Title	TwitterCracy: Exploratory Monitoring of Twitter Streams for the 2016 U.S. Presidential Election Cycle
Authors(s)	Qureshi, M. Atif; Arjumand, Younus; Greene, Derek
Publication date	2016-09-23
Publication information	Proceedings, Part III: Machine Learning and Knowledge Discovery in Databases European Conference, ECML PKDD 2016 (Volume 9853)
Conference details	European Conference on Machine Learning and Principles and Practice of Knowledge þÿDiscovery in Databases (ECML-PKDD 16), Riva del Garda,
Series	Lecture Notes in Computer Science
Publisher	Springer
Item record/more information	http://hdl.handle.net/10197/8367
Publisher's statement	The final publication is available at Springer via http://dx.doi.org/10.1007/978-3-319-46131-1_16.
Publisher's version (DOI)	10.1007/978-3-319-46131-1_16

TwitterCracy: Exploratory Monitoring of Twitter Streams for the 2016 U.S. Presidential Election Cycle

M. Atif Qureshi (✉), Arjumand Younus, and Derek Greene

Insight Center for Data Analytics, University College Dublin,
Dublin, Ireland
{muhammad.qureshi@ucd.ie, arjumand.younus@ucd.ie, and
derek.greene@ucd.ie}

Abstract. We present *TwitterCracy*, an exploratory search system that allows users to search and monitor across the Twitter streams of political entities. Its exploratory capabilities stem from the application of lightweight time-series based clustering together with biased PageRank to extract facets from tweets and presenting them in a manner that facilitates exploration.

1 Introduction

Twitter has established itself as an important medium for online political discourse, as evidenced during events such as the Arab Spring, Barack Obama's 2012 presidential campaign, and India's General Elections in 2014. This has subsequently led to the increased usage of the platform by politicians as a part of their campaign activities [4, 6]. Following this trend, Fortune Magazine has termed the 2016 U.S Presidential Election as the *"social media election"* [1]. The research community has experienced a surge of interest in the analysis of political chatter over Twitter [5]. Much of the current focus lies in the prediction of election outcomes, with relatively few state-of-the-art studies [3, 8] conducted on the analysis of political discussion by general users [5]. Despite the attention given to election predictions in the literature [9], such methods fail to empower the general public in the spirit of "democracy" and "voter empowerment".

The rising prominence of social media as a platform for political discourse has fundamentally altered the way in which candidates conduct election campaign [6]. It is therefore necessary for voters, analysts, and journalists to keep a close eye on the online activity of politicians. We believe such monitoring can help to increase political awareness among the general public, thereby enabling them to make informed choices in electing their representatives. This, in turn, dictates a clear need for analytical tools that can delve into the communication behaviors of politicians on social media.

Towards this end, we have created *TwitterCracy*, a system which aims to facilitate voters and analysts, by keeping them aware of the key agenda issues that are of interest to politicians, as reflected by their ongoing activity on Twitter. The core functionality of the system enables the exploration of various facets of these issues, via the extraction of keywords from politicians' tweets. Our technique for exploratory analysis is based on the application of biased PageRank [2] to a graph of terms, mentions, and hashtags appearing in tweets. In line with the *TweetMotif* tool [7], our system allows a user to

navigate via the extracted keywords and drill down into the data in more depth. However, unlike *TweetMotif*, which only operates on a static corpus, *TwitterCracy* indexes a live stream of tweets and extracts query-specific facets in real-time, while incorporating a light-weight time-series clustering mechanism for the efficient application of the PageRank model. Another novel aspect of *TwitterCracy* is the incorporation of valuable metrics based on theoretical constructs within relational sociology [3] to provide deeper insights into the communication patterns of politicians. To illustrate the use of *TwitterCracy*, we consider the 2016 U.S. Presidential Election as a case study, analyzing the activity of 635 relevant politicians and political organizations on Twitter during the campaign. click

2 TwitterCracy Architecture

In this section, we present an overview of the architecture of *TwitterCracy*, as illustrated in Figure 1. The user, who is central to the system, issues a "query"[1], which is processed by the query module to produce a ranked list of relevant tweets. This ranked list then passes through various components of our processing pipeline: 1) clustering and compression module, 2) facet extraction module, 3) social extraction module and finally, 4) rendering module. We now explain the first three modules in the following sub-sections, as these represent the key system components, while the rendering module simply produces the HTML output. Separately, the crawler module is responsible for back-end data acquisition, continuously collecting from the live stream of politicians' tweets and matching them with the user metadata. This data stream is immediately indexed to provide the user with real-time updates.

2.1 Key Components

Clustering and compression module: This module is responsible for reducing the large, dense graph of terms, mentions, and hashtags into a relatively small, sparse graph for efficient computation of PageRank. First, we apply cost-effective, time-series based clustering to the ranked list of tweets. Based on the assumption that bursts of tweets are likely to indicate significant events [10], we apply k-means clustering over the timestamps of the retrieved tweets to cluster bursts of tweets together. From these clusters, we then pick the top retrieved tweets, in proportion to the size of each cluster. This reduces the full stream to a representative sub-sample of tweets prior to the application of PageRank in the next stage of the processing pipeline.

Facet extraction module: This module extracts various facets[2] from the retrieved tweets by applying biased PageRank. In the graph, the nodes are terms extracted from retrieved tweets, and edges connect pairs of terms that occur together in a tweet. The weight on an edge is the relevance score of the tweet relative to the original query. The biasing of PageRank vector is explained as follows:

[1] Note a query can be a phrase entered by the user or the live stream depicting last 15 minutes.
[2] Facets here are keywords, mentions and hashtags.

