Web Coverage of the 2004 US Presidential Election

Arno Scharl
Know-Center & Graz
University of Technology
Graz, Austria
scharl@ecoresearch.net

Albert Weichselbraun
Vienna University of Economics and
Business Administration
Vienna, Austria
weichselbraun@ecoresearch.net

Abstract

When corporations, news media and advocacy organizations embrace networked information technology, intentionally or unintentionally, they influence democratic processes. To capture and understand the influence of publicly available electronic content, the US Election 2004 Web Monitor[1] tracked the online coverage of US presidential candidates, and investigated how this coverage reflected their position on environmental issues.

1 Introduction

Most attempts to monitor the campaign performance of presidential candidates focus on public opinion, which is influenced by the consumption of media products. Analyzing patterns of political communication, however, should include the consumption as well as the production of content (Howard 2003). Monitoring candidates' coverage on the Web provides a complementary source of empirical data and window into the evolving concept of electronic democracy (Dutton, Elberse et al. 1999).

A recent *Pew/Internet* survey (Horrigan, Garrett et al. 2004) found that four out of ten US Internet users aged 18 or older accessed political material during the 2004 presidential campaign, up 50 percent from the 2000 campaign. For political news in general, more than two thirds of American broadband users and over half the dial-up users seek Web sites of national news organizations. International news sites are the second most popular category at 24 and 14 percent, respectively.

As traditional media extend their dominant position to the online world, analyzing their Web sites should therefore reflect the majority of political content accessed by the average user.

2 Impact of New Media on Public Opinion

Representative democracy offers significant possibilities for exploiting information networks (Holmes 2002), but there is little agreement on their specific impact. Proponents praise the networks' potential to increase the accessibility of information, encourage participatory decision-making, and facilitate communication with policy officials and like-minded citizens. From an advocate's perspective, disseminating environmental information via the Internet, directly or through news media, creates awareness by emphasizing the interdependency of ecological, economic, and social issues (Scharl 2004).

Critics portray McLuhan's global village as a "neofeudal manor with highly fortified and opulent castles (centers of industrial, financial, and media power) surrounded by vast hinterlands of working peasants clamoring for survival and recognition" (Tehranian 1999, p55f.). They argue that information networks polarize society by linking groups with similar political views. Low-overhead forms of personal publishing (Gruhl, Guha et al. 2004) such as Web logs and online discussion forums, for example, might reinforce a group's world view and shun opposing opinions. This reinforcement, amplified by biased media coverage, polarizes groups (Sunstein 2004) and degrades the climate of public discourse (Horrigan, Garrett et al. 2004).

The communication strategies of news media, corporations and advocacy organizations affect democratic processes. Yet they only condition, rather than determine these processes. Assuming

[1] http://www.ecoresearch.net/election2004

deterministic effects of information networks neglects the world's ambivalence, and results in conflicting claims regarding the networks' political impact. News media are free to choose which candidate to emphasize, and how to interpret current events (Wayne 2001). Most Americans prefer unbiased news sources (Horrigan, Garrett et al. 2004), but Web sites tend to reflect their owners' political agenda, and thus contribute to a polarized electorate.

While a narrow margin decided the last two US presidential elections, differences in the candidates' positions became more pronounced in 2004, and the political deliberation more partisan. Partisans tend to perceive mass media content as biased against their point of view. 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). Recent research suggests that selective categorization best explains hostile media effects (Schmitt, Gunther et al. 2004).

3 Methodology

Given an increasingly polarized electorate and hostile media effects that impair partisans' judgment, analyzing political Web content requires objective measures of organizational bias. Yet the volume and dynamic nature of Web documents complicate testing the assumption of organizational bias. To address this challenge, the *US Election 2004 Web Monitor* sampled 1,153 Web sites in weekly intervals. The project drew upon the *Newslink.org*, *Kidon.com* and *ABYZ-NewsLinks.com* directories to compile a list of 42 US news organizations and 72 international sites from four other English-speaking countries: Canada, United Kingdom, Australia and New Zealand. To extend the study, the sample included the Web sites of the Fortune 1000 (the largest US corporations ranked by revenue) and 39 environmental organizations.

