An Automated Approach to Investigating the Online Media Coverage of U.S. Presidential Elections

Arno Scharl
Albert Weichselbraun

ABSTRACT. This paper presents the U.S. Election 2004 Web Monitor, a public Web portal that captured trends in political media coverage before and after the 2004 U.S. presidential election. Developed by the authors of this article, the webLyzard suite of Web mining tools provided the required functionality to aggregate and analyze about a half-million documents in weekly intervals. The study paid particular attention to the editorial slant, which is defined as the quantity and tone of a Web site's coverage as influenced by its editorial position. The observable attention and attitude toward the candidates served as proxies of editorial slant. The system identified attention by determining the frequency of candidate references and measured attitude towards the candidate by looking for positive and negative expressions that co-occur with these references. Keywords and perceptual maps summarized the most important topics associated with the candidates, placing special emphasis on environmental issues.

KEYWORDS. U.S. presidential elections, media monitoring, Web mining, natural language processing, semantic orientation, keyword analysis

Arno Scharl is the Vice President of MODUL University Vienna where he also heads the Department of New Media Technology (www.modul.ac.at/nmt). Prior to this appointment, he was a Key Researcher at the Austrian Competence Center for Knowledge Management, held professorships at Graz University of Technology and the University of Western Australia, and had joined Curtin University of Technology and the University of California at Berkeley as a Visiting Fellow. His current research projects focus on the integration of semantic and geospatial technology, human-computer interaction, computer-mediated collaboration, ontology learning, and the various aspects of environmental online communication.

Albert Weichselbraun is an Assistant Professor at the Department of Information Systems and Operations at Vienna University of Economics and Business Administration (ai.wu-wien.ac.at). After completing two Master's degrees in Economics and Chemical Engineering, his doctoral thesis focused on ontology-based text classification. Dr. Weichselbraun leads the technical development of webLyzard (www.webLyzard.com) and the IDIOM Project (www.idiom.at) with a special focus on the analytical methods involved. His current research focuses on ontology evolution and learning, text mining and the application of semantic technologies to information retrieval.

The IDIOM Research Project (Information Diffusion across Interactive Online Media; www.idiom.at) is funded by the Austrian Ministry of Transport, Innovation & Technology (BMVIT) and the Austrian Research Promotion Agency (FFG) within the strategic objective FIT-IT Semantic Systems (www.fit-it.at). The authors would like to thank Arinya Eller and Jamie Murphy, as well as the special issue's editors and anonymous reviewers for their valuable feedback and suggestions during the preparation of this article.

Address correspondence to: Prof. Arno Scharl, Department of New Media Technology, MODUL University Vienna, Am Kahlenberg 1, 1190, Vienna, Austria (E-mail: scharl@modul.ac.at).

Journal of Information Technology & Politics, Vol. 5(1) 2008
Available online at http://jitp.haworthpress.com

doi:10.1080/19331680802149582

Representative democracy offers significant possibilities for exploiting information networks (Holmes, 2002), but there is little agreement on the specific impacts of these networks. Most agree that new media represent additional channels for political campaigners to project their messages and solicit feedback from citizens (Stromer-Galley & Foot, 2002). Proponents praise the potential of information networks to increase the accessibility of information, encourage participatory decision-making, and facilitate communication with policy officials and like-minded citizens. From the citizens' perspective, interactive Web technologies enable people to find political soul mates and search for their personal truths online (Gibbs, 2004). From the campaigners' perspective, disseminating information directly or through online news media helps create online visibility and influence public opinion.

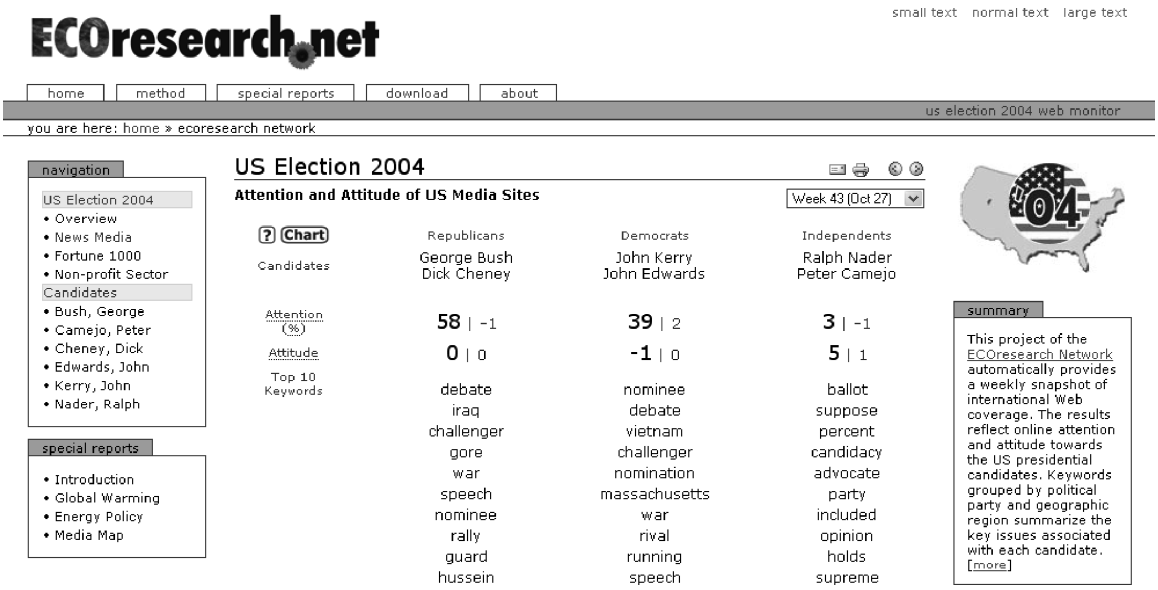
Most attempts to monitor the campaign performance of presidential candidates focus on public opinion, which is influenced by the consumption of media products (Druckman & Parkin, 2005). The analysis of political communication, however, should include both the consumption as well as the production of media products (Howard, 2003). Monitoring information networks to study political campaigns provides a complementary source of empirical data and a window into the evolving concept of electronic democracy (Dutton, Elberse, & Hale, 1999).

