# Mining Web Query Logs to Analyze Political Issues*

**Ingmar Weber**
Yahoo! Research Barcelona
Barcelona, Spain
ingmar@yahoo-inc.com

**Venkata Rama Kiran Garimella**
Yahoo! Research Barcelona
Barcelona, Spain
gvrkiran@yahoo-inc.com

**Erik Borra**
University of Amsterdam
Amsterdam, Netherlands
borra@uva.nl

## ABSTRACT

We present a novel approach to using anonymized web search query logs to analyze and visualize political issues. Our starting point is a list of politically annotated blogs (left vs. right). We use this list to assign a numerical political leaning to queries leading to clicks on these blogs. Furthermore, we map queries to Wikipedia articles and to fact-checked statements from politifact.com, as well as applying sentiment analysis to search results. With this rich, multi-faceted data set we obtain novel graphical visualizations of issues and discover connections between the different variables.

Our findings include (i) an interest in "the other side" where queries about Democrat politicians have a right leaning and vice versa, (ii) evidence that "lies are catchy" and that queries pertaining to false statements are more likely to attract large volumes, and (iii) the observation that the more right-leaning a query it is, the more negative sentiments can be found in its search results.

### Author Keywords

web search logs; political leaning; partisanship; opinion mining and sentiment analysis

### ACM Classification Keywords

H.1.2 Models and Principles: User/Machine Systems; H.3.5 Information Storage and Retrieval: Online Information Services; J.4 Social and Behavioral Sciences: Sociology

### General Terms

Experimentation, Human Factors

## INTRODUCTION

We use logs of anonymized web queries submitted to a commercial search engine to analyze issues and topics in U.S. politics in a quantitative manner. Our point of departure is a list of political blogs, annotated with a leaning. By monitoring queries leading to clicks on these blogs we can label queries according to their leaning: queries leading mostly to clicks on left-leaning blogs are labeled left-leaning and so on. This approach lies behind Political Search Trends[1] and has been described in [3]. We take this idea further by aggregating leanings for Wikipedia entities, by matching queries to fact-checked political statements and by obtaining sentiments for search results. This creates an unprecedented, rich data set of web search information for the political domain. Going beyond a mere description of queries and their leaning, we observe several statistical patterns, including surprising ones.

Opinion mining and sentiment analysis are two related fields of research where opinion, sentiment and subjectivity are computationally derived from a set of (often online) documents [29]. The information gathered as such can be used to construct reviews of products and services. Similarly, our approach could be applied to monitor the impact of political campaigns. Our main contributions and findings are the following.

*Leaning computation:* We assign a political leaning both to queries and Wikipedia articles.

*Topic visualization:* We present techniques to use the leaning as well as mappings of queries to Wikipedia articles and fact-checked statements to give a graphical visualization of topics and apply them to "Health Care", which has been the subject of heated political debate.

*Trending issues:* We describe how to combine our leaning computation with volume-based trend detection to put issues and their leaning on a timeline. Using these techniques we create a visualization for "Newt Gingrich", candidate for the 2012 Republican Party presidential nomination.

*Interest in "the other side":* We show that queries about Republicans generally have a left leaning, whereas queries about Democrats have a right leaning.

*Leaning inertia:* The vast majority of queries have no change in leaning, and those that do are twice as likely as others to correspond to volume bursts.

*Lies are catchy:* We find evidence that queries correspond-

*This research was partially supported by the Torres Quevedo Program of the Spanish Ministry of Science and Innovation, co-funded by the European Social Fund, and by the Spanish Centre for the Development of Industrial Technology under the CENIT program, project CEN-20101037, "Social Media" http://www.cenitsocialmedia.es/.

*WebSci 2012*, June 22–24, 2012, Evanston, Illinois, USA.
Copyright 2012 ACM 978-1-4503-1228-8...$10.00.

[1] `http://sandbox.yahoo.com/Political_Search_Trends`

ing to factually false statements are more likely to have very high volumes than queries for true statements.

*Negative opposition:* There is a statistically significant correlation between whether a query is right-leaning and how many negative sentiments are present in the search results.

## RELATED WORK

### Digital Democracy and the Political Blogosphere

The web has often been regarded as a liberating, deliberative and democratizing place. Empirical research, however, has shown that while there are many new instruments for communication and participation, online politics are but a reflection of the offline political landscape [17]. Online data is increasingly recognized as a rich source of data for studies that normally fall in the domain of social sciences [23, 35, 44]. Many studies have sought to mine public opinion from online data to understand voters as well as consumers, but also to gain business intelligence [29]. In the study of online politics, particular attention has been paid to (political) blogs as many claimed that their ease of use and accessibility would increase political deliberation by citizens.

Political blogs were studied in terms of link structure and content [1, 16, 20], author demographics and reachability [17], technological adoption and use [2], as well as readership [22]. Although initially blogs were hoped to be vehicles to increase political deliberation, amongst other things, all these studies found political blogs to be polarized along opposing partisan lines. Recognizing that political blogs are an important source of political commentary in U.S. politics, we advance such research not by analyzing blogs, but by studying the search engine queries which land on them. By using political blogs annotated with a leaning as our ground truth in order to attribute leaning to search engine queries, we found that grouping the queries by the leaning of the blogs on which they land provides a meaningful description of partisan concern.

### Search as a Measure of Public Attention

Search is necessarily motivated by some degree of interest in, as well as the willingness to spend time with, the topic of the query. Consequently, various social scientists have sought whether changes in search query volume correlate with traditional measures of public attention and agenda setting, such as surveys and media coverage. It was found that measures of public attention based on search volume and media coverage are closely related for the issues Health Care, Global Warming, and Terrorism [34]. Similarly it is observed that changes in query volume for specific (political) issues and candidate names correlate well with fluctuations in news coverage [14, 15], while [37] ascertains that search engine data corresponds quite closely to existing measures of issue salience and religious adherence. [47] conclude that the volume of searches on any given day can act as a quantifiable measure of public interest for salient issues.

Our research is similar in spirit as we also consider queries to be a measure of public attention. Whereas previously discussed studies use Google insights[2], we have access to Yahoo!'s full US search query log and do not have to limit ourselves to few predefined political issues. By looking into all the queries landing on a carefully chosen set of political blogs annotated with partisan leaning we are able to consider a much broader range of issues and 'charge' queries with political partisanship [3].

