Analyzing the Content Emphasis of Web Search Engines

Mohammed A. Alam
Northwestern University
2133 Sheridan Road
Evanston, IL 60208, USA
mohammed.alam@northwestern.edu

Doug Downey
Northwestern University
2133 Sheridan Road
Evanston, IL 60208, USA
ddowney@eecs.northwestern.edu

ABSTRACT

Millions of people search the Web each day. As a consequence, the ranking algorithms employed by Web search engines have a profound influence on which pages users visit. Characterizing this influence, and informing users when different engines favor certain sites or points of view, enables more transparent access to the Web's information.

We present PAWS, a platform for analyzing differences among Web search engines. PAWS measures *content emphasis*: the degree to which differences across search engines' rankings correlate with features of the ranked content, including point of view (e.g., positive or negative orientation toward their company's products) and advertisements. We propose an approach for identifying the orientations in search results at scale, through a novel technique that minimizes the expected number of human judgments required. We apply PAWS to news search on Google and Bing, and find no evidence that the engines emphasize results that express positive orientation toward the engine company's products. We do find that the engines emphasize particular news sites, and that they also favor pages containing their company's advertisements, as opposed to competitor advertisements.

Categories and Subject Descriptors

H.3.3 [**Info. Search and Retrieval**]: Search process

Keywords

Web search engine; search engine bias

1. INTRODUCTION

Web search engines are invaluable tools. Many Web sessions begin with a query to a search engine [10]. Since users are more likely to visit high-ranked URLs [1], search engines influence where users shop and which views disseminate.

English-language search results are delivered today by two primary vendors: Google and Microsoft. Each is a complex business, selling a variety of products and services outside of search. This has led to concern that their search engines may manipulate their result rankings to acquire a competitive advantage or propagate a viewpoint. For example, recent work has shown that engines rank links to their own services (such as e-mail or maps) more highly than links to competing services [6]. Legal scholars have debated whether search engines should be regulated to ensure "neutrality" [3, 7].

In this paper, we do not argue that engines should be free of editorial bias. Our goal is instead to develop methods to *measure* the differences between engine rankings, and then provide these measurements to end users. We present PAWS, a Platform for Analyzing Web Search engines. PAWS measures *content emphasis*, the degree to which differences across search engines' rankings correlate with features of the ranked content. While measuring search engine bias has become a popular task [2, 4, 5, 8, 9], to our knowledge PAWS is the first system to investigate how relative rankings correlate with important attributes of content including orientation (do search engines favor positive news about their company's products?) and advertisement (do search engines drive traffic to their company's sponsored links?).

We describe how PAWS gathers search engine results and analyzes search engines for content emphasis. PAWS collects results each day, and the data is released to the research community via the PAWS Website.[1] A key challenge faced by PAWS is identifying the orientations of result URLs at scale. To this end, we present a new technique for manually ranking results by orientation that minimizes the expected number of human judgments required. We then present PAWS's analysis of content emphasis in news search on Google and Bing.

2. PAWS

PAWS aims to measure how a search engine's rankings correlate with features of the ranked content. As other researchers have observed, there is no "control" engine available to provide a gold standard ranking [2, 9]. Thus, PAWS measures *relative* differences across two primary providers of algorithmic search results today, Google and Bing. PAWS does not explain *why* the differences arise (or which engine is "responsible").

For each pair (q, u) where result URL u is returned by an engine for query q, PAWS calculates a score that indicates whether u tends to be ranked higher for q by Google than by Bing. We refer to the score as $GB(q, u)$, for Google-Bing

[1] http://websail.eecs.northwestern.edu/projects/paws

SIGIR'14, July 6–11, 2014, Gold Coast, Queensland, Australia.

ACM 978-1-4503-2257-7/14/07 ...$15.00.
http://dx.doi.org/10.1145/2600428.2609515.

score. More negative values indicate that Google ranks a result more highly than Bing.

Because the majority of search result clicks occur on the first page of results [1], we consider only these results in our experiments. As search results for a query may change over time, we retrieve results for each query once a day.

