Live Topic Generation from Event Streams

Vuk Milicic, Giuseppe Rizzo,
José Luis Redondo García and
Raphaël Troncy
EURECOM
Biot, France
raphael.troncy@eurecom.fr

Thomas Steiner
Univ. Politècnica de Catalunya
Barcelona, Spain
tsteiner@lsi.upc.edu

ABSTRACT

Social platforms constantly record streams of heterogeneous data about human's activities, feelings, emotions and conversations opening a window to the world in real-time. Trends can be computed but making sense out of them is an extremely challenging task due to the heterogeneity of the data and its dynamics making often short-lived phenomena. We develop a framework which collects microposts shared on social platforms that contain media items as a result of a query, for example a trending event. It automatically creates different visual storyboards that reflect what users have shared about this particular event. More precisely it leverages on: *(i)* visual features from media items for near-deduplication, and *(ii)* textual features from status updates to interpret, cluster, and visualize media items. A screencast showing an example of these functionalities is published at: `http://youtu.be/8iRiwz7cDYY` while the prototype is publicly available at `http://mediafinder.eurecom.fr`.

Categories and Subject Descriptors

H.3.4 [**Information Systems**]: Information Storage and Retrieval—*World Wide Web*

Keywords

Topic Generation, Storyboard Identification, Visual Summarization, Storytelling, Social Media

1. INTRODUCTION

The massive and steady rising amount of heterogeneous data shared on social platforms has attracted the interest of different research communities for the important role to the society of sharing human's 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. Aware of the impossibility of a truly and exhaustive understanding of such phenomena, we are witnessing a proliferation of commercial and research attempts to sample the stream in order to tell stories. Capturing those memes and building narratives using social platforms is for example the goal of Storify[1] that aims to support the creation of stories about events using social networks by: *1)* sorting and organizing the items of an experience similar to the elements of a story, *2)* communicating and discussing strategies on how to guide a user towards an intended experience. The overall storytelling creation is supervised by the user who describes the story as a crafted experience. Generating the big picture from media sharing streams is also the objective of Storyful[2]. This application allows 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 semi-unsupervised summary creation is proposed by Twitter through the photo and video gallery. A user can then create a real-time gallery according to the main hashtags he has searched for where items can be organized by popularity of chronologically. 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 [8], we proposed a generic media collector for retrieving media items that illustrate daily life moments shared on 10 social platforms. We proposed a common schema in order to align the search results of the supported platforms. The outcome of this investigation revealed the importance of creating visual summaries out of filtered search sessions together with the challenge when dealing with content diversity that these platforms intrinsically provide, leaving also insights about the possibility to process the collected results afterward.

The automatic detection of events from social media is also an active research topic. Different approaches for event detection based on users' activities have been proposed in the research literature and several domain-specific methods for event detection showing good accuracy have been proposed, for example, in the sports domain [4]. However, the challenge in this field is to find methods that are content-agnostic. A first category of related work includes research that aims to collect, align, and organize media for trends or events. Liu *et al.* combine semantic inference and visual

[1] `http://storify.com`
[2] `http://storyful.com`

analysis to automatically find media that illustrate events [5]. They interlink large datasets of event metadata and media with the Linking Open Data Cloud[3]. Data reconciliation uses visual, temporal, and spatial similarity measures for attaching photo streams to events [11]. Other ways to collect and order media from social platforms are based on user-driven metadata such as geospatial information [2].

People are used to share content within their social networks when something happens. We often refer to events when we target specific human's activities which involve different people resulting in a lot of activities by the crowd instantaneously shared. Twitter has been largely investigated for detecting events in real-time, because tweets are considered the up-to-date and inclusive stream of information and commentary on current events. Sakaki *et al.* consider tweets as crowd sensors and proposed semantic analysis approach for classifying tweets for understanding the nature of events and their locations [10]. In [7], the authors proposed an approach for open-domain event extraction leveraging on the Twitter stream. Events are filtered according to popularity and importance, and they are categorized according to the time of the event the tweets refer to. Streams are however fragmented and full of noisy information on Twitter which motivate the need for systems that can automatically consume, extract and categorize events. Instead of focusing on the open-domain event detection problem, Mahaya[4] has recently showcase a story about the 12/12/12 concert[5]: the different spikes of the concert when different artists performed are guiding the narrative and images collected from Instagram, and microposts collected from Twitter are used to illustrate the story. Inspired by this approach of making sense of an already filtered stream of heterogeneous data published by the crowd, we process the stream of microposts enclosed with media items coming from Twitter and its ecosystem (TwitPic, TwitterNative, MobyPicture, Lockerz or yfrog), GooglePlus and YouTube, Facebook and Instagram, Flickr and FlickrVideos. Exploiting the live-short phenomena of these microposts, we propose a live approach which first removes near-duplicates of media items grouping the different textual opinions expressed in microposts that belong to the same media item. We further process microposts for automatically generating topics. Finally, we generate different storyboards illustrated by visual galleries.

