Identifying topical content and experts in Twitter using text and visual content

Eleonora Ciceri
Dipartimento di Elettronica, Informazione e Bioingegneria
Politecnico di Milano (Italy)
eleonora.ciceri@polimi.it

Abstract

We present an automatic pipeline for the collection of topic-related content from Twitter. Through this pipeline, we exploit user-generated content, both in textual and multimedia form, so as to identify current user interests as well as influential users.

Introduction

In recent years, microblogging platforms have become a powerful tool to share ideas and news. Twitter has currently 302 million monthly active users and 500 million tweets are created every day[1]. This huge amount of user-generated content reveals key information that can be used for multiple purposes, such as the analysis of user interests and trends in communications (Mathioudakis and Koudas 2010), the detection of events happening in the real world (Wang and Kankanhalli 2015) and the identification of influential users that are experts in a specific topic (Silva and others 2013). The extraction of such information requires the identification of high quality, topic-specific content. However, users commonly produce heterogeneous, untrustworthy information, which, if analyzed without a proper filtering phase, could bring to wrong assumptions on the identification of topical experts, current trends and user interests. Moreover, due to the large amount of produced content, a manual analysis remains infeasible. Works in the state of the art analyze primarily textual information so as to identify automatically topic-related content. This may cause the retrieval of irrelevant data, due to misinterpretations of its meaning. On the other hand, images could contain valuable information which would help the automatic procedure in identifying relevant content. Thus, in this demo we propose a procedure for the automatic analysis of user-generated content, which unveils topic-related information from the huge amount of tweets produced in real time by analyzing them via a multimodal classifier. The collected topic-related content is further analyzed so as to extract *i)* influential users (to identify topical experts), *ii)* sub-topics of interest in the form of words and hashtags (to extract current user interests), and *iii)* the set of most popular tweets for a topic (to single out viral and interesting content).

[1] https://about.twitter.com/it/company

Topic-related content retrieval pipeline

Figure 1 shows the pipeline used to retrieve topic-related tweets from the Twitter live feed. A *crawler* retrieves tweets produced in real-time that either are produced by topic-related users (i.e., *seed users*) or contain a topic-specific keyword/hashtag (i.e., *seed terms*). The selection of an initial set of seed users and terms is conducted by a field expert. Then, the *filtering* module filters non-English tweets, tweets containing inappropriate content and tweets containing topic-specific stopwords. After that, the *multimodal classification* module classifies tweets as related/non-related to the selected topic, by labeling separately the textual and image components and then merging the produced labels. Next, topic-related tweets are stored in a database, and their content is analyzed so as to extract: *i)* influential users; *ii)* the most used keywords/hashtags; *iii)* the most popular tweets. A report of the extracted information is shown to the user via a Web application, called *dashboard*, whose functionalities are presented in more details in the next section. Finally, the most popular terms (i.e., keywords and hashtags) are fed back to the system, so as to keep the crawling always updated with respect to the current conversations and user interests.

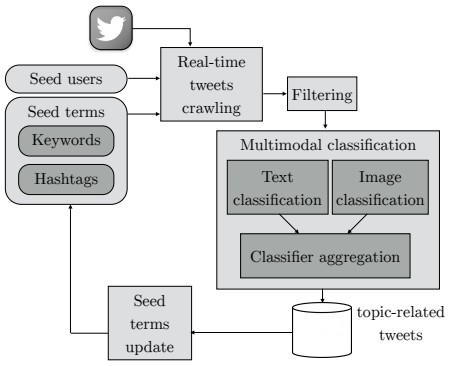
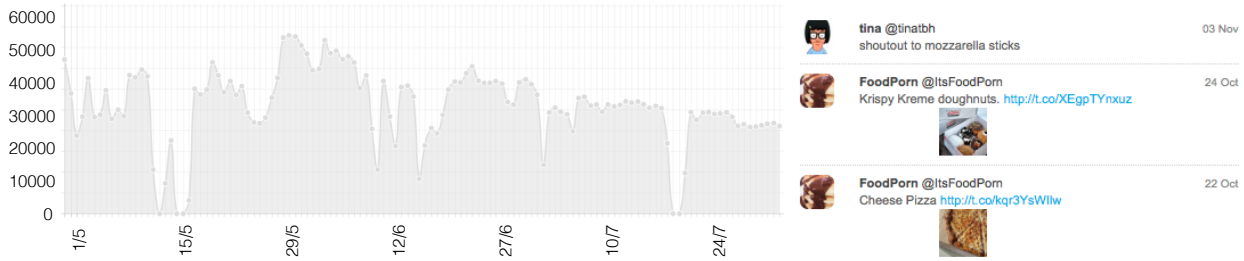