Formally, let $r(d, q, u, e)$ indicate the numeric ranking of each URL u returned on the first page of results for query q on engine e on day d. For URLs u not returned for a given d, q, e, we let $r(d, q, u, e) = \tau$ for a constant τ. Then GB is defined as:

$$GB(q, u) = \sum_{d \in D} r(d, q, u, \text{Google}) - r(d, q, u, \text{Bing}) \quad (1)$$

where the sum is computed over data set D of days d, with each query performed once on both engines each day. GB is computed over only "algorithmic" results, ignoring advertising links on the result page. The constant τ allows GB to account for results returned on the first page by one engine but not the other. In our experiments, we set $\tau = 20$, although altering the parameter by 50% in either direction has negligible impact on our results. In fact, the correlation between the GB scores with $\tau = 20$ and either $\tau = 15$ or $\tau = 25$ is greater than 0.99.

PAWS measures content emphasis by computing the correlation between $GB(q, u)$ and features of the result u. Some features of interest – such as the site u originates from, or whether u contains ads sponsored by the search engine – are relatively straightforward to identify at scale using automated means. However, an additional goal of PAWS is to measure how *orientations* in results vary with GB. Below, we discuss why this task is challenging, and present the novel methods PAWS utilizes to perform the task.

2.1 Orientation Acquisition in PAWS

PAWS attempts to measure if $GB(q, u)$ correlates with positive or negative orientation of document u toward query concept q. For example, we may ask PAWS if an engine is more likely to show documents reporting good news about a political party or expressing negative views about a product.

Given the large size of the document sets we wish to analyze, automated techniques for detecting orientation would be desirable. Although a variety of related work has been performed on automatic sentiment classification, our task is particularly challenging because a document's orientation toward a product may be buried in a single sentence that differs from the rest of the document's orientation, and sometimes obtaining the orientation requires world knowledge. Using a state-of-the-art sentiment analyzer,[2] we obtained only -0.07 correlation with human ratings on our data sets.

For manual acquisition of orientation labels, crowdsourcing on platforms such as Amazon Mechanical Turk (AMT) is a typical approach. However, our controlled experiments show that AMT Workers have difficulty with the task. The responses are of low accuracy even when we ask questions redundantly or restrict to the highest-rated Workers.

Due to the above challenges, PAWS collects pairwise orientation judgments from expert labelers – in our experiments, the first author of this paper. We validated their labels by computing self-agreement and inter-annotator agreement with the second author on 40 rankings of results for two queries, i.e., a total of 380 pairwise comparisons. The Kappa score for self-agreement was 0.617 and for inter-annotator agreement, 0.385. The scores are dramatically better than our sentiment analyzer and AMT baselines, and are considered "fair" agreement, which we believe to be adequate given the subjective nature of our task. The pairwise judgment approach allows for ties in orientation, and produces a partial order of the documents for each query q, avoiding the difficulties of defining a fixed orientation scale.

[2]http://www.alchemyapi.com/products/features/sentiment-analysis/

2.2 Efficient Ranking by Pairwise Judgments

Because the expert judgments that PAWS requires are expensive, we develop a novel approach that ranks documents by orientation while minimizing the expected number of manual judgments. While previous work has considered production of *total* orders from pairwise comparisons (e.g., [11]), these are ill-suited to PAWS because orientations are often indistinguishable. To our knowledge, ours is the first approach for minimizing the expected number of manual judgments needed to produce a partial order.

Formally, we consider placing a new document d at the proper position within a (perhaps empty) partial order O of other documents. Inserting d requires iteratively comparing it to a selected element $i \in O$. If d is of the same orientation as i, the search terminates; otherwise, the search continues in a smaller portion of O, depending on whether d is deemed more positive or more negative than i. Without ties in O, Binary Search is optimal for the insertion task. However, with ties we can sometimes expedite the search by checking larger (i.e., more probable) portions of the partial order first. Our efficient algorithm exploits this intuition.

