Information Research, Vol. 10 No. 3, April 2005

The case of the news search engine: an exploratory empirical analysis of *Google News*

Sandeep Krishnamurthy and Elaine Taniguchi
The University of Washington
Bothell, WA 98011, USA

Abstract

Introduction. News search engines crawl the living Web of news sources to collect, group and distribute timely content. They look at fewer sources than traditional search engines and visit these sources more often. News search engines dynamically group content to create a Web page.
Method. In this paper, we study one news search engine, *Google News* We analyse and classify the sources used by this engine. We also conducted an analysis of two randomly chosen American states, Washington and Kansas, to ascertain the proportion of newspaper Websites featured on *Google News*.
Results. We find that *Google News* is biased towards newspapers in the USA: 73.24% of the sources are American. The top five countries, USA, UK, Canada, Australia and India, account for 91.95% of the sources. The top ten countries account for 94.04% of sources. A total of sixty-five countries contributed 121 sources, which accounted for 2.69% of all sources. '.com' was the most common domain extension with 78.95% of sites using it. Deep linking was used by only 5.46% of sites. Only 41.30% of Washington newspaper Websites and 52.78% of Kansas newspaper Websites were included in *Google News*. This indicates that regional and local news sources may be represented to a lower degree than is widely believed.

Introduction

The news search engine seeks to replace the human editor with a computer algorithm. These search engines collect, group and present content from news sources on the World Wide Web. Examples of such engines include *Google News*, *RocketNews*, *Yahoo! News*, *DayPop*, *AllTheWeb News* and *Ananova*. Two arguments have been proposed in favour of having such search engines: information processing and potential diversity.

The first argument is that a search engine is able to process a vast quantity of information in comparison to a human editor. The World Wide Web is a vast repository of information and it is not possible even for an expert to be able to pay attention to a wide variety of sources. As the number of sources grows, the potential advantage of using an algorithm as an aggregator grows as well. The founder of *Google News*, Krishna Bharat, explains:

> ...we get 100,000 articles a day. A human editor couldn't read that many. We have people who try to create an aggregate of what's been done in the media on a given topic and they write a report about it. Journalists do that all the time, and they do an extremely good job. But imagine doing that for every story in the world, every time. We want to give you speed in addition to timeliness. (Kramer, 2003)

The second argument in favor of news search engines is that they could potentially provide access to more diverse opinions. Once again, the words of the founder of *Google News* are instructive:

> I want this to be a force for a democracy. I want us to be an honest broker, and I want newspapers featured on our site to get traffic from us. There's never been a more controversial time on the planet. I

> think it's great to be a news source at this point because there's so much hunger for news. You see a lot more diversity in the news coverage on our site than on others. I think the diversity is a mirror to the diversity of opinion there is worldwide. One of the things that makes us objective is we show all points of view. Even if you disagree with one, we give you both -- the majority and the minority point of view. The ones you don't agree with are education. It's nice to know what the other side is thinking. You'll see left-leaning ones as much as much as you see right-leaning ones. Frankly, the software doesn't know the difference between left and right, which is good.(Kramer 2003)

Our interest here is in the second argument. Clearly, the value of the news search engine as an intermediary in the content distribution process hinges on the diversity of the content. If the content is diverse, the news search engine is perceived as a legitimate aggregator. If the content is not diverse, the results will appear biased to an interested reader leading to an erosion of credibility. In our view, the diversity of sources directly correlates with diversity of content. Imagine a news search engine that aggregates information from all news sources. Such an engine could potentially become the broker that Bharat envisions. However, if a news search engine simply focuses on a bounded subset of sources(e.g. only college newspapers), the news is likely to be skewed.

