# TweetRank: An adaptation of PageRank algorithm to Twitter world

Joan Puigcerver i Pérez

April 6, 2012

**Abstract**

PageRank is an algorithm presented by Larry Page and Sergey Brin that allows to estimate the importance of a web page using the hyperlinks between them. This algorithm was originally the core of Google's search engine and proved to work very well in the web structure. Our task is to adapt the PageRank to work in the context of Twitter, not the whole web. This paper present the mathematical basis to compute the TweetRank.

## 1 Introduction

Page and Brin defined the importance of a web page as the probability that a person that randomly clicks on links will arrive at that web page. So, PageRank is a probability distribution that represents the likelihood that a person visits some web page by random clicks. This assumption works really well on the World Wide Web since the navigation is driven by hyperlinks among web pages.

Formally, the PageRank is the largest real eigenvector of the matrix $G$ defined as it follows:

$$G = \alpha L + (1 - \alpha)R \tag{1}$$

Each element $G_{i,j}$ represents the probability of visiting the web page $j$ given that the user is in the web page $i$. $G_{i,j}$ is defined as the weighted summation of $L_{i,j}$ and $R_{i,j}$.

$L_{i,j}$ represents the probability of visiting the web page $j$ following one of the links of the web page $i$, and $R_{i,j}$ represents the probability of randomly visiting the web page $j$ from web page $i$.

$R_{i,j}$ is usually set to $\frac{1}{|W|}$, meaning that all the web pages are equally likely to be accessed by a random access. This term is added among other reasons to assert the Perron-Frobenius theorem on the matrix $G$: "a real square matrix with positive entries has a unique largest real eigenvalue and that the corresponding eigenvector has strictly positive components".

This simple definition is not enough in the case of Twitter. PageRank relies on the fact that web pages are visited mainly by following hyperlinks, but this is not true for tweets. Tweets are not usually visited by following the equivalence of hyperlinks (*retweets*), but in many other ways. So, TweetRank adapts and extends the definition of PageRank to consider not only direct links between tweets but other ways to access a tweet.

# 2 Twitter concepts

Twitter is a social network that allows users to send and read short messages called *tweets*. There are many ways of interaction in Twitter: users can *follow* other users and are notified about the posts from these *followed* users, users can *mention* other users on their tweets, they can re-post a tweet from an other user (*retweet*) or they can *reply* to a certain tweet. Given that concepts, we define the following matrices representing the different relationships.

The matrix $RT$ describes the *retweet* relationship among different tweets.

$$RT_{i,j} = \begin{cases} 1 & \text{if tweet } i \text{ is a retweet of tweet } j \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

$RT$ is a $|T| \times |T|$ matrix ($T$ is the set of tweets), where each row has at most one non-zero element (since a tweet $i$ can only be a retweet of a single tweet $j$) and the relationship is asymmetric (if a tweet $i$ is a retweet of $j$, the $j$ is not a retweet of $i$).

The matrix $RP$ describes the *reply* relationship among different tweets, with the same properties of $RT$.

$$RP_{i,j} = \begin{cases} 1 & \text{if tweet } i \text{ is a reply of tweet } j \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

The matrix $MN$ describes the *mention* relationship among tweets

and users. $MN$ is a $|T| \times |U|$, where $U$ is the set of users.

$$MN_{i,j} = \begin{cases} 1 & \text{if tweet } i \text{ mentions user } j \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

Note that $m_i = \sum_{j \in U} MN_{i,j}$ is equal to the number of mentions in the tweet $i$.

The matrix $FW$ describes the *following* relationship among users. $FW$ is a $|U| \times |U|$ matrix.

$$FW_{i,j} = \begin{cases} 1 & \text{if user } i \text{ follows user } j \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

Note that $f_i = \sum_{j \in U} FW_{i,j}$ is equal to the number of users followed by user $i$.

# 3 Algorithm

Our proposed algorithm also defines the relevance of a tweet as the largest real eigenvector of a matrix $G'$, as PageRank did. However, our matrix will take into account not only the direct relations between tweets (retweets and replies) but all the previous concepts.

The idea is that a tweet $j$ can be accessed from a tweet $i$ by:

- Random access, with probability $R_{i,j} = \frac{1}{|T|}$.
- A retweet or reply. $L_{i,j}$ is equal to the probability of access tweet $j$ following a retweet or reply on tweet $i$.
- A user mention. $M_{i,j}$ is equal to the probability of access tweet $j$ by following a mention on tweet $i$.
- A followed user. $F_{i,j}$ is equal to the probability of access tweet $j$ by accessing the owner of tweet $j$ from the following list of the user owner of tweet $i$.

Given that, we define the elements in $G'$ as:

$$G'_{i,j} = \alpha R_{i,j} + \beta L_{i,j} + \gamma M_{i,j} + \delta F_{i,j} \tag{6}$$

The parameters $\alpha$, $\beta$, $\gamma$, $\delta$ must be chosen such that $\alpha+\beta+\gamma+\delta = 1$

**Computing $L_{i,j}$**

$$L_{i,j} = RT_{i,j} + RP_{i,j} \tag{7}$$

It is important to observe that a tweet can be either a retweet, a reply or none of them. So, $RT_{i,j} + RP_{i,j}$ is equal to 0 or 1. An other important detail, already discussed is that there are at most one non-zero element in the $i$-th row of matrix $L$.