Let $E_O[J|LB, UB, i]$ indicate the minimum expected number of judgments needed to place a document d within O if we compare first with $i \in O$, given that the position of d is known to lie between lower bound LB and upper bound UB, inclusive. The expression E_O can be decomposed into a sum over the possible outcomes of the comparison of d to i. We compare d to i (one judgment), and if the two are equal in orientation, no additional judgments are required. If they are unequal, we add the minimum expected number of additional judgments required (in terms of E_O), weighted by the probability of each outcome. Thus, E_O can be expressed recursively as:

$$\begin{aligned} E_O[J|LB, UB, i] = 1+ \\ P(\{LB, .., i-1\}) \min_j E_O[J|LB, i-1, j]+ \\ P(\{i+1, .., UB\}) \min_j E_O[J|i+1, UB, j] \end{aligned} \quad (2)$$

where $P(S)$ is the probability that the query document d belongs within $S \subseteq O$ in the partial order. In our implementation, we approximate these probabilities using the distribution of documents in O.

At any step of the insertion, computing the comparison element i that minimizes the expected number of judgments (assuming correct responses) is straightforward using dynamic programming and Equation 2. We experimentally evaluate our approach, denoted as Min$E[J]$, utilizing several random orderings of the documents we hand-rank in our experiments (see Section 3). The results are shown in Table 1. We allow a variable error rate, where the comparison of documents d_i, d_j belonging to $i, j \in O$ is modeled as a numeric random variable z selected from the density

		Error Rate			
		0	0.25	0.50	1
Judgments	Binary Search	40.60	41.20	42.73	44.60
	MinE[J]	37.13	38.00	40.07	42.80
Accuracy	Binary Search	1.00	0.99	0.98	0.92
	MinE[J]	1.00	1.00	0.98	0.93

Table 1: Pairwise judgments and algorithm accuracy. Min$E[J]$ requires 6.58% fewer judgments on average than Binary Search and is slightly more accurate for all error rates.

$P(z) \propto e^{-|z-i+j|/\sigma}$, where $z > 1/2$ indicates a $d_i > d_j$ response, $z < 1/2$ indicates $d_i < d_j$, otherwise $d_i = d_j$. So, larger errors are less likely than smaller ones, and the error rate increases with the parameter $\sigma \geq 0$, with $\sigma = 0$ indicating perfect responses. The results show that Min$E[J]$ reduces the average number of judgments required by 6.58% compared to Binary Search. We also find that Min$E[J]$ is slightly more accurate (Table 1).

2.3 Data Acquisition

We focus on data from *news* search results. News search is an ideal target for analyzing content emphasis, as the results change frequently and often exhibit orientation toward a concept (e.g., good or bad news, reviews, editorials, etc.). Further, for the query terms we target (described below), news links are often returned prominently even on the primary "Web" search pages of Google and Bing.

The data comprises search results for a total of 165 queries. 34 of these are manually selected, chosen to include controversial queries (e.g., religious and political terms) as well as names of popular products, including several products of the engine companies themselves. We additionally collect results for daily trending queries or "Hot Searches," as reported by Google Trends. The results are collected from both Google and Bing as {header, URL, snippet} triples over timeframe, $T = 138$ days, resulting in 51,634 unique result URLs. Additionally, HTML source code is collected for every Webpage linked to by those URLs. All the collected data is released to the search community (see Section 1).

3. EXPERIMENTS

We present our experimental results, using PAWS to investigate three aspects of result content: orientation toward the engine company's products, presence of the company's advertisements ("ads"), and the site linked to by the result.

3.1 Orientation

We perform orientation measurement on search results for 11 manually selected product-name queries. From the results for each query, we select 20 results to rank by orientation. We do the selection in two ways. In the first, Uniform GB, we select 20 search results that are approximately uniformly spaced in the set of all search results. In the second, Extreme GB, we select 10 search results from each of the two ends of the set (i.e., the results most skewed toward being returned by one engine rather than the other). The intuition behind the second set is that the extremal documents are more likely to reflect content emphasis. We rank each set using Min$E[J]$ as described in the previous section.

The results are shown in Table 2. While the numbers vary between the two result sets, in neither case does GB show that the engines emphasize positive orientations toward their company's own products. The fact that the average correlations are negative across all queries indicates that Google slightly emphasizes negative results in general, on this data set.

