

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 analyze 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 US-centric: 73.24% of the sources are American. Google News is top-heavy. The top 5 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.

Implications: Our work adds to the recent literature on how search engines could potentially reduce diversity (Gerhart, 2003, Hindman, Tsioutsoulis, and Johnson, 2003).

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 favor 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, Tsioutsoulis, and Johnson, \(2003\)](#), have argued that search engines may *reduce* diversity. Gerhart(2003) 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, Tsioutsoulis, and Johnson, \(2003\)](#) 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- e.g. see 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 analyze these sources.

Exploratory findings are presented. While at least 93 countries are represented in Google News, the content is highly US-centric- 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 and placed on a Web site into sub-categories such as World, Business, Technology and Sports. Daypop places news information into these 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 Web site of the news search engine and peruse it just like any other news Web site. 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 30 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 these respects-

Such engines scan a limited number of sources. As an example, News.google.com scans 4,500 sources while Google.com indexes over 4 billion pages. At the time of writing, DayPop seems to look at the largest number of sources with 59,000- these include blogs, 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 per day.

The focus of news search engines is on a particular brand of content- i.e., news. They are, therefore, interested in not only a subset of the sources, but also a subset of the content of the Web. This is the familiar- "what exactly is news?"- question. There is certainly some murkiness around this.

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 15 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](#)). Many news search engines now offer e-mail alert services so that interested users can stay in touch with news stories in a particular category.

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. It also places stories in content categories (e.g. World, Sports, Business). In contrast, traditional search engines simply provide a dump of the search results with the most relevant results on top (this is still true of some news search engines).

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 Googlesmania](#). 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 10 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 Web site. 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 Web site. News.google.com wishes to create a news site as opposed to others who let you search within a few sources. Other news search engines do not group the results of the search to a large degree.

The increased interest in Google among scholars- e.g. [Brooks\(2004\)](#), [Galitsky and Levine\(2004\)](#), [Gerhart\(2003\)](#) and [Hindman, Tsioutsoulis, and Johnson, \(2003\)](#)- allows us to place our work in context.

Google News: a case study

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

news sources. [Google News](#) drew 3.4 million unique visitors in July 2003 ([Glaser, 2003](#)).

The Google News Web site describes itself in this way-

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 Web site 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 Web site. 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 site.

In addition to the main site, Google News runs five country editions- 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.

Methodology

As indicated earlier, Google closely guards its list of news sources for business reasons. Therefore, to gain access to the list of sources, we methodically visited the site and identified the name and Web site 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. 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 5 sources) account for 96.78% of sources. A total of 65 countries contribute 121 sources which account for 2.69% of all sources.

