Interoperability of Social Media Observatories

Karissa McKelvey and Filippo Menczer

Center for Complex Networks and Systems Research
School of Informatics and Computing
Indiana University, Bloomington
cnets.indiana.edu

Abstract. With the broad adoption of social media and micro-blogging platforms such as Twitter, we can observe high-volume data streams of online discourse. Social media observatories, bearing qualitative resemblance to observatories used in the natural and physical sciences, are web-based applications that facilitate the filtering, searching, and observation of this social media data for research purposes. However, best practices for architecture and interoperability are not well defined for building such systems, even for experts in the computational sciences. We outline a road map for Truthy, a social media observatory that has been collecting, filtering, and sharing social media data since August 2010. Our focus is on an architecture that will support querying of historical data and interoperability with other web observatories.

Keywords: Research Data Management; Web Observatory; Social Media Observatory; Web Science

1 Introduction

Web science and computational social science [6] are advancing our understanding of behavioral [4], political [3], and economic [1] phenomena. This has created a spike in interest toward the use of "big data" about people and their interactions as a primary source for interesting questions in these fields [2]. However, this data is often collected with custom toolkits, locked up in various archives, and inaccessible to the public. This emerging trend makes web science research expensive and difficult, if not impossible, to reproduce. *Web observatories* have been proposed as a solution to these problems [9]. Social media have become a major source of data about online social interactions, and therefore we see a *social media observatory* as a key first step toward this goal.

Here we describe our plan to construct a new architecture for the Truthy project, that will support retrieval of historical social media data and statistics through an open standard Application Programming Interface (API). Interoperability between datasets from different social media sources requires common data formats, and — for very large datasets or streaming data — open APIs. This approach has many benefits, including reproducibility and efficiency of web science research. The creation of open standards will require collaboration among

web observatories. For example, the Dataverse Network Project (`thedata.org`) is an observatory that reformats research data with open standards, exchange protocols, and citable digital libraries [5]. One of our future goals is to bring the same type of open standards to social media observatories.

2 The Truthy Project

The Truthy Project (`truthy.indiana.edu`) is a web application originally motivated by the need to detect astroturf, or false grassroots campaigns, in microblogging streams [8]. We have collected a continued stream of compressed data from Twitter since August 2010. The data source is a random sample of public messages obtained via the Twitter Streaming API (`dev.twitter.com/docs/streaming-apis`). In recent extensions, we have created interfaces that allow users to visualize information flows and download analytical data derived from Twitter posts about themes such as politics, social movements, and news [7]. The current infrastructure limits our analytics services to data collected in the previous nine months.

Truthy's data model leverages a unified framework to represent the behavior of users and diffusion of memes. We model a generic stream of social networking data as a series of events that represent interactions between *users* and *memes*. Each event involves some number of users, some number of memes, and interactions among them. For example, a single tweet event might involve three or more actors: the poster, the user she is retweeting, and the people she is addressing. The post might also involve a set of memes consisting of, say, hashtags and URLs referenced in the tweet. Each event can be thought of as contributing a unit of weight to edges in a network structure, where nodes are associated with either users or memes. The timestamps associated with the events allow us to observe the changing structure of this network over time. Our API operates upon the event data model. The specification of methods dealing with memes, users, networks, and timelines can be found online (`truthy.indiana.edu/apidoc`).

3 Future Steps

We have two key goals for our planned social media observatory. First, the common data format based on social media events will allow our system to integrate different data sources. Second, it is important for the computational infrastructure to scale up to analysis of all posts and their metadata since August 2010, with the capability to query this historical data repository with user-defined parameters. We plan to deploy distributed database software such as Solandra or HBase on a new computational cluster, enabling parallelized access to the data through a search API.

Future steps will focus on the creation of a central observatory to store, visualize, and deliver social media data. In Figure 1 we outline an architecture design that utilizes the common data model for social media data. Users will upload data modified to fit the common data standard.

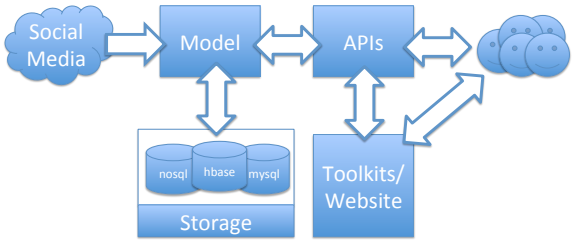

Fig. 1. Architecture of the proposed social media observatory system.

The infrastructure will consist of several components: input adapters for importing external social media data into the data model; toolkits for computing various analytics, such as graph layout and visualization algorithms; an API for programmatic access to derived data and images; and a website for visualizing analysis, saving statistical results, and supporting interactive web tools.

We hope that our project will provide a useful framework to promote interoperability between the many different sources of social media data, both static and streaming. This will be an important first step toward the realization of powerful web observatories.

Acknowledgments. This work is supported in part by NSF grant CCF-1101743. Thanks also to the members of the Truthy group at Indiana University (`cnets.indiana.edu/groups/nan/truthy`) for their many contributions to the Truthy Project.

References

1. J. Bollen, H. Mao, and X. X.-J. Zeng. Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1):1–8, Mar. 2011.
2. D. Boyd and K. Crawford. Six Provocations for Big Data. *SSRN Electronic Journal*, 1926431, 2011.
3. M. D. Conover, J. Ratkiewicz, B. Gonçalves, A. Flammini, and F. Menczer. Political polarization on twitter. In *Proc. 5th Intl. AAAI Conf. on Weblogs and Social Media (ICWSM)*, 2011.
4. N. Eagle, A. S. Pentland, and D. Lazer. Inferring friendship network structure by using mobile phone data. *Proc. Nat. Acad. Sci.*, 106(36):15274–15278, 2009.
5. G. King. An introduction to the dataverse network as an infrastructure for data sharing. *Sociological Methods & Research*, 36(2):173–199, 2007.
6. D. Lazer, A. Pentland, L. Adamic, S. Aral, A.-L. Barabási, D. Brewer, N. Christakis, N. Contractor, J. Fowler, M. Gutmann, T. Jebara, G. King, M. Macy, D. Roy, and M. Van Alstyne. Computational social science. *Science*, 323(5915):721–723, 2009.
7. K. McKelvey, A. Rudnick, M. Conover, and F. Menczer. Visualizing Communication on Social Media: Making Big Data Accessible. In *Proc. 15th ACM CSCW Workshop on Collective Intelligence as Community Discourse and Action*, 2012.
8. J. Ratkiewicz, M. Conover, M. Meiss, B. Gonçalves, S. Patil, A. Flammini, and F. Menczer. Truthy: Mapping the spread of astroturf in microblog streams. In *Proc. 20th Intl. World Wide Web Conf. Companion (WWW)*, 2011.
9. T. Tiropanis, W. Hall, N. Shadbolt, D. D. Roure, N. Contractor, and J. Hendler. The web science observatory. *IEEE Intelligent Systems*, 2013. In press.