An exploratory analysis of news trends on twitter

Konstantinos Bougiatiotis, Anastasia Krithara and George Paliouras
National Center for Scientific Research "Demokritos", Athens, Greece
{bogas.ko, akrithara, paliourg}@iit.demokritos.gr

Abstract

Analyzing social media data can be a rich supplement to the traditional reporting tools of interviews and observation. In this work we proposed a framework for analyzing and exploring twitter streams, by fusing information about the influence of the users, the topics of discussion, the relations and co-occurrences of the named entities. The framework offers expressive visualization tools, enabling users to draw useful conclusions about the data.

1 Introduction

As more people have turned to social-media platforms as a place to gather and share ideas, many journalists have been urged to use these spaces as a place to share their work and indentify important information published. Social media are part of the Big Data paradigm and are characterized by high Velocity, Veracity and Volume ("the 3 Vs"). In Twitter for example, more than 255 million active users publish over 500 million 140-character "tweets" every day[1]. Evidently it has become an important communication medium. More and more people use social media not only to communicate their ideas and thoughts, but also to spread important news. Given the enormous size of information exchange happening every day, it is a rather challenging task for journalists to process these data and filter out the important and relevant information. Analyzing web traffic and social media patterns can be a rich, and increasingly vital, supplement to the traditional reporting tools of interviews and observation. Such data can provide context and clues that also can help better frame and situate stories, as well as furnish new pathways to sources and assess popularity, importance and visibility within the online world.

The topics on Twitter span cross multiple domains from private issues to important public events in the society. Therefore, filtering out the important or relevant to the user information poses the first challenge for automated processing of tweets. For this reason one needs to deploy intelligent methods for focused analysis of all types of content, strongly based on entity and relation extraction techniques, so that information can be clustered around automatically extracted and dynamically evolving entities and relations between them. In addition, another very important issue in social network analysis is the identification of key persons in a social network.

In this work, we propose a framework for analyzing and visualizing the important information from a twitter stream. This framework is based on both natural language processing tools as well as structural analysis in order to identify the important information spread in twitter. It offers a set on expressive visualization tools which can give insights to the user about the analyzed content.

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

2 Proposed Method

2.1 General Workflow

The overall scheme of the framework described in the current work is depicted in figure 1. The main steps involved in extracting knowledge from semantically-diverse sources and fusing it into meaningful visualizations are the following:

- *Text Analysis*: Using natural language processing techniques, our aim is to extract information about important events unfolding in the tweets. To this end, we deploy a Named Entity Extraction scheme based on the assumption that the main protagonists of such events are Named Entities, like persons or organizations and a Relation Extraction module for identifying the events in the stream of news.
- *Structural Analysis*: The goal of this methodology is to pinpoint the influential users of the network, while taking into account the content of the tweets. To do so, we exploit the structural information of the network while still considering the topic content of the tweets, implementing a hybrid method based on both structure of the network and content of the interactions to rank the influence of the users.
- *Visualization Routines*: Finally, fusing information about the influence of the users, the topics of discussion, the relations and co-occurrences of the Named Entities, we can organize and combine the data into expressive visualizations, enabling us to draw conclusions or get insights regarding the data.

The aforementioned methods are explained in detail in the following sections.