Two recent papers, Gerhart(2003) and Hindman, *et al.* (2003), have argued that search engines may *reduce* diversity. Gerhart demonstrates that for controversial topics, search engines are likely to present 'the sunny side' more often, thus, potentially suppressing controversy. Views that not part of the mainstream are not well represented. Hindman, *et al.* argue that, while all Web pages are potentially retrievable, only a few well-linked sites are visible when a user queries a search engine. As a result, they expect that search engines will aid in the creation of a winner-take-all traffic structure with a few sites dominating. This is consistent with the work of Huberman and colleagues; see, for example, Adamic and Huberman (2000). These papers have already drawn some criticism for over-stating the case for diversity and monopoly power. For instance, Brooks (2004) has argued that these papers, '*expect Google to be something other than Google*'. While acknowledging this criticism, it is our position that as information brokers become more pervasive, the mechanism they adopt to generate content deserves academic scrutiny.

Therefore, our goal in this paper is to examine the extent of diversity among the sources used by one news search engine: News.google.com. Google closely guards its list of 4,500 sources for business reasons. Our request for the list of sources was politely declined. Therefore, in this paper, we manually identify, classify and analyse these sources.

Exploratory findings are presented. While at least ninety-three countries are represented in *Google News*, the content is highly US-centred: 73.24% of the sources were American sites. Moreover, sources from a few countries dominate. Sites from five countries: USA, UK, Canada, Australia and India, make up 91.95% of the sources. The vast majority of sources (74.39%) had a .com extension. Moreover, local newspapers are not as well represented as may have been previously thought; only 41.30% of newspaper Websites in Washington state and 52.78% of newspaper Websites in Kansas were represented on *Google News*.

Understanding news search engines

News search engines conduct four types of activities: crawling, indexing, grouping and distribution. The first two steps are similar to other search engines. News search engines crawl several sites to collect information. This information is then added to an index.

News search engines excel at grouping of information. The information in the index is re-organized into sub-categories such as World, Business, Technology and Sports and placed on a Website. *DayPop* places news information into the categories, Top News, Word Bursts (stories are grouped by key words in this section) and News Bursts (stories are grouped by story words in this section).

The distribution of the information from the news search engine to the consumer takes place through many channels. Users could visit the Website of the news search engine and peruse it just like any other news Website. They could also search the index for stories, of course. News search engines do not cache content and typically store content for a short period (e.g., *Google News* retains stories that are thirty days old or less). Alternatively, users could sign up for e-mail alerts in specific categories. For instance, a user could choose to follow all stories that mention the words 'Tiger Woods'. The search engine would then send them an e-mail when there is a new story that mentions these words. News search engines are different from traditional search engines in the following ways:

- Such engines scan a limited number of sources. As an example, *News.google.com* scans 4,500 sources while *Google.com* indexes over four billion pages. At the time of writing, *DayPop* seems to look at the largest number of sources with 59,000, including Weblogs, however. *DayPop* has coined the term *the living Web* to describe this subset of the Web; i.e., the portion that is updated once or more a day.
- News search engines scan their sources multiple times a day in contrast to search engines who may visit a page as infrequently as once a month. The lead story on *news.google.com* changes every fifteen minutes(Kramer 2003). *DayPop* crawls major news sites such as CNN once every three hours and the minor sites once every twenty-four hours (DayPop 2004). As a result, news search engines are excellent ways of searching for information about current events (Sullivan 2003).
- Unlike traditional search engines, news search engines are not simple aggregators of information. They add value by grouping information on a much higher scale. The grouping process is closely analogous to the editorial process. These processes try to emulate the rules a human editor may use. The grouping algorithm decides which story deserves to be on the front page at any given time. .

Our focus

For the purpose of this paper, we focus on *Google News*, the news search engine of *Google.com*. This service scans about 4,500 news sources regularly to gather and organize content. We chose to study this news search engine for the following reasons:

- The place of Google in our culture is on the ascendancy. Googling has become a verb. Wired magazine recently published an entire issue on Googlemania. A parody of News.Google.com now exists.
- Other news search engines scan a limited number of sources while News.google.com has a sufficiently large number. For instance, *Yahoo! News* uses ten sources: AP, Reuters, Agence France Presse (AFP), washingtonpost.com, USATODAY.com, Los Angeles Times, Chicago Tribune, U.S. News & World Report, National Public Radio(NPR) and Reuters Features. *AltaVista News* places a high degree of emphasis on NYTimes.com.
- *Google News* is an award-winning Website. It won the 2003 Webby award in the news category.
- *Google News* is a very popular site. In August 2004, it attracted 5.8 million visitors placing it at number 14 in the list of most popular sites on the Web for current events and global news (Lasica 2004).
- News.google.com groups stories based on the content and also places them in categories such as World, Sports and Technology. Moreover, it decides on the relative placement of stories on the Website.
- The increased interest in Google among scholars, e.g., Brooks (2004), Galitsky and Levine (2004), Gerhart (2003) and Hindman, *et al.* (2003), allows us to place our work in context.