A Pew/Internet survey (Horrigan, Garrett, & Resnick, 2004) found that four out of ten U.S. Internet users aged 18 or older accessed political material during the 2004 presidential campaign, up 50% from the 2000 campaign. For political news, more than two-thirds of American broadband users and over half of dial-up users seek Web sites of national news organizations. International news sites are the second most popular category at 24% and 14%, respectively. This shows that traditional media extend their dominant position to the online world, and that their Web sites reflect the majority of political content that the average Internet user accesses. Therefore, their communication strategies affect democratic processes regardless of whether authors control editorial slant deliberately (Fico & Cote, 1999) or whether readers recognize the slant (Druckman & Parkin, 2005). Nearly one third of respondents in a 2002 Stanford survey regarded bias relevant for judging the credibility of online news (Fogg et al., 2002). Most Americans claim to prefer unbiased news sources (Horrigan et al., 2004), as biased media coverage can polarize groups (Sunstein, 2004) and degrade the climate of public discourse (Horrigan et al., 2004). Consequently, many organizations explicitly hold up fairness and balance in reporting (Fico & Cote, 1999), although they are usually free to choose which candidate to emphasize and how to interpret current events (Wayne, 2001).

News media are the most influential stakeholder groups that report or comment on political campaigns; but, they are not the only one. When corporations and advocacy organizations embrace networked information technology, intentionally or not, they also influence democratic processes. To capture and understand the influence of electronic content published by different stakeholder groups, the U.S. Election 2004 Web Monitor (www.ecoresearch.net/election 2004) gathered and annotated the online coverage of the presidential candidates (Figure 1). It paid particular attention to the editorial slant of news media and the diverging perceptions of environmental advocacy organizations and the most influential US corporations.

This article presents computational methods suited for processing very large document repositories. The sheer volume of information in such repositories and the limits of human cognition preclude manual approaches to analyzing online media coverage and measuring editorial slant. Therefore, the U.S. Election 2004 Web Monitor automatically identified attention and attitude toward each candidate and provided additional means for an in-depth investigation of the underlying document repository: keyword lists summarizing the most important topics associated with a candidate, a query interface to access and sort sentences with candidate references, and special reports that aggregate and visual domain-specific coverage with a special emphasis on environmental issues. After presenting the system's major components and results, the article concludes with an outlook on promising future research avenues, introducing the 2008 edition of the

FIGURE 1. U.S. Election 2004 Web Monitor one week before the election (October 27, 2004).



small text normal text large text
ECOresearch.net
home
method
special reports
download
about
us election 2004 web monitor

you are here: home » ecoresearch network

navigation

US Election 2004
- Overview
- News Media
- Fortune 1000
- Non-profit Sector

Candidates
- Bush, George
- Camejo, Peter
- Cheney, Dick
- Edwards, John
- Kerry, John
- Nader, Ralph

special reports
- Introduction
- Global Warming
- Energy Policy
- Media Map

US Election 2004

Attention and Attitude of US Media Sites

Week 43 (Oct 27)

? Chart	Republicans	Democrats	Independents
Candidates	George Bush Dick Cheney	John Kerry John Edwards	Ralph Nader Peter Camejo
Attention (%)	**58** | -1	**39** | 2	**3** | -1
Attitude	**0** | 0	**-1** | 0	**5** | 1
Top 10 Keywords	debate	nominee	ballot
	iraq	debate	suppose
	challenger	vietnam	percent
	gore	challenger	candidacy
	war	nomination	advocate
	speech	massachusetts	party
	nominee	war	included
	rally	rival	opinion
	guard	running	holds
	hussein	speech	supreme

summary

This project of the ECOresearch Network automatically provides a weekly snapshot of international Web coverage. The results reflect online attention and attitude towards the US presidential candidates. Keywords grouped by political party and geographic region summarize the key issues associated with each candidate. [more]

Election Monitor and outlining how models of information diffusion could help explain the influence of Web content on public opinion during political campaigns.

METHODOLOGY

The volume and dynamic nature of Web documents complicate testing the assumption of organizational bias. To address this challenge, the U.S. Election 2004 Web Monitor mirrored 1,153 Web sites in weekly intervals from September to December 2004. The project drew upon the Newslink.org, Kidon.com, and ABYZNewsLinks.com directories to compile a list of the most influential news media sites—42 from the United States and 72 from four other English-speaking countries: Canada, United Kingdom, Australia, and New Zealand. The study also included 39 environmental advocacy organizations operating either nationwide within the United States or internationally, and all the Web sites of the Fortune 1000 (the largest U.S. corporations ranked by revenue).

A crawling agent mirrored these Web sites by following their hierarchical structure until reaching 50 megabytes of textual data for news media and 10 megabytes for commercial and advocacy sites. These limits helped compare sites of heterogeneous size and reduce the dilution of top-level information by articles found in lower hierarchical levels.

In the literature, such a collection of recorded content used for descriptive analysis is often referred to as a corpus. This research investigated and visualized regularities in such corpora from three different samples of Web sites by applying methods from corpus linguistics and textual statistics (Biber, Conrad, & Reppen, 1998; Lebart, Salem, & Berry, 1998; McEnery & Wilson, 2001). The quantitative textual analysis necessitated three steps to yield a useful machine-readable representation (Lebart et al., 1998):

(a) The first step *converted* hypertext documents into plain text—that is, processing the gathered data and eliminating markup code and scripting elements.

