Search Engine Predilection towards News Media Providers

Leif Azzopardi, Ciaran Owens
Dept. of Comp. Sci., University of Glasgow
Glasgow, United Kingdom
{leif, owensc}@dcs.gla.ac.uk

ABSTRACT

In this poster paper, we present a preliminary study on the predilection of web search engines towards various online news media provider sites using an access based measure.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval: Selection process

Keywords

Search Engine Bias

1. INTRODUCTION

Web search engines are the primary means for millions of users to access online content and as such, they influence much of the information that is consumed on the web [2]. The sites that search engines recommend, and the order in which they are recommended, dictate to a large extent, if not completely, what information is accessed during a user's online experience. This influence has raised many concerns regarding the impartiality of the results presented [2, 4, 5, 6, 7]. In other words, *are search engines biased?* This is a very controversial topic. On one hand, it has been argued that, since search engines are media companies then they will invariably make "editorial" decisions in order to tailor the content for their users [3]. While, this means certain biases will creep into the ranking, it is argued that this is necessary to satisfy the users. On the other hand, underlying biases may exist because of different reasons, e.g. due to commercial or political interests, a misconfiguration of the system, a design feature/flaw of the search algorithm, etc. For example, the use of PageRank has come under criticism because it leads to a "rich getting richer" scenario. Sometimes referred to as the "Googlearchy", where pages with more links receive a higher ranking than newer and less linked sites that are possibly more relevant [2, 5].

While several methods for detecting search engine bias have been already proposed [4, 5, 6, 7], they generally only consider one aspect of the bias: *the coverage of search engines* i.e. how much of the web a search engine covers and how different a search engine's results are from all other search engines. However, these methods require ground truth information which is not available[1]. Instead of prescribing a method which relies on an underlying assumption about what is "ground truth", we employ an access based measure which quantifies the predilection towards the different sites presented to a user in response to a set of queries [1]. This measure enables an intuitive comparison between the predilections of search engines towards particular websites. This preliminary work provides an important starting point and basis for such research, so that:

- search engines can be monitored, and
- users can be provided with information regarding the predilection/bias of search engines.

The remainder of this paper is as follows: in Section 2, we present an empirical study, where we investigate search engine preferences for different news media websites. Then in Section 3, we summarize the contributions of this poster paper and outline directions for future work.

2. EMPIRICAL STUDY AND RESULTS

The main aim of this study was to investigate whether major search engines exhibit any predilection towards particular news media provider websites, or not, and how this varies relatively to each other. In order to determine the predilections of search engines towards particular news media providers, we employed the following methodology which is an extension of the method proposed in [1]. Essentially, a large set of queries are submitted to each search engine, and the web sites of the results returned are recorded. Then a (weighted) count of the number of times a particular web site is retrieved is accumulated. Formally,

$$p(s) = \sum_{d \in s} \sum_{q \in Q} f(d, k) \quad (1)$$

where $p(s)$ is the predilection for site s, which is the sum of the weighted count $f(d, k)$ for all the documents retrieved from site s given the query set Q, where k is the rank of the document in the result list. For simplicity, we employ a cumulative based function with a cutoff of c, such that the function $f(d, k)$ evaluates to one if the rank k of d is less than c, otherwise zero [1]. Intuitively, $p(s)$ is the number of documents from site s that were retrieved in the top c documents given the query set Q.

Experimental Setup: For this study, we choose 4 US based (ABC, CNN, FOX, NY Times) and 5 UK based news

[1]Obtaining ground truth either relies upon a crawl of the entire web, or the assumption of what is "normal".

SIGIR'09, July 19–23, 2009, Boston, Massachusetts, USA.
ACM 978-1-60558-483-6/09/07.

Coverage	Google		MSNLive		Yahoo!	
# Uniq. Sites Ret.	9,377		3,051		3,395	
# Results Ret.	1,466,607		1,004,778		971,030	
Top Sites	Site	%	Site	%	Site	%
1	google.com	6.15	news.bbc.co.uk	3.89	guardian.co.uk	5.98
2	telegraph.co.uk	3.68	telegraph.co.uk	2.69	news.bbc.co.uk	3.30
3	news.bbc.co.uk	2.98	dailymail.co.uk	2.50	telegraph.co.uk	2.85
4	guardian.co.uk	2.43	nytimes.com	2.17	bbc.co.uk	2.56
5	nytimes.com	2.03	washingtonpost.com	2.06	startribune.com	2.49
6	dailymail.co.uk	1.86	independent.co.uk	1.64	timesonline.co.uk	2.09
7	youtube.com	1.86	miamiherald.com	1.42	independent.co.uk	1.93
8	timesonline.co.uk	1.55	lasvegassun.com	1.39	dailymail.co.uk	1.86
9	washingtonpost.com	1.49	iii.co.uk	1.34	blog.wired.com	1.86
10	independent.co.uk	1.25	charlotteobserver.com	1.32	news.cnet.com	1.69

Table 1: Coverage Statistics along with the top ten sites favored according to $p(s)$, and the percentage of times the site was retrieved in the top ten results.