Google News: a case study

At first glance, news.google.com, the Website of *Google News*, appears to be just another news site. In actuality, it is a complicated technological endeavor that seeks to simulate a news Website by gathering information from 4,500 news sources.

The *Google News* Website describes itself as follows:

> *Google News* presents information culled from approximately 4,500 news sources worldwide and automatically arranged to present the most relevant news first. Topics are updated continuously throughout the day, so you will see new stories each time you check the page. Google has developed an automated grouping process for *Google News* that pulls together related headlines and photos from thousands of sources worldwide—enabling you to see how different news organizations are reporting the same story. (*Google News*, 2004)

How does a news source make it to *Google News*? First, a team of reviewers decides which site to crawl(Kramer 2003). Second, sites are able to opt-out of this process. This process requires specific language in a file titled robots.txt. Google will respect the specific meta-tag in this file and not crawl anybody who does not wish to be crawled. As the founder of *Google News* put it:

> We are only able to crawl sites that allow us to crawl. Any news search that tries to link you to new

> content is going to come up against a barrier—either they specify that robots are not allowed to access this site, or they put the content behind registration that the machine cannot get by. This is a fundamental issue. It has to do with how people monetize their content. The news community needs to figure out how they're going to get traffic from us. The *New York Times* has a nice solution. They allow us to connect to the content and send traffic to one page. If people want to browse beyond that then they have to register. If people are really happy with the content they'll register. I think that's a great model. (Kramer 2003)

It may seem strange that publishers are willing and eager to have *Google News* crawl their Websites and collect data several times a day. In contrast to Google.com, at this time, *Google News* does not cache stories. Users are directed to the original Website. Thus, *Google News* directly contributes to the traffic of sites making it an attractive proposition for publishers. Therefore, publishers view the site as beneficial since it increases traffic to their sites.

In addition to the main site, *Google News* runs five country editions for Australia, Canada, France, Germany, India, Italy, New Zealand, Spain, United Kingdom and United States. *Google News* is multi-lingual, the France, Spain, Germany and Italy editions are in French, Spanish, German and Italian respectively. Interestingly, the front page of each edition does not focus on stories exclusively from that country. Rather, it is a mix of sources.

Method

As indicated earlier, Google closely guards its list of news sources for business reasons. Therefore, to gain access to the list of sources, and identified the name and Website of the sources. Data collection started on July 1, 2003 and ended on September 21, 2003. All information was entered into a spreadsheet. As new names were added, we took care to make sure that there was no duplication. As a result of this process, we were able to create a spreadsheet with 4,499 sources. This formed the basis of our analysis.

Results

Country-level analysis

Table 1 provides the number of sources classified by country. The USA tops the list with 73.24% of the sources. The top 5 countries, USA, UK, Canada, Australia and India, account for 91.95% of the sources. The top 10 countries account for 94.04% of sources. The top 28 countries (each of these countries had at least five sources) account for 96.78% of sources. A total of 65 countries contribute 121 sources which account for 2.69% of all sources.

