

Inferring the Public Agenda from Implicit Query Data

Laura Granka
Stanford University
Google, Inc

granka@stanford.edu or
granka@google.com

ABSTRACT

Traditionally, implicit feedback measures are used to evaluate the performance of a particular information retrieval system. This research instead takes a distinctly applied approach to the use of implicit feedback, and extends the inference from aggregate query data to the social and political sciences. Using the three months prior to the 2008 election as a test scenario, the analysis here assesses daily fluctuations in search coverage of candidates and issues as predicted by the amount of news coverage, proximity to election day, and public opinion poll ratings of the candidates. Findings indicate that aggregate shifts in topical search queries may in fact be a useful, inexpensive indicator of political interest.

Categories and Subject Descriptors

H.3.3 [Information Systems]: Information Search and Retrieval – *relevance feedback, search process*; J.4. [Social and Behavioral Sciences]: Sociology.

General Terms

Measurement, Economics, Human Factors.

1. INTRODUCTION

Information retrieval systems have frequently used measures of implicit feedback to evaluate both the performance of a retrieval system and infer searcher satisfaction [1, 2]. Implicit feedback refers to measures that are unobtrusively obtained from a user search session, such as clicks, queries, reading time, session length, and page scrolling. Implicit feedback has been used most frequently to infer result relevance based on user click behavior and reading time [3, 4], and has been validated with eyetracking [3]. To date, little research has applied implicit feedback to situations beyond the actual retrieval system. As usage of online search engines is only increasing [5], it is important to understand how implicit search behavior can be applied to other domains to understand broader user conditions. This research presents results from one such analysis, and discusses additional ways in which implicit feedback can benefit the political and social sciences.

2. AGGREGATE USER FEEDBACK

Much of the work done in the social sciences depends heavily on survey and experimental research. Both of these methods, while extremely desirable when controlling for individual-level variables (e.g., education, age, gender, political affiliation), are

both costly and time-consuming to instrument and analyze. These measures are also susceptible to self-report or self-selection bias, particularly for questions assessing civic engagement or interest in public affairs [6, 7]. The ability to gauge aggregate changes in public opinion and issue awareness, in an immediate and inexpensive manner, is often the most desired alternative. One currently untapped tool for this is publicly available search query data. One specific platform, Google Insights for Search [8], offers users the ability to access daily changes in query volume for specific searches in a specific geography and time period. Existing research using this tool has shown how search volume is both indicative and predictive of external events, from flu outbreaks [9] to economic activity [10].

2.1 Search Queries and Topical Interest

Online search is an active medium, meaning that a user has to explicitly make the effort to acquire information about a given topic by manually typing in a query. Because of this, online searches queries may be a strong behavioral indicator of what issues and topics are at the top of a user's mind. This, coupled with the lack of self-report bias makes search queries an attractive way to implicitly measure fluctuations and changes in political issue interest over time.

Existing political and media research has tracked changes in issue interest over time, though as previously mentioned, through surveys or experiments. Research has repeatedly shown that public perceptions of issue importance are shaped by the amount of news coverage of that issue [11, 12]. In other words, the issues receiving the greatest news coverage are judged to be the most important issues. Our first step is to conduct a systematic evaluation of whether behavioral data obtained via search query volume is also consistent with the conclusions of agenda setting research. In other words, how do real world events and news coverage motivate political search traffic? The issues covered most prominently in the media are typically the issues that people judge to be most important; as such, we would expect to see these perceptions of importance reflected in a greater volume of online searches.

Overall, we hypothesize a strong level of convergence between search queries and news volume. The more interesting insights in our analysis will likely be the deviant cases – instances where the search query volume for a topic or issue exceeds what might be expected by its respective news coverage. Certain issues may be marked by extended periods of search activity, potentially revealing the topics that sustain audience interest enough to pursue additional information past the peak of news coverage.

Copyright is held by the author/owner(s).

SIGIR'09, July 19-23, 2009, Boston, USA.

3. METHODS

Standard surveys gauge public interest in political issues by first assessing issue awareness, and secondly, measuring perceived importance (via a rating scale). Search queries have the advantage of being able directly measure the first dimension – issue awareness. In order to perform a search, an individual already has to know about the topic or individual being queried for. While we don't know the level of detailed knowledge an individual may have about this issue, we do know that the individual knows about the topic and is making an effort to find out more about it.