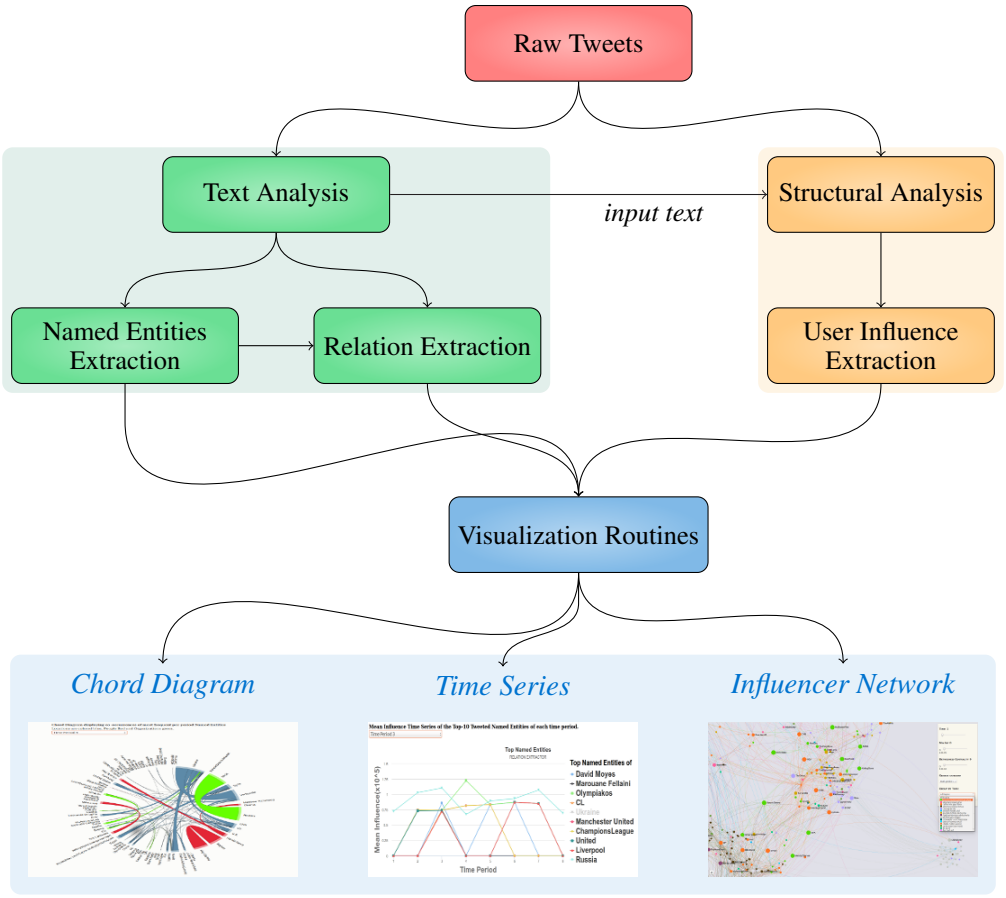

Figure 1: Proposed workflow diagram for visualizing trends in Twitter.

2.2 Text Analysis

The goal of text analysis is to tackle the task of "important" event extraction from the vast stream of Twitter data. To do so, we relied on an innovative work [10], that combines state-of-the-art methods and tools.

In particular, the outline of the pipeline used, is as follows:

1. *Preprocessing*: Cleaning the input data is crucial for accurate Named Entity Recognition and Relation Extraction. This is done through cleaning of the text from residual html tags, tokenization and user name resolution(replacing user mentions with their according twitter username, enabling us to find more relations).
2. *Named Entity Recognition*: Afterwards, we move on to find the named entities in each tweet, as well as their types. For this task, the Stanford Named Entity Recognizer(Stanford NER) [6] was used because it is reported [5] to achieve the highest average precision. We focus on named entities of type Location, Person and Organization that are much more probable to be part of "important" events, where "important" would indicate tweets containing informational value to the user, such as a politician attending a summit, or an organization making a press conference.
3. *Relation Extraction and Selection*: Subsequently, this module aims to extract meaningful relations of the form subject-predicate-object, with subject and object being among the extracted named entities. This is done using ClausIE [4], in conjunction with several modifications tailored for the task. Using the frequency of the occurrences of the named entities, we filter the extracted relations, in order to retrieve the most important news.

2.3 Structural Analysis

The core idea of this part of the work is to identify influential users on an on-line social community like Twitter. In detail, we implemented the novel method of Topic Sensitive-Supervised Random Walks(TS-SRW) [9]. This method utilizes structural information about the users/nodes and interactions/edges, as well as textual content affiliated to each node(the tweets of the user) in the network, to measure topic-sensitive influence of nodes.