(b) The second step *segmented* the textual chain into minimal units by removing coding ambiguities such as punctuation marks, the case of letters, hyphens, and points in abbreviations. With little fluctuation across sampling points between September and December 2004, this process yielded up to a half-million documents in weekly intervals, comprising about 125 million words in 10 million sentences. The system identified and removed redundant segments such as headlines and news summaries whose appearance on multiple pages would otherwise distort frequency counts.

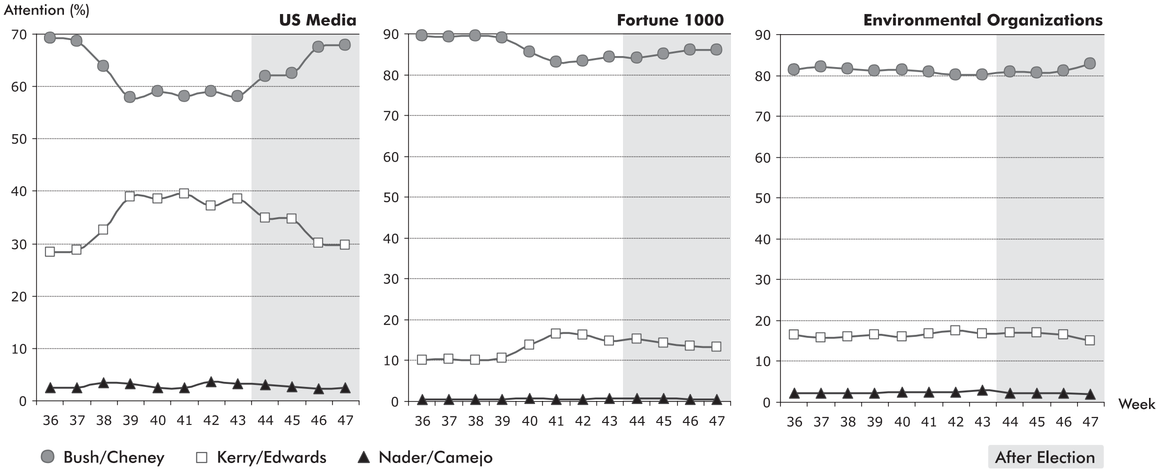
(c) The third step *identified* groups of identical units and counted their occurrences, thereby creating an inventory of words ("word list") or multi-word units of meaning (Danielsson, 2004). The frequency of candidate references presented in the following section is based on such a word list, which typically uses decreasing frequency of occurrence as the primary sorting criterion and lexicographic order as the secondary criterion.

FREQUENCY OF CANDIDATE REFERENCES (ATTENTION)

Media coverage and public recognition go hand-in-hand (Wayne, 2001), with strong correlations between the attention of news media and both public salience and attitudes toward presidential candidates (Kiousis & McCombs, 2004). The U.S. Election 2004 Web Monitor calculated attention as the relative number of references to a candidate. To determine references to candidates or environmental topics, a pattern-matching algorithm considered common term inflections while excluding ambiguous expressions. Only identifying occurrences of *george w. bush*, for example, would ignore equally valid references to *president bush* and *george walker bush*. Yet, a general query for *bush* would fail to distinguish the president's last name from references to wilderness areas or woody perennial plants. A detailed list of the terms used to measure the number of candidate references is available online at www.ecoresearch.net/election2004/candidates.

With regard to the number of documents mirrored, the Web sites of Fortune 1000 companies represented the largest sample with about 330,000 documents, followed by news media (180,000) and environmental organizations (16,000). As expected, the corporate sector had published relatively few articles on the presidential candidates—yielding only 484 relevant documents (0.15% of total documents) as compared to 3,834 documents from news media (2.09%) and 709 documents from environmental advocacy organizations (4.50%). The comparative statistics in Figure 2 highlight the differences between samples. In contrast to the frequently updated coverage of news media, which fluctuated considerably from week to

FIGURE 2. Attention of U.S. News Media, the Fortune 1000, and Environmental Organizations toward the U.S. Presidential Candidates between September and November 2004.



week, the textual data mirrored from the Web sites of Fortune 1000 companies and environmental advocacy organizations show a more static distribution.

In the week before the election, 58% of the media references mentioned George W. Bush and Dick Cheney, down one percentage point from the preceding week (see Figure 1). Thirty-nine percent reported on John Kerry and John Edwards. Across all three samples, the Republican candidates dominated the coverage, while the independent team of Ralph Nader and Peter Camejo garnered less than 5% of the attention. The Republican dominance does not necessarily represent editorial bias. Incumbents who run for office serve a dual role as candidates and government officials. Journalists often do not report on the campaign, but on operative decisions and newsworthy day-to-day activities. The same phenomenon could be observed with data from the Fortune 1000 companies and environmental organizations, which dedicated over 80 percent of their coverage to the incumbent and his running mate. Follow-up studies should devise automated methods to distinguish general from campaign-specific coverage in order to investigate how the dual role of incumbents impacts media attention.

SEMANTIC ORIENTATION (ATTITUDE)

A conclusive measure of organizational bias in media coverage has been elusive (Sutter, 2001), but it is essential when investigating trends and differing perceptions of interest groups. The ever-increasing amount of articles and the limits of human cognition prevent analysts from maintaining a running tally of news media coverage of the candidates during a presidential campaign (Watts, Domke, Shah, & Fan, 1999). Therefore, the US Election 2004 Web Monitor required an automated method to determine the semantic orientation toward a candidate (Scharl, Pollach, & Bauer, 2003). Calculating the frequency of candidate references, as described in the preceding section, only represents a partial answer to this problem since it disregards the specific context of these references (Yi, Nasukawa, Bunescu, & Niblack, 2003).

