Political Polarization of Web Search Queries and Hashtags

Ingmar Weber
Qatar Computing Research Institute

What can user-generated online data tell us about political issues and partisan politics? How polarized is a web search query such as "climategate 2.0" and can we quantify the leaning of the hashtag #climatechange?

We summarize recent work that studies U.S. politics using web search logs and Twitter data from the angle of political left-vs.-right polarization. With a simple yet effective methodology we discover a number of interesting findings. For example, we show a tendency for web search queries with a right leaning to surface more results with a negative sentiment, and we give a description of political hashtag hijacking where one political camp tries to "jam the frequency" of the other one.

We end with a discussion of challenges and opportunities in the area of data-driven political science that might be of interest to researchers looking to apply knowledge management in an interdisciplinary setting.

DOI: 10.1145/2501187.2501191 http://doi.acm.org/10.1145/2501187.2501191

1. INTRODUCTION

From the perspective of many countries the political system in the United States of America is unusual because of its simple two party nature. Whereas countries as diverse as Switzerland, India, Japan, Italy, Canada or recently the United Kingdom have at least occasionally had governing multi-party coalitions in power[1], U.S. politics is defined by a one-dimensional Democrats vs. Republicans, left vs. right[2] bipolar configuration. Through data-driven approaches the degree to which a political system is indeed one-dimensional in a mathematical sense can be quantified using Poole and Rosenthal's seminal work [Poole and Rosenthal 1985]. Observing roll call voting behavior and applying dimensionality reduction techniques, they observed that "a single dimension accounts for about 93 percent of roll call voting choices in the 110th House and Senate"[3].

Given the importance of political opinion formation, researchers have over the last years turned to online data to study the phenomenon of polarization at scale, not only for politicians but for engaged citizens. Adamic et al. [Adamic and Glance 2005] observed two

[1] Switzerland has been governed by a coalition involving all major parties for over 50 years.
[2] For simplicity, we equate the political left with Democrats and the political right with Republicans.
[3] http://voteview.com/polarized_america.htm

strongly separated communities in the political blogosphere, with hyperlinks rarely crossing ideological boundaries. A similar pattern was observed by Conover et al. [Conover et al. 2011] for the case of Twitter, where users would be unlikely to retweet others with a different party affiliation.

In this article, we present a summary of our work on the polarization of (i) web search queries and (ii) hashtags. Using user-generated online data to track political issues and their "leaning" over time not only opens up new possibilities for computational political science but, quite frankly, can also be rather entertaining. The interested reader is invited to experiment with `http://politicalsearchtrends.sandbox.yahoo.com` [Weber et al. 2012b] and `http://politicalhashtagtrends.sandbox.yahoo.com` [Weber et al. 2013] for a data-driven story line of U.S. partisan politics.

In the rest of this article, we will first introduce our methodology used to assign a political leaning to short pieces of text (Section 2). The main part then describes a number of findings arrived at using this approach (Section 3). Finally, we will discuss a number of challenges and promising avenues for future research in this area.

2. METHODOLOGY

Our approach to assign a political leaning to short bits of text, either web search queries or hashtags coming from an ever evolving dictionary is the following. We start with easy-to-establish labels for objects of a different type and then propagate these labels to the bits of text. This is explained for the two cases separately. In both cases, we used mostly data from 2011 and early 2012 for the analyses [Borra and Weber 2012; Weber et al. 2012a; Teka et al. 2013], whereas the online demos visualize data up to late 2012.

2.1 From Blogs to Queries

Though assigning a leaning to a short, general web search query is hard, labeling longer web pages is considerably easier. In particular, partisan blogs, where the authors express their political views, can be relatively easily classified by a human into left or right. Such annotations enabled Adamic et al.'s [Adamic and Glance 2005] study of the interlinkage patterns between the two camps. For our work we used the set of 155 annotated blogs from Shaw and Benkler [Shaw and Benkler 2012] and combined this list with the Wonkosphere Blog Directory[4].