Second, perceived importance is a bit trickier to measure through queries, but can still be done in a more indirect approach. The degree of importance attributed to a given issue can be inferred from overall aggregate changes in query volume for that given topic. Deviation from the norm query volume can be easily exemplified with seasonal examples – for instance, using Google trends, it is clear to see that in the United States, searches such as mittens or gingerbread increase in December. One would expect to see a similar phenomenon for political issues: query volume will reflect the rise and fall of public interest. In sum, overall changes in query volume, or sudden spikes in query volume, are two potential ways to assess how prominent or "important" an issue may be at any given time.

3.1 Data Collection

The data for this research was taken from the three months prior to the 2008 presidential election – the 92 days from August 1, 2008 to October 31, 2008. Overall news coverage was measured by counting instances of issues and political candidates being covered in transcripts from the three major US news networks (ABC, CBS, NBC). Transcripts were obtained from the Vanderbilt transcript database.

Coverage for each candidate and issue was obtained on a daily basis, to ascertain the changing volume of news coverage for every single day during this three-month period. As an additional step, the news coverage data was normalized according to the same normalization scheme as the search query data (as described below), so that when necessary, means could be compared between the two data sources.

Daily query volume data was downloaded from Google Insights for Search [8], which is publicly available online. The range of data collected was over the same three-month period, and limited to US websearch traffic. The purpose of this analysis is to determine the domestic effects of the US presidential campaign, so queries and news coverage were specifically chosen to represent the US market. The query volume does not reflect the actual number of queries that Google received; rather, the data is normalized according to the highest point in the data set, which receives a score of 100 (e.g., if there were 12 million searches for Obama on September 3, that day would receive a score of 100. If there were 6 million searches for Obama on August 1, that day would receive a score of 50). Other normalization factors are used to account for base increases in search traffic over time due to growth in the online population.

The query distributions for individual issues and candidates can then be compared with network news coverage of that issue or candidate. While the query means are not useful points of comparison between issues (each query resides in its own normalized set of data), the standard deviations may be useful, as they are representative of how regular or irregular searches are for

a given term, such as whether certain terms are more severely punctuated with spikes in traffic.

The selected issues varied in degrees of their newsworthiness and sensationalism. As measures of “hard news,” or substantive issues to the US, we tracked occurrences of the terms Iraq, War, Economy, Unemployment, Health Care, Taxes, and Education. To assess more sensationalist or “soft news” coverage, the terms Joe the Plumber, Tina Fey, and Saturday Night Live were analyzed. News coverage and query volume for each candidate's name – Obama, Biden, McCain, and Palin – were also obtained.

4. RESULTS

4.1 Election Proximity, News, and Search

For many campaign issues, the volume of news coverage significantly influenced subsequent search volume. Table 1 presents regression results using news volume and proximity to Election Day as predictors for search query volume. For most issues and candidates, there was a significant relationship between the issues covered in the news and the issues that people were most interested in searching for. However, for the topics War, Unemployment, and Health Care, proximity to the election was more influential than news coverage. In other words, searches for these terms increased as Election Day grew closer, irrespective of news coverage.

Table 1. Predicting Query Volume of Campaign Issues

Issue	Econ	War	Unemp	Taxes	Iraq	Hltcar	Educ
Intercept	16.44 (1.86)	51.61 (1.72)	48.43 (3.41)	48.64 (2.11)	49.39 (2.60)	47.79 (4.21)	60.30 (3.72)
News Trans	1.16** (0.11)	0.24* (0.10)	0.65 (0.79)	0.70** (0.25)	0.98** (0.23)	1.32# (0.66)	2.42** (0.57)
Election Prox	0.19** (0.49)	0.43** (0.02)	0.20** (0.07)	0.24** (0.05)	0.40** (0.04)	0.31** (0.06)	0.14* (0.06)
St. Er. Reg	8.70	5.35	16.12	9.85	9.52	16.17	14.65
R ²	0.81	0.82	0.13	0.48	0.59	0.25	0.23
F Stat	189	200.3	6.47	40.82	64.48	14.8	13.48

