Pattern Recognition Letters 49 (2014) 33–39

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

Feature selection using Principal Component Analysis for massive retweet detection ☆

Mohamed Morchid *, Richard Dufour, Pierre-Michel Bousquet, Georges Linarès, Juan-Manuel Torres-Moreno

Laboratoire Informatique d'Avignon (LIA), University of Avignon, 339, chemin des Meinajaries, Agroparc, BP 91228, 84911 Avignon Cedex 9, France

ARTICLE INFO

Article history:
Received 30 October 2013
Available online 14 June 2014

Keywords:
Massive retweet
Principal Component Analysis
Feature selection
Classification

ABSTRACT

Social networks become a major actor in massive information propagation. In the context of the Twitter platform, its popularity is due in part to the capability of relaying messages (*i.e. tweets*) posted by users. This particular mechanism, called *retweet*, allows users to massively share tweets they consider as potentially interesting for others. In this paper, we propose to study the behavior of tweets that have been massively retweeted in a short period of time. We first analyze specific tweet features through a Principal Component Analysis (PCA) to better understand the behavior of highly forwarded tweets as opposed to those retweeted only a few times. Finally, we propose to automatically detect the massively retweeted messages. The qualitative study is used to select the features allowing the best classification performance. We show that the selection of only the most correlated features, leads to the best classification accuracy (F-measure of 65.7%), with a gain of about 2.4 points in comparison to the use of the complete set of features.

1. Introduction

Twitter is one of the most popular microblogging services which allows real-time traffic, publication, and sharing of user opinions or stories [22]. It also has an impact on economics and politicals [2].

Text messages sent through the Twitter platform, called *tweets*, are composed of up to 140 characters. The power of this microblogging service lies in the fact that users can relay or forward a tweet using the *retweet* paradigm. Thus, information posted by a user can have a very different impact on his environment depending on the number of times it is relayed (*i.e.* retweeted) over a short period of time. For this reason, studying the massive retweet behavior is crucial in information propagation.

Few studies have highlighted the influence of tweet features in the massive information propagation. In this paper, we present a work that aims at detecting the fast and massive spread of short text messages from the Twitter service. We propose to analyze a set of features related to the user environment and to the message content. This study evaluates the impact of each feature in the retweet phenomenon. Finally, two massive retweet classification approaches are compared using the most correlated retweet features chosen from the qualitative study.

The next section enumerates related works about information propagation on social networking services. The experimental protocol for the massive retweet analysis is in Section 3. Then, Section 4 highlights the correlation between tweet features and the massive retweet phenomenon, while Section 5 presents experiments to evaluate the effectiveness of this analysis using two different classification methods. Finally, conclusions and perspectives are given in Section 6.

2. Related work

Although Twitter is a recent information-sharing model, its service mechanisms have been widely studied. In [35,16], the authors analyze Twitter in a general case and study its social impact. Other works have focused on various aspects of Twitter, such as event detection [24,36], user influence [4], sentiment analysis [29], prevention of climate disaster [26], or hashtag analysis [11].

The particular retweet mechanism has been studied in order to predict the number of retweets in a particular range. In [28], the authors proposed to extract a set of features to explain this mechanism. The prediction of potential popular messages was proposed by [34,9].

The behavior of Twitter users is also studied. In [18], the authors analyze the retweet behavior and predict who will retweet a given tweet. Features that influence the retweet probability, such as

☆ This paper has been recommended for acceptance by Qian Xiaoning.

* Corresponding author. Tel.: +33 490 843 577.
E-mail address: mohamed.morchid@univ-avignon.fr (M. Morchid).

http://dx.doi.org/10.1016/j.patrec.2014.05.020
0167-8655/

information freshness, user sending rate or tweet size, have been analyzed by [5]. In [19], the authors analyze the retweet information diffusion while the work by [31] evaluates the impact of the network structure on this phenomenon.

3. Experimental protocol

The study of the massive retweet phenomenon requires a set of relevant features and a large experimental corpus. Table 1 presents the most used features [28,18] for our preliminary qualitative study of the massive retweet mechanism, including tweet content and user features.

A corpus of 6 million tweets was collected using the official Twitter API.[1] The emission dates of the tweets were from April 14th 2006 (d_{start}) to May 13th of 2011 (d_{end}) (lifetime period of 265 weeks which corresponds to about 5 years). Out of this set, 30,903 tweets were retweeted at least once and up to more than 100 times (100+). This tweet set is used to evaluate the correlation between the features described above. Note that, at the time of the corpus extraction, the Twitter service did not allow us to have the exact retweet number for tweets retweeted more than 100 times.

Fig. 1 presents the tweet distribution depending on the number of retweets. We can see that tweets are mostly retweeted either a few times (43% for less than 30) or massively (44% for more than 100 times), while only 13% are retweeted between 30 and 100 times.