Considering the dynamics of Web content in general and presidential campaigns in particular (Howard 2003), 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 help compare sites of heterogeneous size, and reduce the dilution of top-level information by content in lower hierarchical levels (Scharl 2000).

Such a collection of recorded content used for descriptive analysis is often referred to as corpus. This research investigated and visualized regularities in three groups of Web sites by applying methods from corpus linguistics and textual statistics (Biber, Conrad et al. 1998; Lebart, Salem et al. 1998; McEnery and Wilson 2001).

Quantitative textual analysis of Web documents necessitates three steps in order to yield a useful machine-readable representation (Lebart, Salem et al. 1998):

- The first step *converts* hypertext documents into plain text – i.e., processing the gathered data and eliminating markup code and scripting elements.
- The second step *segments* the textual chain into minimal units by removing coding ambiguities such as punctuation marks, the case of letters, hyphens, or points in abbreviations. In the case of the Election Monitor, this process yielded about half a million documents each week, comprising about 125 million words in 10 million sentences. The system then identified and removed redundant segments such as headlines and news summaries, whose appearance on multiple pages distorts frequency counts.
- The third step, *identification*, groups identical units and counts their occurrences – i.e., creating an inventory of words, or multiword units of meaning (Danielsson 2004). The frequency of candidate references presented in the following section is based on such an exhaustive index, which often uses decreasing frequency of occurrence as the primary sorting criterion and lexicographic order as the secondary criterion.

Frequency of References (Attention)

Media coverage and public recognition go hand-in hand (Wayne 2001), documented by strong correlations between the attention of news media and both public salience and attitudes toward presidential candidates (Kiousis and McCombs 2004). The *US 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, ignores equally valid references to *president bush* and *george walker bush*.

Figure 1. US Election 2004 Web Monitor one day after the election (Nov 3, 2004)

Yet a general query for *bush* fails to distinguish the president's last name from references to wilderness areas or woody perennial plants.

After the election, nearly two thirds of the media references mentioned George W. Bush and Dick Cheney, up four percentage points from the preceding week (Figure 1). About one third reported on John Kerry and John Edwards. The Fortune 1000 companies and environmental organizations dedicated over 80 percent of their coverage to the president and his running mate.

Across all three samples, the independent team of Ralph Nader and Peter Camejo garnered less than five percent of the attention.

4 Semantic Orientation of References (Attitude)

Calculating the frequency of candidate references disregards their context (Yi, Nasukawa et al. 2003). Therefore, the system also tracked attitude, the semantic orientation of a sentence towards the candidates (Scharl, Pollach et al. 2003).

The algorithm calculated the distance between the target word and 4,400 positive and negative words from the General Inquirer's tagged dictionary (Stone, Dunphy et al. 1966). Reverse lemmatization added about 3,000 terms to the dictionary by considering plurals, gerund forms, past tense suffixes and other syntactical variations (e.g. manipulate → manipulates, manipulating, manipulated).

Two sentences from news media on November 4 exemplify positive vs. negative coverage of a topic (zero indicates neutral coverage). The underlined words, identified in the tagged dictionary, were used to compute the semantic orientation of sentences with oil price references.

- "US stocks rallied Wednesday, boosted by shares of health and defence companies that are seen benefiting from the re-election of President *George W. Bush*, but higher **oil prices** checked advances" (NEW ZEALAND HERALD). ↑ (+ 4.09)
- "The dollar hit its lowest level in more than eight months against the Euro Thursday, falling sharply on worries about the economic effects of rising **oil prices** and expectations of continued trade and budget deficits in *President Bush's* second term" (ST. PETERSBURG TIMES). ↓ (– 4.03)

Initially, media coverage favored the Republicans, although the Democratic contenders gained ground in September 2004 (Figure 2). Kerry's performance in the first televised debate accelerated these gains in media attitude, followed by a tight race between the two teams in the four weeks preceding the election. The re-election of George W. Bush again widened the gap, understandably considering the positive connotation of terms such as *winning* and *victory*.

