Political insights: exploring partisanship in Web search queries

Borra, E.; Weber, I.

Published in:
First Monday

First Monday, Volume 17, Number 7 - 2 July 2012

PEER-REVIEWED JOURNAL ON THE INTERNET

Political Insights: Exploring partisanship in Web search queries

by Erik Borra and Ingmar Weber

Abstract

We developed Political Insights, an online searchable database of politically charged queries, which allows you to obtain topical insights into partisan concern. In this paper we demonstrate how you can discover such political queries and how to lay bare which issues are most salient to political audiences. We employ anonymized search engine queries resulting in a click on U.S. political blogs to calculate the probability that a query will land on blogs of a particular leaning. We are thus able to 'charge' queries politically and to group them along opposing partisan lines. Finally, by comparing the zip codes of users submitting these queries with election results, we find that the leaning of blogs people read correlates well with their likely voting behavior.

Contents

Introduction

Big corporations and start–ups alike have long since recognized the potential value of the data derived from monitoring and capturing online interactions for marketing and advertising purposes. Recently, scholars have called for an investment in fields such as digital humanities and computational social science, by using the kind of data available in 'big data companies' (Lazer, *et al.*, 2009; Borgman, 2009; Manovich, 2012). This paper takes up these calls, and demonstrates Political Insights, a tool for research into political partisanship, based on nine months of anonymized U.S.–based Yahoo! query logs, which have been found representative of the U.S. population (Weber and Castillo, 2010) [1].

Various studies have heralded query logs as viable alternatives or additions to traditional social science methods for gathering data, such as polls and surveys. In what follows we summarily discuss the pros and cons of employing search engine query logs to gather social and cultural data. Particularly important is the claim that query logs have limited depth because intent is hard to infer (Grimes, *et al.*, 2007). Can query logs then only provide superficial descriptions of social and cultural preferences? We briefly discuss the kind of data encountered in query logs, how they have typically been studied to infer intent, as well as their usefulness for social and cultural research. A short review is provided on how they have already been employed to measure collective preferences; this ranges from grouping queries by the users' geographical location and user demographics to measuring public attention by correlating search queries with patterns of real–world activity. We argue that as big data in general, and Web data in particular, are 'fundamentally networked' (boyd and Crawford, 2011) the consideration of additional variables and data can complement the short queries with a more elaborate description. Marres (in press) more strongly argues that loading (data–)objects with issues can turn them into placeholder objects, where matters of concern and action resonate. Could we then use query logs in combination with carefully chosen additional data in order to 'charge' queries with matters of concern?

In this paper we are specifically interested in how to locate political issues and partisan polarization. To this end, we look at all queries landing on 155 top U.S. political blogs annotated with a political leaning and subsequently assign a leaning to each query, proportional to the number of times it landed on such a blog. We review previous work on political blogs and argue why they are good proxies to infer partisan concern from the queries

landing on them. We demonstrate that the resulting tool consisting of politically charged queries made by hundreds of thousands of users, allows for detailed insights into topical partisan concerns. Consider for instance queries containing [obamacare], which turned out to result much more likely in a click on right–leaning blogs, while queries containing [healthcare bill] much more likely resulted in a click on left–leaning blogs [2]. In order to ground our methodology, we validated our data set with voting polls of the 2010 U.S. mid–term elections: people visiting blogs of a particular leaning are more likely to have a zip code with a higher proportion of voters of that leaning. We conclude the paper with a short summary and provide a number of directions for future research.

Query logs as a source of data for social and cultural research

Traditionally social and cultural data are collected via field studies, user panels, focus groups, interviews, questionnaires and surveys (small or huge like the decennial U.S. census) [3]. Increasingly, the social and cultural interactions passing through or taking place on the Web are considered as valuable sources of data for social and cultural research (Lazer, *et al.*, 2009; Venturini, 2010; Rogers, in press). In this article we focus on the queries submitted to search engines, which have been among the main entry points to the Web (Dodge, 2007).

As a user who submits a search query to a search engine is necessarily motivated by some degree of interest in a particular issue, and has the willingness to invest time in it, recent literature suggests that observations based on analyzing large–scale query logs are viable alternatives, or at least additions, to some of the more traditional methods (Grimes, *et al.*, 2007; Richardson, 2008; Granka, 2009, 2010; Gruszczynski, 2011; Mohebbi, *et al.*, 2011; Ripberger, 2011; Scharkow and Vogelgesang, 2011; Scheitle, 2011; Weber and Jaimes, 2011). Query logs are recommended for their coverage (*e.g.*, the entire U.S.), ease of collection, scope (any entry into a search engine), cost (analyzing text is cheap), and up–to–datedness (almost in real–time). Additionally, studying queries is less prone to the observer effects present in other types of social data collection (Webb, *et al.*, 1972), nor are particular response categories imposed. Query logs seem particularly attractive when alternative data sources are very expensive or not electronically available at all.

These surveys point out that query logs as sources of social and cultural data present difficulties, too. The data are often considered 'noisy' or 'messy' (*e.g.*, because of misspellings, spammers, or a small set of highly biased users), they need to be anonymized (with the risk of tainting, according to some authors), and they are not freely available (query logs generally require commercially negotiated access). Moreover, data must be validated or grounded so that the claims based on Web data in general, and query log data specifically, can be trusted (Thelwall, *et al.*, 2005; Rogers, in press). The biggest difficulty, however, seems to be that as intent is hard to infer, query logs have limited depth.

Advancing such work, this paper introduces the use of query logs to provide insight into partisan concern. We discuss seminal examples of research with query logs, focusing in particular on those which grouped and combined query logs with other data in order to infer collective preference and opinion. Subsequently, charging queries (politically) is proposed as a specific methodology to infer political partisanship. First, we address how search engines have sought to understand the user, and what information can be found in a query log.

