f) Characters by tuit,
   (g) Size of user mentions,
   (h) Size of hashtags.
2. Behavioral Attributes:
   (a) User names in the description,
   (b) Number of hashtags in description,
   (c) URLs used in the description,
   (d) Self-mentions,
   (e) User age,
   (f) Number of tweets,
   (g) Number of followees,
   (h) Number of followers,
   (i) Shared multimedia content,
   (j) Number of favorites,
   (k) Followed by followers,
   (l) Tweets by followers,
   (m) Multimedia content per month,
   (n) URLs used in tweets,
   (o) Number of hashtags in tweets,
   (p) Direct messages,
   (q) Number of retweets,
   (r) Number of favorite tweets,
   (s) Frequency of tweets,
   (t) Standard deviation of the Frencuia by tweet.

To these attributes we add two more attributes *Closeness Centrality* (C_C) and *Betweenness Centrality* (C_B) retrieved from a graph that will be generated by the relationship between users denoted by the mentions that a given user makes other users. In such a way that the users of this graph will be the nodes and the edges will be the representation of the mentions that are between users.

4 Experimentation, Results and Evaluation

In this section we will talk about the obtained results from the training of the classification model and the results of the classification generated by the web application to different users. One of the main approaches of this work is that for the ranking of a Twitter user not only metrics from the profile of users are relevant, we can also use metrics extracted from a graph generated by the relationship between a user and those who are mentioned in their tweets, which will give us greater clarity of how it out the interaction within the network and therefore what is influence within it. The following tests are a result of the models with and without these metrics (C_C and C_B). For the training of the classification model in Weka and for the Web application, we use the Naive Bayes learning algorithm and the 10-fold cross validation test technique, which allows us to reduce the variance in the result. The 10-fold cross-validation consists of taking a test set and dividing it into 10 pieces, starting from one piece the other pieces are used nine to perform the tests, this is done 10 times, one for each one of the pieces and are saving the average of the 10 results. Finally, Weka runs the algorithm for the eleventh time with this data to generate the classification model. The results obtained by the Validation are shown below Crusade of 10 folds compared to the same test set of 250 users that we use for the training of the classification model but without the metrics of the graph.

Table 2. Comparison of 10-fold Cross Validation.

	Tot.Numb. of Instances	Correct Class Instances	Incorrect Class Instances	Rel. abs. error	Prec.	Recall	F-measure
No measures	250	137	113	108.1688	0.656	0.548	0.567
With measures	250	239	11	10.8838%	0.956	0.956	0.956

```
=== Confusion Matrix ===
a      b     a    b  <-- classified as
89    88  | 71    6  |      a = 0
25    48  |  5   68  |      b = 1
Non Graph    Graph
Metrics      Metrics
```

In the table 2 we show three measures evaluated by Weka:

1. **Precision** defined as the fraction of elements that really are classified as positive among all elements whose classification is defined as positive.
2. **Recall** the fraction of elements correctly classified as positive from all elements defined as positive.
3. **F-Measure** is simply the combination of the two previous measures:

$$\frac{2 \times Precision \times Recall}{(Precision + Recall)}.$$

As can be seen, the classification performed with the graph measures is more accurate and less susceptible to errors. On the other hand, the confusion matrix

tells us how they were classified the 250 users, representing the value 0 the non-influent and 1 influent. The table 3 shows a training classification with the same 250 users. The results shown in the table 3 in the precision columns as

Table 3. Training test.

	Tot.Numb. of Instances	Correct Class Instances	Incorrect Class Instances	Average precision	F-measure	Rel. error
No measures	250	117	73	0.501	0.587	100.011%
With measures	250	250	0	1	1	0.006%

absolute relative error indicate the reliability of the classification when the two additional metrics are used. As defined *F-Maesure* also gives us information on the reliability of the model. We can say, from these results that the classification is more precise if we include the two graph measures since as it was possible to observe in the last table, the instances or users classified was higher when graph measures were used than when these metrics were not considered. In a further test, we compared the classification of a user with the classification model. We compare the results obtained with and without the metrics of the graph.

Table 4. Classification of a user.

	Correctly classified of Instances	Precision	Recall	F-measure	Classsified as
No graph measures	1	1	1	1	0
With graph measures	1	1	1	1	0

In this case, there was no difference between the two tests, the user was classified as Non-Influent denoted by a 0. It can be noticed that no difference is detected since only a new set has been classified. This leads us to conclude that although at the moment of classifying a single user there is no difference between using or not the metrics of the graph, at the moment of training the classifier model we can achieve greater precision if we include the graph measures.