To do so, a Latent Dirichlet Allocation [2] model is first used to extract the per topic distributions of each user. Then, the similarity between users based on those topic proportions is computed, as it is needed for the Supervised Random Walk [1]. Finally, exploiting the structure of the network to create weights between edges and the similarity values of the nodes

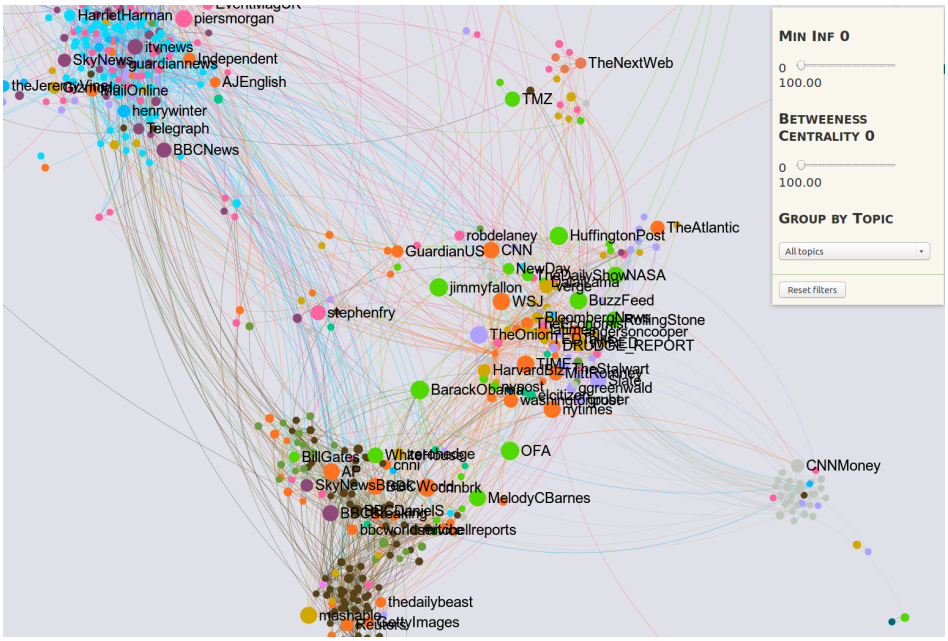

Figure 2: Example of a User Network

regarding topics, a Biased Random Walk is performed on the graph. The intuition is that the higher the topical similarity and the edge weight, the higher the transition probability will be, leading us to more influential nodes.
At the end of this procedure a PageRank-like score is computed for each node, denoting the influence of the user in the network.

3 Experiment and Visualization Tools

The previous stage of the workflow provides us with processed information that can now be combined and subsequently visualized, in order to gain insights about the data. There are many ways to use the information produced and here we will only point out a few of the visualization schemes that can augment the expressiveness of information, enhance user interaction and facilitate knowledge discovery.
In order to showcase the usefulness of the proposed workflow, we conducted an experiment applying it to the Snow 2014 Data Challenge test dataset[2]. This dataset consists of 1.089.909 tweets, by 560.009 users, resulting in 963.685 edges. The keywords used to gather the data, through the Twitter Streaming API, were *Syria*, *terror*, *Ukraine* and *bitcoin*.
As noted before, the main idea was to fuse the knowledge regarding the influence of the users with the relations of the named entities found in the tweets of those users. A few descriptive ways to convey facts about these complex interconnections of users, relations and user-influence are adumbrated below[3].

[2]publicly available at `http://figshare.com/articles/SNOW_2014_Data_Challenge/1003755`

[3]Visit `http://users.iit.demokritos.gr/~bogas.ko/` for live-interactive demos.

User Network

A multiple-aspect browsing of the data is possible through a network representation. In particular, we are interested in how the different users are connected with each other, how influential they are and what is their main topic of conversation. This can be visually expressed through a User Network. It is made feasible by utilizing the results of the Structural Analysis and is depicted in figure 2. Each node is a user, with radius proportional to their influence, color based on their main topic of interest and the links between them are interactions such as mentions, replies, retweets etc.