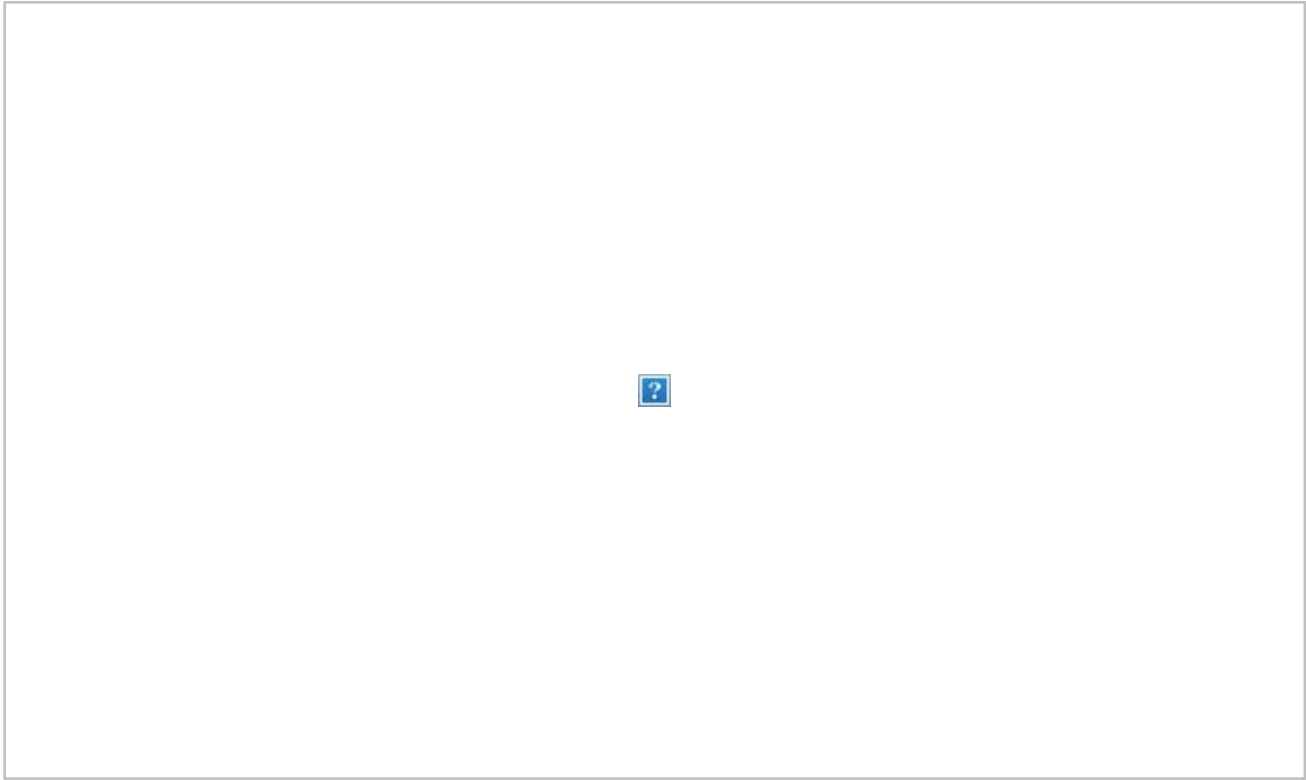


Table 1: Number of Sources, By Country

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 textbook power-law distribution.

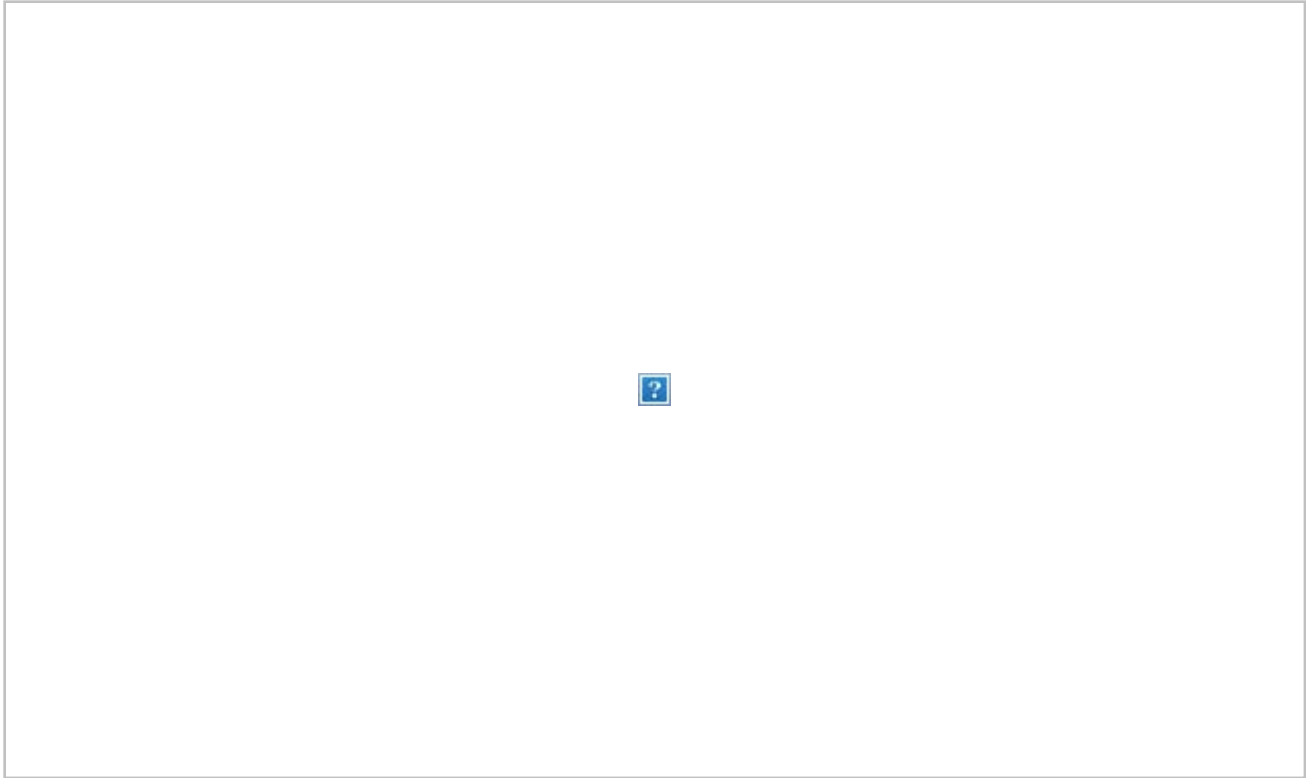


Figure 1: Power-law Distribution of Sources, By Country

We also classified sources based on their geographical region. These results are 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.