Query	Extreme GB	Uniform GB
android	-0.53	-0.47
macbook	-0.34	0.1
nexus 7	-0.3	0.16
microsoft office	-0.29	0.11
xbox	-0.15	-0.33
lumia	-0.03	0.36
kinect	-0.01	-0.01
windows 8	0.1	-0.06
chrome	0.3	-0.17
microsoft surface	0.33	-0.32
gmail	0.46	0.21
Avg. Google Products	-0.02	-0.07
Avg. Microsoft Products	-0.01	-0.05

Table 2: Spearman correlation between GB and orientation rank for product queries. Positive values indicate Google's results favor more positive orientations toward the query. On average, we find no significant evidence of the engines' emphasizing positive orientations toward their company's products.

	Google Ads	Microsoft Ads	Facebook Ads	Other Ads
Average over q	-0.02	0.06	0.05	0.01
(Std. Dev.)	(0.08)	(0.04)	(0.06)	(0.06)
Combined	-0.01	0.05	0.06	0.01

Table 3: Spearman correlation of GB with the presence of ads by the given company. "Average over q" lists the average of 34 correlation values, one for each query. "Combined" lists the correlation when the results from all 34 queries are combined into a single set. When compared against each other, Google and Bing favor content containing their company's own ads, rather than competing ads. The difference between the combined correlation coefficient for Google (-0.01) and that of Microsoft (0.05) and Facebook (0.06) is significant at the $p < 0.001$ level (Fischer r-to-z transformation).

3.2 Advertisements

We investigate whether the presence of ads in a document linked to by a result URL for an engine or one of its competitors influences the URL's position in search results. Engines have a commercial interest in increasing traffic to their ads, which makes ads an important content feature to analyze.

We define *ads* broadly to include not only online text and display advertisements, e.g., Google AdSense and Bing Ads, but also links to the search engine's products and services, e.g., YouTube, Google+, etc. To identify ads in each result in our data set, we manually construct regular expressions for ads by the two major publishers (Google and Microsoft, and their third-party affiliates), and some other companies. We also identify the presence of Facebook *Like* buttons.

We analyze the Spearman correlation between GB and a binary variable indicating the presence of a given company's ads. Table 3 shows the results.[3] We see that the engines

[3]Full results table available on the PAWS Website (see Section 1)

Host	Num. Results	Average Norm. GB	Std. Err.
wnd.com	60	0.22	0.02
joystiq.com	117	0.23	0.02
ign.com	202	0.24	0.02
nationalreview.com	76	0.27	0.02
nbcnews.com	179	0.27	0.02
theverge.com	243	0.27	0.01
hollywoodlife.com	198	0.28	0.01
polygon.com	80	0.28	0.02
slate.com	110	0.28	0.02
siliconvalley.com	110	0.29	0.01
...	...	...	...
societyandreligion.com	56	0.69	0.02
cnn.com	851	0.69	0.01
businessweek.com	328	0.70	0.01
upi.com	124	0.72	0.02
itechpost.com	84	0.72	0.02
i4u.com	57	0.74	0.02
ap.org	54	0.74	0.02
msn.com	333	0.75	0.01
betanews.com	95	0.76	0.02
softpedia.com	371	0.77	0.01

Table 4: Sites emphasis. Google tends to favor smaller news outlets while Bing favors bigger ones.

rank a page significantly higher, relatively speaking, when it contains the engine company's ads, as opposed to competitor ads. Compared to Google, Bing also favors content with Facebook "Like" buttons. The content emphasis on ads seen in this experiment, while not large, may have a non-trivial impact when aggregated over billions of yearly searches.

3.3 Sites

In this section, we describe measurements of news search emphasis across different hosts. We show that GB is, in fact, significantly non-uniform for a large number of hosts, indicating that the two engines often prefer different hosts.

Starting with our complete results data set, we normalize the host names, retaining the suffix. We find 2,990 unique hosts. We focus our analysis on frequent hosts, i.e., those with at least 50 distinct search results in the data.