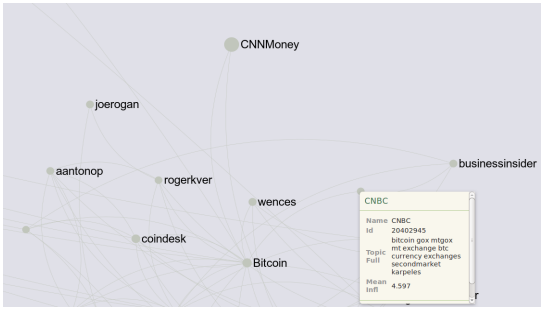

Figure 3: Network filtered by topic about finance.

This representation is very rich, allowing us to view the network from different perspectives. In an example scenario such as viral marketing or opinion propagation, one could be interested in finding the influential users regarding financial matters. We could filter the nodes according to a topic re-

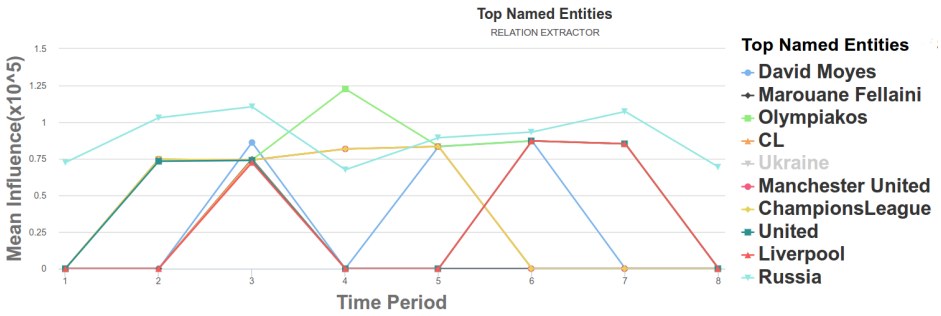

Figure 4: Time Series of the mean influence of named entities over 8 time periods.

garding finance and find the most influential users, based on their mean influence, or the users that act as connecting links between the different communities, based on their betweeness centrality [7]. An example is depicted in figure 3.
We have filtered the network to focus on users discussing about financial topics. It can be seen by the size of the nodes that *CNNMoney* is more influential in this community than *CNBC* or the *Bitcoin* profile. However, using the betweeness centrality filter, one could see that the *Bitcoin* profile is superior in terms of being the interconnecting node between different users.

This view of the users as interconnected nodes, along with filtering capabilities regarding topic, mean influence or other kinds of attributes for each user, allow for diverse and multi-purpose exploratory analysis. Others for example might be more interested in what topics are the most dominant in the network, what are the influential users talking about in a specific time period, which nodes are the main news providers, which users tend to create hubs around them and more.

Chord Diagram

A different view of the data can be offered through a Chord Diagram. A Chord Diagram is a way of exposing the inter-relationships of data [8]. In our case, we focused on the frequency of the named entities found in our corpus, as well, as the co-occurrences of the different pairs, across time. For this purpose, we calculated a co-occurrence matrix regarding all the named entities while taking into account the frequency of the named entities for different time periods. The resulting tool would create a Chord Diagram as the one shown in figure 5.

The insights gained from this type of visualization are multiple. For example, one can see over the different time periods which Named Entities were the most frequent. As expected for our example dataset, entities like Ukraine, Syria, Obama etc. are the most important ones. But looking at the diagrams, someone might be perplexed as to why Ukraine co-occurs so often with Venezuela. This unexpected link is explained by looking at some other top named entities such as *Liubov Yeremicheva*, *Yaryna Pochtarenko*, who are photographers that have taken pictures of Ukraine protesters