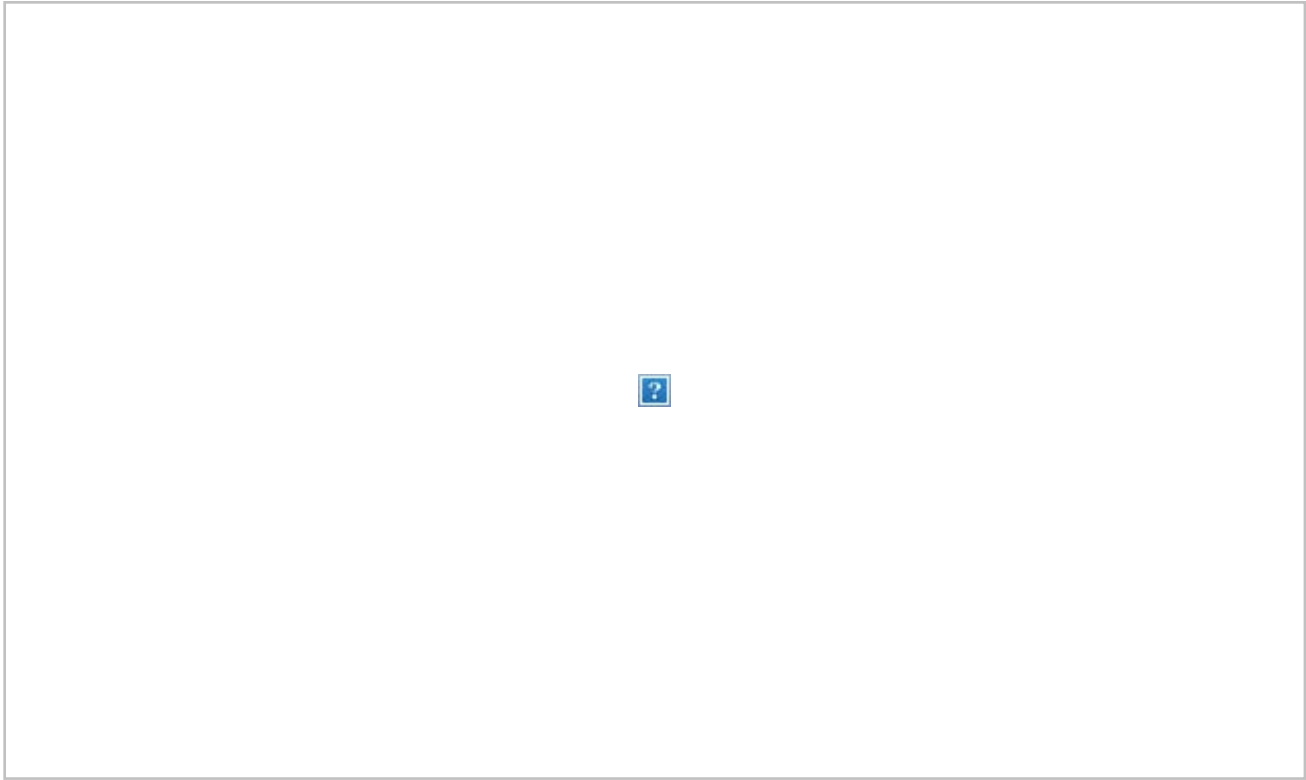


Table 2: Number of Sources, By Region

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.