Table 2. Query Volume for Candidates and Personalities

Issue	McCain	Obama	Palin	Joe the Plumber	Tina Fey
Intercept	-1.22 (46.39)	-74.40 (47.37)	5.09 (2.55)	-0.38 (1.38)	2.67 (3.60)
News Transcripts	0.78** (0.12)	0.45** (0.08)	0.87** (0.09)	2.97** (0.17)	4.59** (0.81)
Election Proximity	0.15** (0.05)	0.28** (0.08)	-0.01 (0.05)	0.02 (0.03)	0.18* (0.07)
Poll Data	0.09 (1.06)	1.69 (1.06)	—	—	6.40
St. Error Regression	11.72	10.09	—	—	17.13
R ²	0.50	0.66	0.57	0.82	0.38
F-Stat	29.81	57.91	57.94	197.4	27.2

Standard errors are reported in parentheses

No observations = 92

Significant p-values are indicated: ** p<.001, * p<.01, # p=.05

This may indicate that searches for these topics are driven by interest or perceived importance, potentially signaling that these issues are important to searchers. An October, 2008 Gallup report indicates that the key issues important to voters were the economy, gas prices, Iraq, healthcare, and terrorism [14]. While gas prices and terrorism were not included in this analysis, the results from this study did compare with the Gallup results, as searches for economy, Iraq and healthcare increased prior to the election (Table 1). Additionally, it was clear that broadcast news did not equally cover the issues of public concern. Figure 1 shows density plots of search queries and news coverage for two issues in our sample: economy and war. Economic news coverage fairly consistently predicts queries for economy; however, a similar trend does not exist when assessing news and queries for war.

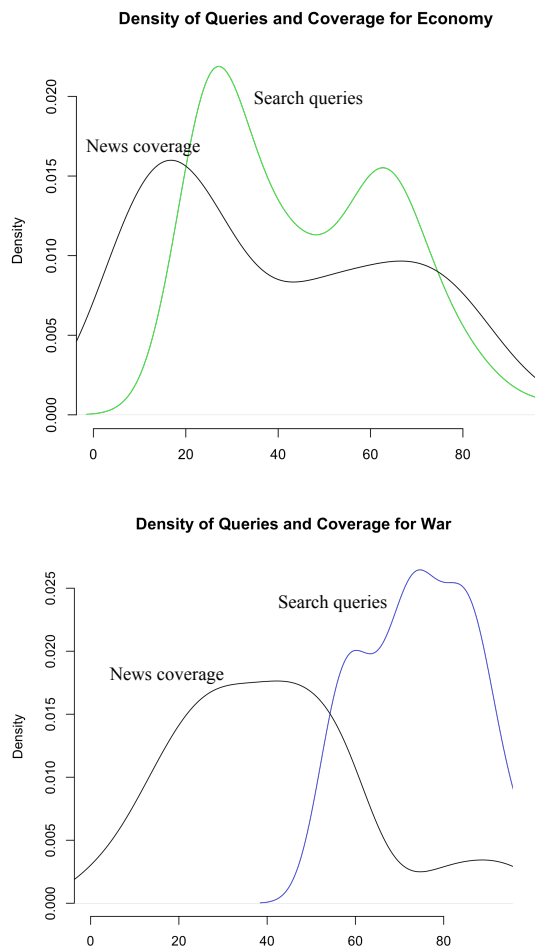


Figure 1. Top: density of news coverage (black) and search queries (green) for economy. Bottom: density of news coverage (black) and search queries (blue) for war. The x-axis represents the 92 days from Aug 1 – October 31, 2009.

4.2 Candidate Queries

Table 2 presents regression results predicting search queries for the candidates (Obama, McCain, Palin) and entertainment personalities (Tina Fey, Joe the Plumber), again using news volume and proximity to the election as predictors. A variable measuring public opinion approval, as assessed through polling data, was also included for the two presidential candidates. This

data was obtained from Pollster.com, which aggregates multiple public opinion polls, and allows users to download data [13].

The regressions show that (as with issue searches), there is a significant relationship between the volume of news queries and the volume of searches for both candidates and entertainment personalities. As might be expected, the proximity to Election Day was only significantly influential for the two presidential candidates. The hypothesis that high approval in public opinion polls might influence search query volume was not supported – external measures of presidential approval (i.e., polls) do not appear to translate into increased search activity. This is particularly interesting, as it hints that political searches may be valence neutral; in other words, while it may be safe to say that queries measure interest, we cannot make the jump to conclude that greater search traffic also leads to support or approval.

