How Diverse are Web Search Results?

Rodrygo L. T. Santos
rodrygo@dcs.gla.ac.uk

Craig Macdonald
craigm@dcs.gla.ac.uk

Iadh Ounis
ounis@dcs.gla.ac.uk

School of Computing Science
University of Glasgow
G12 8QQ Glasgow, UK

ABSTRACT

Search result diversification has recently gained attention as a means to tackle ambiguous queries. While query ambiguity is of particular concern for the short queries commonly observed in a Web search scenario, it is unclear how much diversity is actually promoted by Web search engines (WSEs). In this paper, we assess the diversification performance of two leading WSEs in the context of the diversity task of the TREC 2009 and 2010 Web tracks. Our results show that these WSEs perform effectively for queries with multiple interpretations, but not for those open to multiple aspects related to a single interpretation. Moreover, by deploying a state-of-the-art diversification approach based on query suggestions from these WSEs themselves, we show that their diversification performance can be further improved.

Categories and Subject Descriptors: H.3.3 [Information Storage & Retrieval]: Information Search & Retrieval

General Terms: Experimentation, Performance

Keywords: Web Search, Search Engines, Diversity

1. INTRODUCTION

Queries submitted to a Web search engine (WSE) are typically short and often ambiguous [5]. In such a scenario, an effective approach is to diversify the search results. By doing so, the probability of satisfying potentially different information needs underlying an ambiguous query is maximised [3]. Although search result diversification is a naturally motivated approach in a Web search scenario, it is unclear how much diversity is actually promoted by existing WSEs.

In this paper, we assess the effectiveness of two leading WSEs in the context of the diversity task of the TREC 2009 and 2010 Web tracks [1, 2]. By breaking down this assessment over queries with different levels of ambiguity, we show that these WSEs perform effectively for queries open to multiple interpretations, but not for those open to multiple aspects of a single interpretation. In the latter case, we show that both WSEs can be further improved by leveraging evidence readily available to them—i.e., query suggestions—within a state-of-the-art diversification approach.

In the rest of this paper, Section 2 describes our methodology and the experimental setup that supports our investigations in Section 3. Our conclusions follow in Section 4.

SIGIR'11, July 24–28, 2011, Beijing, China.
ACM 978-1-4503-0757-4/11/07.

2. EXPERIMENTAL METHODOLOGY

Our investigations aim to answer two research questions:

1. How diverse are WSEs' search results?
2. Can we improve WSEs' diversification performance?

We investigate both questions in the context of the diversity task of the TREC 2009 and 2010 Web tracks [1, 2]. These tasks provide a total of 98 queries (50 from 2009, 48 from 2010), each classified by TREC assessors as either *ambiguous* (i.e., open to different *interpretations*) or *underspecified* (i.e., open to different *aspects* of a single interpretation) [3]. Moreover, each query comprises relevance assessments for 1 to 6 sub-topics, identified by TREC assessors as representing different interpretations or aspects of the query. Besides reporting average performance figures over all 98 queries, we break down our evaluation for queries with different types (ambiguous vs. underspecified) as well as with different number of sub-topics (1 to 6).

As WSEs, we consider Bing and Google, henceforth referred to as WSEs A and B, in no particular order, so as to preserve anonymity. In particular, for each of the considered 98 queries, we obtain up to 1000 results from these WSEs using their public APIs.[1] The URLs retrieved by each WSE are then normalised and matched against those in the TREC ClueWeb09 corpus. While this procedure is necessary to enable the reuse of the TREC Web track relevance assessments, the obtained rankings should be seen as a 'lower bound' of what these WSEs could have produced if they constrained themselves to the ClueWeb09 corpus. To provide an ample analysis, we consider both ClueWeb09 A, with 500 million English Web pages, and its first-tier subset, ClueWeb09 B, which comprises 50 million pages.

To investigate whether there is room for improving the diversification performance of these WSEs, we deploy a state-of-the-art diversification framework to re-rank their retrieved results. In particular, the xQuAD framework [4]—one of the top performers at the diversity task of the TREC 2009 and 2010 Web tracks [1, 2]—builds upon an explicit representation of the aspects underlying an ambiguous query. Following Santos et al. [4], we re-rank the results retrieved by each WSE using xQuAD, with query suggestions—provided by the WSE itself—representing different query aspects. As xQuAD requires an estimation of the relevance of each result for every query aspect, we score all the results retrieved for a query with respect to each query suggestion using BM25.