Inferring (collective) preference from query logs

In order to learn from what the user does and wants, search engines will typically keep track of what their users search for and the result they click on. The kind of information collected prompted Battelle (2006) to depict search engines as 'databases of intentions' as they store massive amounts of 'desires, needs, wants, and preferences' [4]. However, queries are typically short (two or three terms) (Jansen and Spink, 2006) and often ambiguous: if a query reads [washington] it is not clear whether the city, newspaper, president or actor is meant. In order to provide the user with the best results, search engines use a variety of techniques to infer user intent. In this respect, information science literature usually distinguishes between three types of queries: navigational (to reach a specific Web site), informational (to find more information about a subject) and transactional (to perform some Web mediated activity) (Brenes, *et al.*, 2009). Increasingly, the user will be offered localized and personalized results too. For example, when a user searches for [restaurant], most likely one close by will be preferred. As location can be, approximately, inferred on the basis of a user's IP address or profile information, search engines offer local domain versions in order to more precisely determine the scope of results returned (Goldsmith and Wu, 2006) [5]. Except for location, a user's past queries turn out to be important in determining a query's intent (Grimes, *et al.*, 2007), hence the push to personalize search results. Additionally your friends' preferences, as expressed through their respective search histories, can be taken into account for the personalization of results (Feuz, *et al.*, 2011).

Except to improve a user's result rankings, query logs have also been used to study the distribution of specific cultural and social preferences, by employing three variables: query

terms (including volume), where the queries were made (location), and the queries' date stamps (Spink, *et al.*, 2009; Rogers, in press). By introducing the searcher's profile information (age, gender, zip code) combined with U.S. census data, one can not only track changes in the distribution of queries along geographic regions but also along demographic dimensions (Weber and Castillo, 2010; Weber and Jaimes, 2010). If we consider that search queries are valid indicators which can be employed for social and cultural research, search engine query histories might thus be used to provide the time and place, as well as the intensity of social and cultural preferences.

Most recently, Seth, *et al.* (2011) examined the full log of one month of queries submitted to the U.S. version of Google's search engine. They grouped the queries on a city–level, based on the user's IP address, and calculated an excess score to focus on those queries which occur either more or less than expected (*e.g.*, in each city [facebook] will be a very frequent query, but this is probably not the most interesting feature to characterize a city by). Based on disparities in query volume they calculated a city–similarity and compared it with a ground truth of city similarity based on census data. They found that 'query logs can be a good representation of the interests of the city's inhabitants and a useful characterization of the city itself' [6]. Weber and Jaimes (2011) came to similar findings based on Yahoo!'s query logs, which they exemplified by noting that the fraction of searches related to actors is about three times higher in the L.A. area, which includes Hollywood, than in any other region considered. Similarly, the fraction of queries related to gambling is highest in Las Vegas and lowest in Salt Lake City. Linking users' zip codes to U.S. census data, Weber and Castillo (2010) found that the Yahoo! query logs provide a good demographic description of the U.S. population and that different segments of the population differ in the topics they search for as well as in their search behavior (see also Weber and Jaimes, 2011).

Various authors recently have sought to appropriate trends of query volume as measures of public attention. Such research found that the search volume of specific (politics and issue) queries often correlates with fluctuations in news coverage (Weeks and Southwell, 2010; Granka 2009, 2010; Ripberger, 2011) and to a certain extent also with polls and surveys (Granka, 2009; Scheitle, 2011).

In a similar vein, other projects sought to match queries, considered as expressions of public interest and concern, to external data. One of the better–known projects, Google Flu Trends, asked whether the frequency of specific search queries (out of the 50 million most recurring U.S. queries) could be used as an indicator of regionally specific seasonal outbreaks of influenza. The project tried to match specific queries to Google with the U.S. Centers for Disease Control and Prevention's historical data on influenza outbreaks. The project concluded that very specific and frequent influenza–related queries can provide topical and geographically precise indicators of such an outbreak (Ginsberg, *et al.*, 2008).

Other work demonstrated that the trends in volume of specific search queries correlates surprisingly well with consumer activities as expressed in economic indicators like retail, automotive, and home sales, travel statistics, and unemployment indicators (Choi and Varian, 2009; Varian and Choi, 2009). Although search terms were found to provide valuable indicators of off–line phenomena, their predictive value often does not exceed simple baseline models (Goel, *et al.*, 2010). Similarly, while query volume of candidate names may reflect topical popularity, query volume is less likely to predict who wins the next election (Lui, *et al.*, 2011).

In many of these projects prior knowledge often influences the choice of queries, so as to match an external baseline. In May 2011 Google released a tool reversing this methodology. Instead of matching specific query trends to external data, on the basis of a pattern of some real world activity submitted, the tool automatically 'surfaces queries which correspond with [that] particular pattern of activity' [7].

In this research, we similarly regard queries as expressions of public interest. However, we do not attempt to match queries to some pattern of off–line activity but take inspiration from research pursued on the high–traffic recipe site allrecipes.com (http://allrecipes.com/). The researchers analyzed the queries and locations of over 750,000 users which searched for a recipe on the site, prior to Thanksgiving 2009 (the U.S. holiday feast), and found that 'regional differences [in taste] and the precise time [of a user's interest] could be pinpointed as never before' (Severson, 2009). The U.S. east coast, for instance, was more interested in recipes on 'sweet potato casserole' and the South and Middle more in 'pecan pie.'

Enriching query logs with other data has made them more useful for social and cultural research. Here, at first, we are not interested in correlating queries with data such as location and demographics. We take advantage of the relations within query logs and look at the queries leading to a click on a specific group of sites: political blogs. Just as allrecipes.com was used as a proxy to measure differences in taste, we have used political blogs as proxies of political intent to charge queries politically.

Political Insights

This study attempts to detect political issues and concerns by looking into the collective

search histories of users querying the U.S. version of the Yahoo! Web search engine, a data set extracted from nine months of anonymized Yahoo! search query logs, from May 2010 to January 2011 [8].