A natural intuition about the retweet behavior would be to consider that older tweets would be more likely to be massively retweeted than recent ones. Our assumption is that the age of a tweet has a limited impact on its number of retweets since information in social media has a tendency to be either massively spread in a short time period or ignored [21]. To evaluate the impact of a tweet lifespan in the retweet behavior, let's define γ as the proportion of seldom ($\gamma_{\text{seldom}}(d)$) or massively ($\gamma_{\text{massive}}(d)$) retweeted tweets knowing the broadcasting date ($d_{created}$) of the tweet t, and δ the difference:

$$\delta(d) = |\gamma_{\text{seldom}}(d) - \gamma_{\text{massive}}(d)|, \quad (1)$$

where d is the number of weeks (*i.e.* a lifetime period) between $d_{created}$ and d_{end} ($d_{created} \geqslant d_{end}$). We chose to consider as *seldom* a tweet retweeted less than 30 times and *massive* a tweet retweeted more than 100 times in order to maintain a comparable dataset size. A straightforward study of the differences between seldom and massively retweeted messages shows that the mean of δ is 0.204% and its standard deviation is 0.206%, which confirms our assumption: the age of a tweet does not influence its low or massive spreading. Fig. 2(a) presents the proportion of $\gamma_{\text{seldom}}(d)$ and $\gamma_{\text{massive}}(d)$ tweets for a lifetime of less than 20 weeks before d_{end}, while Fig. 2(b) presents exactly the same curves but with a log scale for a lifetime d greater than 20 weeks. Fig. 2(b) allows us to point out that difference between the proportion of seldom or massively retweeted tweets follows the same curve even if the proportion of this difference (δ) is small (less than 1%). Moreover, the main proportion of tweets in both subsets, is more or less concentrated during the same epoch (see Fig. 2(a) $0 \leqslant d \leqslant 2$).

4. Analysis of massive retweet factors

In this section, we propose to use this tweet dataset to study the correlation between retweet features described in Table 1. A Principal Component Analysis (PCA) is applied to transform these features from an original space representation of correlated variables, into a set of linearly uncorrelated variables called *principal components* (Factors). Carrying out PCA on the features, including retweet feature, allows to sort and group these variables with respect to the total variability and to assess how the retweet variable affects this variability.

[1] http://dev.twitter.com.

Table 1
Description of tweet features for a user **U**.

	Description	Twitter API names
Content features		
Retweet	# of sharing	*retweet_count*
Hashtag	# of topics in a tweet	*hashtags*
Mention	# of cited usernames	*text*
Url	# of contained URLS	*urls*
User features		
Days	# of days **U** created its account	**U**/*created_at*
Favorite	# of favorite tweets by **U**	**U**/*favourites_count*
Follower	# of users who follow **U**	**U**/*followers_count*
Followee	# of friends of **U**	**U**/*friends_count*
Status	# of tweets wrote by **U**	**U**/*statuses_count*

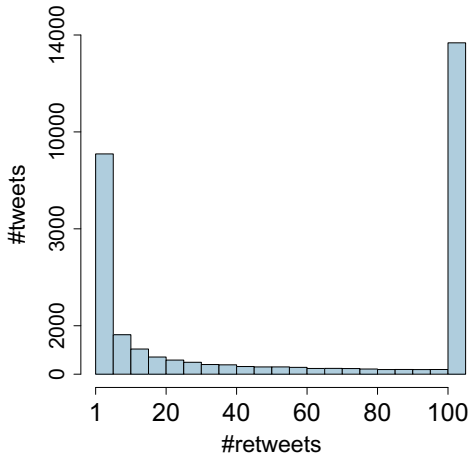

Fig. 1. Tweet distribution depending on the number of retweets.

4.1. Principal Component Analysis

The Principal Component Analysis (PCA) is used by almost all scientific disciplines and is probably the most popular multivariate statistical technique. Introduced for the first time by [3,23], its more recent instantiation was formalized by [10] who also introduced the term *principal component*. The goals of PCA are to:

- extract the most important information from the data table
- compress the size of the dataset by keeping only this important information
- explain and simplify the description of the dataset
- analyze the structure of observations and variables

PCA computes new variables called *principal components* to achieve these goals. These variables are obtained as linear combinations of the original variables. The first principal component is required to have the largest possible variance to "explain" the largest part of the inertia of the dataset. Then, the second component is computed under the constraint of being orthogonal to the first component and to have the largest possible inertia. The other components are computed in the same way. The values of these new variables are called *factor scores* and are interpreted geometrically as the *projections* of the observations onto the principal components.

These are obtained from the Singular Value Decomposition (SVD) of the dataset **X**, with: