How to Measure Public Opinion in the Networked Age: Working in a Googleocracy or a Googlearchy?

Sean Westwood

How to Measure Public Opinion in the Networked Age: Working in a Googleocracy or a Googlearchy?

Sean J. Westwood

Department of Communication, Stanford University,
McClatchy Hall, 450 Serra Mall, Stanford, CA 94305, USA.
seanjw@stanford.edu

Abstract. The rise of the internet has transformed information acquisition from a top-down process originating from media elites to a process of self-selection and searching. This raises a fundamental question about the relationship between information acquisition and opinion formation: do the processes occur in parallel or as part of a self-directed feedback loop? That is, do we look for information to make opinions, do we look for information to support our opinions, or do we do both simultaneously? Analysis using Google search results and polling information from the 2008 US presidential election suggests that public information queries are reflective of polling data and election outcomes. The sheer quantity of search data on political terms also suggests that public information desires may surpass standard assumptions of public political sophistication.

Keywords: Internet, Search, Public Opinion, New Media, Information Acquisition.

1 Introduction

There is a well-documented relationship between the availability of information and public opinion formation: the more information that is available, the more reasoned public opinion is generally believed to be [1]. Traditional models for the process of information acquisition utilize word-of-mouth or directed search. The mediated structure of internet search engines, however, embeds the concept of algorithmic search as an almost intrinsically inseparable component of the process of information access. The prominence of search engines in ranking and filtering data raises a critical question on the relationship between information acquisition and opinion formation: do the processes occur in parallel or as part of a self-directed feedback loop? In a mediated environment, do we search for information to make opinions, do we search for information to support our opinions, or do we do both simultaneously? Moreover, if public opinion is formed through mediated information acquisition does the power of Google and other search engines impose a threat to public knowledge? There are suggestions on the possible threats [2, 3, 4] and advantages [5] to society's reliance on Google and other search engines, but an investigation of the structure of opinion formation in networked societies is needed to justify the threat of a Googlearhcy or the promise of a Googleocracy. This study examined the relationship between Google search records and traditional public opinion measures during the

2008 presidential election, comparing daily measures of both search volume and public opinion. These data do much to highlight the strong relationship between public opinion and information queries, and suggest that in the networked age opinion formation is part of a socially driven feedback loop and not constrained by search information. If anything, the data from this study suggest greater deviation from average opinion preferences in online news searches (i.e., the internet facilitates research on both sides of an issue even for those with set opinions).

The transition from the absolute monopoly of information by top-down media systems to interactive and user-driven information acquisition theoretically increases the amount of available information, but the design of the internet makes locating relevant information difficult without the assistance of search engines and peer recommendation. Although past research shows limited diversity in the websites that users access from search engines, the much more rapid pace at which modern search engines update their indexes, the tendency for Google to promote web pages linked to by highly ranked pages, and the growth of alternative sources such as blogs that frequently update may offer additional breadth of information. At the least, information is more current in search results. Outside controlled laboratory experiments, the first step to gauging how people search for political information is to look at the relative popularity of search terms related to specific measures. For this paper, searches for "Obama" and "McCain" are considered in comparison to polling data and major campaign events.

Prior to formal discussion of search results and the 2008 campaign, there are several important considerations regarding online searches that need further elaboration. Knowledge of how search engines evaluate sites on the internet is general, with the specific implementations of each search engine's algorithm kept as proprietary information. At the simplest level, Google and most modern search engines rank data by relative popularity on the internet. Early analysis of the Google PageRank algorithm by Hindmen et al. [2] suggests that Google limits user exposure to political information on the web to a small set of popular websites. Further work suggests that although search results are algorithmically generated, there is ingrained bias [3] that restricts information access and distribution [4]. This analysis is useful, but as an experimental study it has two modern problems: advancements to the structure of the Google algorithm and the decrease in the interval time between index updates.

First, Google's algorithm prioritizes quality of links over quantity of links in rating websites, so by design Google promotes listings for relatively unpopular websites that are linked to by websites that are highly ranked in Google's existing database [5]. For rapidly changing political information, this means that the Googlearchy may function more like a Googlocracy, where new political information deemed relevant by important peers is promoted in Google's rankings despite a low total number of links to the new information. This may not dilute the dominance currently exerted by the top websites in Google search queries, but the tendency for highly ranked political websites to link to relevant new sources based on content injects new sources into the top of Google by quasi "community action" and not solely through random assessment of internet content by the GoogleBot. Secondly, Google data for many of the most popular areas of the internet is updated in increments less than one hour constantly changing the index to reflect the most recent structure of the internet,

whereas the Hindmen et al. study described a state where Google updated its index every few days or even weekly. While users tend to place great trust in search results [6], the more dynamic and rapid structure of the current Google algorithm may offer great responsiveness than is otherwise assumed in existing literature.

