Identification of Influence Propagation Metrics in Social Networks

Niharika Singh
School of Computer Science
UPES
Dehradun, India
niharika1519@gmail.com

Aakash Malik
Department of Electrical Engineering
Delhi Technological University
Delhi, India
Malikakash22@gmail.com

Oshin Maini
Department of Computer Science
Indira Gandhi Institute of Technology
Delhi, India
osh.maini@gmail.com

Gaurav Rajput
Assistant Professor, Department of CSE
G. L. Bajaj Institute of Technology & Management
Greater Noida, Uttar Pradesh, India
gauravrajput31@gmail.com

***Abstract*—Online Social Networks (OSNs also referred as Online Social Media, OSM) are becoming increasingly popular among Internet users. Often, observed that users rely on the massive information available in these mediums to formulate their opinions and perspectives. In other works, we may say that OSNs "influence" users in a variety of ways in the virtual world, which eventually have an impact on real world. It is observed that executives, marketing agencies, celebrities and other famous personalities try to leverage the medium of OSNs to popularize themselves among Internet users. However, in absence of an effective and tangible set of metrics (parameters) for measuring influence propagation in OSNs, these people try to use these OSNs in an unorganized and ad-hoc manner. Therefore, as part of our work in this paper, we have proposed a set of metrics, which effectively measure the extent of influence propagation in OSNs. In this regard, we have identified five metrics namely in-degree, re-tweets, out-degree, mentions and passivity. In addition, we have used these metrics to find the most influential users (referred as "Influential") in the OSNs among various domains like business, sports, politics etc.**

Keywords— Influence Propagation, On Line Social Network, On Line Social Media

I. INTRODUCTION

Online Social Networks (OSNs) and Online Social Media reason to select Twitter as OSNs is that information available over its network is in public domain and accessible through APIs provided by Twitter.

(OSM) are increasingly becoming de-facto platforms for building social networks and social relationships among Internet users. These networks have been growing at a very rapid pace and henceforth, they have become a very popular area of study and research. According to a survey [1], approximately 1.8 billion users worldwide are connected to these OSNs. In addition to building social connections, the massive information available over these OSNs is being used to formulate their opinions and perspectives. In other works, it can be said that OSNs are "influencing" Internet users in a variety of ways in the virtual world which affect real world. Internet users often use these OSNs to connect to new people, spread awareness regarding social issues, for viral marketing, advertising and many more. It is a matter of common observation that businessmen, marketing agencies, celebrities and other famous personalities are trying to leverage these mediums of OSNs to publicize themselves among Internet users and increase their footprint. But this is not as straightforward as it may appear upfront. In order to obtain maximum results, these campaigns and advertisements have to be spread in a targeted and focused manner, and more so using automated means rather than manual. At present, these publicity building exercises are being done in an unorganized and ad-hoc manner without an effective and tangible set of metrics (parameters) for measuring influence propagation in OSNs. Henceforth, there is a need to measure the influence and its propagation over these OSNs in order to facilitate these people in their publicity drives. Also, from a research perspective, we intend to measure and observe influence propagation patterns over these OSNs to better understand their impact on Internet users.

In our work, we have proposed a set of metrics (parameters) to effectively measure influence propagation in OSNs, five such metrics have been identified namely in-degree, re-tweets, out- degree, mentions and passivity and their impact on the influence propagation studied. Thereafter, we have identified few domains like business, sports, politics, etc. and have used these metrics to find most influential users in these domains in order to test the effectiveness of our metrics.