Blogs as proxies of political concern

Political blogs were chosen as our proxy for gathering queries with political intent as in U.S. politics they are an important source of political commentary. Amongst others, political blogs were studied in terms of link structure and content (Adamic and Glance, 2005; Hargittai, *et al.*, 2008; Kelly, 2010), author demographics and reachability (Hindman, 2008), technological adoption and use (Benkler and Shaw, 2010), as well as readership (Lawrence, *et al.*, 2010). Although initially political blogs where hoped to be vehicles to increase political deliberation, all these studies found these blogs to be polarized along opposing partisan lines [9]. Consequently, we hypothesize that in grouping the queries by the leaning of the blogs on which they land, a meaningful description of partisan concern is provided.

The 155 political blogs for which we gathered the queries landing on them, were listed by Benkler and Shaw (2010) who triangulated seven lists of top blogs, manually coding them as leaning towards the political spectrum's left, center, or right [10]. We considered other sites like those of election candidates, but queries to these sites turned out to be mostly navigational, showing interest in the candidate but not in the candidate's issues. Nor did we use the U.S. House of Representatives' or Congress' sites, not wanting to restrict politics to (official) government information. News sites were rejected as well, as they cover much more than politics alone.

Filtering the query log and retaining only those queries resulting in a click on a predefined set of political blogs' URLs, is based on the assumption that if a specific URL for a particular query is clicked, there is a relation between the two. This relation depends on three elements: the user submitting the query, the URL with content relevant to the query, and the search engine providing a ranked list of results relevant to the query (generally the results contain all the query's words). A search engine will typically return a variety of different (types of) sites, ranging from Wikipedia articles, videos, or news related to the query, to sites run by businesses, NGOs, individuals and authorities. By clicking a certain URL for the query submitted, the user thus not only shows interest in a specific URL but also in a particular type of site [11]. Although in our research the user thus reinforces the political relevance of a query, the focus is not on users but on how queries can be enriched by considering the types of site clicked.

Leveraging search engine results to enrich queries is similar to Goel, *et al.* (2010) who categorized queries as movie or game–related if such a site's URL appeared on the first page of search results. In this study we consider a query to be political if after submitting the query a political blog was clicked. As noted above, we drew inspiration from the work on allrecipes.com, which showed that queries to a specific type of site can provide (regional) characteristics of taste. In this article, however, at first we are not interested in the regional characteristics of queries but whether topical political sites can charge queries politically.

To our knowledge, nobody has yet studied queries landing on political blogs. Mishne and de Rijke (2006) investigated general characteristics of queries submitted to blog search engines. Hindman (2008) compares closest to our study by looking at queries landing on political sites. While Hindman only considered the top twenty queries of one month, we considered all queries landing on political blogs over nine months.

Politically charged queries

We filtered the query log retaining only those queries resulting in a click on the URLs of a predefined set of political blogs [12]. As these blogs were attributed a political leaning too, we could not only politically charge a query, but also determine its partisanship by attributing the query with a value for each leaning, proportional to the number of times the query landed on a blog of that leaning.

Several additional steps ensured that the queries are indeed politically relevant. After aggregating all queries landing on the described political blogs we removed all queries containing personally identifiable information such as credit card numbers, infrequent personal names, social security numbers, or street addresses. In the resulting set, many of the queries landing on political blogs turned out to be navigational (and thus hardly indicative of partisan concern). To filter out these navigational queries we used two complementary techniques. First we looked at the click entropy for each query to find out whether a diverse set of sites was clicked for a particular query. Queries with more than two occurrences but landing mostly on the same site (with an entropy not larger than 1.0), were considered navigational. Additionally, through the use of simple heuristics we tested whether there was a close match between the query and the clicked domain. We first tokenized queries and URLs (based on dots and spaces), stemmed plurals, and alphabetized the words. Subsequently, a query–URL pair is considered navigational if it contains a domain component such as 'www' or '.com,' the domain of the URL is contained in the query (or vice versa), or when the edit distance between queries and the domain is smaller than 2 (for queries with more than four characters). For example, [drudge], [drudge report], and [drudgerport] landing on http://www.drudgereport.com are all considered navigational queries.

To ascertain that our queries had a minimum shared uptake and relevance, we filtered the data to only retain queries resulting in a click–through to at least three political blogs. To

prevent one blog from setting the agenda we also removed queries with a very high query volume but resulting in click–throughs to very few political blogs.

Not all queries are equally frequent, and some might lead to, say, three clicks corresponding to a particular leaning. To address the corresponding sparsity issue and to avoid prematurely marking this query as 'strongly partisan,' we applied Bayesian smoothing which, in practice, means that we evenly distributed a small number of artificial clicks over all leanings, before accounting for the actually incurred clicks. Moreover, certain blogs in our list, *e.g.*, *Huffington Post*, attracted far more traffic than others; in turn making the left attract considerably more click volume than the right. As this potentially tainted the analysis and created a systematic bias towards the left, we normalized each leaning's total click counts by attributing the same total weight to the left, center and right. This might, however, be overly enthusiastic as the Web, for example, might overall be more left–leaning.

Political Insights: A gauge of partisanship

Previously we ascertained that the queries in our dataset are politically relevant; we politically 'charged' each query and assigned partisanship by the fraction of times it landed on a blog with a particular political leaning. We provide a searchable database of such politically charged queries at http://politicalinsights.sandbox.yahoo.com, ranking queries according to their assigned proportion of a particular leaning, *i.e.*, left, center, and right side of the political spectrum. The landing page of Political Insights displays a global ranking based on nine months' data. In fact we built a *political partisanship machine* by sifting out those queries most strongly linked to a particular political ideology.

Our system also allows the user to search for specific queries containing a particular word or phrase. In case of a match, all queries containing the search will be shown and ranked. See Figure 1: when searching for [obama] you will see queries like [obama accomplishments] to be more on the left side, and [obamacare] to be more on the right side of the political spectrum [13].