Queries are then annotated according to which political blogs (if any) are clicked in response to them. For example, the query "climate change" predominantly leads to clicks on left-leaning sites, whereas the query "climategate 2.0" attracts mostly clicks on right-leaning sites. Note that such query-click pairs combine demand (= what is searched for) with supply (= what is being covered on a blog). Details are given in Section 2.3. This where-does-a-query-end-up leaning is aggregated on a weekly basis and is combined with basic trend detection [Subasic and Castillo 2010] to obtain an algorithmic summary of current partisan issues.

To validate that the leaning of a political blog does indeed propagate through the click,

[4] `http://wonkosphere.com/directory.htm`

we used users' ZIP code information and election results from the 2010 midterm elections. Concretely, we showed that users clicking on left (resp. right) blogs in response to web search queries were more likely than random chance to live in a ZIP code that voted Democrat (resp. Republican) in the 2010 midterm elections [Borra and Weber 2012].

2.2 From Politicians to Hashtags

For Twitter, we start with a list of verified accounts belonging to political parties or political leaders such as `@BarackObama` or `MittRomney`. In total, we tracked 14 seed accounts for the left and 19 for the right. The ones with the most followers were Barack Obama and Nancy Pelosi (left) vs. Mitt Romney and Newt Gingrich (right). For each of our seed users we obtained their publicly available tweets and, for each tweet, identified up to 100 retweeters. The combined set of retweeters formed the basis for our analysis. Removing Twitter users from non-U.S. locations left us with 111,813 users to track.

For each week, these users are assigned a *fractional* leaning corresponding to the ratio of their retweets of either left or right seed users. The basic rationale for this is that retweeting in the political domain is a sign of political agreement [Conover et al. 2011].

To obtain a leaning for a hashtag in a given week, we then look at who uses it. Hashtags such as `#climatechange` used predominantly by left-leaning users are assigned a left leaning, whereas hashtags such as `#climategate` are correspondingly assigned a right leaning. The formula, which is identical to the one used for web search queries, is discussed in Section 2.3.

We validated our simple approach of assigning leanings to Twitter users by comparing our classification to (i) `http://wefollow.com/`, (ii) `http://www.twellow.com/` and (iii) `http://persecuting.us/` similar to what has been used in [Pennacchiotti and Popescu 2011]. For an evaluation setting with cases "close to the middle" contributing less and cases "close to the extremes" contributing more the agreement was (i) 98.6%, (ii) 93%, and (iii) 90.4%.

2.3 In Formulas, Please

The formula we use to assign a leaning to both web search and hashtags, jointly denoted by t, involves (i) basic counting to see which side uses the text string more, (ii) normalization of volume bias to avoid that the bigger side owns everything, and (iii) smoothing to ensure that low-volume objects are not prematurely assigned an overly extreme leaning. It is similar to the approach in [Conover et al. 2011], with added smoothing, a generalization to several parties and an inclusion of time.

Concretely, let v_L denote the aggregated user volume of either (i) clicks on left-leaning blogs for a query, or (ii) hashtag usage by left-leaning Twitter users in a given week w. Let V_L denote the total left user volume of all such bits of text t during w. Similarly for v_R and V_R. We compute the leaning of t in w with respect to a party p (either L or R) as

$$\text{Lean}(t, w, p) = \left(\frac{v_p}{V_p} + \frac{2}{\sum\limits_P V_p}\right) \Big/ \left(\sum_P \frac{v_p}{V_p} + \frac{2 * |P|}{\sum\limits_P V_p}\right). \quad (1)$$

Note that for the U.S. setting there are only two parties and Lean(t, w, L)+Lean$(t, w, R) = 1.0$.

2.4 Nitty-Gritty details

Some details left out here can be found in [Weber et al. 2012a; 2012b] and [Teka et al. 2013; Weber et al. 2013]. These include the removal of both apolitical web search queries and hashtags by looking at co-occurrence pattern with known political ones. Certain thresholds were applied to remove low volume objects. Finally, all volumes are over *distinct* users and, for example, a single user using a hashtag in a week 100 times would be counted only once.

3. FINDINGS

Though there is both educational and entertainment value[5] in looking through the list of polarized queries and hashtags in a given week, our ultimate goal is to *quantify* political phenomena and test hypotheses concerning online communication.

3.1 Negative Opposition

One hypothesis we had was that the opposition, currently the Republicans[6], would be more negative in the opinions they voice [Thomas et al. 2006].