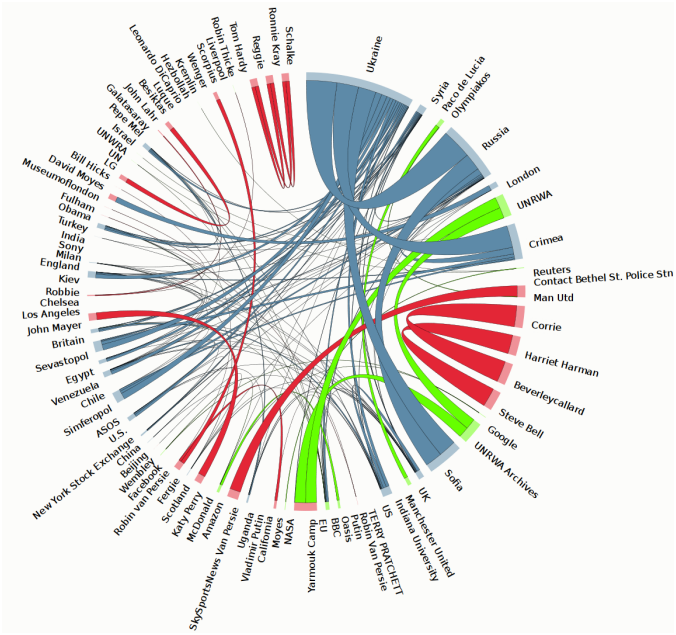

Figure 5: Example of a Chord Diagram

declaring support to their Venezuelan counterparts[4]. The Chord Diagram emphasized this connection, helping us unveil and understand the hidden link between the two events.

Time Series

A more conventional visualization scheme is a Time Series analysis of the influence of the named entities. Specifically, we focus on the top-10 most frequent named entities per time period and calculate the mean influence of each entity, according to the influence of the users that write about them. Then, we can portray how the influence of each term fluctuates between different time periods, as shown in figure 4.

Some entities have generally high influence over all time periods, such as *Russia*(light blue), others spike and disap-

[4]Protests in February, 2014 `http://ireport.cnn.com/docs/DOC-1097482`

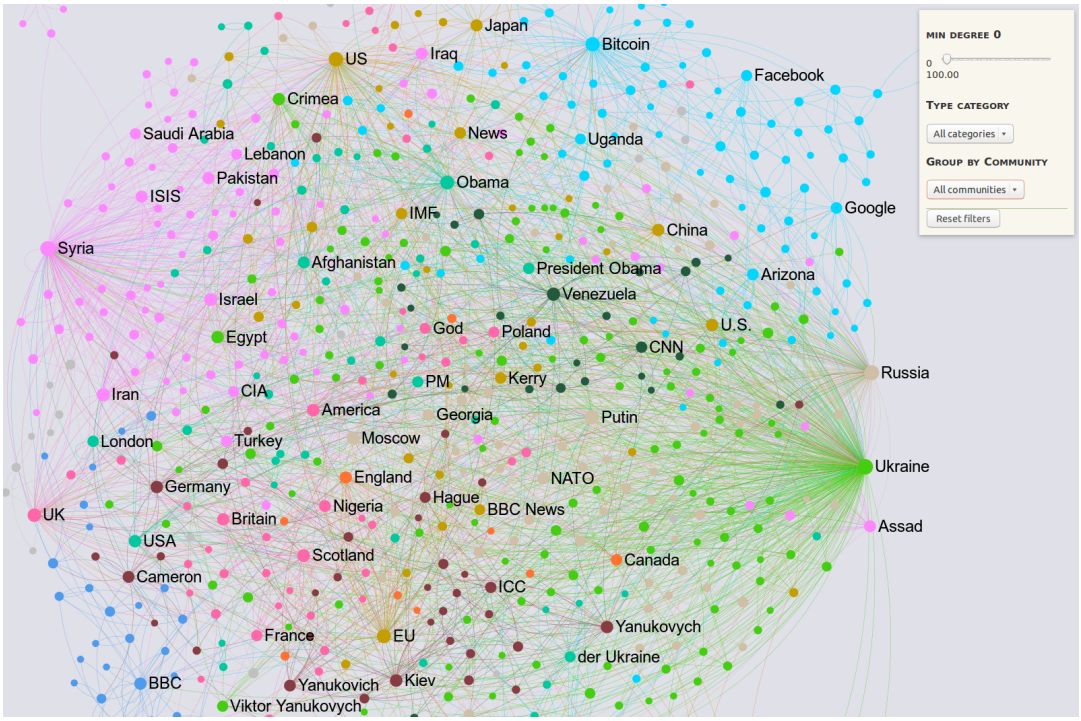

Figure 6: Example of Named Entities Network

pear, such as *Manchester United*(pink), *CL*(orange), because they were triggered by a specific event that ended (a football match).