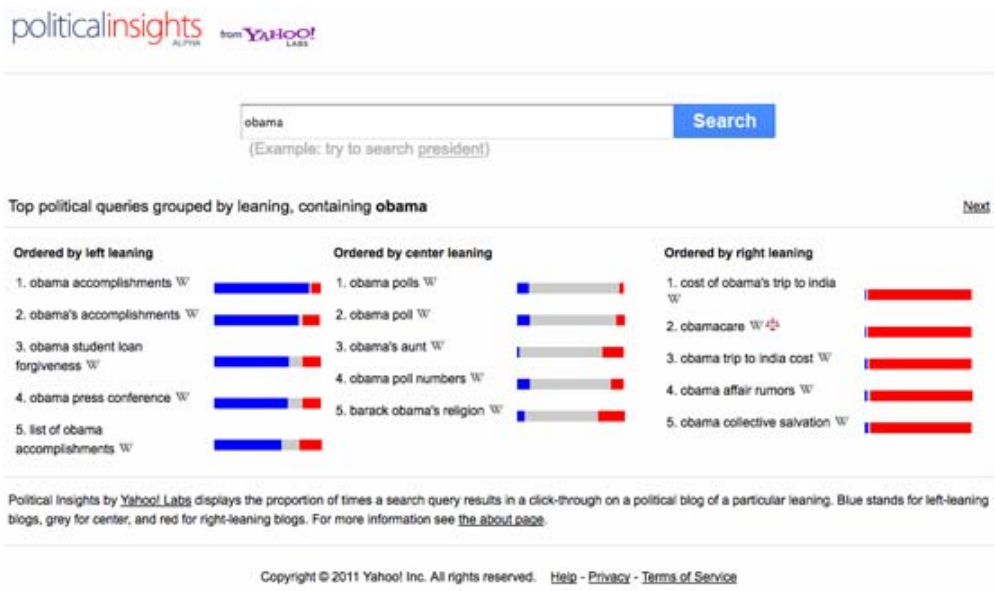

Figure 1: Screenshot of Political Insights. Top results for searches containing [obama]. Available online at http://politicalinsights.sandbox.yahoo.com/index.php?q=obama, accessed 1 July 2012.

In order for the user of our system to get an illustration of the relationship between a query and its politics, we provide the following. Clicking the query itself opens a new window containing its current search results restricted to the blogs of a particular leaning. Furthermore, we mapped all queries to the most relevant Wikipedia articles; by clicking the 'W' next to the query this article is shown, together with the article's categories. Finally, as highly partisan queries might be the result of an effort to introduce slant or spin, we tried to link queries to external 'fact–checking' sites, *i.e.*, http://factcheck.org, http://politifact.com, and http://snopes.com. For instance, when searching for [obama] and clicking on the scales symbol next to the query [obamacare], three results from politifact.com will be shown. For example, Mitt Romney, candidate for the 2012 Republican Party presidential nomination, is found to make a false allegation stating that 'Repealing the health care law would save $95 billion in 2016' [14].

Table 1 shows the top result rankings per leaning for exemplary queries containing particular politicians, issues and stances. As can be seen, in general the results are intuitively correctly aligned. By not only politically charging queries but also charting them along oppositional partisan lines, we actually shifted the notion from *actor partisanship* to *query partisanship*, arguably opening up new ways of research into framing (Entman, 1993). While we made sure

to remove navigational queries (*e.g.*, the names of individual blogs), highly partisan queries might be termed navigational as well, as they predominantly lead to blogs of one particular leaning.

Table 1: Examples of queries ranked by leaning. For clarity's sake, we only included queries which were attributed more than 50 percent of one respective leaning. All examples are available on http://politicalinsights.sandbox.yahoo.com (accessed 11 January 2012).
Note: * This query is included because the system allows partial matches.

Query	Left	Right
[obama] (politician)	[obama accomplishments] [obama student loan forgiveness] [obama press conference]	[cost of obama's trip to india] [obamacare] [obama affair rumours]
[bush] (politician)	[bush deficit] [george w. bush] [president george w bush costs of trip to crawford texas]	[jobs created under bush] [bush vs obama vacation days] [obama extending bush tax cuts]
[lies] (stance)	[glenn beck lies] [fox news lies] [list of republican lies]	[inconvenient truth lies] [lies about obama] [racist signs at tea party rallies]*
[violence] (issue)	[tea party violence] [mexico violence] [right wing violence]	[left wing violence] [liberal violence]
[immigration] (issue)	[immigration] [immigration reform 2010] [arizona immigration news]	[hanson moral implications of illegal immigration] [mexico immigration laws] [az immigrations news]
[gun] (issue)	[gun control] [ergun caner]* [chicago gun ban]	[is the government trying to take away our guns] [eric holder guns] [mexico false claims of us guns causing crime]
[job] (issue)	[jobs bill] [take our jobs]	[jobs created under bush] [1961 bill to send jobs overseas] [epa regulations to cost nearly a million jobs]

In what follows we look at whether our approach using blogs as proxies of political intent resulted in data which can be grounded in voter demographics and to what extent the data are representative of 'off–line' political preference.

Grounding the data

While the differences are insightful, the question remains whether charging queries politically has any relation with the 'off–line.' To what extent do users submitting political queries

represent the U.S. (voting) population? We look at both voter demographics and voting preferences.

(Voter) demographics

By combining the user–provided zip code with U.S. census information we investigated whether gender, age, race and educational level were representative for the U.S. population [15]. Using the U.S. 2000 census data on, where appropriate, a per–zip code level we found that (i) our users were predominantly male (54.7 percent vs. 49.1 percent in the census), (ii) older (median age of 45 vs. 35 in the census), (iii) more white (78.4 percent vs. 75.1 percent in the census) and (iv) more highly educated (27.8 percent vs. 24.4 percent in the census — fraction of population of 25 years and older with at least a B.A. degree) [16].

