A Platform for Supporting Data Analytics on Twitter: Challenges and Objectives[1]

Yannis Stavrakas Vassilis Plachouras

IMIS / RC "ATHENA"
Athens, Greece

{yannis, vplachouras}@imis.athena-innovation.gr

Abstract. An increasing number of innovative applications use data from online social networks. In many cases data analysis tasks, like opinion mining processes, are applied on platforms such as Twitter, in order to discover what people think about various issues. In our view, selecting the proper data set is paramount for the analysis tasks to produce credible results. This direction, however, has not yet received a lot of attention. In this paper we propose and discuss in detail a platform for supporting processes such as opinion mining on Twitter data, with emphasis on the selection of the proper data set. The key point of our approach is the representation of term associations, user associations, and related attributes in a single model that also takes into account their evolution through time. This model enables flexible queries that combine complex conditions on time, terms, users, and their associations.

Keywords: Social networks, temporal evolution, query operators.

1 Introduction

The rapid growth of online social networks (OSNs), such as Facebook or Twitter, with millions of users interacting and generating content continuously, has led to an increasing number of innovative applications, which rely on processing data from OSNs. One example is opinion mining from OSN data in order to identify the opinion of a group of users about a topic. The selection of a sample of data to process for a specific application and topic is a crucial issue in order to obtain meaningful results. For example, the use of a very small sample of data may introduce biases in the output and lead to incorrect inferences or misleading conclusions. The acquisition of data from OSNs is typically performed through APIs, which support searching for keywords or specific user accounts and relationships between users. As a result, it is not straightforward to select data without having an extensive knowledge of related keywords, influential users and user communities of interest, the discovery of which is a manual and time-consuming process. Selecting the proper set of OSN data is important not only for opinion mining, but for data analytics in general.

[1] This work is partly funded by the European Commission under ARCOMEM (ICT 270239).

In this paper, we propose a platform that makes it possible to manage data analysis campaigns and select relevant data from OSNs, such as Twitter, based not only on simple keyword search, but also on relationships between keywords and users, as well as their temporal evolution. Although the platform can be used for any OSN having relationships among users and user posts, we focus our description here on Twitter. The pivotal point of the platform is the model and query language that allow the expression of complex conditions to be satisfied by the collected data. The platform models both the user network and the generated messages in OSNs and it is designed to support the processing of large volumes of data using a declarative description of the steps to be performed. To motivate our approach, in what follows we use opinion mining as a concrete case of data analysis, however the proposed platform can equally support other analysis tasks. The platform has been inspired by work in the research project ARCOMEM[2], which employs online social networks to guide archivists in selecting material for preservation.

2 Related Work

There has been a growing body of work using OSN data for various applications. Cheong and Ray [5] provide a review of recent works using Twitter. However, there are only few works that explore the development of models and query languages for describing the processing of OSN data. Smith and Barash [4] have surveyed visualization tools for social network data and stress the need for a language similar to SQL but adapted to social networks. San Martín and Gutierrez [3] describe a data model and query language for social networks based on RDF and SPARQL, but they do not directly support different granularities of time. Mustafa et al. [2] use Datalog to model OSNs and to apply data cleaning and extraction techniques using a declarative language. Doytsher et al. [1] introduced a model and a query language that allow to query with different granularities for frequency, time and locations, connecting the social network of users with a spatial network to identify places visited frequently by users. However, they do not consider any text artifacts generated by users (e.g. comments posted on blogs, reviews, tweets, etc.).

The platform we propose is different from the existing works in that we incorporate in our modeling the messages generated by users of OSNs and temporally evolving relationships between terms, in addition to relationships between users. Moreover, we aim to facilitate the exploration of the context of data and enable users to detect associations between keywords, users, or communities of users.

3 Approach and Objectives

We envisage a platform able to adapt to a wide spectrum of thematically disparate opinion mining campaigns (or data analysis tasks in general), and provide all the in-

[2] http://www.arcomem.eu/

frastructure and services necessary for easily collecting and managing the data. This platform is depicted in Fig. **1**, and comprises three layers.

