Massive Data Analytics' Project Proposal

Sandro Cavallari, Marco Giglio, Paolo Morettin

March 23, 2014

1 Introduction

Social networks experienced an exponential growth in the last five to ten years. In a few years they became one of the most used communication media: several people, nowadays, tend to spend many hours per day writing on their *walls*, *twitting* etc and basically every big company, important personality, or club, manages accounts on several social networks, using them as its most important communication media.

Given the increasing role of social networks in everyday's lives, researches became interested to them, questioning how they affect our privacy and behavior [1][2] or examining the role they fulfilled during some important recent events, such as the Arab Spring [3][4].

2 Project description

Our interest is to monitor trends on a social network, understand whether people are feeling positive or negative toward a certain topic and correlate this feeling with recent news coming from newspapers. In detail, our project aims in developing a methodology in order to:

1. understand when the common feeling about a certain topic shifts from positive to negative;
2. correlate this shift to news coming from newspapers and news agencies.

The social network we will focus on is Twitter, an online social network that allows users to upload short text messages (*tweets*) of up to 140 characters.

2.1 Work plan

In order to perform our analysis we need a big dataset containing tweets and another one containing the news published by popular newspapers and news agencies during the same temporal interval. Our team is planning in renting a server having low computational power but high bandwidth and availability. The server will be responsible for collecting both tweets and news 24 hours per day for some weeks. All the information collected by the server will be stored on a database.

In a second step, all the information will be downloaded on faster machines and analysed in order to discover trends and correlations.

2.2 Dataset

As over-mentioned the Dataset for this work will be composed by:

1. Twitt downloaded by a server using twitter4j and saved on a MongoDb
2. News collected by the Rss Feed of the most important newspaper websites

Our plan is to download at least 1 month of data and only after start to analyse the correlations.

References

[1] Debatin B. et al., *"Facebook and Online Privacy: Attitudes, Behaviors, and Unintended Consequences"*, Journal of Computer-Mediated Communication, 15, pg. 83-108 (2009)

[2] Acar A., *"Antecedents and Consequences of Online Social Networking Behavior: The Case of Facebook"*, Journal of Website Promotion Vol. 3, N. 1-2, pg. 62-83 (2008)

[3] Howard P. et al., *"Opening Closed Regimes: What Was the Role of Social Media During the Arab Spring?"*, ICT4D Bibliography (2011)

[4] Lotan G. et al., *"The Revolutions Were Tweeted: Information Flows During the 2011 Tunisian and Egyptian Revolutions"*, International Journal of Communication 5 (2011)