Comparing the same census data with the 2010 voting records for registered voters we observed that (i) the gender bias was even more pronounced (54.7 percent vs. 46.6 percent), (ii) our users were slightly younger (median age of 45 vs. approximately 47), (iii) not white enough (78.4 percent vs. 83.4 percent) and (iv) less educated (27.8 percent vs. 32.1 percent with at least a B.A. degree) [17]. However, some caution is appropriate. First, the available census data date back from 2000, the voting records from 2010. So the U.S. population has aged since then, probably eliminating our observed age gap, has become less white, further increasing our observed racial gap, and more highly educated, reducing the educational gap. Concerning the latter, the voting records include people of age 18 and above, while the census definition for educational attainment in this category considers only ages 25 and higher. This would even increase the actual gap. However, people holding a B.A. degree or higher, regardless of the characteristics of their zip code, are more than 10 percent more likely to register to vote than an average citizen. This most likely explains the observed educational difference and we do *not* believe that the users in our sample have a lower educational attainment than the average voter.

2010 U.S. midterm elections

To ascertain whether the online notion of 'left (or right) leaning blog' is linked to the off–line notion of 'voting Democrat (or Republican)' we used per–zip results for the 2010 U.S. House of Representatives elections. We computed the probability that a person who clicked on left blogs ('left–clicking') voted Democrat in the 2010 U.S. elections. To estimate this probability we used the election results for the zip code from the user's profile and assumed that the user was drawn uniformly at random from the voting population. Averaging these over all left–clicking users and all zip codes led to the following equation.

Equation 1:
$$\frac{\sum_z c_z^l \cdot v_z^D}{\sum_z c_z^l}$$

Here c_z^l is the count of left–clicking users in zip code *z*. v_z^D is the fraction of voters voting Democrat in zip code *z*. In a similar manner we can define c_z^r for right–clicking users and v_z^R for Republican voting fractions.

We can thus estimate the probability that a left–clicking user voted Democrat or that a right–clicking user voted Republican. The value above was multiplied by 100 so that the probability estimate lies between 0 percent and 100 percent.

If each zip code voted either 100 percent Democrat or 100 percent Republican, the estimate could theoretically attain 100 percent. However, if each zip code was split 50–50 the maximum would be 50 percent. In fact, across the entire U.S. the Democrat–Republican split was 44.8–51.4 and for zip codes with users clicking on the considered blogs this split was 45.1–49.3, where zip codes were weighted by the number of users.

Given this roughly equal fraction of Democrat and Republican votes, we computed a more realistic bound for our probability estimates. We replaced the number of left–clicking users in a given zip code in Equation 1 by the total number of users in the zip code multiplied by the fraction voting Democrat. This bound corresponds to the case where our assumptions are fully correct and 'left–clicking' equals 'voting Democrat.' This leads to the following equation.

Equation 2:
$$\frac{\sum_z c_z^l v_z^D \cdot v_z^D}{\sum_z c_z^l v_z^D}$$

Again, similar bounds were obtained for Republicans.

If all users clicking at least one political blog are taken into account for this estimate we get Table 2. The upper bounds for this case are 53.1 for left–clicking and Democrat and 56.6 for right–clicking and Republican. Although Table 2 indicates that the trends go into the right direction, *i.e.*, left–clicking users are more likely to vote Democrat, the difference with the

upper bound is still considerable. In an attempt to reduce this gap, we experimented with two ideas. First, we hypothesized a temporal dimension, implying that the match between clicking and voting behavior improved closer to the actual election date. However, overlapping intervals of three months did not reveal any temporal dynamics. Second we hypothesized that users clicking more frequently on a blog of a particular leaning are better indicators for the voting behavior in the corresponding zip code. To test this, we used the top 1,000 users in terms of numbers of clicks for each of the three leanings considered [18]. The results are presented in Table 3. For this set of users the upper bound for left–clicking and Democrat was 54.2 and for right–clicking and Republican 56.2. All the pair–wise differences in Table 3 were found to be significant at a level of 1 percent, using a t–test where each user and the voting estimate of the zip code corresponded to one data point.

Table 2: Estimated voting probability of all users clicking right, center, or left leaning blogs.

Clicked	**Estimated voting probability**	
	Democrat	Republican
Right	43.3	50.8
Center	44.9	49.6
Left	45.5	49.0

Table 3: Estimated voting probability of the top 1,000 users for each leaning.

Clicked	**Estimated voting probability**	
	Democrat	Republican
Right	43.2	51.2
Center	45.9	48.3
Left	49.1	46.4

Overall, our observations indicate that the leaning of the blogs a person clicks on in response to Web search queries correlates with the voting behavior of the area where the person resides. This correlation is stronger for users who repeatedly click on blogs of a particular leaning. The fact that we did not quite attain the bounds for a perfect fit of our model ($49.1 < 54.2$ and $51.2 < 56.2$) can be explained in a number of ways. Not all blogs focus purely on politics; they contain different content as well. This is particularly valid for the *Huffington Post*, which also covers celebrity news; it demonstrates how important good source (proxy) selection is. Another explanation is technical in nature. The election results were retrieved from *USA Today*, which displayed the results per district ID instead of zip code, requiring a not always unambiguous conversion: zip codes, through redistricting, may belong to different legislative districts at different times, in function of population changes [19]. Our information came from the ZCTA to district mapping from the 110th Congress (applicable from 2007 to 2009) [20]. Thus, matching ZCTA to zip codes and using the 110th instead of the 112th legislative district mapping might have introduced errors.

Other explanations include the possibility that the voting and blog–clicking populations are not identical. This could hold on a nation–wide level where, say, older people are less likely to use the Internet or on a per–zip level where voters in a clearly defined state or region do not consult blogs to shape their voting choice.

Summarizing, we find that compared to the average voter our users have about the right age, are predominantly male, not white enough, and have about the right educational background. In addition, we verified that people clicking blogs of a specific leaning are more likely to live in a zip code with a higher proportion of voters with that leaning. This finding is in line with the survey about political blog readership by Lawrence, *et al.* who find that "blog readers gravitate towards blogs that accord with their political beliefs" [21]. It might be argued that politically charged queries disclose partisan concern, as the likely voting behavior of users submitting those queries correlates with the leaning of the blogs they click.