5 Program Runs

The implementation of the Web Application consists of the development of modules that were worked sequentially to cover tasks required. The programs and libraries that were used are the following:

1. Php programming language.
2. Phyton programming language.
3. TwitterAPIExchange library for the connection with Twitter.
4. Weka library for the user classification model.
5. Python library for the treatise and analysis of the text of the tweets.
6. MatLab to identify the relationship between nodes.
7. HTML5 for the user interface.
8. D3js for the visualization of the relationship nodes between users.

With all of the above tests, we began classifications with the web application. In this section we show how is the visualization of the classification of users in the web application. Below is the result of the classification to two users, one was randomly chosen: @JandraSoyYo and a the other is well known political national figure: @EPN.

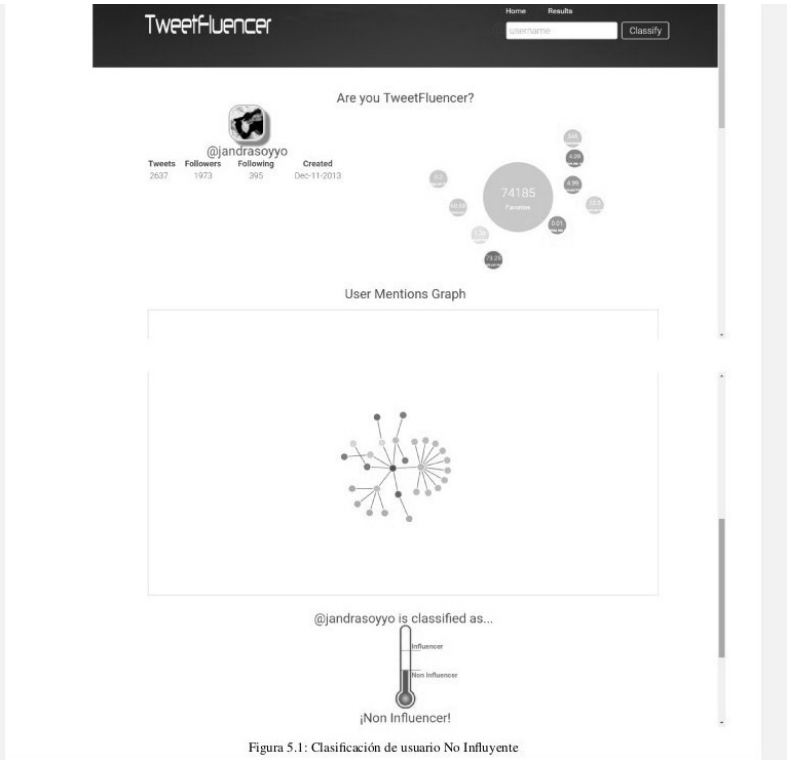

Fig. 2. Twitter influence classification of user @jandrasoyyo as non-influencer.

