DEMO PAPER: STAMAT: A FRAMEWORK FOR SOCIAL TOPICS AND MEDIA ANALYSIS

Giuseppe Serra, Thomas Alisi, Marco Bertini,
Lamberto Ballan, Alberto Del Bimbo

Università di Firenze - MICC
giuseppe.serra, marco.bertini, lamberto.ballan,
alberto.delbimbo @unifi.it - thomas.alisi@gmail.com

Laurent Walter Goix, Carlo Alberto Licciardi

Telecom Italia - Innovation and Industry Relations
- Research & Prototyping
laurentwalter.goix, carloalberto.licciardi
@telecomitalia.it

ABSTRACT

Analysis of trending topics in social networks, based both on textual and multimedia content, can be used by content providers to measure the buzz around their channels, and even to create new material based on the current memes propagating across Twitter, Google+, Facebook etc.

STAMAT is a framework designed to provide a set of tools for social media analysis, and to create, through content curation, personalized web sites and magazines.

***Index Terms*—** Social media analysis, topic detection, social networks.

1. INTRODUCTION

In recent years there has been an explosion of the usage of social networks that provide microblogging services that let users write short updates containing text and links to multimedia content, like Twitter and Google+. These tools allow people to easily produce content, thanks to the short length of texts that does not require an excessive effort in their creation.

Sharing of URLs/information and news reporting [1] are the main intentions among users. User interactions and participation to discussions regarding trending topics show that services like Twitter may be considered as a vehicle for news [2, 3]. Moreover the social aspect of conversations allows to consider human interactions as a channel that can be exploited to identify situations ranging from epidemics to flash mobs, joining social media with additional heterogeneous sources of data [4].

STAMAT (Social Topics and Media Analysis Tool) is a framework for the analysis of real-time social media, like Twitter, Facebook or Google+, taking advantage both of textual and multimedia content. The goal is to let users create an environment that automatically selects the most relevant news, links and images related to a set of topic or news of interest. STAMAT can be used for personal content curation, i.e. creating a personalized magazine from various news sources, or to let content producers and distributors to understand reactions of social networks to current news, such as feeds published by New York Times or TechCrunch.

2. DEMO

The demo shows the functionality of the system, implemented as a web application analyzing a set of data sources and displaying the outcomes of data processing. Users can select different news channels, assigning them to semantic categories; a pre-processing step starting from RSS feeds leads to web scraping, then HTML pages are analyzed to extract topics, named entities (see Fig. 1) and multimedia content. This information is used to retrieve related content on social media, providing both complementary information for news and a measure of the popularity and influence of news content (see Fig. 2).

A CBIR approach is used to evaluate how media embedded in news sites is propagated across social media; this can be used for different tasks, like selecting images that can be considered as representative of popular topics or evaluating if some media content is becoming viral [5].

3. THE SYSTEM

The system has been developed as a web application: the backend and analysis services have been developed in Java using the Play framework, while user facing web applications have been implemented with Backbone.js, Twitter Bootstrap and CodeIgniter, to provide a snappy user experience.

News items from RSS feeds and URLs are processed to determine their language using n-grams and Naïve Bayes classifiers, since this technique has proved to be effective also on short text fragments like tweets, to perform appropriate stemming and stop-word elimination; their content is then summarized using latent topics extracted with LDA [6, 7] and named entities obtained using a mixed approach that employs a set of rules combined with gazetteers [8] and CRFs [9]. Topics and entities are used to perform an initial selection of relevant tweets that could be associated to news. Tweets are used to provide additional social information related to news elements, and also to re-rank news based on their influence within social network: the idea behind our ranking algorithm is that tweets can "vote" for news if they are similar to a news

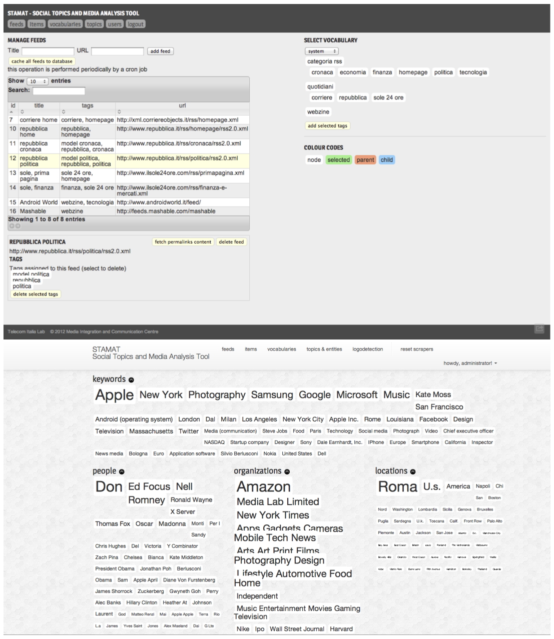
STAMAT - SOCIAL TOPICS AND MEDIA ANALYSIS TOOL
MANAGE FEEDS
SELECT VOCABULARY
COLOUR CODES
REPUBBLICA POLITICA
STAMAT
Social Topics and Media Analysis Tool
keywords
Apple New York Photography Samsung Google Microsoft Music
people
Don Ed Focus Nell Romney
organizations
Amazon Media Lab Limited New York Times
locations
Roma U.S. America

Fig. 1. STAMAT: *top)* managing news feeds; *bottom)* managing topics, entities and concepts.