Conclusion and future work

Query logs have been heralded as an addition, or even alternative, to traditional social science data because they are unprecedented in scope, scale and detail, and the queries are obtained from within their natural environment. However, queries being short they are often

hard to interpret; they seem to have no depth and little associated context. In this paper the careful selection of topical sites, clicked in response to a query, is presented as a proxy with which queries with shared concern can be discovered. The recognition that partisan political blogs can be used as such a proxy to detect political concern, allowed us to charge queries politically and attribute partisanship. This in turn allows us to sift out highly partisan queries to provide detailed insights into political concerns. Subsequently, we found that the leaning of the blogs people read correlate with their likely voting behavior.

The Political Insights tool is based on static data dating from around the 2010 U.S. midterm elections. Ranking the queries according to partisanship made the tool into a *gauge* of query partisanship. In other work we have extended upon this core methodology by using more fresh data and tracking changes over time, permitting us to consider trending queries which can then be ranked by partisanship (Weber, *et al.*, 2012). The resulting *barometer* of political partisanship allows answering questions such as: 'what is trending among the political left?' Additionally, we countered noise and misspellings by grouping similar queries on their stemmed and normalized form. A more extensive use of fact–checking sites, linking the truth–value of queries back to leanings, allowed us to investigate for instance which leaning has the highest query volume in relation to false allegations. We are also considering various other extensions of our application. The 'search the left' or 'search the right' functionality, as in the section describing our application, could be an interesting service in itself, juxtaposing the two leanings' results and queries extracted thereof. Instead of providing a national outlook we could also zoom in on a smaller geographical level. Last but not least, we intend to test our hypothesis that charging queries with shared matters of concern and additionally ranking them by opposing partisanship can be employed to show partisanship in other domains too, such as climate change alarmists versus skeptics.

After Weber and Castillo's work (2010), Yahoo! Clues was made public, a search analysis service allowing 'you to instantly discover what's popular to a select group of searchers — by age or gender [or location] — over the past day, week or even over the past year' (Theodore, 2011). Similarly, we released Political Insights hoping that it will be useful for (re–)searchers. We believe that such online tools offer researchers new horizons to sociological research by simultaneously allowing access to the individual component (queries), as well as the aggregated structure (demographic breakdown and political partisanship respectively).

Whilst our tool cannot predict who will win the next U.S. Presidential elections, it describes which issues resonate most with different sides of the political spectrum and provides insight into political partisanship by placing side by side competing claims. The increasing polarization of (political) discourse, combined with online recommendation cultures suggesting information based on what like–minded have done before, led to warnings for echo chamber effects (Sunstein, 2006) and filter bubbles (Pariser, 2011). We believe that it is beneficial to make these effects insightful. As Lippmann so eloquently stated:

> The individual not directly concerned may still choose to join the self–interested group and support its cause. But at least he will know that he has made himself a partisan, and thus perhaps he may be somewhat less likely to mistake a party's purpose for the aim of mankind. [22]

About the authors

Erik Borra is a Ph.D. candidate and lecturer at the University of Amsterdam's Media Studies department. He is also lead developer for the Digital Methods Initiative, the Ph.D. research program in New Media at the University of Amsterdam.
E–mail: borra [at] uva [dot] nl

Ingmar Weber is a researcher at Yahoo! Research Barcelona. He works on query log analysis, often with a demographic angle, and on large–scale information extraction from Web content.
E–mail: ingmar [at] yahoo-inc [dot] com

Acknowledgements

We would like to thank Venkata Rama Kiran Garimella for the implementation of the fact–checking functionality, and Richard Rogers and Esther Weltevrede for their insightful comments.

Notes

1. In our study, all queries were anonymized by removing personally identifiable information such as telephone numbers, street addresses, social security numbers or infrequent personal names. Yahoo! user names pertaining to queries were replaced by random numbers. All of our data analysis is done in aggregate, without tracking individual users. As of 1 July 2012 Political Insights is still available at http://politicalinsights.sandbox.yahoo.com. However, as the development of this tool has quickly progressed, a more advanced version using the exact same methodology but including search trends is readily available at http://politicalsearchtrends.sandbox.yahoo.com (accessed 1 July 2012). This paper introduces the core methodology on which both tools are built.

2. Whenever we write about a specific query we encapsulate it in brackets to delineate it.

3. Webb, *et al.*, 1972, pp. 1–2; Ferguson, 2000, pp. 21–30.

4. Batelle, 2006, p. 6.

5. Search engines often provide additional services like e–mail for which a user is obliged to fill in a profile. If a user is logged into such a service, logs can be complemented with information, such as a zip code, contained in the profile.

6. Seth, *et al.*, 2011, p. 1.

7. Mohebbi, *et al.*, 2011, p. 2.

8. Note that this period includes the November 2010 U.S. midterm elections.

9. In this study we do not look for the causes or effects of partisanship but take it as a given. For an insightful discussion on the political effects of media choice, see Prior (2007).

10. A 2005 poll of 2,209 U.S. citizens indicates that the labels left and right are generally associated with liberal and conservative respectively (PR Newswire, 2005). We, and our colleagues very familiar with U.S. politics, verified the importance and coding of the blogs.

11. In the studied period results were not personalized; all users were shown the same results.

12. The procedural subsection is also available at http://erikborra.net/blog/2012/04/methods-for-exploring-partisan-search-queries (23 April 2012), accessed 20 May 2012.

13. Note that similar queries might be displayed in the ranked list of one leaning, as in the current version of the application queries are not grouped, but left untouched.

14. http://www.politifact.com/truth-o-meter/statements/2011/nov/04/mitt-romney/mitt-romney-said-repealing-obamacare-would-save-95/, accessed 15 January 2012.

15. To discover relevant queries per leaning we used all available query–click pairs, whereas here we look at those with an associated user profile.