[1] Requests were sent anonymously to the US version of the WSEs, so as to isolate any customisation or localisation effects. All results were retrieved on February 7th, 2011.

				type		number of sub-topics					
			all	amb.	und.	1	2	3	4	5	6
	# queries		98	38	60	2	4	27	31	20	14
ClueWeb09 A	ERR-IA	WSE A	0.2712	$\underline{0.2910}$	0.2586	0.0902	0.1828	0.3574	0.2643	0.2358	0.2217
		+xQuAD	$0.2856^{=}$	$0.2647^{=}$	$0.2988^{\blacktriangle}$	$0.0515^{=}$	$0.2127^{=}$	$0.3584^{=}$	$0.2825^{=}$	$0.2266^{=}$	$0.2905^{=}$
		WSE B	$0.2997^{\triangle}$	$0.2867^{=}$	$0.3080^{\triangle}$	$\underline{0.1803}^{=}$	$0.3204^{=}$	$0.3780^{=}$	$\underline{0.2889}^{=}$	$\underline{0.2595}^{=}$	$0.2412^{=}$
		+xQuAD	$\underline{0.3098}^{\triangle}$	$0.2878^{=}$	$\underline{0.3237}^{=}$	$0.0515^{=}$	$\underline{0.3983}^{=}$	$\underline{0.3971}^{\blacktriangle}$	$0.2807^{=}$	$0.2561^{=}$	$\underline{0.2943}^{\triangle}$
	α-nDCG	WSE A	0.3957	0.4235	0.3782	0.1637	0.2955	0.4971	0.3888	0.3386	0.3590
		+xQuAD	$0.4096^{=}$	$0.3920^{=}$	$0.4208^{\blacktriangle}$	$0.1267^{=}$	$0.3123^{=}$	$0.4915^{=}$	$\underline{0.4131}^{=}$	$0.3407^{=}$	$0.4107^{=}$
		WSE B	$0.4239^{=}$	$0.4195^{=}$	$0.4267^{\triangle}$	$\underline{0.2398}^{=}$	$0.3983^{=}$	$0.5214^{=}$	$0.4038^{=}$	$0.3721^{=}$	$0.3881^{=}$
		+xQuAD	$\underline{0.4402}^{\triangle}$	$\underline{0.4293}^{=}$	$\underline{0.4471}^{=}$	$0.1267^{=}$	$\underline{0.4796}^{=}$	$\underline{0.5412}^{\triangle}$	$0.4062^{=}$	$\underline{0.3792}^{=}$	$\underline{0.4411}^{\blacktriangle}$
ClueWeb09 B	ERR-IA	WSE A	0.2776	0.2927	0.2681	0.1202	0.2258	0.3259	0.2905	0.2451	0.2398
		+xQuAD	$0.3114^{\blacktriangle}$	$0.2859^{=}$	$0.3275^{\blacktriangle}$	$0.0721^{=}$	$0.2014^{=}$	$0.3541^{=}$	$\underline{0.3303}^{=}$	$0.2663^{=}$	$0.3173^{\triangle}$
		WSE B	$0.3075^{\blacktriangle}$	$\underline{0.3106}^{\triangle}$	$0.3055^{=}$	$\underline{0.1803}^{=}$	$0.2546^{=}$	$0.3721^{\triangle}$	$0.3095^{=}$	$0.2773^{\triangle}$	$0.2547^{=}$
		+xQuAD	$\underline{0.3326}^{\blacktriangle}$	$0.3027^{=}$	$\underline{0.3515}^{\blacktriangle}$	$0.0902^{=}$	$\underline{0.3784}^{=}$	$\underline{0.3917}^{\triangle}$	$0.3084^{=}$	$\underline{0.3012}^{=}$	$\underline{0.3381}^{\blacktriangle}$
	α-nDCG	WSE A	0.3977	0.4264	0.3795	0.1900	0.3023	0.4549	0.4129	0.3529	0.3746
		+xQuAD	$0.4311^{\blacktriangle}$	$0.4204^{=}$	$0.4379^{\blacktriangle}$	$0.1470^{=}$	$0.2935^{=}$	$0.4767^{=}$	$\underline{0.4549}^{=}$	$0.3797^{\triangle}$	$0.4439^{=}$
		WSE B	$0.4309^{\triangle}$	$\underline{0.4619}^{\triangle}$	$0.4113^{=}$	$\underline{0.2398}^{=}$	$0.3616^{=}$	$0.5005^{=}$	$0.4224^{=}$	$0.4037^{\triangle}$	$0.4016^{=}$
		+xQuAD	$\underline{0.4557}^{\blacktriangle}$	$0.4560^{=}$	$\underline{0.4556}^{\blacktriangle}$	$0.1637^{=}$	$\underline{0.4766}^{=}$	$\underline{0.5119}^{=}$	$0.4287^{=}$	$\underline{0.4286}^{=}$	$\underline{0.4819}^{\blacktriangle}$

Table 1: Diversification performance (@20) of WSEs across queries of different type or number of sub-topics.