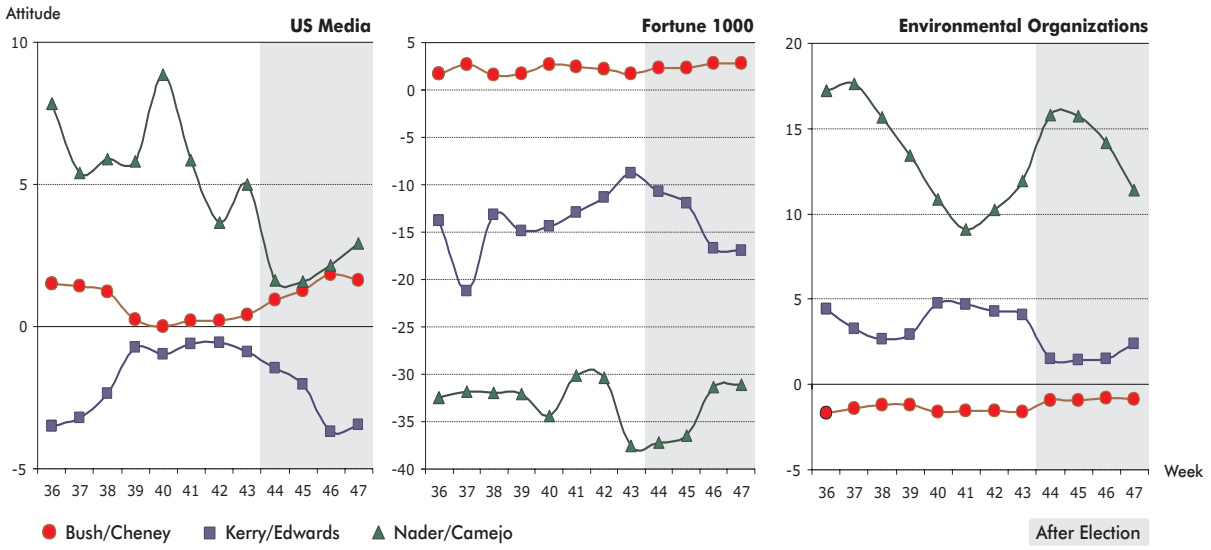

Figure 2. Attitude of the US Media, the Fortune 1000 and environmental organizations towards the US presidential candidates between September and November 2004

Compared to the news media, the other two samples showed more bias. Fortune 1000 companies presented the Republican candidates most favorably, 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.

To investigate these claims, separate analyses related environmental issues to Web sites and candidates. In terms of energy policy, for example, one such analysis investigated Web coverage of renewable energy, fossil fuels and nuclear power – a crucial aspect in light of recent geopolitical events and the global environmental impact of US energy policy decisions. On a micro level, the Election Monitor's Web site allowed users to list sentences containing both candidate references and energy-related terms, and sort these sentences by semantic orientation.

