An empirical approach towards an efficient "whom to mention?" Twitter app

Soumajit Pramanik, Maximilien Danisch, Qinna Wang, Bivas Mitra

An empirical approach towards an efficient "whom to mention?" Twitter app*

Soumajit Pramanik[1], Maximilien Danisch[2,3], Qinna Wang[2,3], and Bivas Mitra[1]

[1]Department of Computer Science & Engineering, Indian Institute of Technology Kharagpur, India, 721302
[2]Sorbonne Universités, UPMC Univ Paris 06, UMR 7606, LIP6, F-75005 Paris, France.
[3]CNRS, UMR 7606, LIP6, F-75005 Paris, France.

Introduction

Twitter is a social microbloging system which has become one of the most important ways for sharing information online. The basic principles of Twitter are (i) a user can follow other users in order to see the short messages (tweets) these users are posting on their profile, (ii) the user can also retweet these messages in order to make them available to his followers and (iii) the user can mention other users in a tweet so that they will receive a notification of the tweet even though they are not one of his followers. Given these principles, how can a particular user increase the visibility of a given tweet? For this task we suggest to mention users selecting them in a smart way.

We suggest to mention users that are (i) popular, as a retweet of a popular user will increase the visibility of the tweet, (ii) using retweets and active on Twitter at the time the tweet is posted and (iii) susceptible of being interested by the content of the tweet. Several users can be mentioned in a tweet, however this number is limited as the tweet should be less than 140 characters. This problems thus leads to finding the right features evaluating these three components and the trade-off between them. Considering features from the user, the tweet and potential users to be mentioned, we designed a Twitter app that aims at recommending the best set of users to be mentioned. Our app associates a utility score to each user based on its popularity and expected probability to retweet the message and then map the problem into the knapsack problem. The application is available at: `http://bit.ly/1BKZURE`. While it is already in a workable state, the app is improving every time it is used. It is indeed collecting data from these real-time experiments checking whether the mentioned users retweets or not and then improves the utility score. On the fundamental research side, the collected data can be used to make the first model of information propagation in online social networks incorporating mentions [2]. Our work is highly inspired by the works of [5, 4, 3]. However, it is different as it aims at (i) building a usable Twitter application to maximize the visibility of a tweet, (ii) follows an empirical approach towards that goal by collecting data and use them to improve the app.

*This work is supported in part by the French National Research Agency contract CODDDE ANR-13-CORD-0017-01. Research conducted within the context of the Joint targeted Program in Information and Communication Science and Technology- ICST, supported by CNRS, Inria, and DST, under CEFIPRA's umbrella

Empirical study - mention dependency

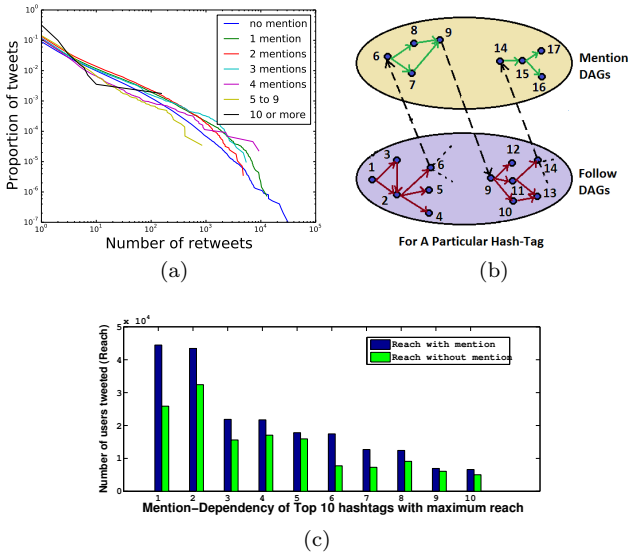
We carried out preliminary studies on a large dataset [1] of tweets in order to evaluate the mention dependency, that is how mention affects the propagation of an information (the causality). We used a dataset of tweets containing hashtags and the underlying follower/friend network. We consider a hashtag as an information. Our methodology relies on a two layer multiplex network representation for each hashtag where the top layer contains directed mention links and the bottom layer contains directed follow links, see Figure 1b. For calculating the mention dependency of a hashtag, we considered only users that tweeted it and tried to evaluate the proportion of these users that would not have tweeted it if there had been no mention links. Figure 1c depicts this mention-dependency of top 10 popular hashtags. It seems that highly popular hashtags are quite heavily dependent on mention links. This result enlightens the potential of using mention to propagate an information and justify our work.

We also tried to evaluate the correlation between mentions and retweets, see the RCDF[2] Figure 1a. We found that generally tweets incorporating many mentions (say 10 or more) have a higher probability of being retweeted few times than tweets having less mentions. However the probability that the tweet leads to a large cascade is much higher for tweets having less mentions and tweets having no mention lead to the largest observed cascades. We also found that the probability that a mentioned user retweets a tweet where he was mentioned was not high (around 4%). Note that the goal of people mentioning users is not always to create cascades, but more to let a user know about a tweet or show to followers that a given user is aware of that information maybe in order to add credibility to the tweet. In that sense, our goal is to twist the common use of mention in order to specifically increase the visibility of a tweet.

