On the Suitability of Diversity Metrics for Learning-to-Rank for Diversity

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

An optimally diverse ranking should achieve the maximum coverage of the aspects underlying an ambiguous or underspecified query, with minimum redundancy with respect to the covered aspects. Although evaluation metrics that reward coverage and penalise redundancy provide intuitive objective functions for learning a diverse ranking, it is unclear whether they are the most effective. In this paper, we contrast the suitability of relevance and diversity metrics as objective functions for learning a diverse ranking. Our results in the context of the diversity task of the TREC 2009 and 2010 Web tracks show that diversity metrics are not necessarily better suited for guiding a learning approach. Moreover, the suitability of these metrics is compromised as they try to penalise redundancy during the learning process.

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

General Terms: Experimentation, Performance

Keywords: Web Search, Learning-to-rank, Diversity

1. INTRODUCTION

An ambiguous query can pose additional difficulties for a search engine, since it may not be clear which *interpretations* or *aspects* underlying this query are of interest to the user [5]. An effective approach for tackling query ambiguity is to diversify the search results, in order to maximise the chance that different users will find at least one relevant result to their particular information need [3].

In this paper, we investigate whether a diverse ranking can be automatically learned from training data by optimising an evaluation metric that rewards diversity. In particular, an optimally diverse ranking should achieve the *maximum coverage* of the aspects underlying an ambiguous query, with *minimum redundancy* with respect to the covered aspects. Hence, it seems intuitive to guide a learning-to-rank algorithm by optimising an evaluation metric that directly accounts for both aspect coverage and redundancy. Although several such metrics have been proposed and shown to effectively reward diversity in the search results [1], it is not clear whether they are any better suited than traditional relevance-oriented metrics for learning a diverse ranking. To investigate this, we contrast two relevance-oriented metrics and their diversity counterparts as objective functions for learning-to-rank for search result diversification.

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

2. EXPERIMENTAL METHODOLOGY

Our investigation is conducted in the context of the diversity task of the TREC 2009 and 2010 Web tracks, henceforth WT09 and WT10 tasks, respectively. In particular, WT09 provides 50 queries, while WT10 provides 48 queries, all with relevance assessments at the sub-topic level. Our experiments are based on the TREC ClueWeb09 (cat. B) corpus, which comprises 50 million English documents, aimed to represent the first tier of the index of a commercial search engine. We index this collection using Terrier,[1] with Porter's weak stemmer and without removing stopwords.

For our learning setup, we produce a sample comprising the top 5,000 documents retrieved for each of the 98 considered queries, using the Divergence From Randomness DPH weighting model, as implemented in Terrier. Besides being effective, DPH is a parameter-free model, and hence requires no training. Based on this initial sample, we compute a total of 61 document features typically used in the learning-to-rank literature [2]. These include standard weighting models (e.g., DPH itself, BM25), field-based (e.g., BM25F) and dependence (e.g., Markov Random Fields) models, spam detection, URL and link analysis (e.g., PageRank) features.

To enable learning by directly optimising an evaluation metric, we use Metzler's Automatic Feature Selection (AFS) listwise learning-to-rank algorithm, which has been shown to perform effectively on a web setting [4]. As optimisation metrics, we consider two standard relevance-oriented metrics: expected reciprocal rank (ERR) and normalised discounted cumulative gain (nDCG). Additionally, we consider their diversity counterparts: ERR-IA and α-nDCG [1]. While ERR and nDCG are established metrics for web search evaluation, ERR-IA and α-nDCG are the primary evaluation metrics in the diversity task of the TREC 2010 Web track. Moreover, besides rewarding aspect coverage and penalising redundancy, both ERR-IA and α-nDCG allow for tuning how much redundancy should be penalised [1]. To assess how penalising redundancy impacts the learning outcome, we consider variants of these two metrics—called ERR-IA* and α-nDCG*, respectively—that do *not* penalise redundancy at all. To enable a deeper analysis, all optimisation metrics are computed at rank cutoffs 20 and 1000. Finally, we report the diversification performance attained by the learned models optimised to each of these metrics in

[1] `http://terrier.org`

Optimisation Metric	Training Performance				Test Performance			
	WT09		WT10		WT09		WT10	
	ERR-IA @20	α-nDCG @20	ERR-IA @20	α-nDCG @20	ERR-IA @20	α-nDCG @20	ERR-IA @20	α-nDCG @20
ERR@20	0.1860	0.2940	0.2760	0.3569	0.1455	0.2377	0.2777	0.3880
ERR-IA@20	$0.2484^{\triangle}$	$0.3577^{\triangle}$	$0.3299^{=}$	$0.4204^{=}$	$0.1389^{=}$	$0.2318^{=}$	$0.2357^{=}$	$0.3442^{=}$
ERR-IA*@20	$0.2790^{\blacktriangle\triangle}$	$0.3942^{\blacktriangle=}$	$0.3410^{\triangle=}$	$0.4423^{\triangle\blacktriangle}$	$0.1747^{=\blacktriangle}$	$0.2835^{=\blacktriangle}$	$0.2373^{\triangledown=}$	$0.3467^{==}$
nDCG@20	0.2095	0.3272	0.3149	0.4305	0.2014	0.3028	0.2666	0.3760
α-nDCG@20	$0.2927^{\blacktriangle}$	$0.4165^{\blacktriangle}$	$0.3503^{=}$	$0.4522^{=}$	$0.1843^{=}$	$0.2889^{=}$	$0.2696^{=}$	$0.3854^{=}$
α-nDCG*@20	$0.2527^{\blacktriangle\blacktriangledown}$	$0.3729^{\triangle\blacktriangledown}$	$0.2902^{=\triangledown}$	$0.4012^{==}$	$0.1632^{==}$	$0.2622^{==}$	$0.2641^{==}$	$0.3728^{==}$
ERR@1000	0.1873	0.2944	0.2802	0.3729	0.1388	0.2176	0.2749	0.3883
ERR-IA@1000	$0.2736^{\blacktriangle}$	$0.3850^{\blacktriangle}$	$0.3283^{=}$	$0.4130^{=}$	$0.1400^{=}$	$0.2400^{=}$	$0.2330^{\triangledown}$	$0.3481^{=}$
ERR-IA*@1000	$0.2727^{\blacktriangle=}$	$0.3799^{\blacktriangle=}$	$0.3347^{\triangle\triangle}$	$0.4295^{\triangle\triangle}$	$0.1614^{=\triangle}$	$0.2673^{\triangle\triangle}$	$0.2736^{==}$	$0.3895^{==}$
nDCG@1000	0.2279	0.3460	0.3121	0.4247	0.2082	0.3290	0.2372	0.3535
α-nDCG@1000	$0.2601^{=}$	$0.3794^{=}$	$0.3423^{=}$	$0.4315^{=}$	$0.1511^{\blacktriangledown}$	$0.2580^{\blacktriangledown}$	$0.2223^{=}$	$0.3326^{=}$
α-nDCG*@1000	$0.2768^{\blacktriangle=}$	$0.4024^{\blacktriangle=}$	$0.3528^{\triangle=}$	$0.4667^{\blacktriangle=}$	$0.2338^{\triangle\blacktriangle}$	$0.3568^{\triangle\blacktriangle}$	$0.2496^{==}$	$0.3662^{=\triangle}$