### Computing $M_{i,j}$

First, we define the probability of visiting the profile of a user $j$ from a mention in a tweet $i$ ($FM_{i,j}$). We use the matrix $MN$ described in equation 4. Remember that we defined $m_i$ as the number of mentions in tweet $i$.

$$FM_{i,j} = \begin{cases} \frac{MN_{i,j}}{m_i} & \text{if } m_i > 0 \\ 0 & \text{otherwise} \end{cases} \tag{8}$$

Given this probability, we simply distribute it among all the tweets of a user. So, the probability of access tweet $j$ by following a mention on tweet $i$ is defined as:

$$M_{i,j} = \frac{FM_{i,u_j}}{|t \in T : u_t = u_j|} \tag{9}$$

Note that $|t \in T : u_t = u_j|$ is just the number of tweets of user $u_j$. Each row $i$ in the matrix $M$ will have a non-zero element on any column $j$ such that the tweet $i$ mentions the user $u_j$.

This means that the user visited a random user profile mentioned by the tweet $i$, and the visited a random tweet of that user.

### Computing $F_{i,j}$

In the first place, we need to define the probability of visiting the profile of user $j$ accessing from the following list of user $i$. We call this probability $FF_{i,j}$. Remember that we defined $f_i$ as the number of users followed by user $i$.

$$FF_{i,j} = \begin{cases} \frac{FWi,j}{f_i} & \text{if } f_i > 0 \\ 0 & \text{otherwise} \end{cases} \tag{10}$$

Given that definition, the probability of visiting tweet $j$ from tweet $i$ is defined as:

$$F_{i,j} = \frac{FF_{u_i,u_j}}{|t \in T : u_t = u_j|} \tag{11}$$

Note that $|t \in T : u_t = u_j|$ is just the number of tweets of user $u_j$. Each row $i$ in the matrix $F$ will have a non-zero element on any column $j$ such that the user $u_i$ follows the user $u_j$.

This means that the user visited a random user profile followed by the owner of tweet $i$, and the visited a random tweet of that user.

# 4 Probability weighting using *hashtags*

An extension of the previous definition of $G'$ could take into account some measure of similarity between users. Our idea is that users that share common interests would rank higher tweets from users with similar interests. We propose to use the *hashtag* concept on Twitter to measure what a user talks about.

### Hashtag similarity

The matrix $H$ represents the relation among users and hashtags, that is, wich hashtags are used (and how many times) by each user. So, $H$ is defined as it follows.

$$H_{i,j} = \text{Number of times that user } i \text{ uses the hashtag } j \tag{12}$$

Each row $i$ in the matrix $H$ represents a feature vector $\vec{h}_i$ of the user $i$. We can define the similarity between two users $i$ and $j$ in terms of *hashtags* as the cosine of the angle formed by the vectors $\vec{h}_i$ and $\vec{h}_j$.

$$d_{i,j} = \frac{\vec{h}_i \cdot \vec{h}_j}{|\vec{h}_i| \cdot |\vec{h}_j|} \tag{13}$$

### $G'$ extension to $G''$

$G'$ can be also expressed in a terms of a matrix $Z$.

$$G' = \alpha R + \beta L + \gamma M + \delta F = \alpha R + (1-\alpha)Z \tag{14}$$

Where $Z$ is defined as:

$$Z = \beta' L + \gamma' M + \delta' F \tag{15}$$

Where $\beta' = \frac{\beta}{1-\alpha}$, $\gamma' = \frac{\gamma}{1-\alpha}$, $\delta' = \frac{\delta}{1-\alpha}$ and $\beta' + \gamma' + \delta' = 1$.

Given that, we extended the definition of $G'$ and $Z$ to take into account the similarity between two users defined before.

$$Z'_{i,j} = \frac{d_{u_i,u_j} Z_{i,j}}{\sum_{j \in T} d_{u_i,u_j} Z_{i,j}} \tag{16}$$

And finally, the weighted-probability of visiting the tweet $j$ from tweet $i$ would be expressed by the matrix $G''$ defined as:

$$G'' = \alpha R + (1-\alpha)Z' \tag{17}$$

# 5 Summary

We adapted the definition of PageRank to fit in the Twitter context where direct references among tweets (retweets or replies) are not the usual way to navigate through Twitter. We proposed to take into account also mentions and followers to determine the importance of a tweet.

Finally, we considered an extension of the TweetRank which considers the similarity between users in terms of common hashtags. We proposed that references between tweets whose owners have similar interests should be more important than references between tweets whose owners do not have so many common interests. The original idea of a random surfer is not valid any more because the probabilities are weighted depending on the users, but can be a good way to take into account the users profiles.

Our task now is to design an efficient algorithm able to compute the largest real eigenvalue for the matrix $G''$. The algorithm should be able to compute the TweetRank dynamically since Twitter is a dynamic environment where millions of tweets are published each second.