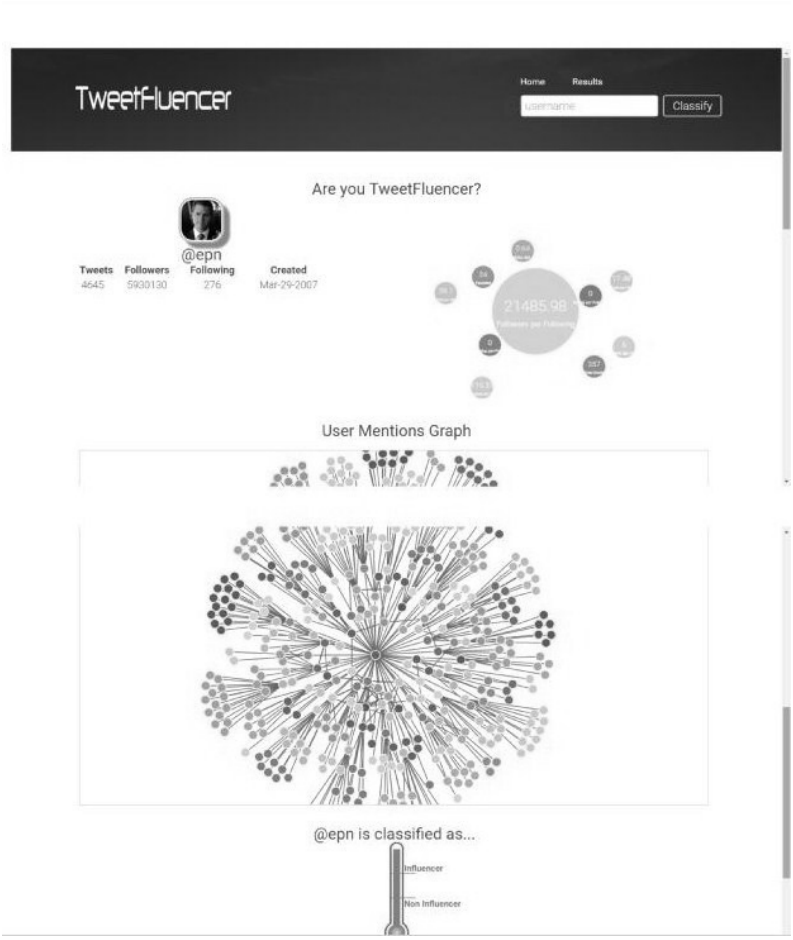

Fig. 3. Twitter influence classification of user @EPN as influencer.

6 Conclusions and Future Work

In the application of classification of influential users in Twitter, managed to classify a user of this social network based on the analysis of the text of your tweets, the relationship you have with other users and the main metrics obtained from your user profile. This gives us a more complete view of how a user is influenced by another within of the social network and the way in which complex relationships are woven among them. In order to carry out this analysis, we used a previously filtered and classified database which contained 2434 twitter users with 600 of their last tweets and information of the profile, from this one trained a classification model which served as base to compare the results obtained from this group with the results obtained from a new user, thus obtaining a

classification based on the model.The classification obtained is not only based on the metrics obtained of a user's public profile, are also based on their writing style and the interaction they have with other users. At the time of training the classifier model we can achieve a lower degree of error if we include the metrics of the graph, therefore we can conclusion that our initial hypothesis where we assume that the analysis of the metrics generated by a relationship graph of user mentions will give us greater elements to classify a user as influential or non-influential is valid. Future work involves the following tasks:

1. Use different algorithms for classification and training model.
2. Add the results of the new classified users to the model of classification.
3. Reduce application process time, code debugging and error handling.
4. Improve graphical representation of the user mention graph, given that alredy the D3js library is robust and flexible, to generate a larger graph depth.

Acknowledgments. We want to acknowledge Carlos Geovany Pérez Velázquez for the programming implementation of the *Tweetfluencer* as part of his final project.

References

1. Aleahmad, A., Karisani, P., Rahgozar, M., Oroumchian, F.: Finding opinion leaders in online social networks Journal of Information Science 42, 1–16 (2015)
2. Anger, I., Kittl, C.:Measuring Influence on Twitter In: i-KNOW '11 Proceedings of the 11th International Conference on Knowledge Management and Knowledge Technologies,ACM N.Y. USA, Graz, Austria(2011)
3. Bakshy,E., Hofman,J.M., Mason,W.A., Watts,D.J.: Everyones an Influencer:Quantifying Influence on Twitter In: WSDM '11 Proceedings of the fourth ACM international conference on Web search and data mining, pp 65–74, ACM N.Y. USA, Hong Kong, China (2011)
4. Brandes, U.: A Faster Algorithm for Betweenness Centrality Journal of mathematical sociology Taylor and Francis 25, 163–177 (2001)
5. Cha, M., Haddadi, H., Benevenuto, F., Gummadi, K.P.: Measuring User Influence in Twitter: The Million Follower Fallacy, In: 4th International AAAI Conference on Weblogs and Social Media, pp 10-18.,Washington D.C. (2010)
6. Cossu, J.V., Labatut, V., Dugue, N.:A Review of Features for the Discrimination of Twitter Users:Application to the Prediction of Offline Influence, Cornell University Library,arXiv:1509.06585v1 [cs.CL] (2015)
7. Latora, V., Marchiori, M.: A measure of centrality based on the network efficiency, Cornell University Library, arXiv:cond-mat/0402050v1 [cond-mat.other] (2004)
8. Ramírez-de-la-Rosa,G.,Villatoro-Tello, E., Jiménez-Salazar, H., Sánchez-Sánchez, C.: Towards Automatic Detection of User Influence in Twitter by Means of Stylistic and Behavioral Features In: MICAI 2014 LNCS, vol. 8856, pp 245–256 Springer, Heildelberg (2014)
9. Riquelme, F.: Measuring user influence on Twitter: A survey, Cornell University Library, arXiv:1508.07951v1 [cs.SI] (2015)