To quantify the sentiment of web search queries we did not use the queries directly as in [Chelaru et al. 2012]. Rather, each query was issued to the Bing API and the content for the top 10 results was fetched. The content was parsed into proper sentences to remove mere navigational items. For each sentence we obtained the sentiment using SentiStrength [Thelwall et al. 2010]. SentiStrength assigns both a positive score p_s and a negative score n_s to every sentence s, ranging from 1 (nothing) to 5 (very strong) in absolute value. We aggregated the sentiments across sentences for a given query into three separate numbers: The fraction of sentences s where $|p_s| \leq 2$ and $|n_s| \leq 2$, i.e., sentences with only weak sentiments. Similarly, we counted the fraction of sentences where $|p_s| > 2$ and the fraction of sentences where $|n_s| > 2$.

Though our simplistic approach comes with lots of noise on a per-sentence basis, we found that overall there were weak ($\approx .1$) but statistically significant Kendall Tau correlations between the leaning and the sentiments. Generally, the more left-leaning a query was the more positive and the less negative its results sentences were.

[5] At an early stage of the development of the Political Search Trends demo we thought that we had a bug in our system when the rather odd query "pizza is a vegetable" was caught as a supposedly political query with a left leaning. After having discovered the reason for this to happen, see `http://www.huffingtonpost.com/2011/11/16/pizza-vegetable-school-lunches-lobbyists_n_1098029.html`, and when mentioning this in a presentation an audience member pointed out that it was not a bug in our analytic system but rather in the U.S. political system.

[6] Even though the Republicans have held a majority in the House of Representatives since 2010, we view the Democrats as being in power due to the presidency and the Senate held by them.

statistics	true	false
count	364	574
mean	1.0	1.37
10%-tile	.08	.07
50%-tile	.27	.27
90%-tile	1.93	1.91
95%-tile	3.58	4.10
99%-tile	14.74	22.01
max	29.73	95.49

Table I. Relative volume for queries pertaining to true vs. false statements. False statements are more likely to attract very large volumes (top 1%), though the typical volumes are identical.

3.2 Lies Potentially Sell

Statements made by politicians can cause media hypes. For example, Michele Bachmans statement concerning the president's trip to India in 2010 and its cost[7] was picked up by numerous bloggers. To analyze if there is an inter-dependence between the party affiliation of the author or its truth value and either the leaning or the volume of queries we obtained a list of all fact-checked statements from `http://www.politifact.com` on January 13, 2012, excluding their "position flops". Statements were indexed by combining the actual statement with its title on Politifact and queries with at least two tokens were matched to corresponding statements. Statements *by* a person were not matched to their name as we were more interested in rumors *about* a person than in all statements by a particular person. In the end 1,598 distinct statements were mapped to 1,069 distinct queries for a total of 3,083 such pairs. Statements on the site have the following truth values which we used without modification: "Pants on Fire!"[8], "False", "Mostly False", "Half-True", "Mostly True" and "True".

We looked at whether there was any correlation between truth values and (i) political leaning and (ii) associated query volume. For the first case, we further conditioned on the authorship to see if side A picks up on lies by side B. We could, however find no correlation between truth value and the leaning of the associated queries. For the second case, we found weak evidence that facts with truth value of either "false" or "pants on fire" were more likely to attract very large query volumes. Table I gives details about the volume distribution. Volumes were normalized with 1.0 corresponding to the average volume of queries with an associated true statement.

3.3 Sudden Leaning Jumps

Political hashtags on Twitter ultimately relate to *framing* and which side is more successful at defining the terms used in a political debate. For this reason, both political camps regularly engage in "hashtag wars"[9], with each side trying to use a hashtags in a context

[7] `http://www.politifact.com/truth-o-meter/statements/2010/nov/04/michele-bachmann/repmichele-bachmann-claims-obamas-trip-india-will/`

[8] Children in the U.S. call out "liar, liar, pants on fire" when they suspect another child of lying.

[9] See, e.g., `http://www.politico.com/blogs/burns-haberman/2012/06/its-official-is-actually-a-twitter-hasthag-war-125986.html` and `http://www.guardian.co.uk/commentisfree/2012/nov/29/obama-white-house-hashtag-wars`.

beneficial for their own cause. To study particular instances of such hashtag wars, we tracked the leaning of hashtags over time.