Of the 150 frequent hosts, 31 have an average GB below 0.35, and 15 have an average GB above 0.65. For frequent hosts, an average GB falling outside of the range [0.35, 0.65] is statistically significant ($p < 0.005$, Monte Carlo simulation).

We see that of the 150 frequent hosts, the 46 (or 31%) discussed above exhibit significantly different ranking behavior in Google than in Bing. Table 4 lists the 20 hosts at the extremes.[3] We see that Google tends to give relatively higher rank to smaller news sites that may be specialized or politically opinionated (whether conservative or liberal), e.g., wnd.com and slate.com. By contrast, Bing ranks larger US media outlets with a wider print and television news presence relatively higher, e.g., ap.com and cnn.com. Also, Microsoft content (msn.com) and that of its search engine partner (yahoo.com) are shown to rank relatively higher in Bing than in Google in this experiment.

4. CONCLUSION

We presented PAWS, a platform for analyzing the content emphasis of Web search engines. We introduced a novel method for obtaining the orientation rankings utilized in PAWS with the minimum expected number of pairwise judgments. We summarized our experimental findings with PAWS, showing no significant emphasis across the target engines on positive orientations toward their company's products. We did find that the engines ranked results with their company's advertisements relatively higher, as opposed to those with competitor advertisements. In future work, we will investigate automated content analysis methods, and make summaries of PAWS's measurements publicly accessible to Web users.

5. ACKNOWLEDGMENTS

This work was supported in part by NSF award CNS-1064595 and DARPA contract D11AP00268.

6. REFERENCES

[1] E. Agichtein, E. Brill, S. Dumais, and R. Ragno. Learning user interaction models for predicting web search result preferences. In *Proceedings of the 29th Annual Int'l ACM SIGIR Conf. on R&D in IR*, SIGIR '06, pages 3–10, New York, NY, USA, 2006. ACM.

[2] L. Azzopardi and C. Owens. Search engine predilection towards news media providers. In *Proceedings of the 32nd Int'l ACM SIGIR Conf. on R&D in IR*, SIGIR '09, pages 774–775, New York, NY, USA, 2009. ACM.

[3] O. Bracha and F. Pasquale. Federal search commission? access, fairness, and accountability on the law of search. *Cornell L. Rev.*, 93:1149, 2008.

[4] S. Chelaru, I. S. Altingovde, S. Siersdorfer, and W. Nejdl. Analyzing, detecting, and exploiting sentiment in web queries. *ACM Trans. Web*, 8(1):6:1–6:28, Dec. 2013.

[5] G. Demartini and S. Siersdorfer. Dear search engine: What's your opinion about...?: Sentiment analysis for semantic enrichment of web search results. In *Proceedings of the 3rd Int'l Semantic Search Workshop*, SEMSEARCH '10, pages 4:1–4:7, New York, NY, USA, 2010. ACM.

[6] B. Edelman and B. Lockwood. Measuring bias in "organic" web search. *Unpublished manuscript*, 2011.

[7] E. Goldman. Search engine bias and the demise of search engine utopianism. In A. Spink and M. Zimmer, editors, *Web Search*, volume 14 of *Info. Science and Knowledge Management*, pages 121–133. Springer Berlin Heidelberg, 2008.

[8] A. Mowshowitz and A. Kawaguchi. Bias on the web. *Commun. ACM*, 45(9):56–60, Sept. 2002.

[9] A. Mowshowitz and A. Kawaguchi. Measuring search engine bias. *Inf. Process. Manage.*, 41(5):1193–1205, Sept. 2005.

[10] C. Spellman. Research from groupm search and comscore sheds light on the role search and social media play in the consumer's path to purchase, February 2011. Retrieved May 10, 2014 from http://groupmnext.com/press-releases/research-from-groupm-search-and-comscore-sheds-light-on-the-role-search-and-social-media-play-in-the-consumer?s-path-to-purchase/?s-path-to-purchase/.

[11] F. Wauthier, M. Jordan, and N. Jojic. Efficient ranking from pairwise comparisons. In *Proceedings of the 30th Int'l Conf. on ML*, pages 109–117, 2013.