Source / Engine	Google	MSNLive	Yahoo!
ABC	0.54	1.23	1.03
CNN	0.94	0.74	1.01
FOX	0.66	0.98	1.21
NYTIMES	2.03	0.00	2.17
BBC	2.98	2.56	3.89
Guardian	2.43	5.98	0.94
Reuters	0.76	1.41	0.00
The Times	2.13	3.05	1.07
Sky	0.57	0.30	0.00

Table 2: % of times each targeted provider was retrieved in the top ten results by each search engine.

media providers (BBC, Guardian, Reuters (UK), The Times, Sky (UK)), as our focus i.e. how do the different search engines treat these sites. From each site, we downloaded 18,000 stories from each media provider's RSS news feeds on business, entertainment, science, sport and world news. The title of each news story in the RSS feed was taken and used as a subsequent query. The queries were extracted over several months - producing approximately 162,000 queries in total. The same number of queries were generated from each site so that stories from one particular site was not queried more than another. Once the RSS feeds were downloaded, we issued the queries one day after they were obtained to three search engines: Google, MSNLive and Yahoo!. The international version of their search service was employed, and the top 10 results returned in results to each query was recorded (i.e $c = 10$). This experimental setup simulates the situation where a user hears about a story and would like to find the related news article from the particular news media provider. Thus, we would expect each of these targeted news providers to be recommended approximately the same number of times, unless their is some predilection by the search engines towards particular news media providers.

Experimental Results: Tables 1 and 2 shows the normalized $p(s)$ for $c = 10$ as a percentage for the top ten sites favored, and for the set of considered news media providers. Also, shown in Table 1 is the coverage statistics. First note that Google provides the greatest coverage. However, from the league table, we can see that Google tends to favor its own news service, and this is almost 2 to 3 times more often than most of the other top ten sites. For the other search engines, we can also see a clear tendency to favor particular sites. For example, Yahoo! tends to favor the Guardian's website. Also of note is in the top ten, few of the targeted news media sites (e.g. those sites the queries are drawn) are present. In Table 2, we show the targeted sites where it is easy to compare the differences between search engines and news media providers. Here, we see that some news media sites are retrieved substantially less often than others, in fact some of the targeted sites are never retrieved in the top ten results (i.e. NY Times by MSNLive and Sky by Yahoo!).

3. SUMMARY AND FUTURE WORK

In this work, we have shown that by employing a relatively simple sampling based methodology, it is possible to shown the relative retrieval tendencies, or predilections, of different search engines towards a set of websites. While we can make no concrete conclusions regarding the nature of the predilections (i.e. whether it is deliberate or incidental), the results from performing such a study are still very interesting and informative. In future work, we aim to examine how search engine predilection varies across topics, over time, and at different cutoffs of c, as well as examining other domains. We also intend to explore how such measures can be used to inform search engine regulators, the general public and search engine designers of the presence of any persistent bias within search engines.

4. REFERENCES

[1] L. Azzopardi and V. Vinay. Accessibility in information retrieval. In *Proceedings of the European Conference in Information Retrieval*, pages 482–489, 2008.

[2] J. Cho and S. Roy. Impact of search engines on page popularity. *ACM World Wide Web*, pages 20–29, 2004.

[3] E. Goldman. Search engine bias and the demise of search engine utopianism. *Yale Journal of Law & Technology*, pages 188–200, 2005–2006.

[4] E.-P. L. Hady Lauw and K. Wang. Bias and controversy: Beyond the statistical deviation. *ACM SIGKDD Knowledge Discovery and Data-Mining*, pages 625–630, 2006.

[5] S. Lawrence and C. L. Giles. Accessibility of information on the web. *Nature*, 11:32–39, 2000.

[6] A. Mowshowitz and A. Kawaguchi. Assessing bias in search engines. *Information Processing & Management*, 38(1):141–156, 2002.

[7] L. Vaughan and M. Thelwall. Search engine coverage bias: Evidence and possible causes. *Information Processing and Management*, 40:693–707, 2004.