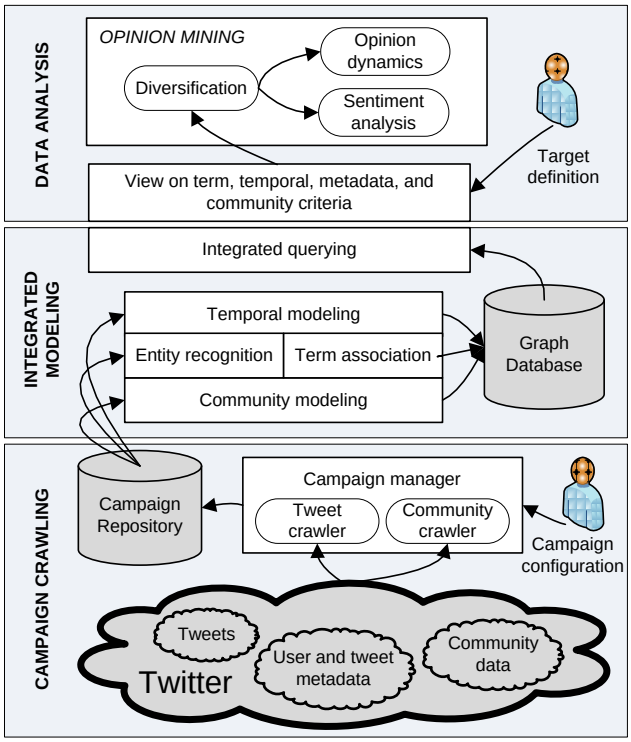

Fig. 1: Platform architecture

The first layer is the *Campaign Crawling* layer. The objective of this layer is to allow for the definition and management of *campaigns*, and to collect all the relevant "raw" data. A campaign is defined by a set of filters and a timespan. For each campaign, the platform monitors the Twitter stream for the duration specified by the campaign's timespan, and incrementally stores in the *Campaign Repository* all the data that match the campaign filters. Stored data fall into three categories: tweets, metadata, and community data. Community data describe the relationships among Twitter users, while metadata may refer either to tweets (timestamp, location, device, language, etc.) or to users (place, total activity, account creation date, etc.). Selecting the correct tweets for a campaign is of paramount importance; therefore, for converging to the proper filters the process of *campaign configuration* follows an iterative approach: a preliminary analysis on an initial small number of tweets gives an overview of the expected results, indicates the most frequent terms, and highlights the most influential users, in order to verify that the campaign is configured correctly. If this is not the case, an adjustment step takes place by modifying the terms that tweets are expected to contain, the user accounts deemed most relevant to the topic of the campaign, or by removing tweets from user accounts that have been identified to be ro-

bots. Those modifications are based on suggestions made by the platform according to the preliminary analysis of tweets.

The second layer is the *Integrated Modeling* layer. The objective of this layer is to support a complex and flexible querying mechanism on the campaign data, allowing the definition of *campaign views*. The significance of campaign views will become apparent in a motivating example that will follow. A prerequisite for querying is the modeling of the campaign "raw" data. The model should encompass three dimensions. First, represent associations between interesting terms contained in tweets. Such terms can be hashtags, or the output of an entity recognition process. Associations between terms can be partly (but not solely) based on co-occurrence of the terms in tweets. Second, represent associations between users of varying influence forming communities that discuss distinct topics. Finally, it should capture the evolution across time of the aforementioned associations, term attributes, and user attributes. This temporal, often overlooked, dimension is of paramount importance in our approach since it enables the expression of time-aware conditions in queries. The resulting model integrates all three dimensions above in a unified way as a graph. A suitable query language is then used to select the tweets that satisfy conditions involving content, user, and temporal constraints.

The third layer is the *Data Analysis* layer. The query language is used to define a "target" view on the campaign data. This "target" view corresponds to a set of tweets that hopefully contain the answer to an opinion-mining question. This set is then fed into a series of analysis processes, like *diversification* for ensuring the correct representation of important attribute values, and *sentiment analysis* for determining the attitude of users towards the question. *Opinion dynamics* is important for recognizing trends across time.

Motivating example. A marketing specialist wants to learn what people say in Twitter about Coca-Cola: (a) in cases when Pepsi-Cola is also mentioned, and (b) during the Olympic Games. The first step is to define a campaign. He launches a preliminary search with the keywords *coca cola*. An initial analysis of the first results reveals other frequent terms and hashtags, and after reviewing them he decides to include the following in the campaign definition: *#cc*, *#cola* and *coke*. Moreover, the platform suggests groups of users whose tweets contain most often the relevant keywords. He decides to include some of them in the campaign definition. Having set the campaign filters, he sets the campaign timespan, and launches the campaign. The crawler downloads data periodically, which are incrementally stored in the Campaign Repository and modeled in the Graph Database. The next important step for our marketing specialist is to define suitable "targets" that correspond to his initial questions. He does that by using the query language of the platform for creating views on the campaign data. An intuitive description of the queries for the two cases follows.

The first query returns the tweets that will hopefully reveal what people say about Coca-Cola in cases when Pepsi is also mentioned, using the following steps: 1) find the terms that are highly associated with Pepsi Cola, 2) return the tweets in which Coca- and Pepsi-related terms are highly associated.

The second query return the tweets that will hopefully reveal what people say about Coca-Cola during the Olympics in the following steps: 1) find the terms that are

highly associated with Olympic Games, 2) find the time periods during which those terms are most often encountered, 3) find the groups of people that most often use those terms during the specific time periods, 4) return the tweets of those people, during the specified time periods, which mention the Coca-Cola search terms.