The size of the internet is exponentially growing, and some filtering and selection occurs in the compression of nearly limitless search results to a smaller and more manageable list. Some argue that this compression artificially constrains the information available to the people [7], so the question of 'where does Google's data come from?' becomes critical in understanding how people interact with information filtered through the Google search engine. Detailed studies have discussed the transition of the print media onto the internet [8], the evolution of internet-based news organizations [9], and the relationships between information sources within the blogosphere [10], but to understand the flow of political information on the internet all three sources are necessary. While sources from each category vary in popularity within Google search results, Google includes all three together in a single index that is available to all users.

2 Methodology

Two datasets were used for this project. The first dataset captures search engine traffic and the second captures polling data for the period under study.

Search data. Data from the three major search engines would be most complete, but at present data is only readily available from Google. Using only Google data does present limitations, but as Google commands well over 50% of the search market and has demographics similar to the other two major search engines [11], Google represents a reasonable sample for internet search activity. The Google data used in this project was gathered through Google Trends. Google Trends is a web applet that provides information on search queries by geographic location and by time. This project used the search terms "Obama" and "McCain" to approximate interest in the two candidates in the Google search engine. In the configuration used for this project, Google Trends includes all queries containing the specified terms, so requests for data on "Obama" include searches such as "Obama's tax policy."

Other search engines have received criticism for result manipulation, and while Google has been criticized for advertisements placing, core search results are unaltered. Google reports data as standardized and normalized values. Data is normalized by the total number of Google queries in the selected period of time and within the selected geographic area, and is standardized by the average number of searches for a term since January 2004. For multiple search terms, the average number of searches for the first search term standardizes the values for the other search terms. Data is therefore easily comparable between search terms as a function of relative popularity.

Relative interest in political candidates is hard to assess from the Google data, as raw search volume is unavailable, but relative popularity was approximated with the inclusion of a well-known reference term. Sex, one the most consistently popular search terms used on Google since 2004 [12], was included in all data extractions from Google to provide a reference point.

Additionally, two sub-sets of data were collected from Google. To show long-term trends in interest in both Obama and McCain throughout the primary process, data from December 2007 to December 2008 was gathered. For a more targeted exploration of the influence of major campaign events on search activity and possible correlations between polling data and daily search activity, a smaller period from August 1, 2008 to November 2, 2008 was selected. The second sample captures most of the major events in the last part of the 2008 campaign and forms the basis for the majority of this paper. Data from November 2nd to November 4th is not included in the sample because the extremely high values for Obama's search popularity compresses the other data points so that visualizing differences between the graphed values is impossible without huge graphs (larger than the printed page).[1]

People also search the wide array of online information in many different ways and with high variance in methodologies [13], but the terms people search by offer insight into general intentions. Search terms including presidential candidate names likely range from searches about perceived conspiracies surrounding each candidate to searches for concrete policy details. Nonetheless, users are still searching for information related to the candidate, so grouping all searches for each candidate provides a general metric for assessing online interest in each candidate.

Polling data. All available polling data from August 1, 2008 to November 2, 2008 was collected from Pollster.com (poll n=436, response n=345,910). The majority of the polls during this period were in the field for several days, so to facilitate a more meaningful comparison between Google data and polling data, daily values were calculated from the polling data. To create daily polling averages, each poll's data was split into daily segments (sample n/days in the field) that weighted the poll's outcome for each day the poll was in the field. All values for each day were subsequently averaged, with the resulting data used exclusively in this paper.

3 Results and Discussion

Global summary: December 2007 to December 2008. Interest in both McCain and Obama grew dramatically between December of 2007 and December of 2008. Throughout this period, Obama remained more popular than McCain, with the exception of the time surrounding the RNC, when McCain surpassed Obama's popularity (scaling and rounding issues obscure this in Figure 1). McCain also remained less popular than the reference search term "sex", while Obama was more popular than sex between October 19th and November 9th. In addition to a larger number of searches for "Obama" during this period than for "sex," the relative interest in sex decreased linearly toward November. To the credit of Americans, people were searching for Obama instead of sexual material during this brief window.

