Data Mining Project:
Mining tweets to predict their popularity

April 11, 2013

1 Problem Description

Twitter is a microblogging system in which users cand send, reply, or forward (retweet), short messages up to 140 characters between them. Typically, users are inter-connected by messages and they form social network communities in which different topics are discussed. Messages which are retweeted several times are thus more likely to represent important topics that people like to discuss.

In this project, you will have to analyze twitter data and build models for solving one of the two following tasks:

1. to predict if a tweet will be retweeted or not, i.e. to predict if a tweet can be considered as popular or not insight a community.

2. to predict if a set of tweets is novel and form a new topic or not according to the topic distribution of past tweets.

To do so, you will have to develop your own data mining tool and propose an approach to solve the problem.

2 Data Description

Your analysis will be made on a recent collection of tweets, all of which are related to the global topic of shampoo. Each tweet instance is characterized by the following features:

1. the id of the tweet

2. the user id from which the tweet has been created

3. the creation date of the tweet

4. the content of the tweet

5. the reply-id of the related tweet if the current tweet is a response to another tweet

6. the user id to which the tweet is a response

In addition, the content of a tweet can have additional information such as hashtags, which are preceded by a #, and urls. If the tweet is a retweet, then it is indicated either in the beginning of the content by ”RT @user”, or at the end of the content by ”via @user”, where @user is the user name of the original author of the tweet.

3 General Approach

Given a collection of tweets, you will have to extract informative features in order to address one of the two above learning tasks. We provide here a general approach that you can make use to build your tool. A first step is to preprocess the content of each tweet in order to build bag-of-words representation of a tweet with TF-IDF. This includes a stemming process plus a stopwords removal process. Be carefull to not stem neither the hashtags, nor the user and url mentions in a tweet, they will be used as additional features for a tweet. Perl script examples will be provided to do basic stemming and stopwords removal.

For the task of predicting retweets, you will need to build the graph of tweets, i.e. how tweets are retweeted between users. An example of such graph is given in Figure 1. From this graph, you will have to extract relevant informations that will be used as feature for a tweet, for instance the average page rank of the users that have retweeted this tweet. The target attribute will be the number of times a tweet has been retweeted. You can transform the target attribute to a binary attribute for classification by separating popular tweets; those that have been retweeted more than x times, to normal tweets. It is your choice to solve this task as a regression or a classification task. An example to build the graph of tweets will be given. We refer to [1] for an example of analysis on retweets.

For the task of novelty detection, you will need to take a look at the book of [2], in particular chapter 9. You can either approach the task as a one-class supervised learning problem, or an unsupervised learning problem. In the latter case, you will have to build topic clusters on past tweets and measure the global distance of a new tweet with the clusters to see if it is abnormal and potentially novel [3].

The goal of the project is to analyze one of the two given tasks, and to provide a comprehensive approach to solve it. You will have to explain and justify your approach during an oral presentation as well as to give your conclusions on this learning problem.

References

[1] Meeyoung Cha, Hamed Haddadi, Fabricio Benevenuto, and Krishna P Gummadi. Measuring user influence in twitter: The million follower fallacy. In *4th international aaai conference on weblogs and social media (icwsm)*, volume 14, page 8, 2010.

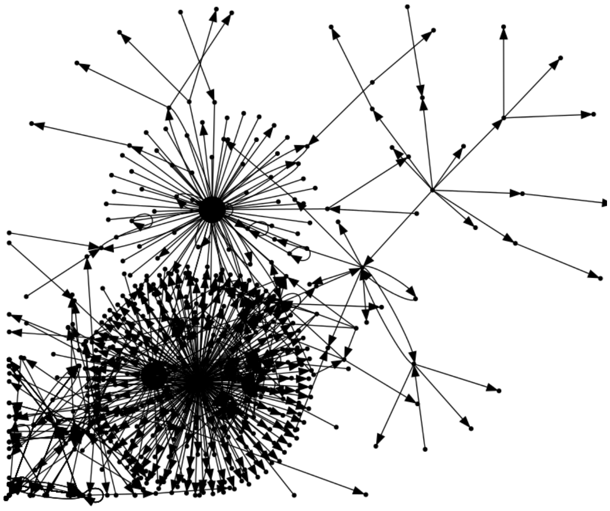

Figure 1: Example of a retweet graph. Nodes are users and edges are retweets.

[2] Joao Gama. *Knowledge discovery from data streams.* Citeseer, 2010.

[3] Eduardo J Spinosa, André Ponce de Leon F de Carvalho, and João Gama. Olindda: A cluster-based approach for detecting novelty and concept drift in data streams. In *Proceedings of the 2007 ACM symposium on Applied computing*, pages 448–452. ACM, 2007.