In our work, we have focused on Twitter which is one of the most widely used OSNs, a micro-blogging social networking site with 200+ million active users and more joining on daily basis. Its popularity can be attributed to its feature of micro blogging, i.e., the messages exchanged over Twitter have to be extremely brief (140 characters or less). Twitter differs from other OSNs in few respects. First, Twitter's ability to form what is referred as "weak ties" (second-order connections). These weak ties bring information to users even from those with whom they share less frequently, thereby increasing information exposure. Second, Twitter allows users to self- organize their data. Third, Twitter users actively use Twitter to gather insight, make recommendations, and lodge public complaints. Besides above, one of the main compelling influence with respect to our work is the property of an OSN user to make others in the network listen and accept the views, beliefs and opinions that he professes. One of the most common mistakes while studying Influence is considering it same as "Homophily" [2]

which is a unique property stating that similar people have similar likings, so tend to follow each other which is not the same as Influence. Traditional Communication theory [3] states that minority of the users can be termed as *Influentials* based on their ability to drive other users' opinions. Hence, these Influentials can be targeted for large scale reactions of Influence. In this paper utilizing the facts from Traditional Communication Theory, we have studied the diffusion properties both, its depth and its range. This is done with respect to particular categories. For example, it is very obvious or natural to believe that if two users, A and B, have vast knowledge in the field of sports and business respectively, then A will have more influence on B in the field of sports whereas B will have more influence in the field of business.

We have proposed an algorithm to find the top influential users in particular categories using various metrics.

In degree: Number of followers of a user.
Out degree: Number of users, the user under consideration is following.
Re-tweets: Measured through the number of rewets containing one's name. It is the ability of the user to generate influential content.
Mentions: Identified by searching for *@username* in the tweet content. It gives the name value of the user.
Passivity: The ability of users to never see or ignore the information shared by other users.

So, using the concepts of Influence propagation that we have discussed we can identify the most influential users and then target only them for broadcasting rather than approaching each and every user which is an impossible as well as an impractical approach.

As discussed the Influence of some users may prove to be very useful. But finding Influence itself is a very mind wrecking task. Since, Influence is an abstract quantity, quantifying it using some empirical formula is not as easy as it seems. This is because there is no empirical formula as such. Even if you do, there is a great possibility of error. Also while categorizing the users, we may find that several users belong to more than one category. We make this decision using the formula as mentioned in [4].

II. Related Works

In [9], the authors are comparing influence on the basis of indegree, retweets and mentions. Spearman's rank correlation coefficient has been used to assign rank to every user.

$\rho = 1 - 6(x_i - y_i)^2 / N3 - N$,Here x_i and y_i are the ranks of users based on two different influence measures in a dataset of N users. Pages of top 20 users were analyzed on the basis of each measure. It was found that in degree is not as important as re tweets and mentions for influence propagation. Also, it was discovered that influence is not gained instantaneously. It needs time and effort.

E.Bakshy et Al. in [2] quantified the influence in Twitter by using various metrics such as number of follower of a user, Page Rank and the retweets of the tweets of seed users. They also used the modified Page Rank measure that accounted for a particular topic and finding the ranking of a user depending on the influence measures. They claimed that the diffusion depends not only on the influential users but also on the content carried by the information. They developed a model that predicts influence using an individual's attributes and past activities along with examining the utility of such model for targeting users.

Another work [4] used the Twitter Follower Graph as the dataset to study influence. Here, users have been classified into "elite" users who are different from "ordinary" users in terms of their:

Visibility

Understanding- their role in introducing information into Twitter as well as how information originating from traditional media sources reaches the masses.

They have used Twitter Lists to quantify the influence. The diffusion of influence is observed category wise because a user having high influence in one category may be amongst the least influential users in another category. Whereas the importance of strong ties and weak ties in Information Propagation has been covered in[5]. According to this work, weak ties play a more dominating role in the dissemination of the information online. The empirical relationship between tie strength and diffusion has also been described.

Tang, Jie, et al. in [6] proposed topical affinity propagation to model topic level social influence on large networks the main focus is on measuring the strength of topic level social influence quantitatively. They emphasized that a user's influence on others not only depends on their own topic distribution but also relies on what kind of social relationship they have with others. Finally they proposed two different propagation rules.

Influence quantization is based on the combination of Page Rank which is an estimate authority of the candidate as well as the language model.