Number	Country	Count	Percent	Cumulative Count	Cumulative Percent
1	USA	3295	73.24%	3295	73.24%
2	UK	412	9.16%	3707	82.40%
3	Canada	233	5.18%	3940	87.58%
4	Australia	147	3.27%	4087	90.84%
5	India	50	1.11%	4137	91.95%
6	South Africa	23	0.51%	4160	92.46%
7	Ireland	20	0.44%	4180	92.91%
8	New Zealand	18	0.40%	4198	93.31%
9	Singapore	17	0.38%	4215	93.69%
10	France	16	0.36%	4231	94.04%
11	Israel	11	0.24%	4242	94.29%
12	Malaysia	10	0.22%	4252	94.51%
13	Netherlands	9	0.20%	4261	94.71%
14	Phillipines	8	0.18%	4269	94.89%
15	Russia	8	0.18%	4277	95.07%
16	Germany	7	0.16%	4284	95.22%
17	Hong Kong	7	0.16%	4291	95.38%
18	Italy	7	0.16%	4298	95.53%
19	Switzerland	7	0.16%	4305	95.69%
20	Czech Republic	6	0.13%	4311	95.82%
21	Iran	6	0.13%	4317	95.95%
22	Japan	6	0.13%	4323	96.09%
23	Malta	6	0.13%	4329	96.22%
24	China	5	0.11%	4334	96.33%
25	Nigeria	5	0.11%	4339	96.44%
26	Norway	5	0.11%	4344	96.55%
27	Pakistan	5	0.11%	4349	96.67%
28	Taiwan	5	0.11%	4354	96.78%
29	Unsure/Broken Link	24	0.53%	4378	97.31%
30	OTHERs	121	2.69%	4499	100.00%

Table 1: Number of sources, by country

This clearly demonstrates the highly concentrated nature of the distribution of sources across countries. To examine this further, we mapped the distribution of sources across countries on a log-log scale after sorting them on the number of sources (Adamic 2000). If the distribution follows the power-law distribution, the expected result will be straight line on this chart (a chart without the log transformation leads to a L-shaped result). The results are shown in Figure 1. This is a typical power-law distribution.

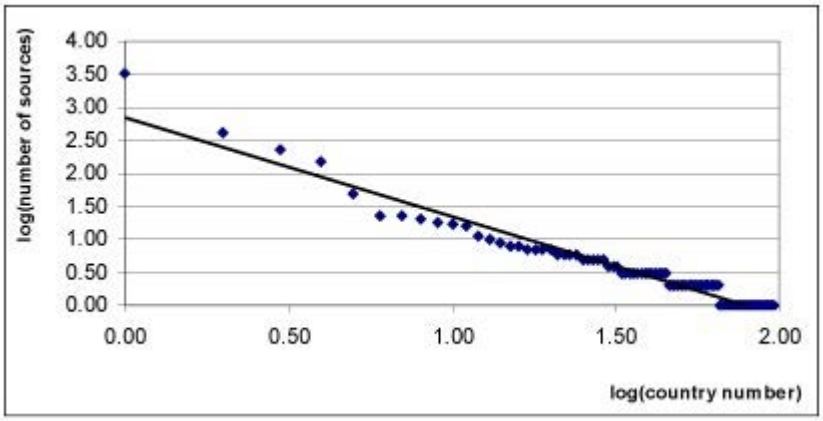

Figure 1: Power-law distribution of sources, by country

We also classified sources by geographical region, with the result shown in Table 2. North America has a clear lead over other regions. North America and Europe jointly account for 90.62% of all sources.

Number	Region	Count	Percent
1	North America	3545	78.80%
2	Europe	532	11.82%
3	Australia	165	3.67%
4	Asia	138	3.07%
5	Africa	52	1.16%
6	Middle East	35	0.78%
7	South America	8	0.18%
8	Unsure/Broken Link	24	0.53%

Table 2: Number of sources, by region

Domain-level analysis

Table 3 provides an analysis of the domain extensions of the news sources used on *Google News*. '.com' was the most common domain extension with 74.39%. We have used the convention '.com/' to indicate deep linking. If this is added, the popularity of '.com' rises to 78.95%. Country-specific domains such as '.co.uk', 'co.au' and 'co.za' were also popular.