Overall, interest in both Obama and McCain grew dramatically throughout the sample period. Searches for Obama grew to 6.6 times their average scaled level at Obama's peak on November 6 from .15 the average level at the sample's starting

[1] Raw Google data ranged in value from the low twenties to the hundreds, so to allow visual comparison between polling data and Google data on a single graph, all Google values were divided by 100.

point. Likewise, interest in McCain peaked at 1.5 (still less than a quarter of Obama's peek popularity) Obama's average scaled search level on November 2 from a starting point of .05 Obama's average scaled search volume[2].

Increases in Google search activity generally corresponded to important events during the campaign, though there were several peaks that are not readily explained by the events of the campaign. Discussion of specific events in each of the final three months of the election follow, but it is important to note the inexact match between campaign events and Google search fluctuations to contextualize the discussion.

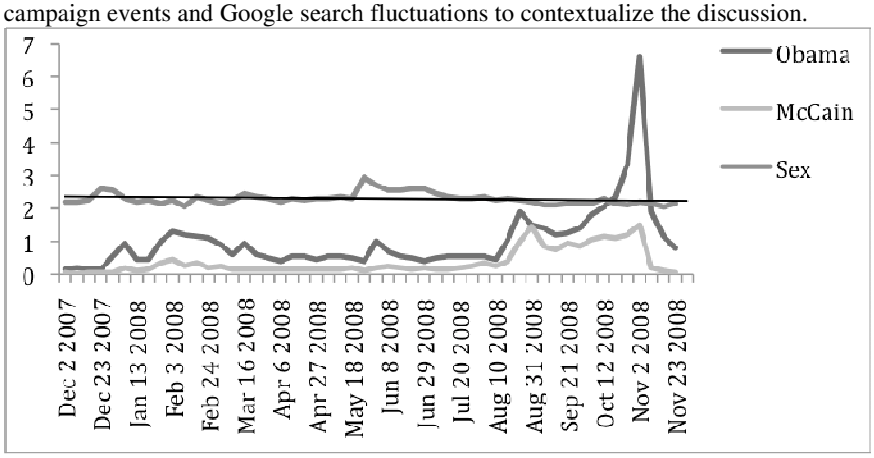

Fig. 1. Google search data December 2007 to December 2008.

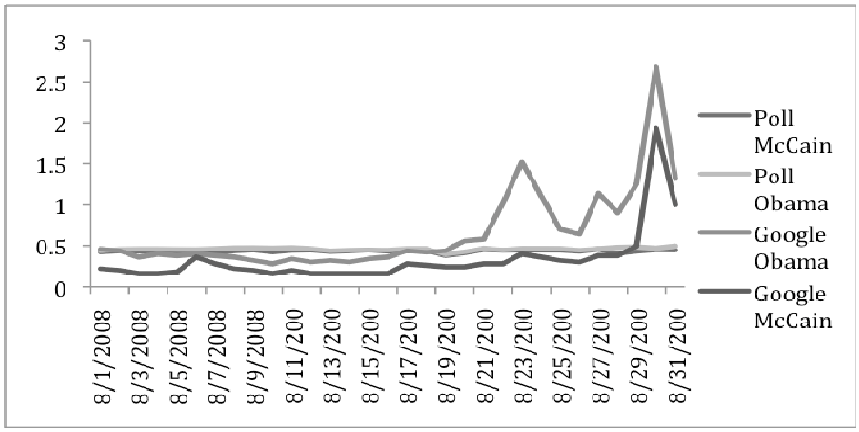

Fig. 2. Google and Polling Data August 2008.

[2] All reported values are scaled by average interest in Obama over time. It is tempting to refer to the units in this paper as "Obama units," but for the sake of readability and clarity the numbers appear without this unit notation.

Table 1. Significant campaign events in August

Date	Event	Difference from Google Average Obama/McCain (Obama-McCain)[3]	Polling Difference %[4]
8/16	Saddleback Debate	-0.64/-0.84 (0.2)	1.68
8/23	Biden Announced as VP	0.52/-0.6 (1.12)	1.74
8/27	Biden Accepts Nomination	0.14/-0.62 (0.76)	0.45
8/28	Obama Accepts Nomination	-0.1/-0.62 (0.52)	7.50
8/29	Palin Announced	0.26/-0.5 (0.76)	5.29

Throughout August, several bumps and drops in search interest for both candidates occurred. Before the Saddleback appearance, search volume for "McCain" was relatively flat, while search volume for "Obama" increased moderately. Following the appearance at Saddleback a sharp rise in search volume for "McCain" and a sharper uptick for "Obama" search volume occurred, although the relative difference between Obama queries before and after Saddleback was low. Search volume for both candidates was relatively flat before the announcement of Biden as Obama's VP, while after the announcement Obama interest tripled while a noticeable jump in McCain interest appeared. Interest in Obama also increased with his formal acceptance of the nomination, although not as significantly as with the announcement of Biden.