The final step is to conduct opinion-mining analysis on the two sets of tweets returned by the queries. In our approach the emphasis is on selecting the most suitable set of tweets, according to the question we want to answer. In our view, selecting the proper set is paramount for opinion mining processes to produce credible results.

4 Technical and Research Challenges

There are several technical and research challenges that need to be addressed in the implementation of the proposed platform, with respect to the acquisition of data, as well as the modeling and querying of data.

Scalable Crawling. In the case that the platform handles multiple campaigns in parallel, there is a need to optimize the access to the OSN APIs, through which data is made available. Typically, APIs have restrictions in the number of requests performed in a given time span. The implementation of the platform should aim to minimize the number of API requests while fetching data for many campaigns in parallel. Hence, an optimal crawling strategy is required to identify and exploit overlaps between campaigns and to merge the corresponding API requests.

Temporal Modeling. A second challenge is the modeling of large-scale graphs, where both nodes and edges have temporally evolving properties. Our approach is to treat such graphs as directed multigraphs, with multiple timestamped edges between two vertexes [6]. Given that the scale of the graphs can be very large both in terms of the number of vertexes and edges, but also along the temporal dimension, it is necessary to investigate efficient encoding and indexing for vertexes and edges, as well as their attributes. We have collected a set of tweets over a period of 20 days using Twitter's streaming API. In this set of tweets, we have counted a total of 2,406,250 distinct hashtags and 3,257,760 distinct pairs of co-occurring hashtags. If we aggregate co-occurring pairs of hashtags per hour, then we count a total of 5,670,528 distinct pairs of co-occurring hashtags. Note that the sample of tweets we have collected is only a small fraction of the total number of tweets posted on Twitter. If we can access a larger sample of tweets and consider not only hashtags but also plain terms, then the corresponding graphs will be substantially larger.

Advanced Querying. A third challenge is the definition of querying operators that can be efficiently applied on temporally evolving graphs. Algorithms such as PageRank, which compute the importance of nodes in a graph, would have to be computed for each snapshot of the temporally evolving graph. While there are approaches that can efficiently apply such algorithms on very large graphs [7], they do not consider the temporal aspect of the graphs that is present in our setting. Overall, the implementation of querying operators should exploit any redundancy or repetition in the temporally evolving graphs to speed-up the calculations.

Data Analysis. The proposed platform allows users to define a body of tweets based on complex conditions ("target definition" in Fig. **1**), in addition to the manual selection of keywords or hashtags. Since the quality of any analysis process is affected by the input data, a challenge that arises is the estimation of the bias of the results with respect to the input data.

5 Conclusions and Future Work

In this paper we have proposed and described in detail a platform for supporting data analytics tasks, such as opinion mining campaigns, on online social networks. The main focus of our approach is not on a specific analysis task per se, but rather on the proper selection of the data that constitute the input of the task. In our view, selecting the proper data set is paramount for the analytics task to produce credible results. For this reason we have directed our work as follows: (a) we are currently implementing a campaign manager for crawling Twitter based on highly configurable and adaptive campaign definitions, and (b) we have defined a preliminary model and query operators [6] for selecting tweets that satisfy complex conditions on the terms they contain, their associations, and their evolution and temporal characteristics. Our next steps include the specification and implementation of a query language that encompasses the query operators mentioned above, and the integration of the implemented components into the platform we described in this paper.

References

1. Y. Doytsher, B. Galon, and Y. Kanza. Querying geo-social data by bridging spatial networks and social networks. 2nd ACM SIGSPATIAL International Workshop on Location Based Social Networks, LBSN '10, pages 39-46, New York, NY, USA, 2010.
2. W. E. Moustafa, G. Namata, A. Deshpande, and L. Getoor. Declarative analysis of noisy information networks. 2011 IEEE 27th International Conference on Data Engineering Workshops, ICDEW '11, pages 106-111, Washington, DC, USA, 2011.
3. M. San Martin and C. Gutierrez. Representing, querying and transforming social networks with rdf/sparql. 6th European Semantic Web Conference on The Semantic Web: Research and Applications, ESWC 2009 Heraklion, pages 293-307.
4. M. A. Smith and V. Barash. Social sql: Tools for exploring social databases. IEEE Data Eng. Bull., 31(2):50-57, 2008.
5. M. Cheong and S. Ray. A Literature Review of Recent Microblogging Developments. Technical Report TR-2011-263, Clayton School of Information Technology, Monash University, 2011.
6. V. Plachouras and Y. Stavrakas. Querying Term Associations and their Temporal Evolution in Social Data. 1st Intl VLDB Workshop on Online Social Systems (WOSS 2012), Istanbul, August 2012.
7. U. Kang, D.H. Chau, C. Faloutsos: Mining large graphs: Algorithms, inference, and discoveries. ICDE 2011: 243-254.