Implementation of our Twitter app

In order to obtain a first step usable app, we designed a score for evaluating the utility of mentioning a user u in

[1]http://bit.ly/15f962f
[2]Y is the proportion of tweets retweeted at least X times



(a) (b) (c)

Figure 1: (1a) Some statistics on the use of mention and number of retweets. (1b) Model on mention/follow network. (1c) Popularity with & without Mention-links for top 10 popular hashtags

the message m. This score is the product of three functions that respectively evaluates the Popularity of the user to mention, its Activity and its Interest to the message. The score is given by: $S(u,m) = f_P(u).(f_A(u)^\alpha.f_I(u,m)^\beta)$.

$f_P(u)$ corresponds to the visibility of the tweet if the user u retweets it, while the second term can be seen as a probability that u will retweet m. The powers α and β are controlling the relative importance of the three functions.

We chose the following functions:

- $f_P(u)$ = the number of followers of u.
- $f_A(u)$ = the number of retweets made recently by u.
- $f_I(u,m)$ = the similarity of the message m (plus some keywords that the user can add to help the app) to the last messages of u. We computed it using classical tools from Natural Language Processing.

The actual rate limitations of the Twitter API are the following, within a 15 minutes time window:

- get 5000 followers (resp. friends) of a user: 15 requests,
- get the basic information of a user (particularly its number of followers): 180 requests,
- get the 200 last tweets of a user: 180 requests.

Given these constraints, we built an app that proceeds as follows:

- Crawl the followers and the friends of the user using the app and wishing to tweet the message m.
- Select the friends of the user which are not his followers and among these users select randomly 180 users.
- For each of these 180 users u, crawl its number of followers and set the value of $f_P(u)$, crawl its 200 last tweets and compute $f_A(u)$ and $f_I(u,m)$ out of it.
- Return the set of users that solves the knapsack problem with the following parameters, total budget: 140 minus the number of characters in m, value of the user: $S(u,m)$, weight of the user: the number of characters in his screen name plus 2. This problem is solvable instantaneously via dynamic programing.
- Tweet the message m followed by a newline character and the screen names of the users selected, each one preceded by the character @ and separated by spaces.

After one day the tweets of the users mentioned are crawled to check which users retweeted. The coefficients α and β are then updated taking all available data to improve the utility score. All data are saved in order to select relevant features and adopt a machine learning approach in the future.

Discussion

We presented a Twitter app to suggest users to mention in a tweet in order to maximise the spread of an information. Users that are popular, active on twitter and interested in the content of the tweet are targeted. The problem is mapped to the knapsack problem, the length of the screen name of a user being an important variable. The collected data will be used to improve the app and theory/models of information spread on OSN.

Let us stress that the tweet could also be posted many times by the app changing the associated mentioned users. However this could be annoying for followers that would then see the same tweet several times and may lead to massive unfollow. Posting the tweet just a few times could be an interesting solution, it would allow to mention more users and be acceptable for followers. Being more extreme, we can imagine that since any Twitter user (and any number of them) can possibly be mentioned, followers are not important. Further studies on that point are needed.

Rather than the tweet with the users solving the knapsack problem mentioned, outputting a ranking of users according to their utility score (possibly normalized by the length of their screen name) should also be examined.

Another point is that on twitter there are spammers and *social capitalists* which try to gather the maximum number of followers by using principles such that: "I follow you, follow me!" to trick classical influence measures such as the number of followers or Klout score. We do not take this into account when measuring the influence of a user, thus other measures such that the one of [1] will be considered in future work.

References

[1] Maximilien Danisch, Nicolas Dugué, and Anthony Perez. On the importance of considering social capitalism when measuring influence on twitter.

[2] Adrien Guille, Hakim Hacid, Cécile Favre, and Djamel A Zighed. Information diffusion in online social networks: A survey. *ACM SIGMOD Record*, 42(1):17–28, 2013.

[3] Kyumin Lee, Jalal Mahmud, Jilin Chen, Michelle Zhou, and Jeffrey Nichols. Who will retweet this? In *Proceedings of the 19th international conference on Intelligent User Interfaces*, pages 247–256. ACM, 2014.

[4] Liyang Tang, Zhiwei Ni, Hui Xiong, and Hengshu Zhu. Locating targets through mention in twitter. *World Wide Web*, pages 1–31, 2014.

[5] Beidou Wang, Can Wang, Jiajun Bu, Chun Chen, Wei Vivian Zhang, Deng Cai, and Xiaofei He. Whom to mention: expand the diffusion of tweets. In *Proceedings of the 22nd international conference on World Wide Web*, pages 1331–1340, 2013.