These are very interesting in terms of finding trending keywords, events or what the topic of interest of the most influential users.

Named Entities Network

Another interesting visualization design displays the named entities found in the tweets as nodes in an interaction network. The connections between nodes denote the appearance of this pair of named entities in a tweet and the size of each node expresses the aggregated user influence of the profiles that talk about this named entity. Moreover, projecting the named entities as nodes in a plane and taking advantage of their connections allows us to find[3] communities of named entities, that is sets of named entities semantically close, based on co-occurrences. An instance of this network is shown in figure 6.

This network view at the granularity level of named entities enables the end user to understand the main topics of discussions through trending keywords at glance. Moreover, one can easily see what influential users talk about and how these named entities interlink in order to form topics of interest.Finally, one can use sophisticated filters in order to find meaningful interconnections between the entities. For example, one could filter based on the type of the named entities looking for connections of organizations or companies with persons, when investigating corruption news.

4 Conclusion and Future work

In this work we proposed a framework for analyzing and exploring twitter streams.This is done through analysis and fusion of information about the influence of the users, the topics of discussion, the co-occurrences of named entities in these topics and the interactions of the users. The results of this process are presented to the end user through expressive visualization tools, enabling users to draw useful conclusions about the data.

As future steps, we would like to incorporate more analysis tools in our framework, such as descriptive statistics regarding topics of discussion or users' interconnections. This, in accordance with new visualization tools, will provide richer information to the end users, such as ways of tracking the source of an event or the probability of it being a false rumor spread between users. Another possible future expansion would be to process the news in a streaming fashion, allowing journalists to monitor multiple story-lines at the same time. That is, journalists would define a few keywords they would like to focus on and using these tools they could delve into specific details about the diffusion of news, the evolution of a topic over time etc., gaining insights and drawing conclusions.

Acknowledgments

This work was supported by REVEAL[5] project, which has received funding by the European Unions 7th Framework Program for research, technology development and demonstration under the Grant Agreements No. FP7-610928.

[5] http://revealproject.eu/

References

[1] L. Backstrom and J. Leskovec. Supervised random walks: Predicting and recommending links in social networks. In *Proceedings of the 4th ACM WSDM Conference*, New York, NY, USA, 2011.

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3, March 2003.

[3] Vincent D Blondel, Jean loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks, 2008.

[4] Luciano Del Corro and Rainer Gemulla. Clausie: Clause-based open information extraction. In *Proceedings of the 22nd WWW Conference*, New York, NY, USA, 2013.

[5] L. Derczynski, D. Maynard, G. Rizzo, M. van Erp, G. Gorrell, R. Troncy, J. Petrak, and K. Bontcheva. Analysis of named entity recognition and linking for tweets. *Inf. Process. Manage.*, 51(2), 2015.

[6] Jenny Rose Finkel, Trond Grenager, and Christopher Manning. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd Annual Meeting on ACL*, 2005.

[7] Linton C. Freeman. A set of measures of centrality based on betweenness. *Sociometry*, 1977.

[8] D. Holten. Hierarchical edge bundles: Visualization of adjacency relations in hierarchical data. *IEEE Transactions on Visualization and Computer Graphics*, 12(5), 2006.

[9] G. Katsimpras, D.s Vogiatzis, and G. Paliouras. Determining influential users with supervised random walks. In *Proceedings of the 24th WWW Conference*, New York, NY, USA, 2015.

[10] G Katsios, S Vakulenko, A Krithara, and G Paliouras. Towards open domain event extraction from twitter: Revealing entity relations. In *Proceedings of the 4th DeRiVE workshop, 12th ESWC 2015, Slovenia*, 2015.