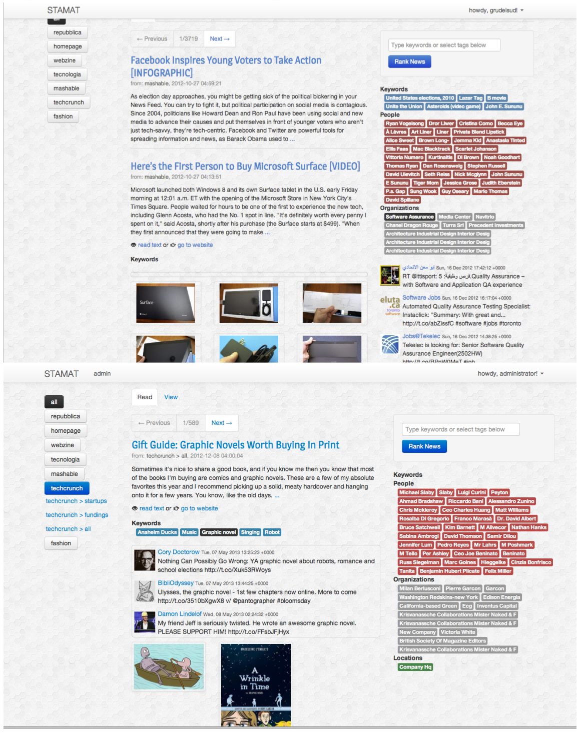
STAMAT
Facebook Inspires Young Voters to Take Action [INFOGRAPHIC]
Here's the First Person to Buy Microsoft Surface [VIDEO]
Keywords
People
Organizations
Rank News
STAMAT admin
Read View
Gift Guide: Graphic Novels Worth Buying In Print
Keywords
People
Organizations
Locations

Fig. 2. STAMAT: *top)* news and social media associated to topics and entities extracted from all the news; *bottom)* news and social media associated with a single news item.

title, by computing its similarity with TF-IDF, after stop-word elimination.

Media elements embedded in news and tweets are indexed using a set of global and local features (CEDD, MPEG-7 descriptors, SIFT BoWs). These features are used to perform CBIR using news and social multimedia; news images are used to search for similar images that propagate across social channels (e.g. to evaluate if some fashion photos hac an impact on a social network, see Fig. 3), while social media are used, similarly to tweets, to "vote" the most representative images of a news topics. To speed up search, images are indexed using an approximate search data structure based on inverted files proposed in [10].

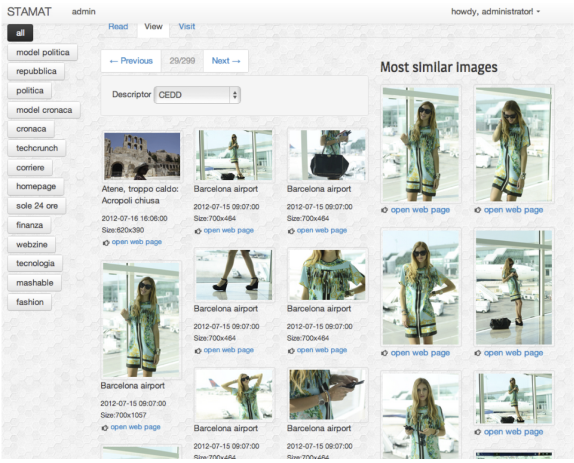

Fig. 3. Example of search of fashion images in websites and social media.

4. REFERENCES

[1] A. Java, X. Song, T. Finin, and B. Tseng, "Why we twitter: understanding microblogging usage and communities," in *Proc. of WebKDD and SNA-KDD*, 2007.

[2] H. Kwak, C. Lee, H. Park, and S. Moon, "What is Twitter, a social network or a news media?," in *Proc. of WWW*, 2010.

[3] K. Lerman and R. Ghosh, "Information contagion: An empirical study of the spread of news on Digg and Twitter social networks," in *Proc. of ICWSM*, 2010.

[4] V. K. Singh, M. Gao, and R. Jain, "Situation recognition: an evolving problem for heterogeneous dynamic big multimedia data," in *Proc. of ACM MM*, 2012.

[5] L. Xie and H. Sundaram, "Media Lifecycle and Content Analysis in Social Media Communities," in *Proc. of ICME*, 2012.

[6] D. M. Blei, "Probabilistic topic models," *Communications of the ACM*, vol. 55, no. 4, pp. 77–84, Apr. 2012.

[7] A. K. McCallum, "MALLET: A Machine Learning for Language Toolkit," http://mallet.cs.umass.edu, 2002.

[8] H. Cunningham, D. Maynard, K. Bontcheva, and V. Tablan, "GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications," in *Proc. of ACL*, 2002.

[9] J.R. Finkel, T. Grenager, and C. Manning, "Incorporating non-local information into information extraction systems by Gibbs sampling," in *Proc. of ACL*, 2005.

[10] G. Amato and P. Savino, "Approximate similarity search in metric spaces using inverted files," in *Proc. of InfoScale*, 2008.