Tracking and Analyzing The 2013 Italian Election

Vuk Milicic, José Luis Redondo García, Giuseppe Rizzo, Raphaël Troncy

EURECOM, Sophia Antipolis, France,
{vuk.milicic, redondo, giuseppe.rizzo, raphael.troncy}@eurecom.fr

Abstract. Social platforms open a window to what is happening in the world in near real-time: (micro-)posts and media items are shared by people to report their feelings and their activities related to any type of events. Such an information can be collected and analyzed in order to get the big picture of an event from the crowd point of view. In this paper, we present a general framework to capture and analyze micro-posts containing media items relevant to a search term. We describe the results of an experiment that consists in collecting fresh social media posts (posts containing media items) from numerous social platforms in order to generate the story of the "2013 Italian Election". Items are grouped in meaningful time intervals that are further analyzed through deduplication, clusterization, and visual representation. The final output is a storyboard that provides a satirical summary of the elections as perceived by the crowd. A screencast showing an example of these functionalities is published at http://youtu.be/jIMdnwMoWnk while the system is publicly available at http://mediafinder.eurecom.fr/story/elezioni2013.

Keywords: Storytelling, Storyboard Creation, Visual Summarization, Topic Generation

1 Introduction

The massive amount and steady increase of heterogeneous data shared on social platforms has attracted the interest of different research communities. Microposts such as status updates or tweets enable people to share their activities, feelings, emotions and conversations, opening a window to the world in real-time. Making sense out this amount of data is an extremely challenging task due to its heterogeneity (media items mixed with textual data) and dynamics making often short-lived phenomena. A growing number of commercial tools and academic research approaches try to partially collect and analyze this data in order to make sense of it. Capturing life moments and building narratives using social platforms is, for example, the goal of Storify[1] where the creators aim to investigate the interaction between event stories and the role of social networks that tell them: *(i)* sorting and organizing the items of an experience

[1] http://storify.com

similar to the elements of a story, *(ii)* communicating and discussing strategies on how to guide a user towards an intended experience. The overall storytelling creation is supervised by the user who composes a story based on streams of news coming through social platforms such as Twitter and YouTube. Generating the big picture from these streams is also the objective of Storyful[2]. This application enables the user to navigate through the story created by other users or to create his own, aggregating content from different social platforms. While these two approaches position the role of a social platform as a container of fresh and breaking news items, they are leveraging on the user interaction that defines the summary creation as a supervised task. A disruptive innovation has been recently revealed by Mahaya[3] which proposes an automatic crowd storyfication of the 12/12/12 concert[4]. In this example, the highlights of the concert corresponding to social media spikes when performers appeared are emphasized with images collected from Instagram, and microposts collected from Twitter. Inspired by the idea of automatic summarization through visual galleries, we focus more on the automatic sorting and clustering of media items for topic visualization. In [2], we proposed a generic media collector for retrieving media items that illustrate daily life moments shared on social platforms. In particular, we proposed a common schema in order to align the search results of numerous social platforms. This demonstration extends [1], adding to the codebase, the temporal feature to the cluster operations.

2 Use case: The 2013 Italian Election

The 2013 Italian Election will be remembered as the *messy*[5] Italian political election because of the absence of a clear winner. Such a result triggered lots of discussion on numerous Italian and international newspapers, and especially on different social platforms. We have tracked and analyzed media posts tagged as *elezioni2013* from 2013-02-26 to 2013-03-03. In particular, we have automatically extracted the main named entities in these posts and perform different automatic clustering operations. We have then compared those results with a baseline made of the facts occurring during those six days, taking as input the news appearing in the daily Italian and international newspapers.

2.1 Event Streaming Tracking

We performed an event stream processing from all media posts shared on the following social platforms: Twitter and its ecosystem (TwitPic, TwitterNative, MobyPicture, Lockerz or yfrog), GooglePlus and YouTube, Facebook and Instagram, Flickr and FlickrVideos. Leveraging on the search API of those platforms, we applied a cron job composed of different searches using the term *elezioni2013*.

[2] http://storyful.com

[3] http://mahaya.co

[4] http://121212.mahaya.co

[5] http://online.wsj.com/article/SB10001424127887323384604578325992879185934.html

Each search operation proceeds as follows: first, each social platform is queries and the resulting microposts containing at least one media item (image or video) are collected. The output of those search queries result in a heterogeneous collection of items, varying in terms of serialization formats, schemas, data types, and topics (hidden or declared), that have been harmonized to a common schema. We applied a near-deduplication process to group microposts which contain the same image. By using the Hamming Distance over the discrete cosine similarity image fingerprints, we created a collection of items where each one is attached to a list of microposts that contain this media resource. Finally, for each micropost, we extract named-entity using the NERD framework [3]. A multi-lingual entity extraction is performed and the output result is a collection of entities annotated using the NERD Ontology[6] and attached to each micropost. We repeated this search operation every 30 minutes for the 6 days.

2.2 Temporal grouping

One of the key aspect of an event is the time dimension. Hence, we have sliced the *elezioni2013* timeline in 24 hours slots where the different social media posts were aligned to. Our assumption is that people react and share opinions and feelings following the daily newspapers reporting. Even though it is a simple assumption, it fits well (as we have observed in our use case) this particular political scenario. The period studied is therefore divided into six slots (one per day), and each search is aligned to the corresponding interval according to the time it was queried. We have then calculated the union between all the searches belonging to the same time interval and we have discarded the microposts published outside the boundaries of the interval. At the end of this step, social media posts were grouped in consecutive intervals of times which were relevant in the context of the current event.