The chosen method is based on the notion that there is a conceptual connection between words and their adjacent text (Giora, 1996). The semantic orientation toward a target term within a sentence is calculated by measuring the distance (d_{ts}) between the target term (t) and a predefined list of sentiment words (s) known to have positive or negative connotations (Scharl et al., 2003). This list was taken from the tagged dictionary of the *General Inquirer* containing 4,400 positive and negative sentiment words (Stone, Dunphy, Smith, & Ogilvie, 1966). To a large extent, the validity of this approach depends on the size of the tagged dictionary. It is therefore essential that *all* instances of the sentiment words in the corpus are included in the analysis and not just their base forms. Therefore, reverse lemmatization was used to add about 3,000 words to the dictionary by considering plurals, gerunds, past tense suffixes, and other syntactical variations (for example, manipulate → manipulates, manipulating, manipulated).

Calculated for each sentence separately, the semantic orientation (SO) represents the sum of the tagged words' semantic charges (SO_s) divided by their distance to the target term based on the following equation—the exponent λ determines the influence of the distance on calculating the semantic orientation:

$$SO = \sum \frac{SO_s}{d_{ts}^{\lambda}}$$

Two sentences from the *New Zealand Herald* and the *St. Petersburg Times* containing references to oil prices (below) exemplify the difference between positive and negative coverage. The sentiment charges from the underlined words of the tagged dictionary are aggregated to identify each sentence's semantic orientation. In this example, a λ of zero (no influence) implies that the semantic orientation equals the sum of the sentiment charges according to the tagged dictionary.

- "US stocks *rallied*$_{(+1.0)}$ Wednesday, <u>boosted</u>$_{(+1.0)}$ by <u>shares</u>$_{(+1.0)}$ of health and defense$_{(-1.0)}$ companies$_{(+0.22)}$ that are seen benefiting$_{(+1.0)}$ from the re-election of President *George W. Bush*, but higher **oil prices** checked advances$_{(+\ 0.87)}$" (NEW ZEALAND HERALD). ↑ (+ 4.09)

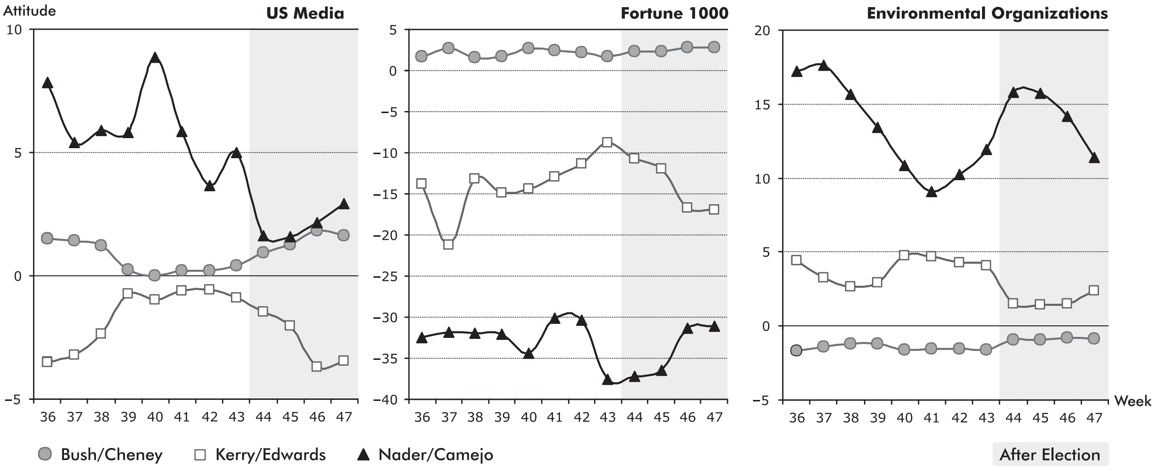
- "The dollar hit$_{(-1.0)}$ its lowest$_{(-0.98)}$ level in more than eight months against the Euro Thursday, falling$_{(-0.05)}$ sharply on worries$_{(-1.0)}$ about the economic effects of rising **oil prices** and expectations of continued trade and budget *deficits*$_{(-1.0)}$ in *President Bush's* second term" (St. Petersburg Times). ↓ (–4.03)

The presented method focuses on the lexis of text and is therefore most appropriate for large document archives. Full grammatical parsing of the textual data may deliver superior results for individual sentences—for example, in the case of sarcasm or complex sentence structures—but suffers from poor scalability and thus was not computationally feasible for the U.S. Election 2004 Web Monitor.

Acknowledging the lack of an objective standard by which to assess bias, previous research recommends focusing on relative comparisons of coverage (Druckman & Parkin, 2005). In line with these recommendations, the US Election 2004 Web Monitor did not interpret a semantic orientation of zero as neutral coverage, but rather investigated the differences between candidates. The weekly statistics of Figure 3 show that media coverage initially favored the Republican Party candidates, but that their Democratic contenders gained ground in September 2004. Kerry's performance in the first televised debate accelerated these gains in media attitude and was followed by a tight race between the two teams in the four weeks preceding the election. The re-election of George W. Bush understandably widened the gap, considering the positive connotation of terms such as *winning* and *victory*. Sentiment toward Ralph Nader remained remarkably positive up until the election. Since he received very little attention, it can be concluded that the few outlets reporting about him were supporting his environmental and consumer causes, and that his opponents did not deem his candidacy relevant enough to divert campaign resources from confronting their main competitors.

While a narrow margin decided the U.S. presidential elections in 1996 and 2000, differences in the candidates' positions became more pronounced in 2004 and the political deliberation more partisan (Weisberg, 2004). Partisans tend to perceive mass media content as biased against their ideological vantage point (Schmitt, Gunther, & Liebhart, 2004). Explanations for this *hostile media effect* range from selective recall (preferentially remembering hostile content), selective categorization (perceiving the same content differently), and conflicting standards (considering hostile content as invalid or irrelevant).