Figure 1: Topic-related content retrieval pipeline



(a) Tweet throughput graph (b) Tweet sample

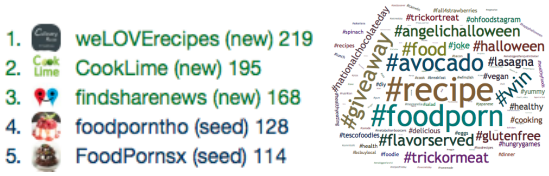
Figure 2: Tweet throughput graph (a) and tweet feed (b)

Dashboard functionalities

In this section we introduce the dashboard functionalities that allow users to easily inspect the collected data. As an example, we ran the pipeline on the topic *food*, although it can be easily deployed for every topic of interest.

Influencers retrieval

An influencer is defined as a person that is focused on the selected topic, very active, well connected with others and talkative. The system analyzes the profiles of the Twitter users whose tweets were captured and classified as topic-related, to identify the top-25 influencers (Figure 3(a)). The interface allows us to visualize either the current influencers (i.e., the most influential users within the last 24 hours) or a past view of the same list. Influencers are tagged as either *seed influencers* (if they were initially identified by the field expert) or *new influencers* (if the machine discovered them automatically). Finally, we can require the visualization of the influencers map (Figure 4): each influencer location is read from the textual user location field (if present) and translated in geo coordinates via the ArcGIS API[2].



(a) Influencers list (b) Hashtag tag cloud

Figure 3: A sample of retrieved influencers (a) and tag cloud reporting the most popular hashtags (b)

Popular keywords and hashtags selection

The most popular keywords and hashtags are those that appear frequently in topic-related tweets and rarely in topic-unrelated tweets. The system automatically identifies the most popular keywords and hashtags appearing in the collected topic-related tweets, and shows them in the form of a tag cloud (Figure 3(b)). The tag clouds can be filtered either to show the most used words in the last 10 minutes or to show past data.

[2] https://geocode.arcgis.com

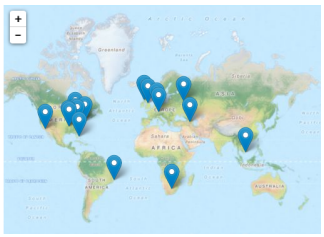


Figure 4: Map of influencers

Tweet stream analysis

To analyze the incoming real-time tweet feed, the dashboard provides us with: *i)* **Tweet throughput graph**: a graph reporting the amount of retrieved tweets (Figure 2(a)); *ii)* **Tweet feed**: a sample of the most popular tweets retrieved by the system (Figure 2(b)). Both the graph and the tweet feed can be filtered to show either real-time data (i.e., an analysis of the tweets collected during the last 2 minutes, which is updated every 10 seconds), or past data.

Conclusion

We presented our tool for the extraction of significant information from user-generated content published via microblogging platforms. Specifically, we highlighted the characteristics of the content retrieval pipeline and the functionalities of the dashboard users can exploit to inspect the retrieved data. In the future, we would like to expand the tool so as to infer automatically the geolocation of tweets whose coordinates are not provided, exploiting the information contained in the user profile, in the text of the produced tweets and in the visual content of the posted images.

References

Mathioudakis, M., and Koudas, N. 2010. Twittermonitor: trend detection over the twitter stream. In *SIGMOD*, 1155–1158. ACM.

Silva, A., et al. 2013. Profilerank: finding relevant content and influential users based on information diffusion. In *SNMA Workshop*, 2. ACM.

Wang, Y., and Kankanhalli, M. S. 2015. Tweeting cameras for event detection. In *WWW*, 1231–1241.