	Domain	Label	Count	Percent	Cumu. Perc.
1	.com	Commercial	3347	74.39%	74.39%
2	.co.uk	United Kingdom	243	5.40%	79.80%
3	.com/	Commercial(deep link)	205	4.56%	84.35%
4	.org	Organization	158	3.51%	87.86%
5	.com.au	Australian Commercial Site	124	2.76%	90.62%
6	.net	Network	122	2.71%	93.33%
7	.ca	Canada	38	0.84%	94.18%
8	.edu	Educational Institutions	30	0.67%	94.84%
9	.co.za	South Africa	16	0.36%	95.20%
10	.ie	Ireland	14	0.31%	95.51%
11	.co.nz	New Zealand	12	0.27%	95.78%
12	.info	Information	9	0.20%	95.98%
13	IP Address	IP Address	9	0.20%	96.18%
14	.co.uk/	United Kingdom(deep link)	7	0.16%	96.33%
15	.gov	Government	7	0.16%	96.49%
16	.mil	Military	7	0.16%	96.64%
17	.de	Germany	6	0.13%	96.78%
18	.edu/	Educational Institutions(deep link)	6	0.13%	96.91%
19	.com.sg	Singapore	5	0.11%	97.02%
20	.no	Norway	5	0.11%	97.13%
21	OTHER		129	2.87%	100.00%
	TOTAL		4499	100.00%	

Table 3: Summary of domain extensions

Finally, we look at the deep linking practices of *Google News* in Table 4. A total of 244 news sources or 5.44% were

deep-linked and '.com/' was the most common deep-linked domain with 4.56%.

Number		Count	Percent of Deep Linkers	Percent of all sites
1	.com/	205	84.02%	4.56%
2	.co.uk/	7	2.87%	0.16%
3	.edu/	18	7.38%	0.40%
4	.net/	4	1.64%	0.09%
5	.org/	4	1.64%	0.09%
6	.net.au/	3	1.23%	0.07%
7	.co.zw/	1	0.41%	0.02%
8	.ie/	1	0.41%	0.02%
9	.tw/	1	0.41%	0.02%
	TOTAL	244	100.00%	5.42%

Table 4: Overview of deep linking practices

State newspaper analysis

To understand the proportion of the news universe captured by *Google News*, we procured a list of regional and local newspapers Washington State. This list was obtained from the Newspaper Association of America's Website. A total of forty-six newspapers was listed for Washington State. Every newspaper in this list had a Website and, therefore, was a potential candidate for *Google News*.

The results of our analysis is shown in Table 5. Only 41.30% of newspapers was listed on *Google News*. This was a surprise to us since we expected almost the entire list of newspapers to be represented.

Newspapers Not Listed in News.Google.Com			Newspapers Listed in News.Google.Com		
	Newspaper - Home Page	City		Newspaper - Home Page	City
1	Bainbridge Island Review	Bainbridge Island	28	Chronicle, The	Centralia
2	Bellingham Herald, The	Bellingham	29	Columbia Basin Herald	Moses Lake
3	Bothell/Kenmore Reporter	Bothell	30	Daily News, The	Longview
4	Columbian, The	Vancouver	31	Daily Record	Ellensburg
5	Daily Sun News	Sunnyside	32	Herald, The	Everett
6	Daily World	Aberdeen	33	King County Journal	Bellevue
7	East County Journal	Morton	34	News Tribune, The	Tacoma
8	Federal Way News	Federal Way	35	Olympian, The	Olympia
9	Goldendale Sentinel	Goldendale	36	Omak-Okanogan County Chronicle	Omak
10	Issaquah Press	Issaquah	37	Peninsula Daily News	Port Angeles
11	Mercer Island Reporter	Mercer Island	38	Puget Sound Business Journal	Seattle
12	Methow Valley News	Methow Valley	39	Seattle Daily Journal of Commerce	Seattle
13	Monroe Monitor and Valley News	Monroe	40	Seattle Post-Intelligencer	Seattle
14	Nisqually Valley News	Yelm	41	Seattle Times, The	Seattle
15	Northern Kittitas County Tribune	Cle Elum	42	Skagit Valley Herald	Mt. Vernon
16	Peninsula Gateway, The	Gig Harbor	43	Snoqualmie Valley Record	Snoqualmie
17	Sequim Gazette, The	Sequim	44	Sun, The	Bremerton
18	South Whidbey Record	Langley	45	Tri-City Herald	Pasco
19	Spokesman-Review, The	Spokane	46	Walla Walla Union-Bulletin	Walla Walla
20	Stanwood/Camano News	Stanwood			
21	Star, The	Grand Coulee			
22	The Puyallup Herald	Puyallup			
23	The South County Sun	Royal City			
24	Vashon-Maury Island Beachcomber	Vashon			
25	Wenatchee World, The	Wenatchee			
26	Whidbey News Times	Oak Harbor			
27	Yakima Herald-Republic	Yakima			

Table 5: Washington State Newspapers

We found that 52.78% of newspapers was included. Data for this are shown in Table 6.