FIGURE 3. Attitude of U.S. News Media, the Fortune 1000, and Environmental Organizations towards the U.S. Presidential Candidates between September and November 2004.



Claims of a liberal bias in news coverage have become quite common in recent U.S. presidential campaigns. In 1996, for example, Republican Party candidate Bob Dole blamed news media coverage as the reason for lagging behind Democratic incumbent Bill Clinton (Watts et al., 1999). The empirical evidence of the US Election 2004 Web Monitor contradicts this notion of "liberal media" and demonstrates that the study of organizational bias requires a more differentiated approach. It does not suffice to classify any view that does not comport with a conservative ideological viewpoint as "liberal" (Alterman, 2003). Clearly favoring the Republican candidates, the attention and attitude data on the 2004 U.S. election allows two interpretations: (a) Either mainstream media coverage is more sympathetic to conservative causes, or (b) the incumbent benefits from a journalistic inclination toward reporting about government officials and their "spin" on events (Sutter, 2001). Analyzing elections with liberal incumbents is a promising area for future research to investigate these interpretations.

Compared to the news media, industry and environmental advocacy organizations showed a more pronounced articulation of their political views. While environmental organizations tended to criticize the environmental record of George W. Bush (particularly abandoning the Kyoto Treaty ratification and reducing air pollution controls through the Clear Skies Act), Fortune 1000 companies presented the Republican Party candidates most favorably. Given the growth of corporate ownership of news media and their reliance on revenues from corporate advertising, the conservative position of Fortune 1000 companies might influence the fundamental rules of news production and help explain the observed media bias.

Organizational bias is often specific to a certain topic or domain. Many journalists hold liberal social views, for example, but conservative economics views (Croteau, 1998). The following two sections demonstrate how automated Web site analysis can account for these differences and provide a more nuanced view of candidate coverage—either on the aggregated level by using keywords to summarize the unfolding of events, or by providing effective means to query individual sentences that report on candidate activities.

KEYWORD ANALYSIS

Complementing measures of attention and attitude toward a candidate, keywords grouped by political party and geographic region highlight the most popular topics in a particular week. (Keywords are computed by locating them in a target corpus and comparing their frequency with a reference distribution from a usually larger corpus of text.) To identify the topics that each stakeholder group associated with the candidates, the U.S. Election 2004 Web Monitor compared the term frequency distribution in sentences mentioning a candidate (target corpus) with the term frequency distribution in the entire sample (reference corpus). The system used a chi-square test with Yates' correction for continuity (Yates, 1934) to analyze the significance of co-occurrence by comparing term frequencies between target and reference corpus. The ratio between both corpora in conjunction with a term's total counts in the reference corpus provides the expected frequency (E_i). Comparing this value with the observed frequency (O_i) and applying the chi-square test with Yates' correction yields the significance of the deviation between expected and observed frequencies.

$$\chi^2_{\text{Yates}} = \sum_{i=1}^{N} \frac{(|O_i - E_i| - 0.5)^2}{E_i}$$

Table 1 exemplifies the process for a hypothetical target corpus one-fourth the size of the reference corpus, containing all the sentences with at least one reference to *George W. Bush.* Common words such as prepositions, articles, and conjunctions occur frequently, but in line with expectations assuming equal term distributions in the reference corpus and the target corpus. The term *Iraq*, by contrast, appears three times as often as expected and thus appears to be a good descriptor of candidate coverage. Term frequency thresholds prevent the inclusion of

TABLE 1. Example of Keyword Calculation Including Word Frequency in the Reference Corpus, Expected and Observed Word Frequencies in the Target Corpus, and Chi-Square Significance

Word	AND	THE	WAR	IRAQ	WASHINGTON	WAHINGTON [SIC]
Reference corpus	300	596	200	100	300	4
Target corpus (expected)	75	149	50	25	75	1
Target corpus (observed)	72	158	127	75	200	2
Chi-square significance	0.09	0.44	123.86	113.48	197.73	0.33

rare expressions or typographical errors such as *Wahington* [sic].

Table 2 summarizes keywords that U.S. news media associated with the candidates and their running mates in the week preceding the election—that is, those terms that were significantly over-represented in sentences reporting on a specific candidate. The list ranks keywords by decreasing significance as computed by the chi-square test. To avoid outliers, the list only considers nouns with at least 100 occurrences in the reference corpus.

Keyword analysis distills information from gigabytes of raw textual data and tells the reader a concise story of what happened on the campaign trail in a given week. Table 2 shows that the television *debates* between the major candidates and their *running* mates remained topical up until the election. The *war* on *terrorism* and persistent problems in dealing with insurgents in *Iraq* dogged Bush, while his *challenger's* service in *Vietnam* continued to occupy the media. Vice-President and former CEO of *Halliburton,* Cheney, was busy traveling to *Pensacola, Wyoming,* and *Wilmington* and addressing media questions about his wife *Lynne* and his *lesbian* daughter Mary. A *speech* by former President *Clinton,* joining Kerry in his first appearance after undergoing heart surgery, reminded undecided voters of more prosperous times. At the same time, actor *Ashton* Kutcher hit the campaign *trail* for John

TABLE 2. Candidate Keywords Based on U.S. News Media Sites as of October 27, 2004