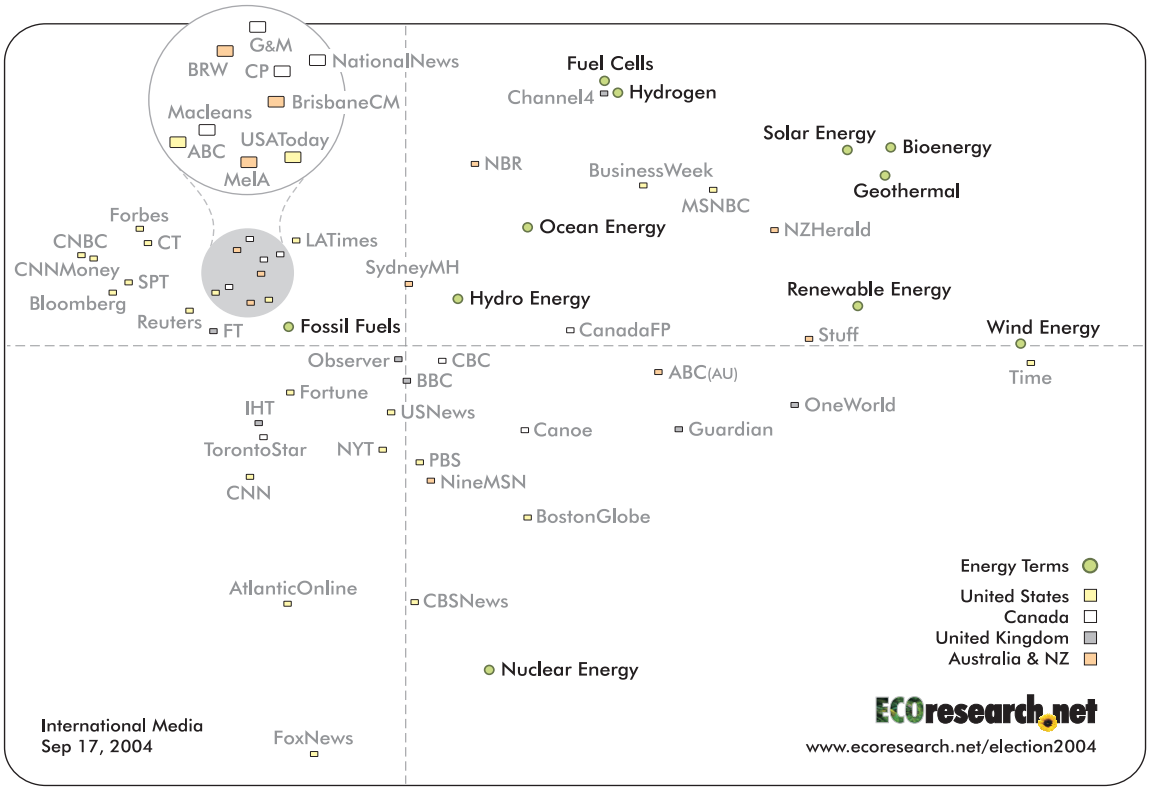

Figure 3. Perceptual map of energy terms among international news media

On a macro level, the perceptual map of Figure 3 summarizes the prominence of energy terms among international media sites. The diagram is based on correspondence analysis, which processed the table of term frequencies as of 17 September 2004. Related concepts and organizations 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 term frequency, not media attitude, determines the position of data points.

The distribution of data points 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. 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.

5 Refining Attitude Measures

Lexically identical occurrences with differing or even opposite meanings, depending on the context, represent an inherent problem of automatically determining the semantic orientation of Web content (Wilson, Wiebe et al. 2005). Word-sense ambiguity, for example, is a common phenomenon. *Arrest* as a noun takes custody by legal authority, for example, while *arrest* as a verb could mean to catch or to stop. Similarly, in economics the noun *good* refers to physical objects or services. As an adjective, *good* assigns desirable or positive qualities. Part of speech tagging considers this variability by annotating Web content and distinguishing between nouns, verbs, adjectives and other parts of speech.

Besides differences in word-sense, analysts also encounter other types of ambiguities – e.g., idiomatic versus non-idiomatic term usage, or various pragmatic ambiguities involving irony, sarcasm, and metaphor (Wiebe, Wilson et al. 2005).

Given the considerable size of the corpus and the need to publish results in weekly intervals, the system was designed to maximize throughput in terms of documents per second. A comparably simple approach restricted to single words and without sentence parsing ensured the timely completion of the weekly calculations.

Planned extensions will add multiple-word combinations to the tagged dictionary to discern morphologically similar but semantically different terms such as *fuel cell* and *prison cell*. Yet the lexis of Web content only partially determines its semantic orientation, despite using multi-word units of meaning instead of single words or lemmas (Danielsson 2004). Prototypical implementations such as the *OpinionFinder* (Wilson, Hoffmann et al. 2005) have demonstrated that grammatical parsing can successfully address this limitation by identifying ambiguities and capturing meaning-making processes at levels beyond lexis – correctly identifying, annotating and evaluating nested expressions of various complexity (Wiebe, Wilson et al. 2005).