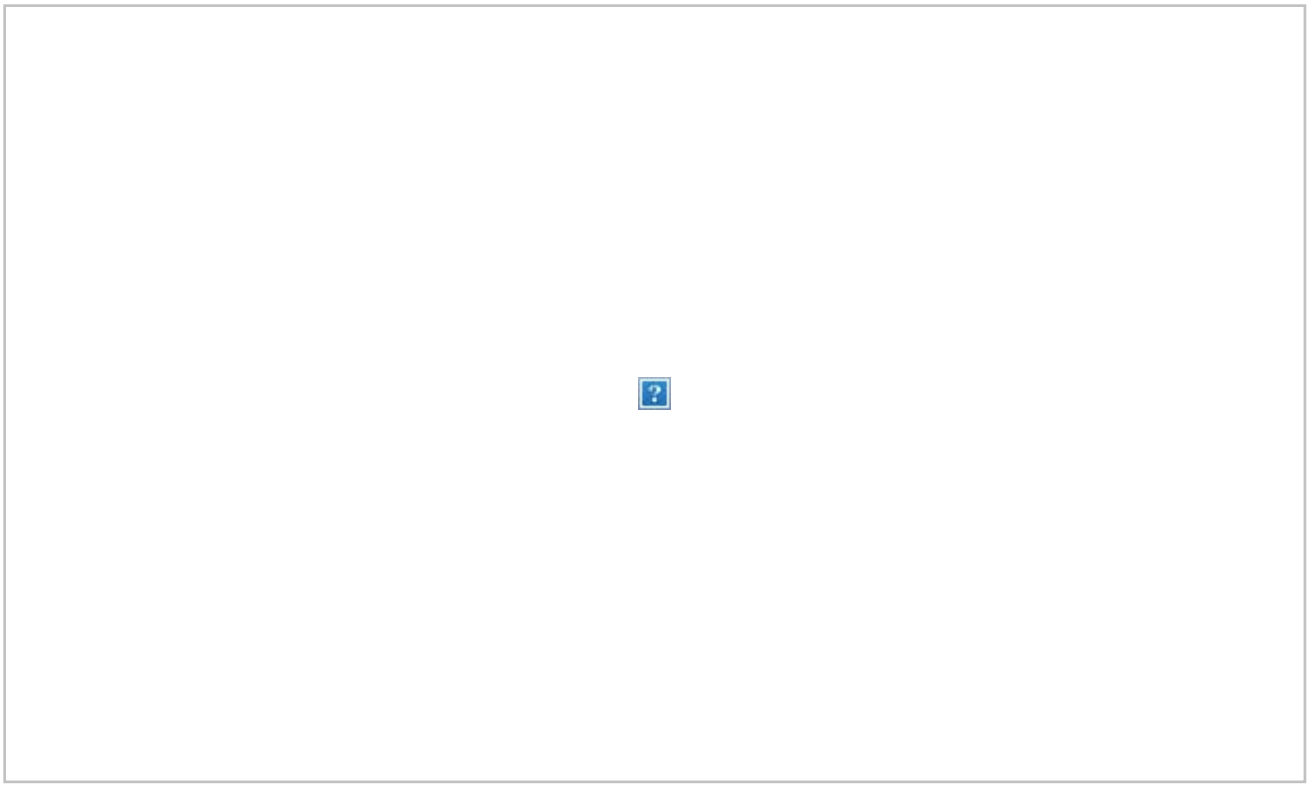

Table 6: Kansas state newspapers

Conclusion

In this paper, we present an empirical analysis of a news search engine, *Google News*. Our results show that, while there is some evidence for diversity among news sources, *Google News* is US-centred. Most of the sources have a '.com' extension. To our surprise, we found that only 41.30% of Washington and 52.78% of Kansas newspaper Websites were represented.