While some studies have used search logs for prediction tasks such as flu trends [11] or box office sales [12], greater search traffic does not necessarily mean support or approval of e.g. candidates in US elections [15]. Studies considering the volume of tweets mentioning names of candidates or political parties to predict elections have similarly found that greater volume does not necessarily mean greater approval [25, 18, 10]. While we recognize that search queries can be used to measure interest, in this article we are not interested in predicting elections but in quantifying political opinion and sentiment as measured through query logs.

### Searche(r)s' Demographic Profiles

In the current work we are indirectly assigning profiles in terms of party affiliation to queries. Related is the idea of assigning demographic profiles, such as gender or age [45, 46]. Other properties such as income levels can be estimated from the user's ZIP code. Here we we did not apply the same methodology with per-ZIP election results, as this approach would not allow us to distinguish the leaning for different issues (queries) in the same area. However, in [3] we showed that users clicking on left- (or right-) leaning blogs were more likely than random to live in a ZIP code that voted Democrat (or Republican) in the 2010 U.S. midterm elections. Similarly, user demographics were used to re-confirm knowledge about likely voter profiles and users clicking on left-leaning blogs were, e.g., younger and more female.

### Assigning a Political Leaning to Content or Users

Our work classifies queries according to political partisanship, proportional to the amount of clicked blogs with a particular leaning. Other work pursuing a political classification of content and users, utilizes, among other things, different sets of labeled training data or hyperlink structures to propagate leaning. Word frequencies have been extracted from labeled sets of political texts (e.g. party programs) in order to classify unknown texts or as an estimate of party positions [21, 39]. We believe that such classification methods might benefit from our data set as it could help in the construction of a dictionary of highly partisan terms, similar to using sentiment dictionaries for sentiment analysis.

Other work similarly starts from a set of labeled examples but exploits the hyperlink structure of web documents. The BLEWS system uses a combination of text processing and link analysis to annotate news articles with political orientation and emotional charge [9]. Co-citation analysis has been applied to estimate the political orientation of hypertext documents [6] and users [26], while others propagate political leaning to both content and users [49].

[2] `http://www.google.com/insights/search/`

Various studies have employed user content and information from Twitter's social graph to classify social media users and predict political alignment [32, 4]. Others have computed the political preferences for Twitter audiences of major news media, based on the leaning of Congresspeople with shared Twitter followers [13].

While sentiment detection on Twitter has been found to replicate Presidential approval rates and consumer confidence [28], recognizing sentiment is not enough for political opinion classification [48]. Focussing on the behavior of individual commenters, however, sentiment patterns in comments on political news articles have been employed to indicate the articles' political orientation [31]. Most fruitful seems the employment of sentiment analysis to determine support or opposition to specific legislation [42].

Recent work by Fang et al. [8] does not build a classifier but uses annotated senate speeches (by Republicans and Democrats) to build models and to derive different sub-issues for both sides on topics such as immigration. They use the same approach to show that the United States (represented by the New York Times) and China (represented by Xinhua News) have different views for the topic "Dalai Lama". Our approach achieves a similar goal and different aspects on the same issue are presented in Table 2.

### Political Fact-checking and Rumor Spreading Online

Our analysis on the statistics of true vs. false statements and their leanings or volumes is conceptually related to rumor spreading. Examples of rumors spreading in newsgroups, even though they long have been debunked, are discussed in [36]. In [5] the authors investigate rumor spreading in Twitter and find that false rumors are questioned more often than true ones. This could potentially be applied to automatic fact checking.

The idea of automatically identifying disputed content and claims for online content has been previously explored. [7] describe a browser plug in to identify and annotate disputed claims. This idea was also discussed in the Blogosphere[3]. Similarly, different viewpoints are mined from online sources and presented to users in [27] and [19].

## DATA SET AND METHODOLOGY

### Query Set Used

For our study we used anonymized search queries submitted to the U.S. web front end of the Yahoo! Search Engine from July 4, 2011 to January 8, 2012. Queries containing personally identifiable information such as social security numbers, full street addresses or rare person names were removed. Furthermore, we only retained (query, clicked URL) pairs where the clicked URL belonged to a set of 1,099 political blogs described further down. Queries that were easily detectable as navigational for the clicked URL using string similarity, e.g. "drudge" for `http://www.drudgereport.com` or a mistyped "altrenet" for `http://www.alternet.org/`, were removed as navigational queries do not help to identify political issues.

We split the 27-week period into three 9-week periods and computed leanings, described in detail later, both for the whole period and, to track changes over time, for each period separately. Queries were normalized and only queries which surpassed certain thresholds described further down were kept. Table 1 shows basic statistics of the queries used.

| Date range | Distinct queries | volume |
|---|---|---|
| Jul 4, '11 - Sep 4, '11 (P1) | 3,296 | 28% |
| Sep 5, '11 - Nov 6, '11 (P2) | 4,067 | 34% |
| Nov 7, '11 - Jan 8, '12 (P3) | 4,513 | 38% |

**Table 1. Data statistics after query normalization. For all three periods combined there were 8,694 distinct queries, 2064 of which were present in all three periods.**

### Query Normalization and Filtering

Queries went through the following steps of normalization and filtering. First, all queries were mapped to lower case. Second, we applied an automatic spelling correction such that "barrack obama" was mapped to "barack obama". Third, we used a list of 49 personal names, obtained by post-filtering the list at `http://www.politifact.com/personalities/`, with entries such as "barack obama ⇒ obama", "sarah palin ⇒ palin" and so on. This list was applied to remove redundant first names where present to conflate otherwise identical queries. We avoided ambiguous cases such as "hillary/bill clinton". Next, we applied Porter stemming [33]. Additionally, stop words were removed and all remaining tokens were uniqued and sorted. Though all aggregations were done for the *normalized* queries, for the visualization we always show *raw* forms. Finally, we manually removed a few dozen queries with adult content terms. However, queries such as "republican sex scandals" or "egypt nude protests" were kept. The remaining (query, click) pairs were then uniqued on a per-week basis for each user to reduce the effect of individual users repeatedly issuing the same query and clicking the same results.

Queries that passed all these steps could still correspond to a single user or an apolitical article on a single blog. To remove such queries we required that each query had to be issued by a certain minimum number of distinct users, led to clicks to several distinct blogs and had to have a minimum volume. We also applied upper bounds on the (volume for query)/(number of distinct blogs for query) to remove queries such as "facebook" or "youtube", which had a high volume but only due to a comparably small number of blogs.