Finally, in the final days leading up to the election, a number of searches increased. Searches for Obama spiked, as did searches for taxes. Prior to this, spikes in issue-based query traffic were limited to only one or two days, but immediately prior to the election, searches for these queries showed an increasing trend for multiple days. Recognizing how search volume changes directly before an electoral event could indicate the public's attached importance to the particular issue.

4.3 Differences in News and Query Volume

Figures 2 and 3 present graphical differences between the news coverage and query volume of the presidential candidates and entertainment personalities. From these graphs, it is clear that search volume and news coverage are punctuated by key events in the campaign, such as political announcements and conventions. For some of these instances, particularly with individuals such as Sarah Palin and Joe the Plumber, who were previously unknown, the surge in query volume can also likely be attributed to novelty and curiosity – when a relative unknown comes on the scene, we may expect unsustainable spikes in query volume to learn about the newcomer.

It is also evident that news coverage of issues does not always generate equivalent spikes in search traffic, and furthermore, sometimes the spikes in query volume last longer than the increases in news coverage. Specifically, on October 16th (the day following Joe the Plumber's mention in the 3rd presidential debate), searches for Joe the Plumber surpassed online search activity for Obama and McCain, as people turned to the Internet to find out about this previously unknown individual.

To quantitatively compare the difference between news coverage and query volume for each candidate and entertainment persona, we conducted Welch two-sample t-tests between the normalized transcripts and normalized query volume. There was a comparable amount of news coverage and query volume for **Sarah Palin** (transcripts = 17.21, queries = 17.36, $t=-0.06$, $p=.955$) and **Joe the Plumber** (transcripts = 3.76, queries = 4.73, $t=-0.37$, $p=.71$). The same was true for **Obama** (transcripts = 35.32, queries = 30.68, $t=1.76$, $p=.08$).

There were significant differences between the amount of news coverage and the level of query volume for John McCain and Tina Fey. While **McCain** received significantly fewer online searches than what his news coverage might predict (news = 40.31, queries = 28.06, $t =4.55$, $p <.001$), **Tina Fey** generated significantly more online searches than what her news coverage might indicate (news = 7.61, queries = 16.46, $t=-3.19$, $p=.002$).

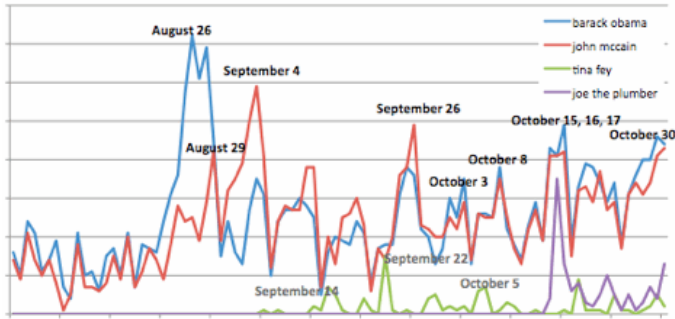


Figure 2. News transcripts for Obama (blue), McCain (red), Fey (green), Joe the Plumber (purple)

Why might this be? For the case of Tina Fey, it is likely that search activity surpassed news coverage because individuals wanted to watch (or re-watch) her SNL skits online. Searchers were not simply seeking out information, but additional media content in the form of videos and comedy clips from the show. McCain may have generated fewer queries than news coverage because he was already an established Senator (whereas Obama was largely an unknown), and individuals felt they needed to learn less about him.

5. FUTURE RESEARCH

The larger scope of this research effort is to take the first step at assessing how implicit feedback from the search process can effectively be applied towards the social sciences. The present study analyzed how fluctuations in query volume may be influenced by news coverage and external events. The degree of media influence on subsequent search activity is quite high, though in several cases (unemployment, war, healthcare), searches increased near Election Day irrespective of news coverage.

A logical next step is to gather real world data (e.g., unemployment claims/ layoffs, Dow Index/ interest rates) to compare changes in query volume with actual conditions. It will also be useful to gather public opinion data from National Election surveys to understand how search queries may fluctuate with survey data about issue importance.