Republicans		Democrats		Independents	
George Bush	Dick Cheney	John Kerry	John Edwards	Ralph Nader	Peter Camejo
DEBATE	**LYNNE**	**NOMINEE**	**CAROLINA**	**BALLOT**	**OPINION**
5510 | 1650.44	205 | 1935.06	623 | 3516.29	1363 | 1842.84	1333 | 10080.58	1939 | 1893.62
CHALLENGER	**DAUGHTER**	**VIETNAM**	**RUNNING**	**PERCENT**	**RUNNING**
547 | 1558.19	2090 | 1624.17	1475 | 2652.35	3462 | 1701.16	16270 | 2359.83	3462 | 400.81
IRAQ	**HALLIBURTON**	**CHALLENGER**	**DEBATE**	**CANDIDACY**	**ELECTORS**
19156 | 1528.19	377 | 1262.01	547 | 2266.91	5510 | 925.14	197 | 1212.04	116 | 72.77
GORE	**DEBATE**	**DEBATE**	**NOMINEE**	**ADVOCATE**	**COMMONWEALTH**
1403 | 1497.44	5510 | 1150.93	5510 | 2109.72	623 | 892.68	356 | 811.17	133 | 63.34
WAR	**LESBIAN**	**MASSACHUSETTS**	**GEPHARDT**	**SUPREME**	**BALLOT**
14957 | 1288.99	404 | 117.84	1260 | 1602.77	154 | 406.09	1665 | 574.54	1333 | 54.72
SPEECH	**RUMSFELD**	**NOMINATION**	**IOWA**	**GORE**	**RESPONDENTS**
2069 | 1076.86	1278 | 451.90	585 | 1373.27	1489 | 390.60	1403 | 429.32	225 | 37.04
NOMINEE	**PENSACOLA**	**WAR**	**ASHTON**	**PARTY**	**ENDORSEMENT**
623 | 839.79	98 | 446.17	14957 | 1133.62	154 | 289.11	8624 | 386.75	323 | 25.50
GUARD	**RALLY**	**RIVAL**	**NORTH**	**PETITION**	**NOMINEE**
1771 | 695.87	1342 | 392.09	676 | 889.56	5789 | 281.62	139 | 382.75	623 | 12.75
HUSSEIN	**WYOMING**	**SPEECH**	**OPTIMISM**	**COURT**	**BATTLEGROUND**
2609 | 578.92	201 | 365.91	2069 | 592.17	293 | 278.36	5376 | 377.01	681 | 11.59
TERRORISM	**WILMINGTON**	**CLINTON**	**TRAIL**	**PENNSYLVANIA**	**BALANCE**
2613 | 515.38	115 | 269.22	3499 | 498.88	1372 | 237.17	1503 | 347.22	1437 | 5.00

The values indicate the total number of occurrences in the references corpus and the chi-square significance.

Edwards, senator from *North Carolina* and *running* mate of John Kerry. Although the *Supreme Court* refused his *candidacy* in *Pennsylvania* over invalid nominating petitions, Ralph Nader was on the ballot in more than 30 states. Articles about him reiterated controversies over vote-splitting in the previous election and the *Supreme Court's* decision to end the Bush vs. *Gore* recounts in December 2000.

The results support claims that personalities, strategies, and campaign events dominate over substantive policy and governance issues—a possible reason for the average voter's limited interest in and knowledge about political processes (Haswell, 1999; Watts et al., 1999; Wayne, 2001). The news media's ongoing hunt for currency and novelty helps explain this observation, as the candidates' positions tend to remain constant throughout a campaign and therefore lack newsworthiness. The proliferation of polling (Traugott & Lavrakas, 2000) and market forces that demand media formats with significant entertainment value (Iyengar, Norpoth, & Hahn, 2004) also contribute to the lack of substantive information.

COVERAGE OF ENVIRONMENTAL ISSUES

The Special Reports section of the U.S. Election Monitor 2004 allowed users to investigate the candidates' positions on environmental issues and shows how these positions were portrayed by the media. This component used an annotated archive of Web data also referred to as *contextualized information space* (Scharl, 2007). Annotation services added several types of metadata (topic, time stamp, semantic orientation) to the documents stored in this archive. A topic hierarchy with three distinct layers defined the report type (global warming and energy policy), relevant subtopics (renewable energy, fossil fuels, nuclear energy, greenhouse gases, climate effects), and the respective text patterns (regular expressions) that indicate the topics.