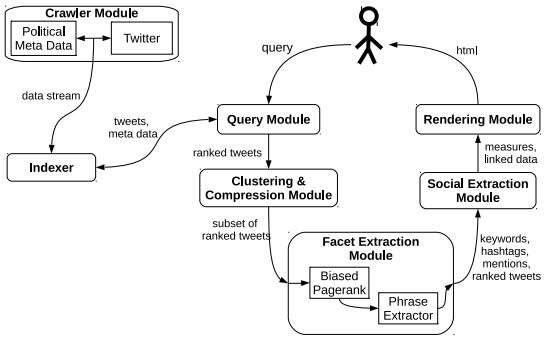

Fig. 1. Architecture of the *TwitterCracy* system.

- The terms in retrieved tweets are biased in proportion to the amount of their significance calculated by chi-square test of independence.
- The named entities in retrieved tweets are biased in proportion to their correlation with an event where the correlation is calculated by means of their document frequencies in retrieved tweets.

Finally, we merge single terms identified by biased PageRank to extract longer keywords as facets[3]. To achieve this, we add the individual PageRank scores of the co-occurring terms according to their probability of co-occurrence. This means that sets of terms with high PageRank scores and that co-occur frequently are extracted as facets, and appear in the exploratory search interface (see Figure 2)

Social extraction module: This module applies theoretical measures from relational sociology to quantify various aspects of online conversational practices of politicians. More specifically, we make use of three measures introduced by Lietz et al. [3]: cultural similarity, cultural focus, and cultural reproduction. The level of similarity between the stances of political parties (e.g. Democrats and Republicans) in relation to various issues is measured by means of cultural similarity. The stability of a political party's ideology can be quantified by both cultural focus and cultural reproduction.

3 Case Study: 2016 U.S. Presidential Election

To illustrate the use of *TwitterCracy*, we consider the 2016 U.S. Presidential Election as a case study, analyzing the activity of 635 relevant politicians and political organizations on Twitter during the campaign. The dataset contains 1,473,514 number of tweets (from 3 June 2008 to 11 May 2016) and it is still growing. A video demonstrating the system can be accessed at `http://mlg.ucd.ie/twittercracy`. A query such as "guns" can reveal significant insights (see Figure 2): we observe the low level of cultural similarity between parties, while aspects like "gun sales", "gun violence", and "gun legislation" highlight various facets within this topic which the user can navigate

[3] Note that we restrict this extraction to bigrams as tweets are short.

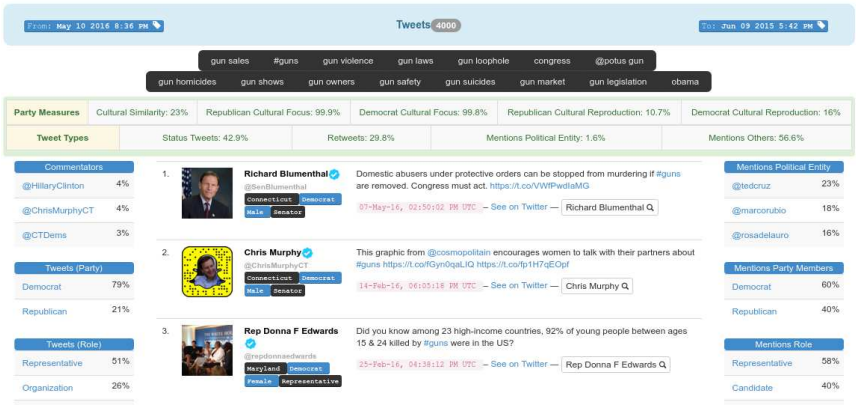

Fig. 2. *TwitterCracy* user interface showing results for a sample query "guns". Identified facets include "gun violence" and "gun legislation", which can be explored in more detail.

for further exploration. Together with the various insights from theoretical measures, these facets help uncover various issues of U.S. politics that may concern the voter. Three further examples are: 1) the different facets evident between the parties for the query "abortion", 2) the high level of cultural similarity between parties on matters of foreign policy, such as "Israel" and "Syria", 3) the low level of cultural similarity between parties on matters of domestic policy such as "drugs".

Acknowledgments: This publication has emanated from research conducted with the support of Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289.

References

1. This is why social media will decide the 2016 election. http://fortune.com/2015/12/01/social-media-2016-election/, December 2015.
2. T. H. Haveliwala. Topic-sensitive pagerank. In *WWW*, 2005.
3. H. Lietz, C. Wagner, A. Bleier, and M. Strohmaier. When politicians talk: Assessing online conversational practices of political parties on twitter. In *ICWSM*, 2014.
4. Y. Mejova, P. Srinivasan, and B. Boynton. Gop primary season on twitter: popular political sentiment in social media. In *WSDM*, 2013.
5. Y. Mejova, I. Weber, and M. W. Macy. *Twitter: a digital socioscope*. Cambridge University Press, 2015.
6. P. Nulty, Y. Theocharis, S. A. Popa, O. Parnet, and K. Benoit. Social Media and Political Communication in the 2014 Elections to the European Parliament. 2015.
7. B. O'Connor, M. Krieger, and D. Ahn. Tweetmotif: Exploratory search and topic summarization for twitter. In *ICWSM*, 2010.
8. E. J. Schweitzer. Normalization 2.0: A longitudinal analysis of german online campaigns in the national elections 2002–9. *European Journal of Communication*, 26(4):310–327, 2011.
9. A. Tumasjan, T. O. Sprenger, P. G. Sandner, and I. M. Welpe. Predicting elections with twitter: What 140 characters reveal about political sentiment. *ICWSM*, 10:178–185, 2010.
10. J. Weng and B.-S. Lee. Event detection in twitter. In *ICWSM*, 2011.