### Assigning a Leaning to Queries

To assign a political leaning to queries we used a set of political blogs that had been assigned a left, center, or right leaning. The list of blogs was obtained by combining the lists from [2] and the Wonkosphere Blog Directory[4]. The combination was then cleaned to correct re-directs and remove abandoned blogs. In the end, it comprised 1,099 blogs (644

[3] `http://techcrunch.com/2011/11/28/true-or-false-automatic-fact-checking-coming-to-the-web-complications-follow/`

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

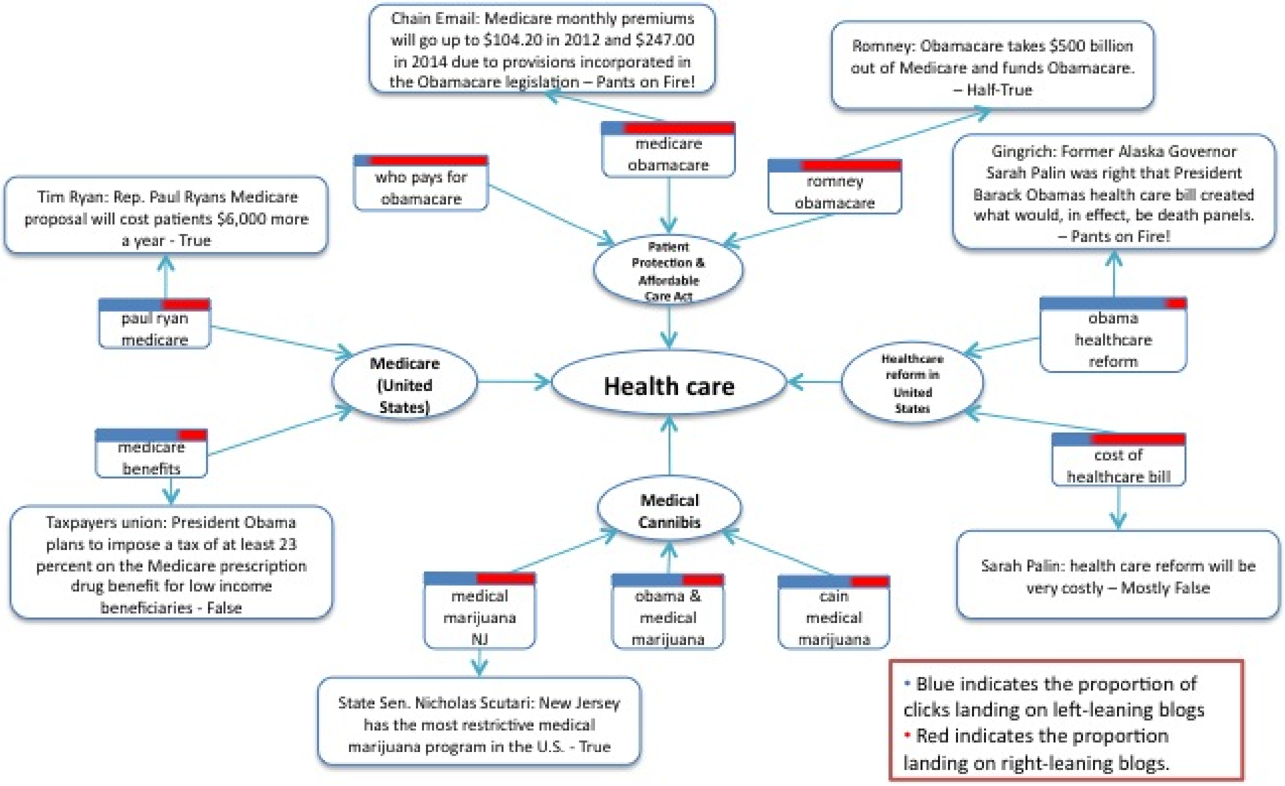
Chain Email: Medicare monthly premiums will go up to $104.20 in 2012 and $247.00 in 2014 due to provisions incorporated in the Obamacare legislation – Pants on Fire!
Romney: Obamacare takes $500 billion out of Medicare and funds Obamacare. – Half-True
medicare obamacare
who pays for obamacare
romney obamacare
Gingrich: Former Alaska Governor Sarah Palin was right that President Barack Obamas health care bill created what would, in effect, be death panels. – Pants on Fire!
Tim Ryan: Rep. Paul Ryans Medicare proposal will cost patients $6,000 more a year - True
Patient Protection & Affordable Care Act
paul ryan medicare
obama healthcare reform
Medicare (United States)
Health care
Healthcare reform in United States
medicare benefits
cost of healthcare bill
Medical Cannibis
Taxpayers union: President Obama plans to impose a tax of at least 23 percent on the Medicare prescription drug benefit for low income beneficiaries - False
Sarah Palin: health care reform will be very costly – Mostly False
medical marijuana NJ
obama & medical marijuana
cain medical marijuana
State Sen. Nicholas Scutari: New Jersey has the most restrictive medical marijuana program in the U.S. - True
• Blue indicates the proportion of clicks landing on left-leaning blogs
• Red indicates the proportion landing on right-leaning blogs.

**Figure 1. A visualization of the queries pertaining to a Wikipedia category about "Health care". Queries are depicted with the associated (colored) proportion of leaning. They are mapped on, and point to, fact-checked statements as well as Wikipedia articles.**

right, 387 left, 68 center). The blogs spanned several orders of magnitude concerning volume of queries landing on those blogs, with `http://www.huffingtonpost.com/` (left) leading the pack. Though we used the center blogs to improve recall in finding politically charged queries, they did not contribute to the left-right leaning computation. The methodology to assign leanings is the same as in [3] and is given for completeness. In [3] we validated the leaning of the blogs by showing that users clicking on left (resp. right) blogs were more likely than random chance to live in a ZIP code that voted Democrat (resp. Republican) in the last midterm elections.

Let $v_L$ denote the volume of clicks on left leaning blogs for a given query in a given time period. $V_L$ denotes the total volume of clicks on left leaning blogs for all queries in that week. We define $v_R$ and $V_R$ similarly. We compute the leaning of a query using the following Equation.

$$\text{leaning} = \frac{\frac{v_L}{V_L} + \frac{2}{V_L+V_R}}{\frac{v_L}{V_L} + \frac{v_R}{V_R} + \frac{4}{V_L+V_R}} \tag{1}$$

We used smoothing factors $\frac{2}{V_L+V_R}$ and $\frac{4}{V_L+V_R}$ to ensure that queries with, say, a single click on a left leaning block do not automatically have an extreme leaning. The leaning score is between 0 and 1, where 0.0 corresponds to an extreme right leaning and 1.0 to left leaning. Table 2 shows examples of highly partisan queries.

