TWINE: A real-time system for TWeet analysis via INformation Extraction

Debora Nozza, Fausto Ristagno, Matteo Palmonari,
Elisabetta Fersini, Pikakshi Manchanda, Enza Messina
University of Milano-Bicocca / Milano
{debora.nozza, palmonari, fersini,
pikakshi.manchanda, messina}@disco.unimib.it
f.ristagno@campus.unimib.it

Abstract

In the recent years, the amount of user generated contents shared on the Web has significantly increased, especially in social media environment, e.g. Twitter, Facebook, Google+. This large quantity of data has generated the need of reactive and sophisticated systems for capturing and understanding the underlying information enclosed in them. In this paper we present TWINE, a **real-time** system for the **big data** analysis and **exploration** of information extracted from Twitter **streams**. The proposed system based on a Named Entity Recognition and Linking pipeline and a multi-dimensional spatial geo-localization is managed by a scalable and flexible architecture for an interactive visualization of micropost streams insights. The demo is available at *http://twine-mind.cloudapp.net/streaming*[1,2].

1 Introduction

The emergence of social media has provided new sources of information and an immediate communication medium for people from all walks of life (Kumar et al., 2014). In particular, Twitter is a popular microblogging service that is particularly focused on the speed and ease of publication. Everyday, nearly 300 million active users share over 500 million of posts[3], so-called tweets, principally using mobile devices.

Twitter has several advantages compared to traditional information channels, i.e. tweets are created in real-time, have a broad coverage over a wide variety of topics and include several useful embedded information, e.g. time, user profile and geo-coordinates if present.

Mining and extracting relevant information from this huge amount of microblog posts is an active research topic, generally called Information Extraction (IE). One of the key subtask of IE is Named Entity Recognition and Linking (NEEL), aimed to first identify and classify named entities such as people, locations, organisations and products, then to link the recognized entity mentions to a Knowledge Base (KB) (Derczynski et al., 2015).

Although several Information Extraction models have been proposed for dealing with microblog contents (Bontcheva et al., 2013; Derczynski et al., 2015), only few of them focused on the combination of these techniques with big data architecture and user interface in order to perform and explore real-time analysis of social media content streams. Moreover, the majority of these research studies are event-centric, in particular focusing on the tasks of situational awareness and event detection (Kumar et al., 2011; Leban et al., 2014; Sheth et al., 2014; Zhang et al., 2016).

In this paper we propose TWINE, a system that visualizes and efficiently performs real-time big data analytics on *user-driven* tweets via Information Extraction methods.

TWINE allows the user to:

- perform real-time monitoring of tweets related to their topics of interest, with unrestricted keywords;
- explore the information extracted by semantic-based analysis of large amount of tweets, i.e. (i) recognition of named entities and the information of the correspondent

[1] At the moment, the application is deployed on Azure client service with traffic and storage limits given by the provider.

[2] The TWINE system requires Twitter authentication, if you do not want to use your twitter account you can try the demo at *http://twine-mind.cloudapp.net/streaming-demo*.

[3] http://www.internetlivestats.com/

Proceedings of the EACL 2017 Software Demonstrations, Valencia, Spain, April 3-7 2017, pages 25–28

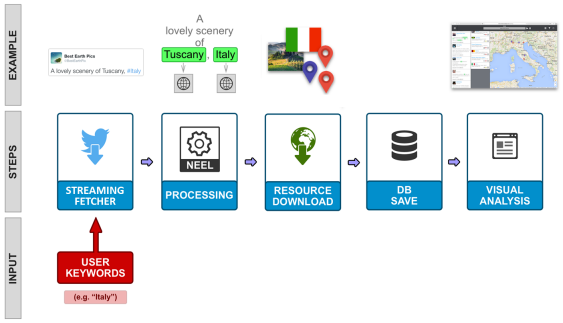

Figure 1: TWINE system overview.

KB resources, (ii) multi-dimensional spatial geo-tagging for each tweet, including the geo-localization of the named entities identified as locations and (iii) two semantic-driven interactive visualization interfaces.

The following section will present the details of the architecture for supporting real-time tweets analysis and the description of the conceived graphical user interface.

2 TWINE system

The proposed system TWINE, acronym for TWeet analysis via INformation Extraction, is a real-time system for the analysis and exploration of information extracted from Twitter data. Figure 1 outlines its macro steps coupled with corresponding examples.

Given a set of keywords provided by the user (e.g. "Italy") as input query, the system fetches the stream of all the tweets (text and tweet's author information) matching the keywords using Twitter APIs. Next, each tweet text is processed by the NEEL pipeline. This step provides as output a set of recognized named entities linked to the correspondent KB resource (e.g. *Tuscany* - *http://dbpedia.org/page/Tuscany*). After this elaboration, the system retrieves all the additional relevant information needed for the exploration: from the KB we extract the resource data, i.e. image, text description, type and coordinates if the entity is a location, and from Twitter we extract the tweet author account's profile location, that is resolved wih a georeferincing system.

This information are subsequently stored in a database that incrementally enriches information generated by the precedent phases. Then, the TWINE web interface fetches the stored data from the DB for populating two different interactive visualisation interfaces.