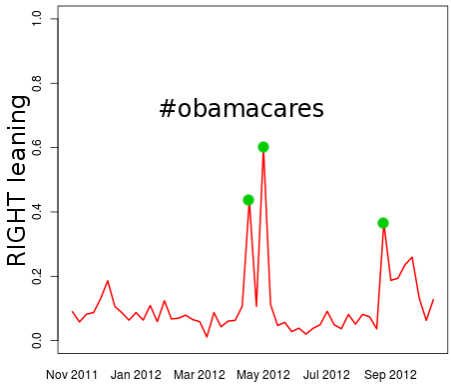

Fig. 1. An example of a left-jumping-towards-right hashtag. Identified change points are highlighted in green.

Figure 1 shows an example of a hashtag with sudden changes in leaning towards the right. We will refer to such outliers as *change points*, which we investigate further. We restrict our focus to change points corresponding both to (i) *upwards* jumps and (ii) cases where the party is "usually inactive", meaning an average leaning across all weeks of $< 1/2$. These interesting cases are directly caused by an unusually high level of hashtag usage by a given leaning, rather than by the absence thereof. To detect change points, we tried different algorithms [Hodge and Austin 2004] against a rule-based heuristics. In the end, we used the following rule-based approach as it gave the most consistent results.

(1) Total_number_of_weeks ≥ 4
(2) Change_from_previous_week $>$ std
(3) Change_from_previous_week > 0.20
(4) Current_value $-$ Average_value $>$ std

If the leaning value of a hashtag at a given week meets all the criteria it is marked as a "change point". We observed that change points of both leanings happen for low volume, untrending weeks. This makes sense as high volume hashtags such as `\#tcot` would be hard to hijack. When, however, only hashtags with change points are considered the volume is slightly higher than for the non-change points and, in particular, the hijacking side puts in about 5-8 times its usual (normalized) volume, again indicating a concerted effort. Finally, change points do not occur sooner or later after the introduction of new hashtags than non-change points.

3.4 Hashtag Hijackers

After identifying change points (h, w), we looked at users from the "other leaning" (compared to h's overall normal leaning) using h in week w and looked at their usage frequencies of h during that week. A user was awarded a "hijacker point" if they used h during a change point. Then, we ranked users according to the number of hijacker points they collected for all identified change points and considered the top 1,000 users for each leaning, which we refer to as "hashtag hijackers". To give a qualitative impression of how

these hijackers differ from normal users, Figure 2 shows term clouds for the four sets of user profiles (left vs. right and normal vs. hijacker). Hijacking users can be seen to use party-related terms more often.