16. http://factfinder.census.gov/home/saff/main.html?_lang=en, accessed 15 December 2011.

17. Page 4 on http://www.census.gov/prod/2010pubs/p20-562.pdf lists voter demographics. The exact median age is not reported and must be inferred from results reported for age buckets, accessed 10 December 2011.

18. For all other parts of our analysis many more than 1,000 users contributed.

19. The election results were scraped from http://projects.usatoday.com/news/politics/2010/elections/, accessed 15 December 2011.

20. http://www.census.gov/geo/www/cd110th/natl_code/zcta_cd110_natl.txt, accessed 15 December 2011.

21. Lawrence, *et al.*, 2010, p. 141.

22. Lippmann, 1993, p. 104.

References

L.A. Adamic and N. Glance, 2005. "The political blogosphere and the 2004 U.S. election: Divided they blog," *LinkKDD '05: Proceedings of the Third International Workshop on Link Discovery*, pp. 36–43.

J. Battelle, 2006. *The search: How Google and its rivals rewrote the rules of business and transformed our culture*. New York: Portfolio.

Y. Benkler and A. Shaw, 2010. "A tale of two blogospheres: Discursive practices on the left and right," Harvard University Berkman Center for Internet and Society, Research Publication, number 2010–6 and Harvard Public Law Working Paper, number 10–33, at http://cyber.law.harvard.edu/publications/2010/Tale_Two_Blogospheres_Discursive_Practices_Left_Right,

accessed 23 June 2012.

C.L. Borgman, 2009. "The digital future is now: A call to action for the humanities," *Digital Humanities Quarterly*, volume 3, number 4, at http://digitalhumanities.org/dhq/vol/3/4/000077/000077.html, accessed 6 February 2010.

d. boyd and K. Crawford, 2011. "Six provocations for big data," *A Decade in Internet Time: Symposium on the Dynamics of the Internet and Society* (September), at http://ssrn.com/abstract=1926431, accessed 14 January 2012.

D.J. Brenes, D. Gayo–Avello and K. Pérez–González, 2009. "Survey and evaluation of query intent detection methods," *WSCD '09: Proceedings of the 2009 Workshop on Web Search Click Data*, pp. 1–7.

H. Choi and H. Varian, 2009. "Predicting initial claims for unemployment benefits," at http://research.google.com/archive/papers/initialclaimsUS.pdf, accessed 13 December 2011.

D. Dodge, 2007. "Search engines are the Start page for the Internet," *Don Dodge on The Next Big Thing* (13 December), at http://dondodge.typepad.com/the_next_big_thing/2007/12/search-engines.html, accessed 14 January 2012.

R.M. Entman, 1993. "Framing: Toward clarification of a fractured paradigm," *Journal of Communication*, volume 43, number 4, pp. 51–58.

S.D. Ferguson, 2000. *Researching the public opinion environment: Theories and methods*. London: Sage.

M. Feuz, M. Fuller and F. Stalder, 2011. "Personal Web searching in the age of semantic capitalism: Diagnosing the mechanisms of personalization," *First Monday*, volume 16, number 2, at http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/3344/2766, accessed 23 June 2012.

J. Ginsberg, M.H. Mohebbi, R.S. Patel, L. Brammer, M.S. Smolinski and L. Brilliant, 2008. "Detecting influenza epidemics using search engine query data," *Nature*, volume 457, number 7232, pp. 1,012–1,014.

S. Goel, J.M. Hofman, S. Lahaie, D.M. Pennock and D.J. Watts, 2010. "Predicting consumer behavior with Web search," *Proceedings of the National Academy of Sciences of the United States of America*, volume 107, number 41, pp. 17,486-17,490.

J.L. Goldsmith and T. Wu, 2006. *Who controls the Internet? Illusions of a borderless world*. New York: Oxford University Press.

L.A. Granka, 2010. "Measuring agenda setting with online search traffic: Influences of online and traditional media," *Annual Meeting of the American Political Science Association*, at http://ssrn.com/abstract=1658172, accessed 23 June 2012.

L.A. Granka, 2009. "Inferring the public agenda from implicit query data," *SIGIR '09: Understanding the User–Logging and Interpreting User Interactions in Information Search and Retrieval*, at http://laura.granka.com/publications/granka_SIGIR09paper.pdf, accessed 23 June 2012.

C. Grimes, D. Tang and D.M. Russell, 2007. "Query logs alone are not enough," *WWW 2007: Workshop on Query Log Analysis*, at http://www2007.org/workshop-W6.php, accessed 23 June 2012.

M. Gruszczynski, 2011. "Examining the role of affective language in predicting the agenda–setting effect," *APSA 2011 Annual Meeting Paper*, at http://ssrn.com/abstract=1902270, accessed 23 June 2012.

E. Hargittai, J. Gallo and M. Kane, 2008. "Cross–ideological discussions among conservative and liberal bloggers," *Public Choice*, volume 134, numbers 1–2, pp. 67–86.

M. Hindman, 2008. *The myth of digital democracy*. Princeton, N.J.: Princeton University Press.

B.J. Jansen and A. Spink, 2006. "How are we searching the World Wide Web? A comparison of nine search engine transaction logs," *Information Processing & Management*, volume 42, number 1, pp. 248–263.

J. Kelly, 2010. "Parsing the online ecosystem: Journalism, media, and the blogosphere," In: G. Einav (editor). *Transitioned media: A turning point into the digital realm*. New York: Springer, pp. 93–108, at http://www.springerlink.com/content/q836x0472284076k/, accessed 4 January 2012.

E. Lawrence, J. Sides and H. Farrell, 2010. "Self–segregation or deliberation? Blog readership, participation, and polarization in American politics," *Perspectives on Politics*, volume 8, number 1, pp. 141–157.

D. Lazer, A.S. 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, 2009. "Life in the network: The coming age of computational social science," *Science*, volume 323, number 5915, pp. 721–723.