This work [7] basically documents the key attributes such as information flow, actor types involved, user participation etc. Each flow is then broken down into sub- flows and studied to identify actor types. Here, Lotan, Gilad, et al. basically followed three steps

- Data collection
- Information flow identification
- Actor type classification

Dataset is obtained using twitter API querying tweets having keywords "#sidibouzid" or "Tunisia." for first and "#sidibouzid" or "Tunisia." for the second. The main agenda her is to concentrate on the flow of communication. Information Flow Identification includes classifying tweets into alike bins, sort them by size, chose top 10% and them chose random 6% of them. Actor Type Classification divides users into MSM, Media, Non-Media, and Bloggers etc. Influence doesn't include just popularity but passivity as well.

Passivity defines the resistivity of user to the tweets posted by other users. This work [8] formulates IP algorithm to find Influence and Passivity. This algorithm finally finds the users that are:

- Most influential
- Most passive
- Least influential with many followers
- Most influential with less followers

III. Proposed Approach

Dataset: The proposed approach uses the twitter dataset.

Metrics: To find the influential users of a particular category, we used five metrics to quantify each other's influence. Each metric has its own contribution to the final calculations of user influence and this contribution is quantified using constants k_1, k_2, k_3, k_4, k_5 which will be determined experimentally and signifies the importance of the metric.

In-degree: It refers to the number of the followers a user has, i.e. the size of audience her tweets are exposed to. We expect higher the number of followers a user has, more is the probability of her tweets influencing others.

Adjusted In-degree of the i^{th} user = $k_1 * I_i$ where I_i is the in-degree of the i^{th} user.

Retweets: They are the tweets posted by some user who is not a seed user. Retweeting is a chain process, the user who originally writes the tweet is called a seed user and this tweet may or may not be posted by her followers. If this tweet is reposted by one of her follower, then this reposted tweet is called as retweet and this chain may keep on growing. The longer the chain, the deeper it will be its propagation and more audience it will face. In this paper, exploring of this one aspect of the retweet metric. The other aspect is the width of the influence of a user found through re-tweeting. Larger will be her influence if more number of her tweets are getting re-tweeted by her followers.

Depth: Number of times a tweet is reposted again and again in a chain / depth of tweet propagation

Width: Number of tweets retweeted / Total number of tweets posted

Adjusted retweet of the i^{th} user = $k_2 *((c_1 *\text{Depth}) + (c_2 *\text{width}))$

$c_1+c_2=1$

where c_1, c_2 are the constants to determine the relative significance of Depth and Width of the retweets.

Outdegree: It constitutes the number of people a user follows. Since, the number of people a user follows is directly proportional to the number of tweets he receives on his home page. So, this is how he gets the domain with which he can spread the influence. One of the major challenges while calculating the out degree is that spammers have high out degree because of the fact that they keep on following people at a high rate so that they come to the notice other people and others may also follow them. So, to compensate for it, we have calculated the mean of the out degrees of all the users under consideration.

$M_0=\sum^{n}_{i=1} o_i$ where M_0 is the mean of the out degree of all the users under considerations, o_i is the outdegree of the i^{th} user and n is the total number of users under consideration. Utilizing this mean value we can find the deviation of the out degree of a particular user from the mean and as this deviation increases, It is clear that lesser will be the influence because if the out degree is less than the mean out degree then domain of the user will be less whereas if the out degree is more than the mean out degree then there are more chances of the user being a spammer which is undesirable.

Normalized out degree=$d_1 * o_{i,,}$

Adjusted Out degree of the i^{th} user = $k_3 * d_1 * o_i$ where d_1 is the proportionality constant.

Mentions: It indicates the ability of a user to involve others in a conversation. It is identified by searching for @username in the tweet contents excluding the retweets. It gives the name value of the user. More is the name value, higher is the influence.

Adjusted mentions of the i^{th} user =$k_4 * m_i$ where m_i is the total number of mentions of the users.

Passivity: It can be defined as the ability of a user to ignore the posts on their homepages. There are many users who neither post many tweets nor retweet others'.We calculate passivity as suggested by the IP Algorithm. If a passive user is influenced, i.e. if a passive user retweets or mentions a username, his influence should be considered as more than the other ordinary users. Now, we will be calculating influence propagated by all the metrics. Influence of each of them is multiplied by a constant lying between 0 and 1. And the total influence is given as the sum of all these influences.

