Massive Data Analytics' Project Proposal

Sandro Cavallari, Marco Giglio, Paolo Morettin

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, spend many hours per day writing on their *walls* or *twitting* 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 in 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;
3. extract words or entities that may help a human reader in understanding why people change their feeling.

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.

3 Work plan

A Java program has been provided in order to detect some labeled sentiment shift. Given this program and a dataset of tweets collected in 2009 we'll be able to identify when users on twitter changed their sentiment toward a certain topic.

Then, we'll have to look in news papers and news agency for news about that topic and in that temporal period. We'll need to collect news which has been published both before and after the *shift* in order to be able to detect what had

changed. Unfortunately, many news papers only provide old news for paying users, hence we might be forced in subscribing and paying small fees in order to retrieve all the information we need. Of course some noise may be introduced in this phase due to uncorrelated news, ambiguity or people having the same name.

In the next phase we'll try to detect which news caused the shift and to extract some word or event which might help a human reader in understanding its cause. In order to solve this problem, different type of approach are taken in consideration:

- a simple approaches such as the one depicted in [5] based on word frequency analysis allow to detect the difference in the most used word of the document
- a small variation of the similarity measurements of the N-gram graph proposed in [6] can find difference in documents and highlight interesting events
- a more complex approach based on *event extraction* methodologists proposed in [7] and [8] can produce good performance, but require a high training time

One big difficulty we'll probably have to cope with in this final phase is to synchronize the shift of the tweets with the news: it is not possible to establish from a peek in the sentiments when the news which caused it had been published! Some news have a direct and fast impact, causing a fast response by the social media; others are characterized by a slower penetration and are commented on twitter only days after they have been published; some other events may be discussed on social networks even before they are published!

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)

[5] Albert B. et al., *"Detecting Sentiment Change in Twitter Streaming Data"*, JMLR: Workshop and Conference Proceedings 17 (2011) 5-11

[6] G. Giannakopoulos, *"Automatic Summarization from Multiple Documents"*,Department of Information and Communication Systems Engineering, University of the Aegean 2009

[7] Hristo Tanev, Jakub Piskorski and Martin Atkinson, *"Real-time News Event Extraction for Global Monitoring Systems"*, Joint Research Center of the European Commission Web and Language Technology Group of IPSC 2008

[8] M. Naughton , N. Kushmerick and J. Carthy, *"Event Extraction from Heterogeneous News Sources"*, School of Computer Science and Informatics, University College Dublin 2006