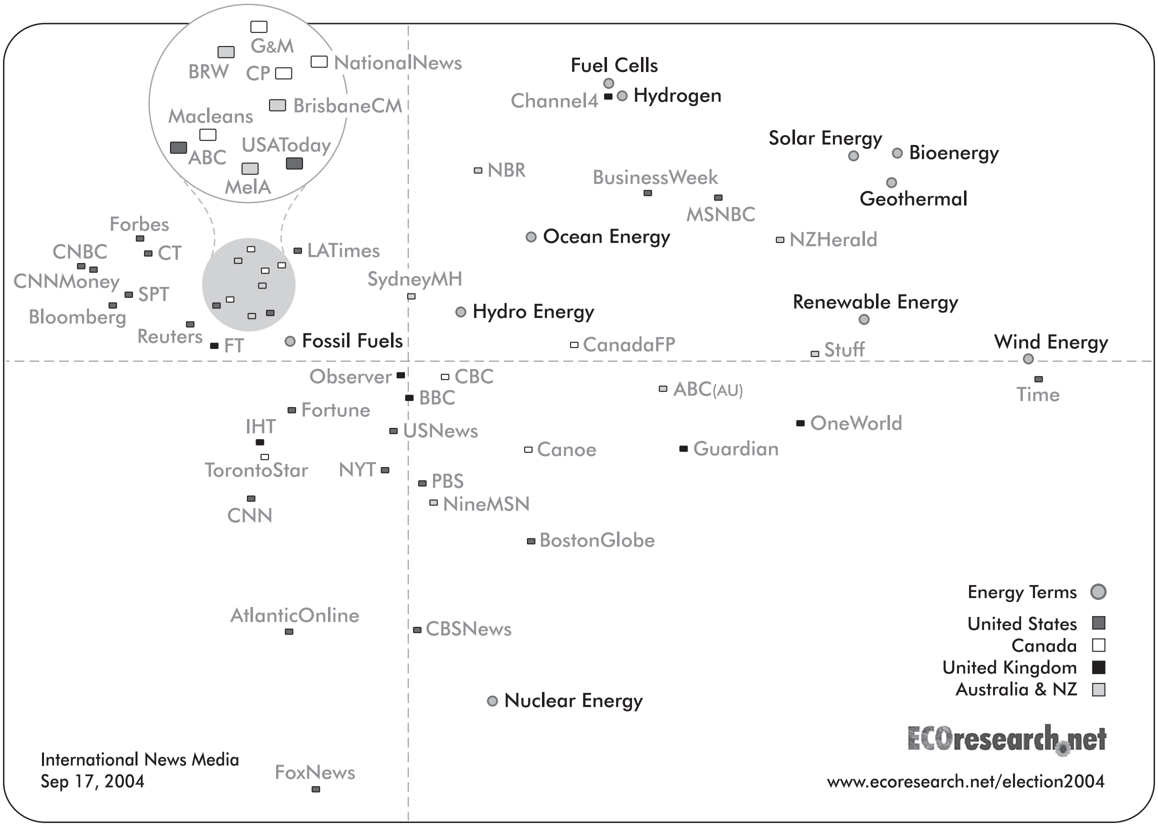
Focusing on energy policy, one such analysis investigated Web coverage of renewable energy, fossil fuels, and nuclear power—crucial issues considering the global environmental impact of U.S. energy policy decisions. The Election Monitor's Web site allows users to list sentences containing both candidate references and energy-related terms and to sort these sentences by semantic orientation. It also provides aggregated data by summarizing the prominence of energy terms among international news media sites in the perceptual map of Figure 4. Such maps aim to provide a high-level view on news media coverage of related topics within a specific domain, shedding light on the relative importance of these topics as perceived by the media. Such a condensed representation of the information space is a valuable tool to classify organizations. Generated repeatedly over time, perceptual maps allow analysts to track trends in online media coverage and quickly identify unusual developments (Scharl, 2004).

The diagram's distribution of data points was calculated by using the correspondence analysis module of SPSS (Statistical Package for the Social Sciences; www.spss.com) to process the table of term frequencies as of September 17, 2004. In this snapshot analysis, co-occurring terms appear close to each other in the computationally created two-dimensional space. The circular and rectangular markers represent energy-related concepts and media sites, respectively.

When interpreting the diagram, it is important to note that (a) term frequency rather than media attitude determines the position of data points, (b) axes of artificially generated spaces have no units of measurement, and (c) cross-proximities between row and column points (that is, terms and media sites) originate from different initial spaces. Thus, the position of a particular point may be interpreted only with respect to the other dimension's set of points, but not with a single point of the other dimension (Lebart et al., 1998).

The distribution of data points in Figure 4 shows a tripolar structure: fossil fuels in the upper left, renewable energies in the upper right, and nuclear energy near the bottom of the diagram. The diagram illustrates news organizations with distinct content—Fox News and Time, for example, with their coverage of *nuclear energy* and *wind energy*, respectively.

FIGURE 4. Perceptual map of energy terms and international news media.



Geographic differences are also apparent. *Fossil fuels* align with many Australian media, reflecting the country's richness in mineral resources. Most business publications congregate around fossil fuels as well, Business Week being a notable exception. References to *fuel cells* and *hydrogen* often appear together. Their potential use with various energy sources explains their slightly isolated position. Combining the energy carrier hydrogen with fuel cell conversion technology yields high efficiency and low pollution in applications such as zero-emission vehicles, energy storage, and portable electronics.

While the 2004 edition of the Election Monitor only provided a limited number of reports suggested by domain experts, the 2008 edition introduced in the following section uses a more flexible framework based on automated ontology extension (Liu, Weichselbraun, Scharl, & Chang, 2005) that identifies relevant topics and their relations based on unstructured textual data stored in the knowledge repository.

CONCLUSION AND OUTLOOK

The U.S. Election 2004 Web Monitor provided weekly snapshots of international Web coverage. It revealed the editorial slant of Web sites by measuring attention and attitude toward the U.S. presidential candidates. The empirical evidence contradicts the notion of liberal media, as the coverage between September and December 2004 clearly favored the Republican candidates. Since organizational bias is often specific to certain topics or domains, the second part of this paper introduced more nuanced representations of candidate coverage: keyword lists and perceptual maps that aggregate semantic associations within the knowledge base, and query interfaces for accessing this knowledge base on a granular level.

As the coverage of the 2004 U.S. presidential election might be atypical in several ways, future research should replicate and extend this study in the context of other campaigns. For this purpose, a revised Web portal available at www.ecoresearch.net/election2008 investigates the 2008 U.S. presidential election. Aside from major revisions of the underlying annotation services, the new portal adds a social aspect to the analysis of aggregated content by inviting users to cast their votes online. The portal also features just-in-time information retrieval agents, as well as visual navigational aids such as information landscapes, ontology views, tag clouds, and geospatial maps (Scharl et al., 2007).

The U.S. Election 2008 Web Monitor is part of the IDIOM research project [Information Diffusion across Interactive Online Media (www.idiom.at)], which aims to model the production, propagation, and consumption of electronic content to address four research questions: How redundant are online news media, and what are the major determinants of information flows within electronic networks? Can existing models of information propagation such as hub-and-spoke, syndication, and peer-to-peer adequately explain these information flows? How does Web content influence public opinion during political campaigns, and what are appropriate methods to measure and model the extent, dynamics, and latency of this process? Finally, which content placement strategies increase the impact on the target audience and support self-reinforcing propagation among individuals?

REFERENCES

Alterman, E. (2003). *What liberal media? The truth about bias and the news*. New York: Perseus Books.