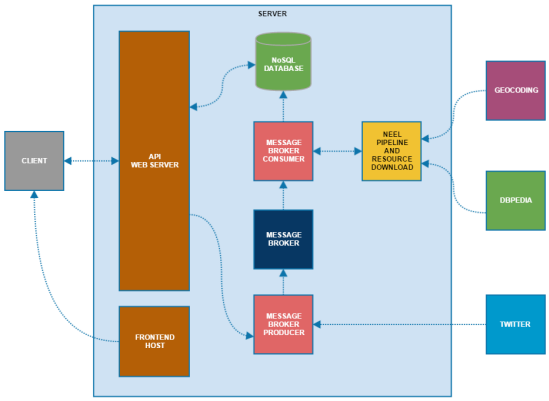

Figure 2: TWINE system architecture.

2.1 System Architecture

The proposed system is implemented using a centralized system architecture, as shown in Figure 2. The main requirement was to develop a system able to process in real-time large incoming of data streams.

In TWINE, all the aforementioned processes are triggered by the user from the client and elaborated on the server-side, i.e. the streaming fetching phase, the NEEL processing, the KB resources retrieval, the geo-codification of the locations and the database storing.

With this design implementation all the computations are performed on the server. This improves the independence on the client technical specifications, preventing different problems such as slow loading, high processor usage and even freezing.

The system architecture, presented in Figure 2, is composed of several independent modules:

External Services. The system makes use of Twitter APIs for fetching the streaming of tweets given an input query, a SPARQL endpoint over the DBpedia data set for the retrieval of the KB resource information and a georeferencing system, OpenStreetMap[4], to obtain the geographic coordinates from the tweet author account's profile location.

NEEL pipeline. This module uses the NEEL pipeline proposed by Caliano et al. (2016) on the tweets.

Message Broker system. This module is necessary to build pipelines for processing streaming data in real time, in such a way that components

[4]https://www.openstreetmap.org/

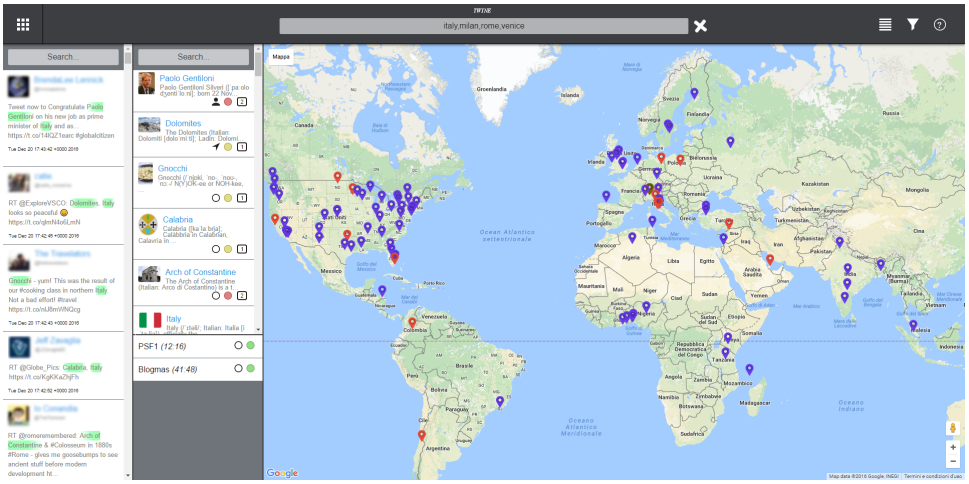

Figure 3: TWINE Map View snapshot.

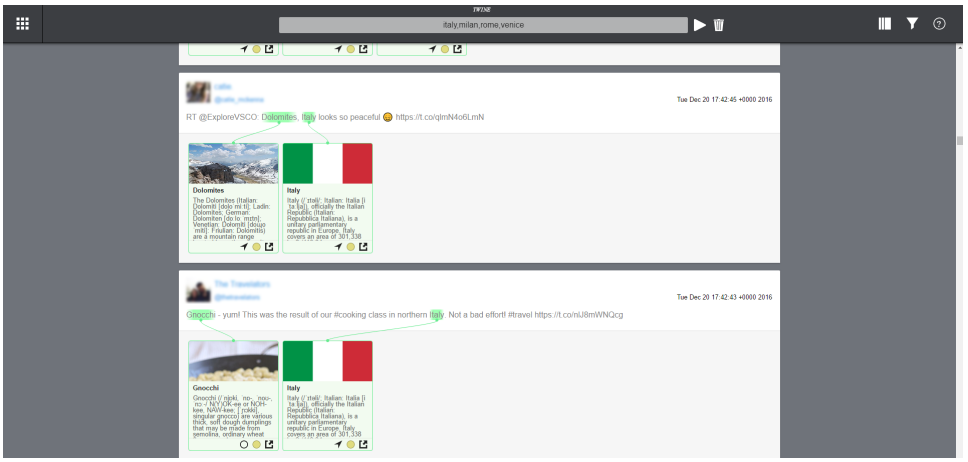

Figure 4: TWINE List View snapshot.

can exchange data reliably. The Apache Kafka platform[5] permits us to store and process the data in a fault-tolerant way and to ignore the latency due to the Information Extraction processing.