The remainder of this paper is organized as follows. The architectural overview of our approach is presented in Section 2. In Section 3 we detail a demonstration of the proposed framework, and in Section 4 we propose an outlook of insights and on the ongoing research.

2. ARCHITECTURE OVERVIEW

Figure 1 shows the main components of the proposed framework. In the following subsections, we detail each component.

2.1 Live Stream

Twitter and its ecosystem (TwitPic, TwitterNative, MobyPicture, Lockerz or yfrog), GooglePlus and YouTube, Facebook and Instagram, Flickr and FlickrVideos constantly record heterogeneous data shared by their users. To ease the access to this amount of data, these social platforms enable search programmatically through their APIs. Those search functions, however, provide results that vary according to the time the query has been triggered, covering a window of time which ranges from only the recent past to many years ago. In addition, they offer different parameters that enable to customize search queries (e.g. filtering by location). The system detailed in this paper triggers different searches, one per each social platform supported, filtering the results to microposts that contain a media item (i.e. an image or a video), hosted on the social network on a third party service.

2.2 Data Model Harmonization

Inputs of the harmonization component are a heterogeneous collection of items and metadata, varying in terms of serialization formats, schemas, data types and topics (hidden or declared). This component harmonizes the results and projects them to a common schema. This component performs also a cleansing process, discarding items which are older than seven days ago, in order to keep only fresh media items. The unified output is a collection of microposts that each includes a media item together with its textual description.

2.3 Near De-duplication

The same media item can be found in different microposts, typically where re-tweet or re-share operations are performed. This component performs a crucial step for avoiding redundant information to be displayed while at the same time counting the number of times an item can be found as an additional measure of popularity. It accomplishes this objective by using state of the art techniques in content-based image retrieval: we compute a signature using the Discrete Cosine Transform (DCT) for images, while for video, the signature is computed on the poster thumbnail. Pair of signatures are then mutually compared using the Hamming distance. Finally, the output of this process is a collection of media items where a media item is attached to a list of microposts. This approach has the drawback to detect only exact duplicates media items, but it has the important strength of being computationally inexpensive time (real-time process), and it is based on the empirically proven assumption that within social networks, users are used to post the same media item as a result of re-sharing it without any further editing.

2.4 Named Entity Extraction

For each micropost, we perform named-entity recognition using the NERD framework [9]. Microposts can vary in terms of content and also in terms of language. A multilingual entity extraction is performed and the output result is a collection of entities attached to each micropost. Then, the entities are annotated using the NERD Ontology v0.5[6]. A set of candidate entities are then used as inputs for the topic generation component.

2.5 Topic Generation

Diversity has a central role in our approach and even for a simple query, the data collected from social platforms may be very diverse in terms of visual and textual distance. This component mitigates such a diversity, identifying top-

[3] `http://lod-cloud.net`
[4] `http://mahaya.co`
[5] `http://121212.mahaya.co`
[6] `http://nerd.eurecom.fr/ontology/nerd-v0.5.n3`