As news search engines become more popular, their claims of diversity and objectivity deserve careful scrutiny. Our work suggests that *Google News* may not be as diverse as previously thought. *Google News* is sure to expand its list of sources as it grows. Future research must examine how content is grouped and the impact of grouping algorithms on diversity.

Some readers may look at our findings and conclude that *Google News* is as diverse as it should be. Is it reasonable to expect a news search engine to capture the universe of sources for a defined set (e.g., all local newspaper Websites in the US)? Perhaps there is a minimum level of diversity one would expect from an impartial broker.

Much is unknown about how exactly Google groups content on *Google News*. Therefore, simply making it to the list of 4,500 sources will not be adequate to shine.

Acknowledgements

The author wishes to thank his student, Karen Tellevik, for her contributions, the suggestions of the anonymous referees.

References

- Adamic, L.(2000). *Zipf, power-laws, and Pareto-a ranking tutorial.* Retrieved 18 March, 2004 from http://www.hpl.hp.com/shl/papers/ranking/ranking.html.
- Adamic, L. & Huberman, B. (2000), The nature of markets in the World Wide Web. *Quarterly Journal of Electronic Commerce*, **1**(1), 5-12.
- Brooks, T.A. (2003). Web search: how the Web has changed information retrieval. *Information Research*, **8**(3), paper 154. Retrieved 18 March, 2004 from http://informationr.net/ir/8-3/paper154.html.
- Brooks, T.A. (2004), The nature of meaning in the age of Google. *Information Research*, **9**(3), paper 180. Retrieved 18th April, 2005 from http://informationr.net/ir/9-3/paper180.html
- *DayPop* (2004). *DayPop technology in detail*. Retrieved March 18, 2004 from http://www.DayPop.com/info/technology.htm
- Galitsky, B. & Levene, M. (2004). On the economy of Web links: simulating the exchange process. *First Monday*, **9**(1). Retrieved 18 March, 2004 from http://firstmonday.org/issues/issue9_1/galitsky/index.html.
- Gerhart, S.L. (2004). Do Web search engines suppress controversy? *First Monday*, **9**(1). Retrieved 18 March, 2004 from http://firstmonday.org/issues/issue9_1/gerhart/index.html.
- Glaser, M. (2003, September 3). *Google News* finally makes the grade, *Online Journalism Review*. Retrieved 18 March, 2004 from http://www.ojr.org/ojr/glaser/1062114819.php.
- *Google News* (2004). *A novel approach to news*. Retrieved 18 March, 2004 from http://news.google.com/intl/en_us/about_google_news.html.
- Hindman, M., Tsioutsiouliklis, K. & Johnson, J.A. (2003). 'Googlearchy': how a few heavily-linked sites dominate politics on the Web. Paper presented at the Annual Meeting of the Midwest Political Science Association, Chicago, IL., April 4, 2003. Retrieved March 18, 2004 from http://www.princeton.edu/~mhindman/googlearchy--hindman.pdf.
- Kramer, S.D. (2003, September 25). *Google News* creator watches portal quiet critics with 'Best News' webby, *Online Journalism Review*, Retrieved March 18, 2004 from http://www.ojr.org/ojr/kramer/1064449044.php.
- Lasica, J.D. (2004, September 24). Balancing act: how news sites serve up political stories. *Online Journalism Review*. Retreived 11 November, 2004 from http://www.ojr.org/ojr/technology/1095977436.php.
- Sullivan, D. (2003, September 10). News search engines. *Search Engine Watch*. Retrieved March 18, 2004 from http://searchenginewatch.com/links/article.php/2156261.

How to cite this paper:

Krishnamurthy, S. and Taniguchi, E. (2004) "The case of the news search engine: an exploratory empirical analysis of *Google News*" *Information Research*, **10**(4) paper 235 [Available at http://InformationR.net/ir/10-4/paper233.html]