Database. All the source and processed data are stored in a NoSQL database. In particular, we choose a MongoDB[6] database because of its flexibility, horizontal scalability and its representation format that is particularly suitable for storing Twitter contents.

Frontend host and API web server. The presence of these two server-side modules is motivated by the need of make the TWINE user-interface independent on its functionalities. In this way, we improve the modularity and flexibility of the entire system.

[5]https://kafka.apache.org/
[6]http://www.mongodb.org/

2.2 User Interface

TWINE provides two different visualisations of the extracted information: the Map View, which shows the different geo-tags associated with tweets in addition to the NEEL output, and the List View, that better emphasizes the relation between the text and its named entities.

The Map View (Figure 3) provides in the top panel a textual search bar where users can insert keywords related to their topic of interest (e.g. *italy, milan, rome, venice*). The user can also, from left to right, start and stop the stream fetching process, clear the current results, change View and apply semantic filters related to the geo-localization and KB resource characteristics, i.e. type and classification confidence score.

Then, in the left-hand panel the user can read the content of each fetched tweet (text, user information and recognized named entities) and

directly open it in the Twitter platform.

The center panel can be further divided into two sub-panels: the top one shows the information about the Knowledge Base resources related to the linked named entities present in the tweets (image, textual description, type as symbol and the classification confidence score), and the bottom one provides the list of the recognized named entities for which it does not exist a correspondence in the KB, i.e. NIL entities.

These two panels, the one that reports the tweets and the one with the recognized and linked KB resources, are responsive. For example, by clicking on the entity *Italy* in the middle panel, only tweets containing the mention of the entity *Italy* will be shown in the left panel. Respectively, by clicking on a tweet, the center panel will show only the related entities.

In the right-hand panel, the user can visualize the geo-tag extracted from the tweets, (i) the original geo-location where the post is emitted (*green marker*), (ii) the user-defined location for the user account's profile (*blue marker*) and (iii) the geo-location of the named entities extracted from the tweets, if the corresponding KB resource has the latitude-longitude coordinates (*red marker*).

Finally, a text field is present at the top of the first two panels to filter the tweets and KB resources that match specific keywords.

The List View is reported in Figure 4. Differently from the Map View, the focus is on the link between the words, i.e. recognized named entities, and the corresponding KB resources. In the reported example, this visualisation is more intuitive for catching the meaning of *Dolomites* and *Gnocchi* thanks to a direct connection between the named entities and the snippet and the image of associated KB resources.

3 Conclusion

We introduced TWINE, a system that provides an efficient real-time data analytics platform on streaming of social media contents. The system is supported by a scalable and modular architecture and by an intuitive and interactive user interface.

As future work, we intend to implement a distributed solution in order to faster and easier manage huge quantity of data. Additionally, current integrated modules will be improved: the NEEL pipeline will be replaced by a multi-lingual and more accurate method, the web interface will include more insights such as the user network information, a heatmap visualization and a time control filter.

References

Kalina Bontcheva, Leon Derczynski, Adam Funk, Mark A Greenwood, Diana Maynard, and Niraj Aswani. 2013. Twitie: An open-source information extraction pipeline for microblog text. In *Proceedings of International Conference on Recent Advances in Natural Language Processing*, pages 83–90.

Davide Caliano, Elisabetta Fersini, Pikakshi Manchanda, Matteo Palmonari, and Enza Messina. 2016. Unimib: Entity linking in tweets using jaro-winkler distance, popularity and coherence. In *Proceedings of the 6th International Workshop on Making Sense of Microposts (# Microposts)*.

Leon Derczynski, Diana Maynard, Giuseppe Rizzo, Marieke van Erp, Genevieve Gorrell, Raphaël Troncy, Johann Petrak, and Kalina Bontcheva. 2015. Analysis of named entity recognition and linking for tweets. *Information Processing & Management*, 51(2):32–49.

Shamanth Kumar, Geoffrey Barbier, Mohammad Ali Abbasi, and Huan Liu. 2011. Tweettracker: An analysis tool for humanitarian and disaster relief. In *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media*.

Shamanth Kumar, Fred Morstatter, and Huan Liu. 2014. *Twitter data analytics*. Springer.

Gregor Leban, Blaz Fortuna, Janez Brank, and Marko Grobelnik. 2014. Event registry: learning about world events from news. In *Proceedings of the 23rd International Conference on World Wide Web*, pages 107–110.

Amit Sheth, Ashutosh Jadhav, Pavan Kapanipathi, Chen Lu, Hemant Purohit, Gary Alan Smith, and Wenbo Wang. 2014. Twitris: A system for collective social intelligence. In *Encyclopedia of Social Network Analysis and Mining*, pages 2240–2253. Springer.

Xiubo Zhang, Stephen Kelly, and Khurshid Ahmad. 2016. The slandail monitor: Real-time processing and visualisation of social media data for emergency management. In *Proceedings of the 11th International Conference on Availability, Reliability and Security*, pages 786–791.