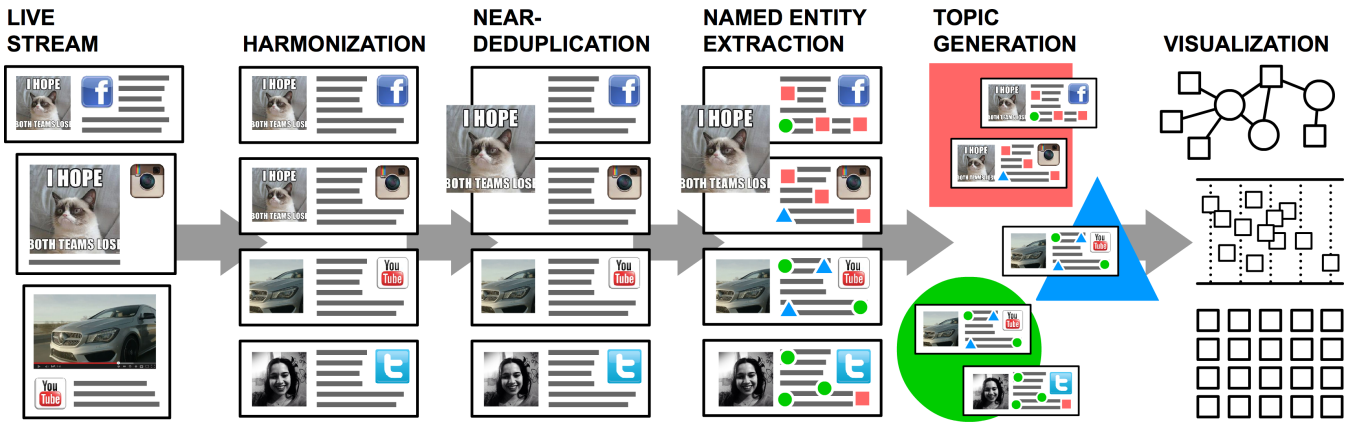

Figure 1: A stream of heterogeneous data from social platforms is collected. After the harmonization process, visual and textual analysis processes are run and results are inputs for the the topic generation component. Finally, the topics found are illustrated with media galleries.

ics extracted in the collected results using clustering operations. The proposed approach implements four clustering methods working exclusively on 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 (an additional distance is computed to measure the similarity between entity labels); *ii) named entity type* based cluster where microposts are grouped according to the predominant types of the named entities extracted, such as Thing, Amount, Animal, Event, Function, Location, Organization, Person, Product and Time; *iii) generative model* that extracts hidden topics from the large set of microposts collected, using the Latent Dirichlet allocation (LDA); *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 operations are completed, the entity which best represents a set is selected. We call this process *topic generation*. For each cluster, a Bag of Entities (BOE) is computed and the most representative entity is selected to be the cluster topic. We disambiguate this topic using a DBpedia URI[7]. Consequently, the output of this component is a set of clusters (limited to ten for visualization optimization) that corresponds to topics extracted from the data collected.

2.6 Visualization

The visualization emphasizes the different aspects of storyboards. The graph view shows the relationships between microposts and topics, while the timeline view emphasizes the time dimension. The user can watch and interact with the summarized view of all the topics or select a particular one with the additional details. The visualization exploits the data-join concept [1] which enables direct manipulation of a native representation (the DOM model) improving expressiveness and allowing animated transitions that significantly improve graphical perception [3]. In addition, the states of different views are persistent through the URLs which makes easy sharing possible.

3. DEMONSTRATION

[7]`http://dbpedia.org`

This section illustrates the proposed framework with the 2013 Superbowl event. We show that the framework is not only able to extract fresh media items about the event, but it can also be used to identify insights happening during the event. After the user has launched the search operation, microposts containing media items are collected. Focusing on the collected media items, we can see the contenders' logos, pictures about relevant plays in the game, promotional videos, etc. By simply looking at those items, the user can have a first idea of what happened during the event. But if those results are organized in a more meaningful manner, it is more intuitive for the user to see which were the most relevant topics and to relive the hidden parts of the event. While the Grid View is displayed, the entity extraction process has annotated each microposts and the clustering operations are run in the background. Figure 2 *a)* shows the output of this process where media items are organized in regions and link to different circles which represent the topics generated. One of those regions (labeled as "Beyonce") brings the attention of the user. From the set of media items belonging to that cluster, he immediately knows that Beyoncé has been performing during the Superbowl. As he is a fan of this artist, he wants to have more details. He can go further by simply clicking on the cluster itself. Figure 2 *b)* shows the Beyonce cluster where the illustrative narrative description is mixed with the textual microposts shared by the crowd.

From the continuous flow of pictures about the game, and their diversity, the proposed framework has successfully highlighted a particular topic inside the event. This information, which has been proven to be relevant for the user in the context of that event, was hidden in the initial set of media items. This inferred knowledge is also connected with other topics. This represents an innovative way of finding relevant insights of an event by not solely spotting that something relevant has happened, but relating it with other important topics that are also inside the umbrella of the event.

4. CONCLUSION AND FUTURE WORK

The diversity of the collected results is due to the best effort of the search services provided by the social platforms. The same search, performed in a short period of time af-