3. EXPERIMENTAL RESULTS

Clarke et al. [1, 2] have shown that rankings produced by a WSE outperform all systems in the diversity task of the TREC 2009 and 2010 Web tracks. While this result suggests that WSEs indeed seek to diversify their search results, it is unclear how much diversity they actually promote under different scenarios. To investigate this, Table 1 shows the diversification performance of WSEs A and B in contrast to each other and as input to the xQuAD diversification framework. In particular, we break down this evaluation across two orthogonal dimensions: document corpora and query ambiguity level. The first dimension covers a large sample of the Web (ClueWeb09 A) and a high-quality, first-tier subset of this sample (ClueWeb09 B). The second dimension groups all 98 available queries from the TREC 2009 and 2010 Web tracks according to their type (ambiguous or underspecified) or their number of sub-topics. Diversification performance is given by ERR-IA and α-nDCG, both at rank cutoff 20. Significance is verified for WSE B compared to WSE A, as well as for xQuAD compared to WSEs A and B, according to the Wilcoxon signed-rank test. In particular, the symbols $\blacktriangle$ ($\blacktriangledown$) and $\triangle$ ($\triangledown$) denote a significant increase (decrease) at the $p < 0.01$ and $p < 0.05$ levels, respectively, while the symbol $=$ denotes no significant difference.

Focusing on the ClueWeb09 A results, we note that WSE B outperforms WSE A on the whole set of 98 queries (significantly for ERR-IA). Looking at the breakdown by type, we observe that the higher performance of WSE A comes mostly from underspecified queries, for which it significantly outperforms WSE B. Indeed, when ambiguous queries are considered, the two WSEs do not differ significantly. In terms of the number of sub-topics underlying a query, WSE B consistently outperforms WSE A regardless of the number of sub-topics, although not significantly. When xQuAD is deployed on top of these WSEs (using each WSE's own query suggestions), no significant difference is observed for ambiguous queries. Regarding our first question, this suggests that Web search results are highly diverse, at least for ambiguous queries. For underspecified queries, however, both WSEs can be markedly improved (significantly for WSE A) with xQuAD. Regarding our second question, this shows that there is room for improving these WSEs' diversification performance, by leveraging their own query suggestions.

The results for ClueWeb09 B confirm these observations. In particular, WSE B significantly outperforms WSE A on the entire query set. Besides still markedly outperforming WSE A for underspecified queries, we observe that WSE B is now also significantly superior for ambiguous queries. Once again, the number of sub-topics does not play a significant role in explaining the relative performance of these two WSEs, with WSE B consistently outperforming WSE A. Lastly, we once more observe that xQuAD significantly improves both WSEs, with gains coming from underspecified queries, but not from ambiguous queries.

4. CONCLUSIONS

We have thoroughly investigated the diversification performance of two leading WSEs in the context of the diversity task of the TREC 2009 and 2010 Web tracks. Our results expand on initial observations from the literature by analysing how diverse Web search results are for queries with different ambiguity levels. In particular, we have shown that the considered WSEs are already very effective at diversifying the results for ambiguous queries. However, when underspecified queries are considered, we have shown that there is room for significantly improving these WSEs' performance by leveraging their own query suggestions.

5. REFERENCES

[1] C. L. A. Clarke, N. Craswell, and I. Soboroff. Overview of the TREC 2009 Web track. In *TREC*, 2009.

[2] C. L. A. Clarke, N. Craswell, I. Soboroff, and G. V. Cormack. Preliminary overview of the TREC 2010 Web track. In *TREC*, 2010.

[3] C. L. A. Clarke, M. Kolla, G. V. Cormack, O. Vechtomova, A. Ashkan, S. Büttcher, and I. MacKinnon. Novelty and diversity in information retrieval evaluation. In *SIGIR*, pages 659–666, 2008.

[4] R. L. T. Santos, C. Macdonald, and I. Ounis. Exploiting query reformulations for Web search result diversification. In *WWW*, pages 881–890, 2010.

[5] K. Spärck-Jones, S. E. Robertson, and M. Sanderson. Ambiguous requests: Implications for retrieval tests, systems and theories. *SIGIR Forum*, 41(2):8–17, 2007.