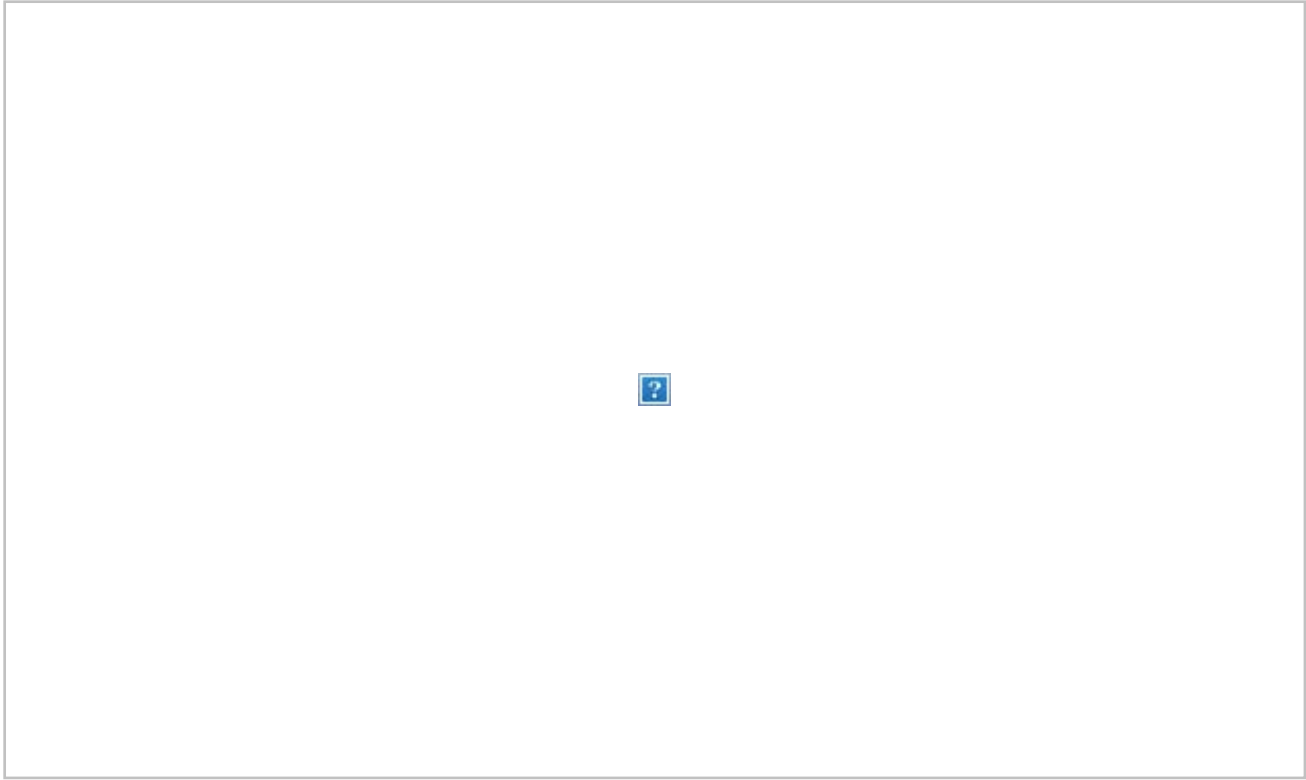


Table 3: Summary of Domain Extensions

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

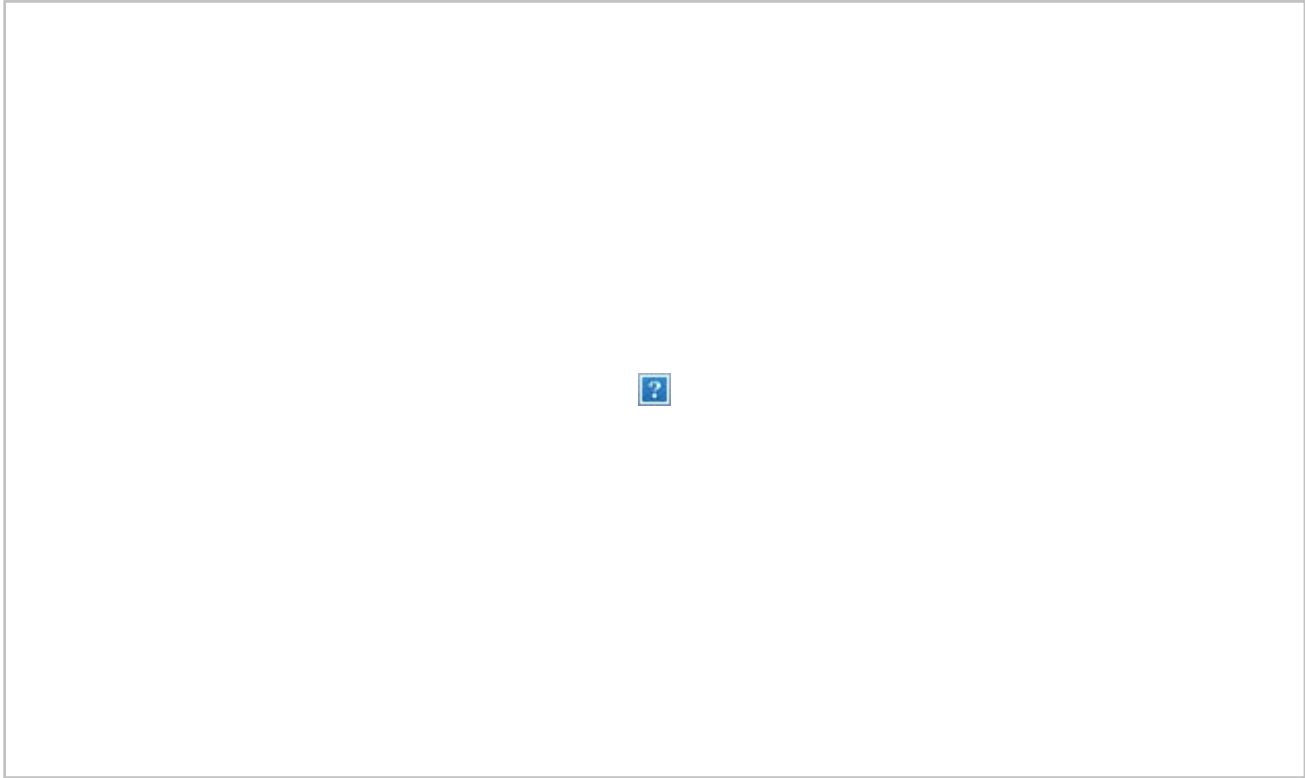


Table 4: Overview of Deep Linking Practices

State Newspaper analysis

To understand the proportion of the news universe captured by Google News, we procured a list of regional and local newspapers in one randomly picked state- Washington State. This list was obtained from the [Newspaper Association of America](#)'s Web site. A total of 46 newspapers were listed for [Washington State](#). Every newspaper in this list had a Web site and was hence, a potential candidate for Google News.

The results of our analysis is shown in Table 5. Only 41.30% of newspapers were listed on Google News. This was a surprise to us since we expected almost the entire list of newspapers to be represented on Google News. Initial analysis did not indicate any particular geographic trends.

--

Table 5: Washington State Newspapers

We conducted a similar analysis for another state- Kansas. We found that 52.78% of newspapers were included. Data for this is 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-centric and top-heavy. 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. However, this does not necessarily increase its objectivity. 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 Web sites in the US)? This may be a matter of some debate. 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. Examining the grouping practices adopted by Google News will be of great importance.

Acknowledgements

The author wishes to thank his student, Karen Tellevik, for her contributions.

References

- Adamic, L.(2000). [Zipf, power-laws, and Pareto-a ranking tutorial](http://www.hpl.hp.com/shl/papers/ranking/ranking.html). 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. (2004), [The nature of meaning in the age of Google](http://informationr.net/ir/9-3/paper180.html). *Information Research*, **9**(3), paper 180. Retrieved 18th April, 2005 from <http://informationr.net/ir/9-3/paper180.html>