To fully assess the impact of news, a more specific time-series analysis comparing news volume to changes in search query volume could be particularly informative: does media coverage always precede queries? What is the lag time before a news item becomes popularized in search volume? A multimodal analysis would also be interesting to more rigorously compare the spikes in query volume against the spikes in news coverage – for instance, what is it about some media events or news that causes query volume to increase much more than one would expect given the amount of news coverage. While this paper used network news transcripts as the predictor for news, future analyses may attempt to show whether different news sources, such as newspapers or web blogs, show stronger or weaker agenda setting effects.

Finally, the only form of implicit data used in this paper was aggregate query data. Subsequent analysis should also incorporate other typical measures of implicit feedback, such as reading time (to assess interest), clicks (from what sites did users acquire information), and query reformulation patterns. These additional measures, combined with a better understanding of how the voting electorate is represented in online search traffic will be useful in for making predictions about voter behavior or election results.

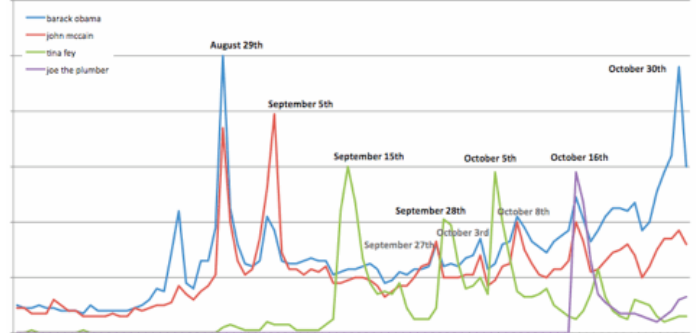


Figure 3. Search query volume for Obama (blue), McCain (red), Fey (green), Joe the Plumber (purple)

6. ACKNOWLEDGMENTS

Thank you to Shanto Iyengar, Solomon Messing, and Hilary Hutchinson who provided valuable feedback on earlier drafts of this paper.

7. REFERENCES

- [1] Kelly, D. 2005. Implicit Feedback: Using Behavior to Infer Relevance. In eds, Spink, A & Cole, C. *New directions in cognitive information retrieval*, Springer: Netherlands.
- [2] Fox, S., Karnawat, K., Mydland, M., Dumais, S., White, T. 2005. Evaluating Implicit Measures to Improve Web Search. *ACM Transactions on Information Systems*, 23, 2, 147-168.
- [3] Joachims, L. Granka, B. Pan, H. Hembrooke, F. Radlinski, G. Gay. 2007. Evaluating the Accuracy of Implicit Feedback from Clicks and Query Reformulations in Web Search, *ACM Transactions on Information Systems (TOIS)*, Vol. 25, No. 2.
- [4] Radlinski, F., Kleinberg, R., Joachims, T. 2008. Learning Diverse Rankings with Multi-Armed Bandits. *International Conference on Machine Learning*, Helsinki, Finland.
- [5] Fallows, Deborah. *Search Engine Use*. Pew. Aug 2008.
- [6] Krosnick, J., 1999. Survey Research. *Annual Review of Psychology*, 50: 537-67.
- [7] Hovland, C.I. 1959. Reconciling conflicting results derived from experimental and survey studies of attitude change. *American Psychologist*, 14, 8-17.
- [8] Google Insights for Search. <http://www.google.com/insights/search/>
- [9] Ginsberg, Mohebbi, Ginsberg, J., Mohebbi, M., Patel, R., Brammer, L., Smolinski, M., Brilliant, L. 2008. Detecting influenza epidemics using search engine query data. *Nature*.
- [10] Choi, H. & Varian, H. 2009. Predicting the Present through Google search queries. April 2.
- [11] McCombs and Shaw, 1972. The Agenda-Setting Function of Mass Media. *Public Opinion Quarterly*, 26, 176-187.
- [12] Iyengar, S. & Kinder. 1984. *News that Matters: Television and American Opinion*. Chicago: U. of Chicago Press.
- [13] Pollster.com <http://pollster.com>. Retrieved Dec 5, 2009.
- [14] Newport, F. 2008. Obama has key edge on key election issues. Gallup Poll, June 24. Retrieved: <http://www.gallup.com/poll/108331/Obama-Has-Edge-Key-Election-Issues.aspx>