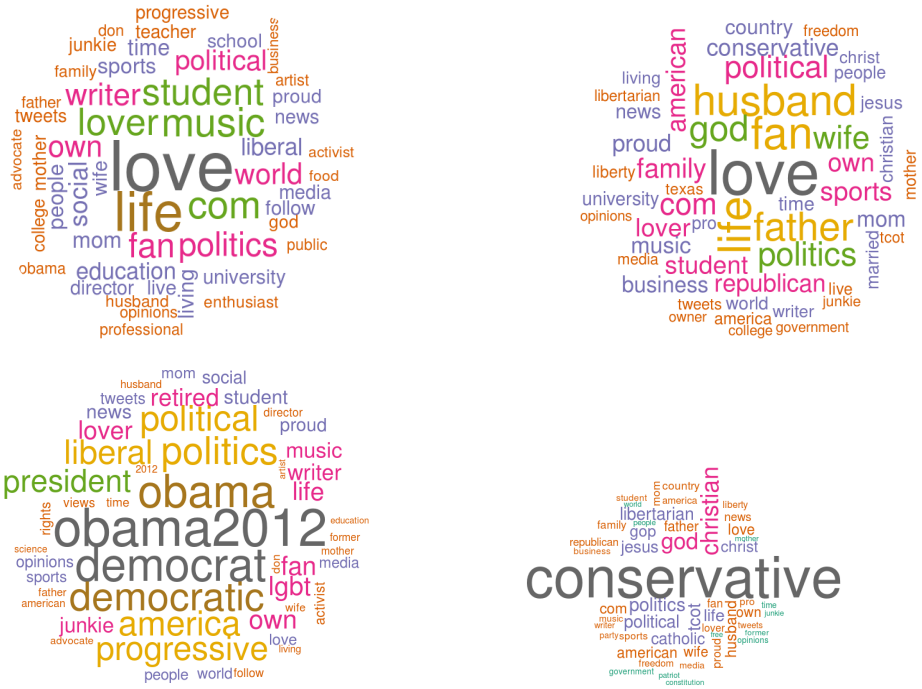

Fig. 2. Clouds of terms used in Twitter profiles. Left users on the left, right users on the right; normal users on top, hijackers on the bottom.

To find quantitative differences, we looked at their respective (i) number of tweets, the fraction of (ii) retweets in general and (iii) of seed users in particular, and the number of (iv) their followers and (v) users they are following. All of those statistics indicate that hashtag hijackers are more active than other politically interested users. However, this could be explained by the fact that we rank users by their number of hijacking points, thereby creating a bias towards more active users. To rule out this explanation, we computed the (i) focus of a user's activity during change points as the fraction of a user's hashtag volume that occurs during change points, (ii) fraction of tweets by these users containing political hashtags, (iii) average leaning of these users. All three features are higher for hijackers, indicating a non-random, conscious decision of active partisan users.

4. LOTS OF OPPORTUNITIES AND CHALLENGES

Selection Bias: It would be naïve to assume that online data in general and Twitter in particular is representative of the offline population. However, when considering polarization this problem can be largely side-stepped by claiming "X is popular among the political left" rather than a more ambitious "X is popular among the overall offline population". Of

course, for challenges such as election prediction [Tumasjan et al. 2010; Gayo-Avello et al. 2011] it is exactly these latter claims that are of interest, which brings selection bias to the forefront.

Not All Political Worlds are Flat: Analyzing political systems with several parties and higher dimensionality comes with interesting challenges. The assignment of "ownership" generalizes easily to several parties (see Equation 1) but things such as studies of hashtag wars become more complex as the interaction network now has more than a single edge. So party A could be "at war" with party B but not with party C. Similarly, the visualization of positions of objects in a higher dimensional space poses challenges, in particular if this position changes over time. Still, we deem the study of dynamic, complex political interaction networks a worthwhile challenge and some progress in this direction has been made [Feller et al. 2011].

Changes Over Time: Twitter can more and more be used not only as a tool to study the here and now but also as a time machine to study changes over time. For example, in the political setting it would be interesting to analyze the network or cascade effect [Leskovec et al. 2006] of users changing their leaning for a large number of users. Such changes are unlikely to happen over the period of a month or even a year but, over several years, a statistically meaningful number of such events is surely observable. Can online data give hints at the *cause* of such changes by providing additional context?

Twitter is More than a Passive Data Source: Twitter is a communication tool but, as far as knowledge management is concerned, it is mostly used as "passive" data source. It would be fascinating to combine a data mining approach with a survey-based one. For example, the hashtag hijackers identified in Section 3.4 could be contacted and asked about their motivation. Such an approach could be benefit both political analysts and data journalists, as long as a certain freak factor can be avoided.

Mashup Please: Most research projects look at one isolated data source at a time. A grand project would face lots of data management and integration challenges but the depth of analysis that a combination of search logs, Twitter data, political donation records, voter registrations, Wikipedia edits and in Senate/House voting records would allow surely warrants such efforts.

Political Persecution: Identifying and analyzing political ideology comes with considerable risks for abuse and political persecution. In fact, the project Persecuting.Us[10], which offers one million Twitter handles with inferred political affiliation for download, was launched to raise awareness of such dangers. Any researcher working with ideologically sensitive data has to be aware of the potential of abuse and should guard against it.

Reminiscent of the 17th century, when astronomers were first looking at the sky through telescopes, we now get to put on the data goggles to look at society. What an amazing time to be a data scientist!