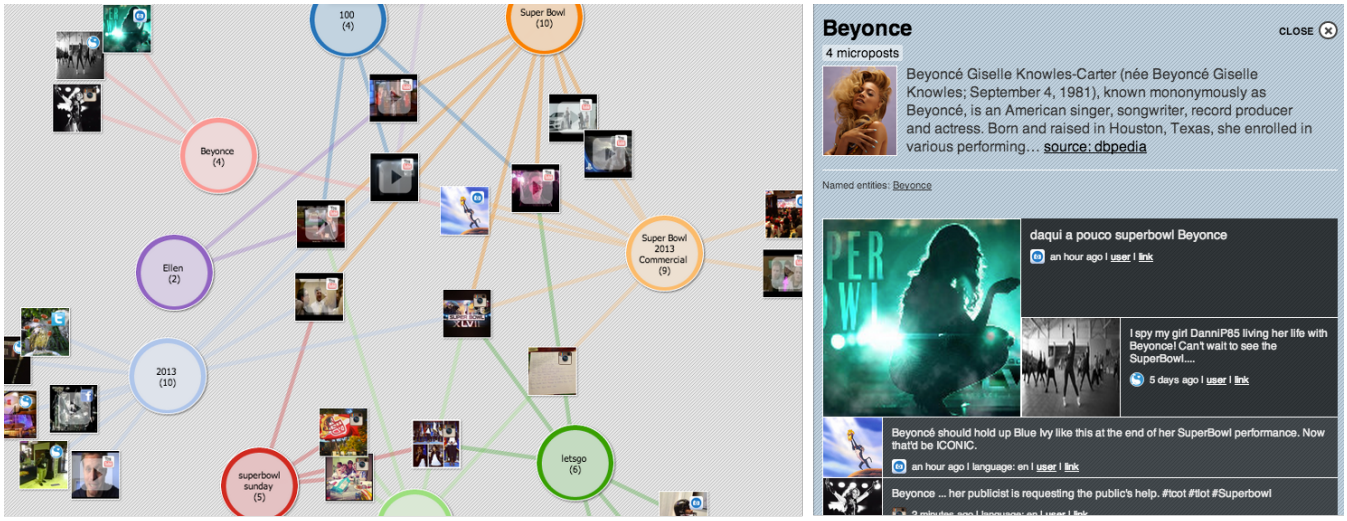

Figure 2: ***a)*** **Graph View of the** ***named entity*** **based clustering operation on the collected items.** ***b)*** **Visual summary of the cluster "Beyonce": the introduction is fetched through disambiguating the selected topic on DBpedia and it is followed by the storyfication of microposts and media items.**

ter, may not retrieve the same items collected in previous attempts. Part of the ongoing work is using searches performed in different points of time for better understanding the phenomena of the targeted event. Hence, MediaFinder has a repeated search option where the user can input a time interval where his search query will be re-issued. This results in a more fine grained topic generation that can be further illustrated using visual transitions [6]. The time dimension allows to relive the story about an event, moving towards a crowd-based visual storyfication. Microposts and media items differ in terms of popularity. This dimension is used as input for the topic generation and visualization. An important part of our future work consists in evaluating the relevance of the clusters generated. Next, we would like to use storytelling techniques in order to generate interactive multimedia stories.

Acknowledgments

This work was partially supported by the European Union's 7th Framework Programme via the projects LinkedTV (GA 287911) and I-SEARCH (GA 248296).

5. REFERENCES

[1] M. Bostock, V. Ogievetsky, and J. Heer. D3 Data-Driven Documents. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2301–2309, 2011.

[2] D. J. Crandall, L. Backstrom, D. Huttenlocher, and J. Kleinberg. Mapping the world's photos. In *18th International Conference on World Wide Web (WWW'09)*, Madrid, Spain, 2009.

[3] J. Heer and G. Robertson. Animated Transitions in Statistical Data Graphics. *IEEE Transactions on Visualization and Computer Graphics*, 13(6):1240–1247, 2007.

[4] B. Li and M. Sezan. Event detection and summarization in American football broadcast video. In *Storage and Retrieval for Media Database*, pages 202–213, 2002.

[5] X. Liu, R. Troncy, and B. Huet. Finding Media Illustrating Events. In *1st ACM International Conference on Multimedia Retrieval (ICMR'11)*, Trento, Italy, 2011.

[6] V. Milicic, J. L. Redondo García, G. Rizzo, and R. Troncy. Tracking and Analyzing The 2013 Italian Election. In *10^10 Extended Semantic Web Conference (ESWC'13), Demo Session*, Montpellier, France, 2013.

[7] A. Ritter, Mausam, O. Etzioni, and S. Clark. Open domain event extraction from Twitter. In *18th International Conference on Knowledge Discovery and Data mining (KDD'12)*, Beijing, China, 2012.

[8] G. Rizzo, T. Steiner, R. Troncy, R. Verborgh, J. L. Redondo García, and R. V. 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.

[9] 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.

[10] T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake shakes Twitter users: real-time event detection by social sensors. In *19th International Conference on World Wide Web (WWW'10)*, Raleigh, North Carolina, USA, 2010.

[11] J. Yang, J. Luo, J. Yu, and T. S. Huang. Photo stream alignment for collaborative photo collection and sharing in social media. In *3rd ACM International Workshop on Social Media (WSM'11)*, Scottsdale, Arizona, USA, 2011.