2.3 Topic Generation

Once the entire set of social media posts is divided into chronologically ordered time intervals, we have further analyzed the details of every of those generated groups in order to automatically find the main highlights. For each temporal group, our approach identifies the topics that best describe the result set using clustering operations. Specifically, it implements four clustering methods working exclusively on the textual features from the microposts: *(i) named entity* based cluster algorithm which groups microposts according to the most frequent named entities extracted in the microposts; a further distance process is applied to compute the label similarity among the extracted entity. *(ii) named entity type* based cluster where microposts are grouped according to the dominant types of extracted named entities, such as Thing, Amount, Animal, Event, Function, Location, Organization, Person, Product, Time. *(iii) generative model* that extracts hidden topics from the large set of microposts collected, using the

[6] http://nerd.eurecom.fr/ontology/nerd-v0.5.n3

latent Dirichlet allocation (LDA) model. *(iv) density* based cluster that exploits micropost proximity based on several micropost features such as temporal distance, text similarity and entity label similarity. Once the clustering operation is completed, we generate a Bag of Entities (BOE) that best describes the cluster while the most representative entity is disambiguated using a DBpedia URI[7] and chosen to be the label of this cluster. Ultimately, the final output of this processing step is a set of clusters (limited to ten for visualization purpose) that best describes the topics extracted from one time interval. In our case study, we identified the following clusters at the end of this step: *Monti*, *Bersani*, *Italia*, *Berlusconi*, *Grillo*, and *Stelle* which basically correspond to the main entities and actors of the elections.

2.4 Visualization and Storyboarding

Once computed, these clusters are displayed according to a chronological timeline. The visualization of those results is crucial since the user has to be able to identify what are the main facts for a period of time for an event. We generate a stacked bar chart where the horizontal axis contains the time intervals (six days) and the vertical axis depicts the number of social media posts available. Each bar is decomposed into different portions that corresponds to a cluster inside this particular time range. To easily differentiate them, they are labeled and displayed using different colors. The user can, at anytime, show or hide a particular cluster in the diagram by interacting with the total list of clusters available in the left panel. There are two main indicators that help users to decide about the importance of a particular cluster: (i) minimal background knowledge about the context of the event attached to each portion (or cluster); (ii) the number of social media posts inside the cluster, and intuitively, if a cluster contains a higher number of items, it is because this topic is trending and it is therefore most probably important for the topic generation. Finally, by clicking on a particular cluster, a more detailed representation with the corresponding media items are shown, giving the user an insight about the relevance of that particular fact. This way of displaying an event can go further, being possible not only to figure out what is the most relevant fact for a particular day, but also to track the way this topic has evolved over the time of the event.

3 Discussion

During the first day (February 26th), no party was a clear winner of the election, but Monti's defeat stood out. International newspapers reported about it, fueling the discussion on social platforms. Similarly, the 27th is the day after Bersani's speech. His words about the dramatic situation became popular and viral. Another example of the usefulness of our automatic approach is the day 28th, when the world started to investigate about *The cult of Silvio Berlusconi* and the reaction from the crowd (catched by our storyboard) is immediately depicted. On

[7] http://dbpedia.org

March 1st, the Italian president visited Germany and met the German chancellor Merkel to discuss, amongst others, the Italian elections. On March 2nd, Grillo claimed no collaboration with any party. Grillo is the *Movimento 5 Stelle* leader, incorrectly identified by our approach as *Stelle*. This is due to the lack of knowledge of our entity spotter. In the same day, both a famous Italian talk show and an Economist article triggered a lot of discussion about the current Italian situation, Grillo and Berlusconi. Again, Grillo and Berlusconi were trending. On March 3rd, the kick-off meeting of the Beppe Grillo's party took place. The main action lines were rejecting possible alliance with any parties and especially with the Bersani's party.

The election use case showed that our approach provides useful insights about events while leveraging from the crowd. In a wide sense, it provides a snapshot of how the crowd is experiencing an event. We focus on generating visual summaries in order to illustrate those phenomena. We argue that the need of human intervention in the collection and interpretation processes has been drastically reduced. The different clustering operations and the user interface give to the user a tool for interpreting the story, even if some background knowledge about the event is required. Although the crowd talked about the same main entities reported by daily newspapers, we observe that different satirical expressions for talking about politics are used. Actually, numerous images and posts are biased towards being satirical. Tracking and understanding an event from social data should also include the popularity of microposts. We have planned to use this measure as an additional mean for weighting the importance of a social peak. We will further demonstrate how this tool can been used to generate visual summaries of highly viral memes such as *Harlem Shake*.

Acknowledgments

This work was partially supported by the European Union's 7th Framework Programme via the project LinkedTV (GA 287911).

References

1. V. Milicic, G. Rizzo, J. L. Redondo Garcia, R. Troncy, and T. Steiner. Live Topic Generation from Event Streams. In *22nd International World Wide Web Conference (WWW'13)*, Rio de Janeiro, Brazil, 2013.
2. G. Rizzo, T. Steiner, R. Troncy, R. Verborgh, J. L. Redondo García, and R. Van de Walle. What Fresh Media Are You Looking For? Retrieving Media Items from Multiple Social Networks. In *International Workshop on Socially-aware multimedia (SAM'12)*, Nara, Japan, 2012.
3. G. Rizzo and R. Troncy. NERD: A Framework for Unifying Named Entity Recognition and Disambiguation Extraction Tools. In *13th Conference of the European Chapter of the Association for Computational Linguistics (EACL'12)*, Avignon, France, 2012.