W. Lippmann, 1993. *The phantom public*. With a new introduction by W.M. McClay. London: Transaction Publishers.

C. Lui, P.T. Metaxas and E. Mustafaraj, 2011. "On the predictability of the U.S. elections through search volume activity," *Proceedings of the IADIS International Conference on e–Society*, at http://cs.wellesley.edu/~pmetaxas/e-Society-2011-GTrends-Predictions.pdf, accessed 23 June 2012.

L. Manovich, 2012. "Trending: The promises and the challenges of big social data," In M.K. Gold (editor). *Debates in the digital humanities*. Minneapolis: University of Minnesota Press, pp. 460–475.

N. Marres, in press. "The environmental teapot and other loaded household objects: Re–connecting the politics of technology, issues and things," In: P. Harvey, E. Casella, G. Evans, H. Knox, C. McLean, E. Silva, N. Thoburn and K. Woodward (editors). *Objects and materials: A Routledge companion*. London: Routledge.

G. Mishne and M. de Rijke, 2006. "A study of blog search," In: M. Lalmas, A. MacFarlane, S.M. Rüger, A. Tombros, T. Tsikrika and A. Yavlinsky (editors). *Advances in information retrieval, 28th European Conference on IR Research, ECIR 2006, London, U.K., April 10–12, 2006, Proceedings. Lecture Notes in Computer Science*, number 3936, pp. 289–301.

M. Mohebbi, D. Vanderkam, J. Kodysh, R. Schonberger, H. Choi and S. Kumar, 2011. "Google Correlate Whitepaper" (9 June), at https://www.google.com/trends/correlate/whitepaper.pdf, accessed 13 December 2011.

E. Pariser, 2011. *The filter bubble: What the Internet is hiding from you*. London: Penguin Press.

M. Prior, 2007. *Post–broadcast democracy: How media choice increases inequality in political involvement and polarizes elections*. New York: Cambridge University Press.

PR Newswire, 2005. "Political labels: Majorities of U.S. adults have a sense of what conservative, liberal, right wing or left wing means, but many do not" (9 February), at http://www.prnewswire.com/news-releases/political-labels-majorities-of-us-adults-have-a-sense-of-what-conservative-liberal-right-wing-or-left-wing-means-but-many-do-not-54020207.html, accessed 15 January 2012.

M. Richardson, 2008. "Learning about the world through long–term query logs," *ACM Transactions on the Web*, volume 2, number 4, article number 21.

J.T. Ripberger, 2011. "Capturing curiosity: Using Internet search trends to measure public attentiveness," *Policy Studies Journal*, volume 39, number 2, pp. 239–259.

R. Rogers, in press. *Digital methods*. Cambridge, Mass.: MIT Press.

M. Scharkow and J. Vogelgesang, 2011. "Measuring the public agenda using search engine queries," *International Journal of Public Opinion Research*, volume 23, number 1, pp. 104–113.

C.P. Scheitle, 2011. "Google's insights for search: A note evaluating the use of search engine data in social research," *Social Science Quarterly*, volume 92, number 1, pp. 285–295.

R. Seth, M. Covell, D. Ravichandran, D. Sivakumar and S. Baluja, 2011. "A tale of two (similar) cities: Inferring city similarity through geo–spatial query log analysis," In: J. Filipe and A.L.N. Fred (editors). *KDIR 2011: Proceedings of the International Conference on Knowledge Discovery and Information Retrieval*, pp. 179–189.

K. Severson, 2009. "Butterballs or cheese balls, an online barometer," *New York Times* (25 November), at http://www.nytimes.com/2009/11/26/dining/26search.html, accessed 5 June 2011.

A. Spink, B. Jansen and I. Taksa, 2009. "Web log analysis: Diversity of research methodologies," In: B.J. Jansen, A. Spink and I. Taksa (editors). *Handbook of research on Web log analysis*. Hershey, Pa.: Information Science Reference, pp. 506–522.

C.R. Sunstein, 2006. *Infotopia: How many minds produce knowledge*. New York: Oxford University Press.

M. Thelwall, L. Vaughan and L. Björneborn, 2005. "Webometrics," *Annual Review of Information Science and Technology*, volume 39, number 1, pp. 81–135.

B. Theodore, 2011. "New Yahoo! Clues launches," *Yahoo! Search Blog* (29 June), at http://www.ysearchblog.com/2011/06/29/new-yahoo-clues-launches/, accessed 30 June 2011.

H.R. Varian and H. Choi, 2009. "Predicting the present with Google Trends," *Google Research Blog* (2 April), at http://ssrn.com/abstract=1659302, accessed 5 December 2011.

T. Venturini, 2010. "Diving in magma: How to explore controversies with actor-network theory," *Public Understanding of Science*, volume 19, number 3, pp. 258–273.

E.J. Webb, D.T. Campbell, R.D. Schwartz and L. Sechrest, 1972. *Unobtrusive measures:*

Nonreactive measures in the social sciences. Chicago: Rand McNally.

I. Weber and A. Jaimes, 2011. "Who uses Web search for what? And how?" *WSDM '11: Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, pp. 15–24.

I. Weber and C. Castillo, 2010. "The demographics of Web search," *SIGIR '10: Proceedings of the 33rd international ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 523–530.

I. Weber and A. Jaimes, 2010. "Demographic information flows," *CIKM '10: Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, pp. 1,521–1,524.

I. Weber, V. R. K. Garimella and E.K. Borra, 2012. "Mining Web query logs to analyze political issues," *ACM Web Science 2012: Conference Proceedings*, pp. 479–488.

B. Weeks and B. Southwell, 2010. "The symbiosis of news coverage and aggregate online search behavior: Obama, rumors, and Presidential politics," *Mass Communication and Society*, volume 13, number 4, pp. 341–360.

Editorial history

Received 20 May 2012; accepted 23 June 2012.

Political Insights: Exploring partisanship in Web search queries
by Erik Borra and Ingmar Weber
First Monday, Volume 17, Number 7 - 2 July 2012
http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/rt/printerFriendly/4070/3272