[10] `http://persecuting.us/`

ACKNOWLEDGMENTS

This article bridges a body of work done in collaboration with Erik Borra [Borra and Weber 2012; Weber et al. 2012b; 2012a], Kiran Garimella [Weber et al. 2012b; 2012a; Weber et al. 2013; Teka et al. 2013] and Asmelash Teka [Teka et al. 2013; Weber et al. 2013]. The majority of research was performed while the people involved were at Yahoo! Research Barcelona. Some inspiration for further thoughts came from conversations with Ana-Maria Popescu, Marco Pennacchiotti and Yelena Mejova during the organization of related workshops, tutorials and special journal issues.

REFERENCES

ADAMIC, L. A. AND GLANCE, N. 2005. The political blogosphere and the 2004 u.s. election: divided they blog. In *LinkKDD@KDD*.

BORRA, E. AND WEBER, I. 2012. Political insights: Exploring partisanship in web search queries. *First Monday 17,* 7.

CHELARU, S., ALTINGÖVDE, I. S., AND SIERSDORFER, S. 2012. Analyzing the polarity of opinionated queries. In *ECIR*. 463–467.

CONOVER, M., RATKIEWICZ, J., FRANCISCO, M., GONCALVES, B., FLAMMINI, A., AND MENCZER, F. 2011. Political polarization on twitter. In *ICWSM*.

FELLER, A., KUHNERT, M., SPRENGER, T. O., AND WELPE, I. M. 2011. Divided they tweet: The network structure of political microbloggers and discussion topics. In *ICWSM*.

GAYO-AVELLO, D., METAXAS, P. T., AND MUSTAFARAJ, E. 2011. Limits of electoral predictions using twitter. In *ICWSM*.

HODGE, V. AND AUSTIN, J. 2004. A survey of outlier detection methodologies. *Artificial Intelligence Review 22,* 2.

LESKOVEC, J., SINGH, A., AND KLEINBERG, J. M. 2006. Patterns of influence in a recommendation network. In *PAKDD*. 380–389.

PENNACCHIOTTI, M. AND POPESCU, A.-M. 2011. Democrats, republicans and starbucks afficionados: user classification in twitter. In *KDD*. 430–438.

POOLE, K. T. AND ROSENTHAL, H. 1985. A spatial model for legislative roll call analysis. *AJPS*, 357–384.

SHAW, A. AND BENKLER, Y. 2012. A tale of two blogospheres discursive practices on the left and right. *ABS 56,* 4, 459–487.

SUBASIC, I. AND CASTILLO, C. 2010. The effects of query bursts on web search. In *WI*. 374–381.

TEKA, A., GARIMELLA, V. R. K., AND WEBER, I. 2013. Political hashtag hijacking in the u.s. In *WWW*. 55–56.

THELWALL, M., BUCKLEY, K., PALTOGLOU, G., CAI, D., AND KAPPAS, A. 2010. Sentiment strength detection in short informal text. *61*, 25442558.

THOMAS, M., PANG, B., AND LEE, L. 2006. Get out the vote: Determining support or opposition from congressional floor-debate transcripts. In *EMNLP*. 327–335.

TUMASJAN, A., SPRENGER, T. O., SANDNER, P. G., AND WELPE, I. M. 2010. Predicting elections with twitter: What 140 characters reveal about political sentiment. In *ICWSM*.

WEBER, I., GARIMELLA, V. R. K., AND BORRA, E. 2012a. Mining web query logs to analyze political issues. In *WebSci*. 330–339.

WEBER, I., GARIMELLA, V. R. K., AND BORRA, E. 2012b. Political search trends. In *SIGIR*. 1012.

WEBER, I., GARIMELLA, V. R. K., AND TEKA, A. 2013. Political hashtag trends. In *ECIR*. 857–860.

Ingmar Weber is a senior scientist in the Social Computing Group at the Qatar Computing Research Institute (QCRI). His recent work focuses on how user-generated online data can be used to answer questions about

society at large and the the offline world in general. Ingmar is co-organizer of the "Politics, Elections and Data" (PLEAD) workshop at CIKM 2012 and 2013, contributor to a WSDM 2013 tutorial on "Data-driven Political Science" and is co-editor of a Social Science Computing Review special issue on "Quantifying Politics Using Online Data". He loves chocolate, enjoys participating in the occasional ultra marathon or triathlon and tweets at @ingmarweber.