Table 1: Training and test diversification performances of ranking models learned by optimising for relevance- (ERR, nDCG) or diversity-oriented (ERR-IA, α-nDCG, ERR-IA*, α-nDCG*) evaluation metrics.

terms of ERR-IA@20 and α-nDCG@20 under training and test scenarios. In particular, the latter scenario uses the WT09 and WT10 queries as separate training and test sets (i.e., we train on WT09 and test on WT10, and vice versa).

3. EXPERIMENTAL RESULTS

In this section, we investigate the suitability of relevance- and diversity-oriented metrics for guiding a listwise learning-to-rank approach in order to learn a diverse ranking. Table 1 shows the diversification performance of the models learned by AFS by optimising the previously described metrics. Statistical significance is verified using the Wilcoxon signed-rank test. The symbols $\blacktriangle$ ($\blacktriangledown$) and $\triangle$ ($\triangledown$) denote a significant increase (decrease) at the $p < 0.01$ and $p < 0.05$ levels, respectively, while $=$ denotes no significant difference. A first instance of these symbols denotes the significance (or lack thereof) of ERR-IA and ERR-IA* compared to ERR, as well as of α-nDCG and α-nDCG* compared to nDCG. A second instance denotes the significance of ERR-IA* and α-nDCG* compared to ERR-IA and α-nDCG, respectively.

From the training results in Table 1 (left half), we observe that diversity-oriented metrics generally lead to an improved diversification performance compared to relevance metrics. However, these improvements are not always significant, particularly for metrics that penalise redundancy. Indeed, on the WT10 queries, neither ERR-IA nor α-nDCG can significantly outperform their relevance-oriented counterparts, and are generally worse than ERR-IA* and α-nDCG* (except for α-nDCG*@20), which do not penalise redundancy. These results show that, although models that promote diversity can be learned via listwise learning-to-rank, metrics that penalise redundancy do not seem to be particularly suited for guiding the learning process.

The test scenario in Table 1 (right half) complements these observations. In particular, the test performance attained by using ERR-IA and α-nDCG as optimisation metrics is not significantly better than that attained using ERR and nDCG, respectively. In fact, α-nDCG@1000 (for WT09) and ERR-IA@1000 (for WT10) significantly underperform compared to their relevance-oriented counterparts. On the other hand, when ERR-IA and α-nDCG are compared to their variants that do not penalise redundancy, we observe a significantly superior performance of the latter, particularly for ERR-IA*@1000 and α-nDCG*@1000. Nevertheless, although generalising better across different query sets in the test scenario, these variants still cannot consistently and significantly improve compared to relevance metrics.

4. CONCLUSIONS

We have investigated the suitability of diversity metrics as objective functions for learning effective models for search result diversification. Our results show that, contrarily to our intuition, deploying a diversity metric does not necessarily help produce learned models with superior diversification performance compared to those produced using standard relevance metrics as objective functions. This highlights the need to develop features that better indicate not only the relevance, but also the diversity of individual documents.

Our results also show that models learned by optimising for metrics that penalise redundancy seem to overfit to the training queries, with a poor generalisation across different query sets. This observation corroborates related research that shows that the target evaluation metric is not necessarily the best suited to guide a learning-to-rank approach [6]. Indeed, since standard listwise learning-to-rank approaches assume that the relevance of a document is independent of other documents, optimising for metrics sensitive to redundancy (i.e., metrics that consider the relevance of a particular document in light of the other documents) may actually introduce noise to the learning process.

5. REFERENCES

[1] C. L. A. Clarke, N. Craswell, I. Soboroff, and A. Ashkan. A comparative analysis of cascade measures for novelty and diversity. In *WSDM*, 2011.

[2] T.-Y. Liu. Learning to rank for information retrieval. *Found. Trends Inf. Retr.*, 3:225–331, 2009.

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

[4] R. L. T. Santos, C. Macdonald, and I. Ounis. Intent-aware search result diversification. In *SIGIR*, 2011.

[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.

[6] E. Yilmaz and S. E. Robertson. On the choice of effectiveness measures for learning to rank. *Inf. Retr.*, 13(3):271–290, 2010.