- Brooks, T.A. (2003). [Web search: how the Web has changed information retrieval](http://informationr.net/ir/8-3/paper154.html). *Information Research*, **8**(3), paper 154. Retrieved 18 March, 2004 from <http://informationr.net/ir/8-3/paper154.html>.
- Daypop (2004). [Daypop technology in detail](http://www.daypop.com/info/technology.htm). 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](http://firstmonday.org/issues/issue9_1/galitsky/index.html). *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?](http://firstmonday.org/issues/issue9_1/gerhart/index.html) *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](http://www.ojr.org/ojr/glaser/1062114819.php), *Online Journalism Review*. Retrieved 18 March, 2004 from <http://www.ojr.org/ojr/glaser/1062114819.php>.
- Google News (2004). [A novel approach to news](http://news.google.com/intl/en_us/about_google_news.html). Retrieved 18 March, 2004 from http://news.google.com/intl/en_us/about_google_news.html.
- Hindman, M., Tsioutsoulis, K. & Johnson, J.A. (2003). ['Googlearchy': how a few heavily-linked sites dominate politics on the Web](http://www.princeton.edu/~mhindman/googlearchy--hindman.pdf). 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](http://www.ojr.org/ojr/kramer/1064449044.php), *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](http://www.ojr.org/ojr/technology/1095977436.php). *Online Journalism Review*. Retrieved 11 November, 2004 from <http://www.ojr.org/ojr/technology/1095977436.php>.
- Sullivan, D. (2003, September 10). [News search engines](http://searchenginewatch.com/links/article.php/2156261). *Search Engine Watch*. Retrieved March 18, 2004 from <http://searchenginewatch.com/links/article.php/2156261>.