The most interesting observation from August coincides with McCain announcing Palin as his nominee for vice president. Interest in McCain increased to its highest point for the month, but interest in Obama also increased to its highest level of the month. Obama, even during the period of Palin's announcement, was a more popular search term throughout the month than McCain. While it is reasonable to argue that Obama's popularity during the Palin announcement resulted from the buildup and bleed from his highly televised acceptance speech, interest in Obama as a search term immediately dropped after his speech before increasing at the same time as Palin's announcement. It is hard to explain this observation, but it is possible that interest in Obama increased during this time because internet users waiting for the McCain vice presidential nomination were not satisfied with Palin and wanted to look at Obama as an alternative. It is also possible than Obama loyalists were interested in the Obama response to the Palin announcement. Both explanations, while not exhaustive, might also occur because of the younger and more liberal internet audience.

[3] Values show the difference between search volume on each day compared to average search volume from January 2004. The parenthetical shows the difference between the relative search popularity of Obama and McCain. Positive values show an Obama advantage.

[4] Difference in calculated daily polling numbers. Positive values show an Obama advantage. Source of event dates: New York Times.

During August, search interest was roughly comparable to polling data, though the increases in Obama search traffic far exceed his polling gains around the Democratic National Convention.

Table 2. Significant Campaign events in September 2008

Date	Event	Difference from Google Average Obama/McCain (Obama-McCain)[1]	Polling Difference %[2]
9/3	McCain Accepts Nomination	-0.13/-0.2 (0.07)	3.10
9/4	Palin Accepts Nomination	0.35/0.21 (0.14)	5.32
9/5	Fannie Mae and Freddie Mac Bailout	0.36/0.98 (-0.62)	-1.44
9/15	Collapse of Lehman Brothers	-0.22/-0.46 (0.24)	0.20
9/23	Palin Couric Interview	-0.28/-0.56 (0.28)	2.80
9/25	McCain Suspends Campaign	-0.2/-0.28 (0.08)	1.24
9/26	WaMu Fails/Presidential Debate	-0.09/-0.17 (0.08)	6.82

Although Google interest in McCain did not surpass Obama with the initial announcement of Palin as the Republican vice presidential nominee, during the period of the Republican National Convention Google recorded more searches for "McCain" than for "Obama." This blip is the only time McCain searches outpaced Obama searches and corresponds to a brief inversion of McCain and Obama in polling data.

The period of the financial crisis between the placement of Fannie Mae and Freddie Mac into government conservatorship and the collapse of Lehman Brothers saw a minor increase in searches for Obama and McCain as compared to the period immediately before and after the events. The upswing during this time was, however, very small and relatively even between the two candidates except for September 5th, when interest in the McCain campaign was high but decreasing. The relative differences between Obama searches and McCain searches during this period were approximately consistent at around .24.

As was the case with the initial announcement of Palin, the suspension of McCain's campaign and the financial sector bailout increased interest in both candidates on Google, but while the increase was roughly equal among the candidates Obama remained more popular than McCain.

For the majority of October the relative number of Obama and McCain searches rose and sank in unison and roughly in line with poll numbers, with the number of searches for Obama increasing significantly throughout the month and the number of searches for McCain remaining relatively flat. Searches for both candidates rose during the time of the third presidential debate, but other newsworthy events such as the cost of Palin's wardrobe and the relative buzz of the Al Smith dinner do not appear to show any significant change in search numbers.

Table 3. Significant Campaign events in October and November 2008

Date	Event	Difference from Google Average Obama/McCain (Obama-McCain)[1]	Polling Difference %[2]
10/3	Stimulus Package Passed	0.12/-0.30 (0.42)	6.25
10/16	Final Debate	0.88/0.16 (0.72)	5.15
10/18	Al Smith Dinner	0.4/-0.22 (0.62)	6.81
10/21	Palin's Wardrobe	0.54/-0.24 (0.78)	7.56
10/30	Obama Infomercial	2.38/-0.06 (2.44)	6.38

Interest in Obama started to rise before the Obama infomercial and peaked during the pre-election period on the night of the broadcast. Unlike large events in the McCain campaign, where interest in Obama rose with McCain, interest in McCain in the period around the Obama infomercial did not significantly change.