6 Keyword Analysis

Keyword Analysis locates words in a given text and compares their frequency with a reference distribution from a usually larger corpus of text. To complement measures of attention and attitude towards a candidate, keywords grouped by political party and geographic region highlighted issues associated with each candidate. For that purpose, the system compared the term frequencies in sentences mentioning a candidate (target corpus) with the term frequencies in the entire sample of 1,153 Web sites from media organizations, the Fortune 1000, and environmental organizations (reference corpus).

The results suggest that personalities and campaign events dominate over substantive policy issues, a possible reason for the average voter's limited interest in and knowledge about political processes (Wayne 2001). Table 1 summarizes keywords that US news media associated with the presidential candidates and their running mates in the week preceding the election. The list ranks keywords by decreasing significance, computed via a chi-square test with Yates' correction for continuity. To avoid outliers, the list only considers nouns with at least 100 occurrences in the reference corpus.

Republicans		Democrats		Independents	
George Bush	**Dick Cheney**	**John Kerry**	**John Edwards**	**Ralph Nader**	**Peter Camejo**
debate	lynne	nominee	carolina	ballot	opinion
challenger	daughter	vietnam	running	percent	running
iraq	halliburton	challenger	debate	candidacy	respondents
gore	debate	debate	nominee	advocate	electors
war	lesbian	massachusetts	gephardt	party	commonwealth
speech	rumsfeld	nomination	iowa	supreme	ballot
nominee	pensacola	war	ashton	gore	endorsement
guard	rally	rival	optimism	petition	nominee
hussein	wyoming	speech	north	court	battleground
terrorism	wilmington	clinton	trail	pennsylvania	balance

Table 1. Keywords of US news media (Oct 27, 2004)

The keywords document 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* of 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 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.

7 Conclusion and Future Research

The *US Election 2004 Web Monitor* provided a weekly snapshot of international Web coverage, measuring attention and attitude towards the US presidential candidates. Keywords grouped by political party and geographic region summarize issues associated with each candidate.

Compared to the Web sites of news media, campaign managers have less control of spin and impact in media that rely on citizenry for message turnover (Howard 2003). Extending the current system will allow measuring information propagation, not only among corporate Web sites but also via Web logs, online discussion forums and other forms of personal publishing. Investigating the propagation of political content in such environments requires large samples to measure spatial effects, and frequent monitoring to account for temporal effects.

For measuring information propagation, Gruhl et al. (Gruhl, Guha et al. 2004) suggest distinguishing between internally driven, sustained discussions (chatter) and externally induced sharp rises in activity (spikes). Occasionally, spikes result from chatter through resonance when insignificant events trigger massive reactions. Resonance occurs when individual interactions generate large-scale, collective behavior, often showing a sensitive dependence on initial conditions. Social network analysis attempts to explain such macroscopic propagation of information between people, groups and organizations (Kumar, Raghavan et al. 2002). By disseminating information via their social networks, individuals create strong peer influence that often surpasses exogenous influences.

Efforts to create a more responsible electorate (Dutton, Elberse et al. 1999) can leverage this peer influence to trigger self-reinforcing content propagation among individuals. Relationships between these individuals determine the paths of information dissemination. It is along these paths that inter-individual communication multiplies the impact of spikes and creates widespread attention. Knowledge on the structure and determinants of these paths could help promote issue-oriented voting. This in turn would lead to a better-informed electorate aware of its leadership choices, and able to hold decision-makers accountable.

Modeling the production, propagation and consumption of political Web will help address four research questions: How redundant is Web content, and what technical and organizational factors influence information flows within the network? 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, 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?