We assigned a leaning to queries for (i) each of the three 9-week periods separately and (ii) for the 27 weeks together. The first was done to study changes in leaning (which are rare) and the second was used for the remaining analysis to have a single leaning for queries.

Table 3 shows that there are more queries with a left leaning than with a right leaning, even though (i) there were more right-leaning blogs in our list and (ii) Equation 1 normalizes for volume. This bias is stronger for queries with above median volume (Vol$^+$) and largely disappears for the others. It appears to be caused by a number of sites such as `http://www.huffingtonpost.com` which not only attract a large volume - which we normalize - but also attract a large number of distinct queries.

| group | # | mean | 10% | median | 90% |
|---|---|---|---|---|---|
| All | 8,694 | 0.59 | 0.05 | 0.72 | 0.98 |
| Vol$^+$ | 4,335 | 0.67 | 0.07 | 0.84 | 0.99 |
| Vol$^-$ | 4,359 | 0.51 | 0.04 | 0.60 | 0.94 |
| D polit. | 984 | 0.44 | 0.03 | 0.38 | 0.94 |
| R polit. | 1,269 | 0.54 | 0.05 | 0.59 | 0.95 |
| D statem. | 552 | 0.58 | 0.08 | 0.65 | 0.96 |
| R statem. | 741 | 0.58 | 0.06 | 0.67 | 0.97 |

**Table 3. Basic leaning stats for different groups of queries. Values range from 0.0 (fully right) to 1.0 (fully left).**

| Query Group | Top Left Queries | Top Right Queries |
|---|---|---|
| obama<br>(string) | obama speech video (0.97)<br>obama new mortgage program (0.88)<br>obama accomplishments (0.86) | is obama the worst president ever (0.01)<br>obamacare waivers (0.01)<br>obama gaffes list (0.01) |
| Pat. Prot. & Afford. Care Act<br>(Wikipedia article) | obama healthcare bill text (0.91)<br>obama health care privileges (0.83)<br>is affordable care act unconstitutional (0.78) | who pays for obamacare (0.04)<br>obamacare reaches the supreme court (0.09)<br>is obamacare constitutional (0.16) |
| Occupy<br>(Wikipedia category) | who started occupy wall street (0.94)<br>we are the 99% (0.91)<br>occupy movement supporters (0.78) | occupy wall street rape (0.09)<br>occupy movement violence (0.25)<br>crime in occupy movement (0.44) |

**Table 2. Examples of partisan queries for queries grouped by (i) string, (ii) Wikipedia article, and (iii) Wikipedia category. The leaning values range from 0.0 (fully right) to 1.0 (fully left).**

Given a numerical leaning for each query and a mapping from queries to Wikipedia articles, see the following section for details, we compute a leaning for each article by macro-averaging across all queries mapping to the article. Table 4 shows the extreme cases. As several of the left-leaning blogs in our set also treat topics other than politics, entries such as "IPhone" appear high in this list. "Gibson Guitar Corporation" on the other hand corresponds to a political controversy.[5] The relative position of Dick Cheney (Republican - left-leaning article) and Joe Biden (Democrat - right-leaning article) agrees with our findings discussed later.

| Top left | Top right |
|---|---|
| IPhone (.94) | Gibson Guitar Corporation (.11) |
| Oklahoma (.94) | ATF gunwalking scandal (.12) |
| Immigration reform (.93) | The Falling Man (.12) |
| IPad (.93) | Joe Biden (.16) |
| Troy Davis (.90) | Redistricting (.17) |
| Keystone Pipeline (.90) | OWS (.17) |
| Jon Stewart (.88) | Sharia (.20) |
| Bank of America (.88) | Pat. Prot. & Aff. Care Act (.25) |
| Japan (.86) | Islam (.27) |
| Dick Cheney (.86) | Chris Matthews (.28) |

**Table 4. Top partisan Wikipedia articles and their macro-averaged leaning. Only articles with at least 10 distinct queries were considered.**

### Mapping Queries to Wikipedia Pages

To obtain additional meta information for an issue and to group semantically related queries we mapped queries to Wikipedia articles using two approaches in parallel. First, we applied the machine learning approach described in [30] (Paranjpe) and implemented in a scalable manner in a proprietary software library. Second, we issued each query (before stemming) to the Bing Search API[6] (Bing). A query was mapped to a Wikipedia page if the top 20 search results contained a result from `http://en.wikipedia.org/wiki/`. In this case, the first such result was used, ignoring results for special pages such as "User:", "Talk:" or "Template:". Both approaches have strengths and weakness related to recent articles, web lingo and the detection of multiple articles within a single queries. Table 5 lists a few illustrative examples. The detected articles were combined by OR-ing giving a many-to-many mapping. Note that 12% of distinct queries, e.g. "who is to blame for the economy", are not mapped to any article.

### Obtaining Sentiments for Queries

To quantify the sentiment of web search queries we did the following. Similar to [38], each query was issued to the Bing API and the content for the top 10 results was fetched. The content was parsed into sentences using a simple heuristics: splitting on delimiters such as "." or "?", looking for groups with 10-50 tokens, and using such a group if it contained at least two stop words. These steps were applied to ensure we are dealing with proper sentences and not mere navigational items. For each detected phrase we obtained the sentiment using SentiStrength [41]. SentiStrength assigns *both* a positive score $p_s$ *and* a negative score $n_s$ to every sentence $s$, both ranging from 1 (nothing) to 5 (very strong) in terms of absolute value. We combine the sentiments across sentences for a given query into three separate numbers: The fraction of sentences $s$ were $|p_s| \leq 2$ and $|n_s| \leq 2$. These are sentences in which none or only weak sentiments were found. Similarly, we counted the fraction of sentences where $|p_s| > 2$ - strong positive sentiments - and the fraction of sentences where $|n_s| > 2$ - strong negative sentiments. Table 6 shows examples of queries with their pertaining sentiment scores. Though our simplistic approach comes with lots of noise[7], statistically significant correlations with other measures are discussed in a later result section.

### Mapping Queries to Statements on politifact.com

Statements made by politicians can cause a media hype. For example, Michele Bachman's statement concerning the president's trip to India and its cost[8] 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. We used only actual statements and not the "position flops" listed on the site. Statements were indexed by combining the actual statement with its title on

[5] `http://en.wikipedia.org/wiki/Gibson_Guitar_Corporation#Recent_criticism_and_controversy`

[6] `http://www.bing.com/toolbox/bingdeveloper/`

[7] The query "missing teen" has a strong positive sentiment as several search results were for "miss teen" pageant competitions.