The data from this project show the relative popularity of political search terms increased over time and were biased in favor of Obama. The approximate demographics of the internet favor those younger than 65 (predominantly those under 29) and those with higher incomes [11]. In this environment, the higher frequency of searches for Obama over McCain makes sense given the higher level of Democratic support among younger and more financially secure Americans. Although it is also possible that the general movement in support of Obama influenced the popularity of searches for Obama, the inability for McCain to gain more searches than Obama during any period other than the RNC suggests that demographics play a significant role in explaining search popularity. This is most probable given that when interest in McCain rose, interest in Obama usually rose to a level above McCain.

While "sex," one of the most popular search queries on the internet, was included in the initial analysis as a reference term to gauge an approximation of how frequent searches for Obama and McCain were, the selection of the term showed an interesting swing in popularity between Obama and sex. For the last week of the campaign, possibly due to the Obama infomercial and heavy press coverage suggesting an Obama victory, searches for Obama outnumbered searches for sex by a significant amount, and searches for sex decreased as compared to previous points in the year. Although no statistical test was performed to gauge the magnitude of the swap, it is possible that people who would otherwise search the internet for sex searched for information related to Obama, or that searches for Obama increased broadly enough to decrease other types of searches during the period.

4 Correlation between Polling Data and Google Searches

The observational conclusions in the previous sections of the paper are important in contextualizing how Google searches and campaign events are related. Graphs for Google data from August, September, and October included polling data as a reference, though the scale of the Google search data flattens the polling data and limits useful comparison. For the purposes of the following analysis, two graphs appropriately scaled for Google data and polling data are most useful.

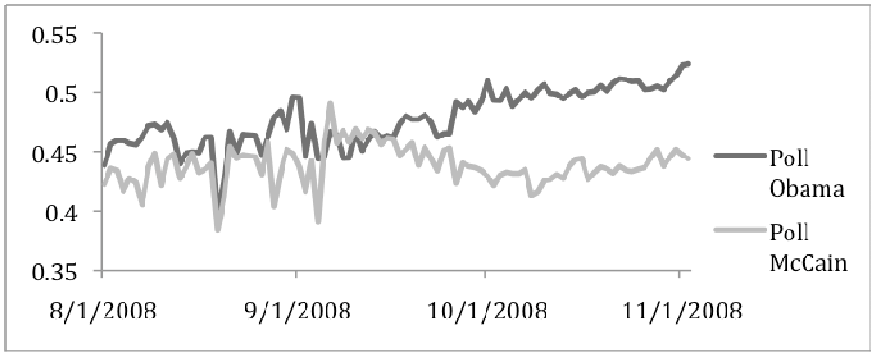

Fig. 3. Averaged daily Obama and McCain polling data.

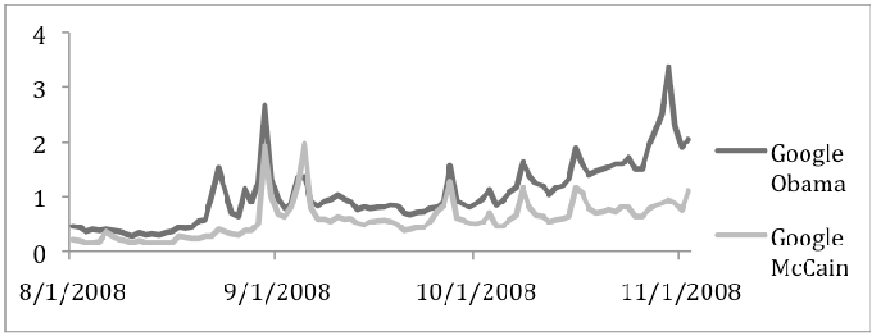

Fig. 4. Daily Obama and McCain Google searches.