IV. Algorithm

The Twitter Follower Graph is used. This graph is like any other graph has nodes and edges where nodes represents the users and edges are the social links between users. Using this graph, we will calculate the indegree, outdegree, retweet, mentions, passivity of the users. Our approach is to find the top influential users in a particular category. So, for that we have found the important and common categories.

- Business
- Science
- Technology
- Art
- Fashion

To segregate tweets into their respective categories, a simple topic modelling approach is followed. We have used certain popular key words of each category and by measuring the frequencies of these keywords in the tweets we decide the scope of the tweet. For eg, the keywords for business category are stock, export, import, shares, Sensex, Nifty, budget, etc. There is a possibility of a particular tweet belonging to two categories. To resolve such conflicts, we will find the weight of each category in that tweet denoted by w_{cj}.

$$W_{cj} = \frac{number\ of\ the\ words\ belonging\ to\ the\ category\ c}{total\ number\ of\ words\ in\ the\ tweet}$$

where w_{cj} is weight of the category c in the j^{th} tweet of the user. After segregating the tweets into categories we will apply the same algorithm for all the categories.

The algorithm is:

First, we will calculate the value of each metrics using the Twitter Follower Graph. After this, we will plot the graph between nodes on y- axis and on x- axis, we will plot the respective value of the metrics for that particular node / user. We will find number of followers of each user in our refined dataset to know their indegrees. For the depth aspect of our retweet influence, we will use the Twitter-Follower graph as well. We will initialize the graph with zero and for every tweet, add one to each user's node for every user (follower or not) who retweets that tweet. As far as the width angle of retweet influence is concerned, we can find out the number of tweets a user retweets in the dataset we have obtained and refined. Similar to the method of obtaining indegree, outdegrees found as well. Mentions can be found by searching all the tweets for @username for each user. Lastly, to find passivity, we will be using the IP algorithm of [8] papers follows:

$$Pi = \sum vjiIj \text{ where } vji = 1 - wji / \sum (1 - wjk)$$

Then, we will select a particular area (from x_1 to x_2) in the graph where the density of the points of the metrics is high as compared to other areas. Using this dense area of plot, we will find the values of constant k_1, k_2, k_3, k_4, k_5 as

$$K_i = \frac{number\ of\ points\ of\ the\ i^{th}\ matrics\ in\ the\ area}{total\ number\ of\ points\ in\ the\ area}$$

After this we have to find the adjusted values of all the influence metrics for each user then we will add the adjust metrics for each user to find the value for her quantified influence. This quantified influence will help us to find the top influential user to each category.

V. CONCLUSION

In this paper, an algorithm is proposed for social networks metrics. The implementation of the twitter data analysis on the proposed algorithm would be presented in the next paper. The different metrics parameter was discussed and are applied in the efficiency evaluation of the algorithm. These metrics parameters are selected for better efficiency of the algorithm.

REFERENCES

[1] "Statistics and facts about Social Networks", 2014; URL: http://www.statista.com/topics/1164/social-networks/.

[2] Bakshy, Eytan, et al. "Everyone's an influencer: quantifying influence on twitter." Proceedings of the fourth ACM international conference on Web search and data mining. ACM, 2011.

[3] Rogers, E.M. "Diffusion of Innovations", 1962.

[4] Wu, Shaomei, et al. "Who says what to whom on twitter." Proceedings of the 20th international conference on World wide web. ACM, 2011.

[5] Bakshy, Eytan, et al. "The role of social networks in information diffusion."Proceedings of the 21st international conference on World Wide Web. ACM, 2012.

[6] Tang, Jie, et al. "Social influence analysis in large-scale networks."Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2009.

[7] Lotan, Gilad, et al. "The Arab Spring| the revolutions were tweeted: Information flows during the 2011 Tunisian and Egyptian revolutions." International Journal of Communication 5 (2011): 31.

[8] Romero, Daniel M., et al. "Influence and passivity in social media." Machine learning and knowledge discovery in databases. Springer Berlin Heidelberg, 2011. 18-33.

[9] Cha, Meeyoung, et al. "Measuring User Influence in Twitter: The Million Follower Fallacy." ICWSM 10 (2010): 10-17.