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

| Query | Paranjpe | Bing |
|---|---|---|
| obama's aunt deported | Barack Obama | Zeituni_Onyango |
| obamacare | Pat. Prot. & Afford. Care Act | Pat. Prot. & Afford. Care Act |
| dukes vs walmart | Duke; Wal-Mart | Dukes vs. Wal-Mart |
| atheist christopher hitchens | Christopher Hitchens; Atheism | Christopher Hitchens |
| ndaa 2012 obama signed | Barack Obama | Nat. Def. Auth. Act for Fiscal Year 2012 |

**Table 5. Examples of Web search queries being mapped to Wikipedia pages using the approach by Paranjpe and Bing Search.**

| fraction neutral | fraction positive | fraction negative |
|---|---|---|
| housing news (0.98) | newt gun control (.46) | ** (.51) |
| 2012 polls for republicans nomination (0.95) | jack the cat (.40) | ** (.42) |
| contract for the american dream (.94) | missing teen (.37) | war on terror (.42) |
| 2012 primaries calendar (.93) | amazon tablet rumors (.37) | will casey anthony be found guilty (.41) |
| peer to peer lending (.93) | arnold schwarzenegger (.35) | iran israel (.41) |

**Table 6. Top five queries for each sentiment class. The blank ** refer to a violent crime and its victims.**

the web page. We matched queries to these statements as follows. We used only queries with at least two tokens. If the query was a name such as "barack obama" then statements *by* the person were not matched, but statements *about* the person were used. This was done as we were more interested in rumors than in all statements by a particular person. In the end 1,598 distinct statements were mapped to 1,069 distinct queries where both one statement can be associated with several queries and vice versa. In total there were 3,083 such pairs. Statements on the site have the following truth values which we used without modification: "Pants on Fire!", "False", "Mostly False", "Half-True", "Mostly True" and "True".

### Weekly Aggregation and Trend Detection

Whereas the aggregation into 9-week periods has the advantages that (i) low volume queries have more time to pass minimum volume thresholds and (ii) leaning estimates are more robust, we also aggregated things on a weekly basis to detect trending issues. The same thresholds as for the 9-week periods were applied, resulting in fewer queries. For the corresponding queries in a given week we then computed a trending score as follows.

Let $v_w^q$ be the volume of a query $q$ in week $w$. Let $n^q$ be the number of weeks in which query $q$ occurred and $V^q$ be the total volume of the query $q$ in all these weeks. Then, the trending score of a query $q$ in week $w$ is defined as:

$$\text{trending}(q, w) = \frac{v_w^q}{\frac{w_n \cdot \frac{v_w^q}{r} + (V - v)}{n - 1 + w_n}}, \qquad (2)$$

where, $w_n$ and $r$ are parameters to control the number of new (= previously unseen) and old (= rebounding) queries that are trending. For our experiments, we used values $w_n = 0.1$ and $r = 100$. The trending formula is a modification of the "burst intensity" formula in [40] to not only have new queries trending.

For each week we ranked queries according to this score and marked the top 20 left queries (leaning $\geq 0.5$) and top 20 right queries (leaning $\leq 0.5$) as trending. This approach can be used to visualize political search trends at a given point in time. This is shown in Figure 2 and discussed in a later section.

## RESULTS

### Changes in Leaning

As we computed a separate leaning for each of the three 9-week periods (P1, P2 and P3) we could identify queries that changed their leaning. According to the size in leaning difference we label a query as *constant* (change $\leq .25$), *change* ($.25 <$ change $\leq .50$) and *flip* (change $> .50$). Only queries with a minimum volume of 50 occurrences in consecutive periods were considered. Table 8 shows examples for queries with leaning flips. We labeled a query for a pair of periods as bursty if its volume satisfied max/min $\geq 5.0$.

| | Query | Leaning |
|---|---|---|
| P1→P2 | social security increase 2012 | right (0.33) → left (0.99) |
| | joe biden | right (0.12) → left (0.98) |
| | ann coulter | right (0.09) → left (0.86) |
| | gop polls | left (0.95) → right (0.38) |
| | fullerton police beating | left (0.98) → right (0.42) |
| P2→P3 | ron paul | left (0.82) → right (0.27) |
| | bill maher | left (0.83) → right (0.12) |
| | iowa polls | right (0.14) → left (0.93) |
| | new hampshire primary | right (0.33) → left (0.99) |
| | joel osteen | right (0.40) → left (0.99) |

**Table 8. Examples of queries with leaning flips.**

Table 9 shows that among the "flipping" queries volume bursts are twice as likely as among the rest. The Wikipedia category "Living People" correlated strongest with whether a query would change or flip.

### The Other Side

We compiled a list of 93 Democrat and 114 Republican politicians by intersecting the Wikipedia Category "Living People" with any category for Democrat* or Republican*. Interestingly, queries mapping to a Democrat politician have a

| Query | Statement | Truth value | Authorship |
|---|---|---|---|
| obama health care | Mitt Romney once supported President Obamas health care plan but now opposes it. | Mostly False | D. Nat. Com. (D) |
| obama health care | Those who fail to buy health insurance under Obamacare face the threat of jail time. | Pants on Fire! | Marshal (R) |
| obama health care | On the night of the Iowa caucuses, Obama promised the nation that he would do health care reform focused on cost containment, he opposed an individual mandate, and he said he was going to do it with Republicans. | Mostly True | Pawlenty (R) |
| taxes and obama | Fifty percent of Speaker Gingrichs tax plan goes to the top 1 percent. | Mostly True | Obama (D) |
| rick perry immigration | Rick Perry wrote a newspaper item saying he was open to amnesty for illegal immigrants in the United States. | Half-True | Romney (R) |
| obama vs. perry | President Barack Obama is a socialist. | Pants on Fire! | Perry (R) |
| obama socialist | President Barack Obama is a socialist. | Pants on Fire! | Perry (R) |

**Table 7. Some example of Web search queries being mapped to statements on politifact.com.**

| Period | Group | Constant | Change | Flip |
|---|---|---|---|---|
| P1→P2<br>(618) | All | 462 (100%) | 109 (100%) | 47 (100%) |
| | Burst | 39 (8.4%) | 11 (10.1%) | 10 (21.3%) |
| | L→R | 0 (0%) | 98 (89.9%) | 42 (89.4%) |
| | R→L | 0 (0%) | 11 (10.1%) | 5 (10.6%) |
| | Living | 189 (40.9%) | 66 (60.6%) | 25 (53.2%) |
| P2→P3<br>(791) | All | 651 (100%) | 106 (100%) | 34 (100%) |
| | Burst | 75 (11.5%) | 11 (10.4%) | 10 (29.4%) |
| | L→R | 0 (0%) | 32 (30.2%) | 6 (17.6%) |
| | R→L | 0 (0%) | 74 (69.8%) | 28 (82.4%) |
| | Living | 251 (38.6%) | 52 (49.1%) | 14 (41.2%) |

**Table 9. Statistics about leaning changes for queries with sufficient volume. Most queries do not change their leanings, and among flipping queries bursty queries are more common.**

stronger *right* leaning than other queries. Similarly, queries for Republican politicians were more *left* leaning, indicating that both sides tend to discuss politicians of the other side. However, these queries were still less left leaning than random queries. See entries for "D/R polit." in Table 3 for details.