Biber, D., Conrad, S., & Reppen, R. (1998). *Corpus linguistics—investigating language structure and use*. Cambridge, England: Cambridge University Press.

Croteau, D. (1998). Challenging the “liberal media” claim. *Extra!, 11*(4), 4–9.

Danielsson, P. (2004). Automatic extraction of meaningful units from corpora. *International Journal of Corpus Linguistics, 8*(1), 109–127.

Druckman, J. N., & Parkin, M. (2005). The impact of media bias: How editorial slant affects voters. *The Journal of Politics, 67*(4), 1030–1049.

Dutton, W. H., Elberse, A., & Hale, M. (1999). A case study of a Netizen’s guide to elections. *Communications of the ACM, 42*(12), 48–54.

Fico, F., & Cote, W. (1999). Fairness and balance in the structural characteristics of newspaper stories on the 1996 presidential election. *Journalism and Mass Communication Quarterly, 76*(1), 124–137.

Fogg, B. J., Soohoo, C., Danielson, D., Marable, L., Stanford, J., & Tauber, E. R. (2002). *How do people evaluate a Web site’s credibility?* Palo Alto, CA: Stanford University, Consumer Web Watch and Sliced Bread Design.

Gibbs, N. (2004). Blue truth, red truth. *Time, 164*(13), 24–33.

Giora, R. (1996). Discourse coherence and theory of relevance: Stumbling blocks in search of a unified theory. *Journal of Pragmatics, 27,* 17–34.

Haswell, S. (1999). The news media’s role in election campaigns: A big audience or a big yawn? *Australian Journalism Review, 21*(3), 165–180.

Holmes, N. (2002). Representative democracy and the profession. *Computer, 35*(2), 118–120.

Horrigan, J., Garrett, K., & Resnick, P. (2004). *The Internet and democratic debate*. Washington, DC: Pew Internet & American Life Project.

Howard, P. N. (2003). Digitizing the social contract: Producing American political culture in the age of new media. *The Communication Review, 6,* 213–245.

Iyengar, S., Norpoth, H., & Hahn, K. S. (2004). Consumer demand for election news: The horserace sells. *The Journal of Politics, 66*(1), 157–175.

Kiousis, S., & McCombs, M. (2004). Agenda-setting effects and attitude strength—political figures during the 1996 presidential election. *Communication Research, 31*(1), 36–57.

Lebart, L., Salem, A., & Berry, L. (1998). *Exploring textual data* (vol. 4). Dordrecht, Holland: Kluwer Academic Publishers.

Liu, W., Weichselbraun, A., Scharl, A., & Chang, E. (2005). Semi-automatic ontology extension using spreading activation. *Journal of Universal Knowledge Management, 0*(1), 50–58.

McEnery, T., & Wilson, A. (2001). *Corpus linguistics* (2nd ed.). Edinburgh, Scotland: Edinburgh University Press.

Scharl, A. (2004). Web coverage of renewable energy. In A. Scharl (Ed.), *Environmental Online Communication* (pp. 25–34). London: Springer.

Scharl, A. (2007). Towards the Geospatial Web: Media platforms for managing geotagged knowledge repositories. In A. Scharl & K. Tochtermann (Eds.), *The geospatial Web—How geobrowsers, social software and the Web 2.0 are shaping the network society* (pp. 3–14). London: Springer.

Scharl, A., Pollach, I., & Bauer, C. (2003). Determining the semantic orientation of Web-based corpora. In J. Liu, Y. Cheung, & H. Yin (Eds.), *Intelligent data engineering and automated learning, 4th international conference, IDEAL-2003, Hong Kong (Lecture Notes in Computer Science, Vol. 2690*) (pp. 840–849). Berlin: Springer.

Scharl, A., Weichselbraun, A., Hubmann-Haidvogel, A., Stern, H., Wohlgenannt, G., & Zibold, D. (2007). *Media watch on climate change: Building and visualizing contextualized information spaces*. Paper presented at the 6th International Semantic Web Conference (ISWC-2007), Busan, Korea.

Schmitt, K. M., Gunther, A. C., & Liebhart, J. L. (2004). Why partisans see mass media as biased. *Communication Research, 31*(6), 623–641.

Stone, P. J., Dunphy, D. C., Smith, M. S., & Ogilvie, D. M. (1966). *The General Inquirer: A computer approach to content analysis*. Cambridge, MA: MIT Press.

Stromer-Galley, J., & Foot, K. A. (2002). Citizen perceptions of online interactivity and implications for political campaign communication. *Journal of Computer-Mediated Communication, 8*(1), http://jcmc.indiana.edu/vol8/issuel/.

Sunstein, C. R. (2004). Democracy and filtering. *Communications of the ACM, 47*(12), 57–59.

Sutter, D. (2001). Can the media be so liberal? The economics of media bias. *Cato Journal, 20*(3), 431–451.

Traugott, M. W., & Lavrakas, P. J. (2000). *The voter's guide to election polls*. New York: Chatham.

Watts, M. D., Domke, D., Shah, D. V., & Fan, D. P. (1999). Elite cues and media bias in presidential campaigns: Explaining public perceptions of a liberal press. *Communication Research, 26*(2), 144–175.

Wayne, S. J. (2001). *The road to the White House 2000—the politics of presidential elections*. New York: Palgrave.

Weisberg, H. F. (2004). The U.S. presidential and congressional elections. *Electoral Studies, 24*, 777–784.

Yates, F. (1934). Contingency table involving small numbers and the $\chi 2$ test. *Journal of the Royal Statistical Society, 1*, 217–235.

Yi, J., Nasukawa, T., Bunescu, R. C., & Niblack, W. (2003). *Sentiment analyzer: Extracting sentiments about a given topic using natural language processing techniques*. Paper presented at the 3rd IEEE International Conference on Data Mining, Florida, USA.