Acknowledgements. Our first word of appreciation goes to Jamie Murphy for his ongoing support throughout the project. We would also like to thank Astrid Dickinger, Wilhelm Langenberger, Wei Liu, Antonijo Nikolic, Maya Purushothaman, Dave Webb and Mark Winkler for their valuable help and suggestions. The US Election 2004 Web Monitor represents an initiative of the Research Network on Environmental Online Communication (www.ecoresearch.net), cooperating with the University of Western Australia, Graz University of Technology, Vienna University of Economics and Business Administration, and the Know-Center, which is funded by the Austrian Competence Center program Kplus under the auspices of the Austrian Ministry of Transport, Innovation and Technology (http://www.ffg.at) and by the State of Styria.

References

Biber, D., S. Conrad, et al. (1998). *Corpus Linguistics – Investigating Language Structure and Use.* Cambridge, Cambridge University Press.

Danielsson, P. (2004). "Automatic Extraction of Meaningful Units from Corpora", *International Journal of Corpus Linguistics,* 8(1): 109-127.

Dutton, W. H., A. Elberse, et al. (1999). "A Case Study of a Netizen's Guide to Elections", *Communications of the ACM,* 42(12): 48-54.

Gruhl, D., R. Guha, et al. (2004). Information Diffusion Through Blogspace. *13th International World Wide Web Conference,* New York, USA, ACM Press.

Holmes, N. (2002). "Representative Democracy and the Profession", *Computer,* 35(2): 118-120.

Horrigan, J., K. Garrett, et al. (2004). *The Internet and Democratic Debate.* Washington, 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.

Kiousis, S. and M. McCombs (2004). "Agenda-Setting Effects and Attitude Strength – Political Figures during the 1996 Presidential Election", *Communication Research,* 31(1): 36-57.

Kumar, R., P. Raghavan, et al. (2002). "The Web and Social Networks", *Computer,* 35(11): 32-36.

Lebart, L., A. Salem, et al. (1998). *Exploring Textual Data.* Dordrecht, Kluwer Academic Publishers.

McEnery, T. and A. Wilson (2001). *Corpus Linguistics.* Edinburgh, Edinburgh University Press.

Scharl, A. (2000). *Evolutionary Web Development.* London, Springer. http://webdev.wu-wien.ac.at/.

Scharl, A., Ed. (2004). *Environmental Online Communication.* London, Springer. http://www.ecoresearch.net/springer/.

Scharl, A., I. Pollach, et al. (2003). Determining the Semantic Orientation of Web-based Corpora. *Intelligent Data Engineering and Automated Learning, 4th International Conference, IDEAL-2003* (Lecture Notes in Computer Science, Vol. 2690). J. Liu, Y. Cheung and H. Yin. Berlin, Springer: 840-849.

Schmitt, K. M., A. C. Gunther, et al. (2004). "Why Partisans See Mass Media as Biased", *Communication Research,* 31(6): 623-641.

Stone, P. J., D. C. Dunphy, et al. (1966). *The General Inquirer: A Computer Approach to Content Analysis.* Cambridge, MIT Press.

Sunstein, C. R. (2004). "Democracy and Filtering", *Communications of the ACM,* 47(12): 57-59.

Tehranian, M. (1999). *Global Communication and World Politics – Domination, Development, and Discourse.* Boulder, Lynne Rienner.

Wayne, S. J. (2001). *The Road to the White House 2000 – The Politics of Presidential Elections*. New York, Palgrave.

Wiebe, J., T. Wilson, et al. (2005). "Annotating Expressions of Opinions and Emotions in Language", *Language Resources and Evaluation* 39(2-3): 165-210.

Wilson, T., P. Hoffmann, et al. (2005). OpinionFinder – A System for Subjectivity Analysis. *Human Language Technology Conference / Conference on Empirical Methods in Natural Language Processing (HLT-EMNLP 2005),* Vancouver, Canada.

Wilson, T., J. Wiebe, et al. (2005). Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis. *Human Language Technology Conference / Conference on Empirical Methods in Natural Language Processing (HLT-EMNLP 2005),* Vancouver, Canada.

Yi, J., T. Nasukawa, et al. (2003). Sentiment Analyzer: Extracting Sentiments about a Given Topic using Natural Language Processing Techniques. *3rd IEEE International Conference on Data Mining,* Florida, USA.