### Statement Authorship

Using the mapping of queries to statements and the subset of statements that were made by either Democrats ("D statem.") or Republicans ("R statem."), rather than journalists or organizations, we could investigate whether statements by, say, Republicans attract a right leaning. Surprisingly, we could not find any difference in leaning according to this authorship, indicating that statements made by one side were just as likely to be picked up by the other side. Details are in Table 3.

### Impact of Truth

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 true value of either "false" or "pants on fire" were more likely to attract very large query volumes. The highest volume queries, all of which have only false and no true statements associate to them, were "government shutdown", "michelle obama", "glenn beck"[9] and "obama approval rating". Table 10 gives details about the volume distribution. Volumes were normalized with 1.0 corresponding to the average volume of queries with an associated true statement.

| 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 10. Relative volume for queries pertaining to true vs. false statement. False statements are more likely to attract very large volumes (top 1%), though the typical volumes are identical.**

### Leaning vs. Sentiments

We used assigned sentiments to investigated whether there was a link between the topic of the query, say, Democrat vs. Republican politicians, and the leaning and the sentiment. We found that overall there were weak (≈ .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. These correlations were stronger for queries referring to Democrat politicians. For queries referring to Republican politicians we could not find any significant correlation between leaning and sentiments.

## CASE STUDY: HEALTHCARE REFORM

[9] Recall that this query would *not* be mapped to statements by Glenn Beck but only to statements *about* Glenn Beck. Ditto for Michelle Obama.

| Group | frac. neutral | frac. positive | frac. negative |
|---|---|---|---|
| All | - | $.12^{**}$ | $-.09^{**}$ |
| $\text{Vol}^-$ | $.03^*$ | $.10^{**}$ | $-.10^{**}$ |
| $\text{Vol}^+$ | - | $.13^{**}$ | $-.08^{**}$ |
| D politic | - | $.15^{**}$ | $-.12^{**}$ |
| R politic | - | - | - |

**Table 11. Statistically significant ($^*$ at 5%, $^{**}$ at 1%) Kendall tau correlations between leaning and sentiments. More left-leaning correlated with more positive and less negative.**

In previous sections queries were assigned a leaning and mapped both to (i) Wikipedia articles and to (ii) fact-checked statements. These techniques now allow us to generate a controversy map [43] for any issue, as measured through search queries. Figure 1 shows Wikipedia articles, as well as the queries mapped to them, for the Wikipedia categories matching "Health care*". Each query and its leaning are visualized in this figure as the proportion of left blogs (blue) and right blogs (red) on which this query lands. It can be seen that queries containing 'obamacare' land more on right-leaning blogs, while queries mentioning 'medical marijuana' are more likely to lead to left-leaning blogs. Figure 1 also displays facts for those queries which we could map on fact-checking sites. The selection of the items shown is automatic and volume-based. The same methodology could be applied to visualize a topic such as "Occupy*", though the matching items would still need to be manually arranged.

## CASE STUDY: NEWT GINGRICH

We described earlier how our assignment of leanings can be combined with volume-based trend detection to find trending topical queries in a given week. As an example, we applied this approach to all queries containing the string "gingrich". Figure 2 shows an excerpt of the trending queries identified in this manner. They generally followed media events such as interviews or debates. This is in line with [24] who found close relations between the temporal patterns of identical phrases in news media and blogs. Additionally, it is well known that the volume of news coverage often correlates with spikes in aggregate search [47]. We discuss a few examples in detail.

*gingrich scott pelley* - trending in week Nov 14-20. During a Republican presidential debate in South Carolina on November 13, CBS Evening News anchor Scott Pelley argued with Newt Gingrich whether killing American born terrorists overseas complied with "the rule of law". Gingrich replied that enemy combatants have no right to civil liberties[10]. Right leaning blogs were quick to pick up on the story.

*gingrich janitor* - trending in week Nov 21-27. On November 18, Newt Gingrich proposed during an event at Harvard that certain "schools ought to get rid of the unionized janitors, have one master janitor and pay local students to take care of the school"[11]. The query "gingrich child labor" is about the same topic.

*gingrich arresting judges* - trending in week Dec 19-25. On December 18, in CBS's "Face the Nation" Gingrich responded affirmatively on the question whether he would arrest judges who make controversial decisions [12].

## CONCLUSIONS

Alhough query logs have been used before to study politics and its perception in the media, their use has been generally limited to simple volume information. We used search result clicks on political blogs annotated with a leaning to assign a numerical leaning (left vs. right) to queries. The combination of this leaning information with additional information such as correspondence to Wikipedia articles, sentiments of web search results and correspondence to fact-checked statements created a dataset of unprecedented richness. The analysis of this data set led to a number of both expected ("the more left-leaning the more positive are the sentiments in search results about Democrats") and surprising ("queries about Democrats have a right leaning") results. We also showed how this data can be used to visualize political issues both in terms of covering various sub-issues (see Figure 1) and in terms of identifying trending issues in time (See Figure 2).

## ACKNOWLEDGMENTS

The authors would like to thank Vicente-Arturo Romero-Zaldivar for his help with parts of the implementation during an early phase of this project, Albert Bifet for obtaining a large list of politically annotated blogs and Ethan Zuckerman for insightful discussion.

## REFERENCES

1. Adamic, L. A., and Glance, N. The political blogosphere and the 2004 u.s. election: divided they blog. In *LinkKDD@KDD* (2005), 36–43.

2. Benkler, Y., and Shaw, A. A tale of two blogospheres: Discursive practices on the left and right, 2010. `http://cyber.law.harvard.edu/publications/2010/Tale_Two_Blogospheres_Discursive_Practices_Left_Right`.

3. Borra, E., and Weber, I. Methods for exploring partisan search queries, 2012. `http://erikborra.net/blog/2012/04/methods-for-exploring-partisan-search-queries/`.