Both graphs look approximately similar in shape and show similar general trends (when different scaling is accounted for), with both graphs showing a clear Obama advantage with the exception of the period surrounding the RNC. To compare the two data sets, differences between Obama and McCain were calculated for polling data and Google data. The correlation between the difference between Obama and McCain's polling numbers and the difference between the number of Google searches for Obama and McCain is highly statistically significant with $r=.612$ and $p<.001$. Despite relatively high increases in Google search volume and relatively low decreases in search volume as compared to fluctuations in polling data, this

correlation makes sense. Polling data has an upper and lower bound defined by partisan voters, while the desire for information logically crosses ideological bounds that constrain more dramatic shifts in polling data.[5]

5 Implications

The proportion of Google searches conducted for each of the candidates is highly correlated to poll numbers. Considering the limitations of online search data, it appears as if the swings in search volume simultaneous to polling data illustrate a relationship between campaign events, public opinion, and search volume. Further research is required to investigate the causal forces for this correlation, but these findings are suggestive of a public that forms opinions and searches for information in similar ways and volumes. The relative freedom in information acquisition compared to opinion change perhaps also magnifies jumps in search data as compared to jumps in polling data, which explains the much more significant increases in search volume compared to polling data around important campaign events. It is also possible that people are actually motivated to investigate major campaign events regardless of ideology, although since we do not know who is searching it is conceivable that political junkies are simply following the news as it happens in Google. However, it is probably less likely that the truly informed and motivated would actively search out information on Google instead of visiting trusted news websites or blogs, while it is more reasonable to assume that the less technologically skilled and political innocents would start information searches with Google.

These data further suggest that public information queries—a lá Google—are reflective of polling data and election outcomes, and that public information desires may surpass standard assumptions of public political sophistication. The high correlation between search queries and standard public opinion measures also suggest that although Google filters information, it provides information within a Googleocracy of rapidly changing information networks. That is, the information indexed by Google responds to the rise of new information networks and that these networks are driven—or at least—related to the opinion of the actual public. For rapidly changing political information, this suggests that in such a Googlocracy new political information deemed relevant by the public is presented at a higher position in Google's results[6], and that this becomes a recursive process of information demand and information provision.

---

[5] As an interesting aside, a correlation between election outcome and Google search volume difference (Obama – McCain) by state is significant r=.293 (p=.018). States where Obama is more popular are more likely to support Obama in the general election. Although, this suggests that McCain victory in states is a function of the extent to which he was less popular in each state and not a result of actual superiority in search quantities as Obama had a clear advantage in all states and Washington D.C.

[6] The Page Rank algorithm is, roughly defined, a means of ranking based on ties between network actors (websites). More central websites, as defined network measures are more "important" in assessing content. In this model, new information relayed by highly ranked peers is promoted in rankings. Simultaneously, the calculation of network importance is a result of public action. Rankings are therefore derived from public action and used to filter future public queries.

References

1. Mill, J.S.: On liberty. Broadview Press, Ontario (1999).
2. Hindman, M., Tsioutsiouliklis, K., Johnson, J.A.: "Googlearchy:" How a Few Heavily-Linked Sites Dominate Politics on the Web. In: Annual Meeting of the Midwest Political Science Association, pp. 1-33. Chicago (2003).
3. Goldman, E.: Search Engine Bias and the Demise of Search Engine Utopianism. In: Spink, A., Zimmer, M. (eds.) Web Search: Multidisciplinary Perspectives, pp. 121-133. Springer, Berlin, Heidelberg, Dordrecht (2008).
4. Diaz, A.: Through the Google goggles: Sociopolitical bias in search engine design. In: Spink, A., Zimmer, M. (eds.) Web Search: Multidisciplinary Perspectives, pp. 11-34. Springer, Berlin, Heidelberg, Dordrecht (2008).
5. Menczer, F., Fortunato, S., Flammini, A., Vespignani, A.: Googlearchy or Googlocracy? In IEEE Spectrum Online (2006).
6. Fallows, D.: Search Engine Users. Pew Internet and American Life Project (2008).
7. Adamic, L., Huberman, B.: The Web's Hidden Order. Communications of the ACM, 44(9), 55-60 (2001).
8. Boczkowski, P.J.: Digitizing the News: Innovation in Online Newspapers. MIT Press, New Bakerville (2004).
9. Blumler, J.G., Gurevitch, M.: The new media and our political communication discontents: democratizing cyberspace. Information, Communication and Society, 4, 1-13 (2001).
10. Adamic, L., Glance, N.: The political blogosphere and the 2004 U.S. election: divided they blog. In: International Workshop on Link Discovery. ACM, Chicago (2005).
11. Quantcast (2008) [cited 2008 November 11], http://www.quantcast.com.
12. Google Trends (2008) [cited 2008 November 14], http://www.google.com/trends.
13. Hargittai, E.: Informed Web Surfing: The Social Context of User Sophistication. In: Howard, P. and Jones, S. (eds.) Society Online: The Internet in Context. Sage Publications, Thousand Oaks (2004).