4. Boutet, A., and Yoneki, E. Member classification and party characteristics in twitter during uk election. In *DYNAM* (2011).

5. Castillo, C., Mendoza, M., and Poblete, B. Information credibility on twitter. In *WWW* (2011), 675–684.

[10] `http://www.cbsnews.com/8301-505103_162-57323734/cbs-news-nj-debate-transcript-part-1/?pageNum=10&tag=contentMain;contentBody`

[11] `http://www.huffingtonpost.com/2011/12/01/newt-gingrich-janitors-students-child-labor_n_1123800.html`

[12] `http://www.cbsnews.com/8301-3460_162-57344818/face-the-nation-transcript-december-18-2011/?pageNum=2`

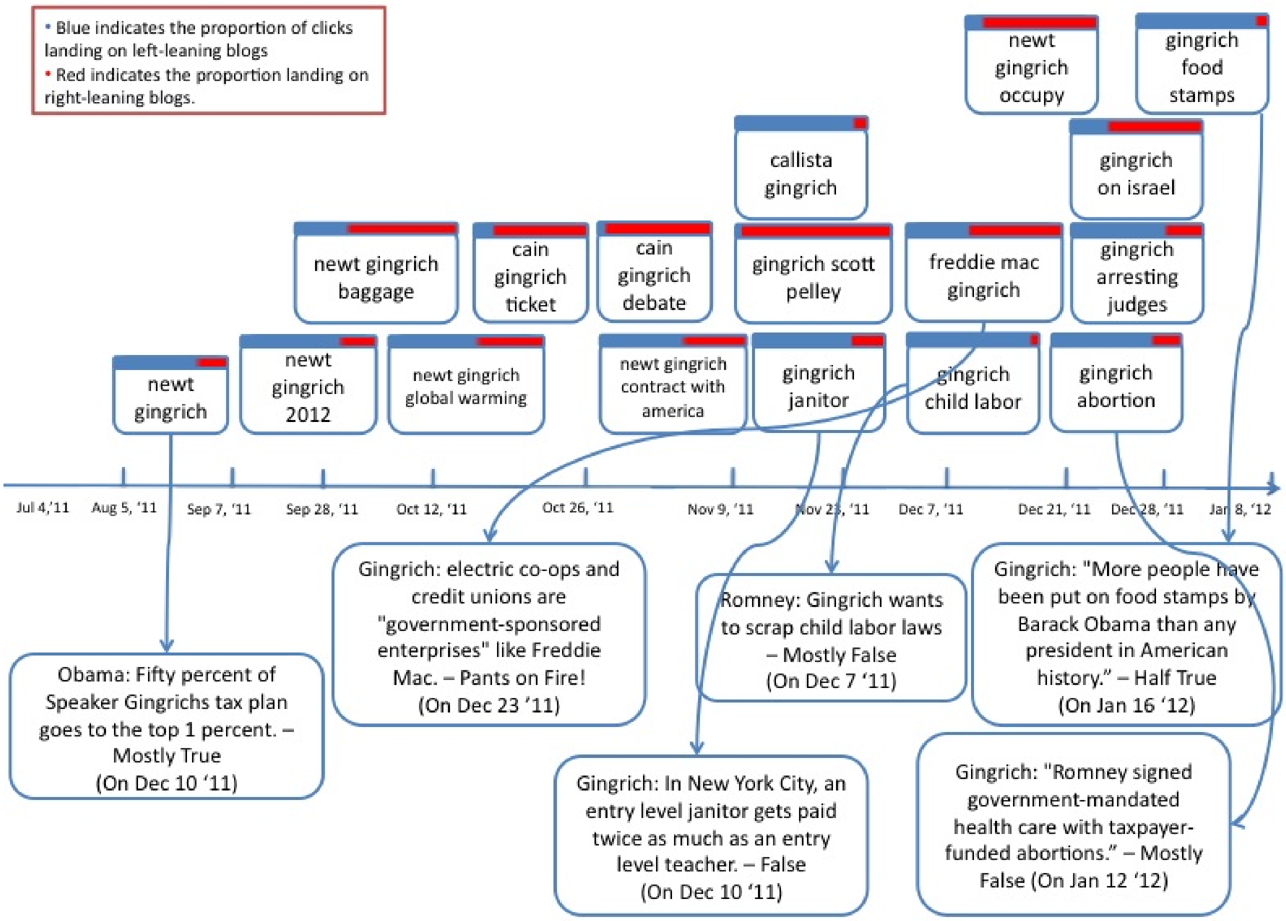
• Blue indicates the proportion of clicks landing on left-leaning blogs
• Red indicates the proportion landing on right-leaning blogs.
newt gingrich occupy
gingrich food stamps
callista gingrich
gingrich on israel
newt gingrich baggage
cain gingrich ticket
cain gingrich debate
gingrich scott pelley
freddie mac gingrich
gingrich arresting judges
newt gingrich
newt gingrich 2012
newt gingrich global warming
newt gingrich contract with america
gingrich janitor
gingrich child labor
gingrich abortion
Jul 4,'11
Aug 5, '11
Sep 7, '11
Sep 28, '11
Oct 12, '11
Oct 26, '11
Nov 9, '11
Nov 23, '11
Dec 7, '11
Dec 21, '11
Dec 28, '11
Jan 8, '12
Gingrich: electric co-ops and credit unions are "government-sponsored enterprises" like Freddie Mac. – Pants on Fire! (On Dec 23 '11)
Romney: Gingrich wants to scrap child labor laws – Mostly False (On Dec 7 '11)
Gingrich: "More people have been put on food stamps by Barack Obama than any president in American history." – Half True (On Jan 16 '12)
Obama: Fifty percent of Speaker Gingrichs tax plan goes to the top 1 percent. – Mostly True (On Dec 10 '11)
Gingrich: In New York City, an entry level janitor gets paid twice as much as an entry level teacher. – False (On Dec 10 '11)
Gingrich: "Romney signed government-mandated health care with taxpayer-funded abortions." – Mostly False (On Jan 12 '12)

**Figure 2. Timeline about issues pertaining to Newt Gingrich. Queries are shown for the week in which they were trending. The dates for the facts or those of the statement.**

6. Efron, M. The liberal media and right-wing conspiracies: using cocitation information to estimate political orientation in web documents. In *CIKM* (2004), 390–398.

7. Ennals, R., Trushkowsky, B., and Agosta, J. M. Highlighting disputed claims on the web. In *WWW* (2010), 341–350.

8. Fang, Y., Si, L., Somasundaram, N., and Yu, Z. Mining contrastive opinions on political texts using cross-perspective topic model. In *WSDM* (2012), 63–72.

9. Gamon, M., Basu, S., Belenko, D., Fisher, D., Hurst, M., and König, A. Blews: Using blogs to provide context for news articles. In *ICWSM* (2008), 60–67.

10. Gayo-Avello, D., Metaxas, P. T., and Mustafaraj, E. Limits of electoral predictions using twitter. In *ICWSM* (2011).

11. Ginsberg, J., Mohebbi, M., Patel, R., Brammer, L., Smolinski, M., and Brilliant, L. Detecting influenza epidemics using search engine query data. *Nature 457*, 7232 (2008), 1012–1014.

12. Goel, S., Hofman, J., Lahaie, S., Pennock, D., and Watts, D. Predicting consumer behavior with web search. *PNAS 107*, 41 (2010), 17486–17490.

13. Golbeck, J., and Hansen, D. Computing political preference among twitter followers. In *CHI* (2011), 1105–1108.

14. Granka, L. Measuring agenda setting with online search traffic: Influences of online and traditional media. In *APSA* (2010).

15. Granka, L. A. Inferring the public agenda from implicit query data. In *UIIR@SIGIR* (2009), 28–31.

16. Hargittai, E., Gallo, J., and Kane, M. Cross-ideological discussions among conservative and liberal bloggers. *Public Choice 134*, 1 (2008), 67–86.

17. Hindman, M. *The myth of digital democracy*. Princeton University Press, 2008.

18. Jungherr, A., Jrgens, P., and Schoen, H. Why the pirate party won the german election of 2009 or the trouble with predictions: A response to tumasjan, a., sprenger, t. o., sander, p. g., & welpe, i. m. "predicting elections with twitter: What 140 characters reveal about political sentiment". *SSCR* (2011).

19. Kawahara, D., Inui, K., and Kurohashi, S. Identifying contradictory and contrastive relations between statements to outline web information on a given topic. In *COLING* (2010), 534–542.

20. Kelly, J. Parsing the online ecosystem: Journalism, media, and the blogosphere. In *Transitioned Media*. Springer New York, 2010, 93–108.

21. Laver, M., Benoit, K., and Garry, J. Extracting policy positions from political texts using words as data. *APSR 97*, 02 (2003), 311–331.

22. Lawrence, E., Sides, J., and Farrell, H. Self-Segregation or deliberation? blog readership, participation, and polarization in american politics. *POP 8*, 01 (2010), 141–157.

23. Lazer, D., Pentland, A. S., Adamic, L., Aral, S., Barabasi, A. L., Brewer, D., Christakis, N., Contractor, N., Fowler, J., Gutmann, M., et al. Life in the network: the coming age of computational social science. *Science 323*, 5915 (2009), 721.

24. Leskovec, J., Backstrom, L., and Kleinberg, J. Meme-tracking and the dynamics of the news cycle. In *KDD* (2009), 497–506.

25. Lui, C., Metaxas, P. T., and Mustafaraj, E. On the predictability of the us elections through search volume activity. In *e-Society* (2011).

26. Malouf, R., and Mullen, T. Taking sides: user classification for informal online political discourse. *Internet Research 18* (2008), 177–190.

27. Murakami, K., Nichols, E., Mizuno, J., Watanabe, Y., Masuda, S., Goto, H., Ohki, M., Sao, C., Matsuyoshi, S., Inui, K., and Matsumoto, Y. Statement map: reducing web information credibility noise through opinion classification. In *AND* (2010), 59–66.

28. O'Connor, B., Balasubramanyan, R., Routledge, B. R., and Smith, N. A. From tweets to polls: Linking text sentiment to public opinion time series. In *ICWSM* (2010).

29. Pang, B., and Lee, L. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval 2*, 1-2 (2008), 1–135.

30. Paranjpe, D. Learning document aboutness from implicit user feedback and document structure. In *CIKM* (2009), 365–374.

31. Park, S., Ko, M., Kim, J., Liu, Y., and Song, J. The politics of comments: predicting political orientation of news stories with commenters' sentiment patterns. In *CSCW* (2011), 113–122.

32. Pennacchiotti, M., and Popescu, A.-M. Democrats, republicans and starbucks afficionados: user classification in twitter. In *KDD* (2011), 430–438.

33. Porter, M. An algorithm for suffix stripping. *Program 14*, 3 (1980), 130–137.

34. Ripberger, J. T. Capturing curiosity: Using internet search trends to measure public attentiveness. *PSJ 39*, 2 (2011), 239–259.

35. Rogers, R. The end of the Virtual-Digital methods. *Inaugural speech* (2009).

36. Rosengren, E. Patterns in prevalence and debunking of three rumors online, 2011. `http://hdl.handle.net/2027.42/85319`.

37. Scheitle, C. P. Google's insights for search: A note evaluating the use of search engine data in social research. *SSQ 92* (2011), 285–295.

38. Shen, D., Sun, J.-T., Yang, Q., and Chen, Z. Building bridges for web query classification. In *SIGIR* (2006), 131–138.

39. Slapin, J., and Proksch, S. A scaling model for estimating time-series party positions from texts. *AJPS 52*, 3 (2008), 705–722.

40. Subasic, I., and Castillo, C. The effects of query bursts on web search. In *WI* (2010), 374–381.

41. Thelwall, M., Buckley, K., Paltoglou, G. Cai, D., and Kappas, A. Sentiment strength detection in short informal text. *JASIST 61* (2010), 2544–2558.

42. Thomas, M., Pang, B., and Lee, L. Get out the vote: Determining support or opposition from congressional floor-debate transcripts. In *EMNLP* (2006), 327–335.

43. Venturini, T. Diving in magma: How to explore controversies with actor-network theory. *PUS* (2010).

44. Venturini, T. Building on faults: how to represent controversies with digital methods. *PUS* (2012).

45. Weber, I., and Castillo, C. The demographics of web search. In *SIGIR* (2010), 523–530.

46. Weber, I., and Jaimes, A. Who uses web search for what: and how. In *WSDM* (2011), 15–24.

47. Weeks, B., and Southwell, B. The symbiosis of news coverage and aggregate online search behavior: Obama, rumors, and presidential politics. *HMCS 13*, 4 (2010), 341–360.

48. Yu, B., Kaufmann, S., and Diermeier, D. Exploring the characteristics of opinion expressions for political opinion classification. In *dg.o* (2008), 82–91.

49. Zhou, D. X., Resnick, P., and Mei, Q. Classifying the political leaning of news articles and users from user votes. In *ICWSM* (2011).