A Query Performance Analysis for Result Diversification

Jiyin He[1], Marc Bron[2], and Maarten de Rijke[2]

[1] CWI, Science Park 123, 1098 XG Amsterdam
[2] ISLA, University of Amsterdam, Science Park 904, 1098 XH Amsterdam
j.he@cwi.nl, m.m.bron@uva.nl, derijke@uva.nl

Abstract. Which queries stand to gain or loose from diversifying their results? Some queries are more difficult than others for diversification. Across a number of conceptually different diversification methods, performance on such queries tends to deteriorate after applying these diversification methods, even though their initial performance in terms of relevance or diversity tends to be good.

1 Introduction

Result diversification is a retrieval strategy for dealing with ambiguous or multi-faceted queries; the system makes an educated guess as to the possible facets of the query and presents documents pertaining to different facets to the user [1, 3, 5, 7]. However, diversification is not a universal solution from which all queries stand to gain. Some queries benefit, while others get hurt, e.g., non-relevant documents may be promoted to the top of a ranked list because of their "diversity." A mehod addressing this issue balances relevance and diversity with a trade-off parameter on a per query basis, which leads to improved diversification effectiveness [7]. However, we are interested in what properties of a query make it suitable for diversification. More generally, how can diversification methods without a "trade-off parameter" benefit from these insights?

We investigate query diversification performance in a more general setting, aiming to provide a better understanding of when a query is (un)suitable for diversification. Let's call a query "difficult" when diversification is ineffective or deteriorates performance (in terms of relevance or diversity) across multiple types of diversification method. We use result diversification methods that are conceptually different and seek answers to the following research questions: *RQ1. Are some queries more difficult than others for diversification across different diversification methods?* and *RQ2. What properties of a query make it difficult?* There are many avenues to explore here, e.g., the ambiguity of a query, the facets associated with a query covered by the collection, etc.; we focus on the relation between diversification effectiveness and the initial performance of queries in terms of relevance and diversity.

2 Method

Diversification methods. We employ three diversification methods: MMR [3], IA-select [1] and Round Robin (RR) [5] that diversify a ranked list via re-ranking. By doing so we expect to identify query properties that hold across diversification methods with different underlying assumptions. MMR determines the value of a document for

diversification through a linear combination of its similarity to the query (relevance) and the smallest similarity to the documents already returned (diversity), where the trade-off between relevance and diversity is controlled by a parameter λ: $score_{d,q} = \lambda Rel_{d,q} + (1-\lambda) Div_{d,q}$. Unlike MMR, IA-select explicitly models the facets associated with a query. Documents are selected based on their initial retrieval scores, weighted by the probability that the selected document covers the underlying facets given that previously selected documents failed to do so. In RR, facets are modeled via clustering and ranked according to their estimated relevance to the query. Documents in each cluster keep the order of their original retrieval scores; then, documents in different clusters are selected in a round robin fashion. While IA-select aims to cover the *most important* facet of a query in the top ranked documents, RR seeks to cover *different* facets.

Analysis. For RQ1, we analyse the correlation among different diversification methods. Let m be a diversification method, $Q = q_1, \ldots, q_n$ a list of queries and $S_m = s_{q_1}, \ldots, s_{q_n}$ the per-query evaluation scores of the diversification results for Q, in terms of an evaluation metric. We calculate Pearson's linear (ρ) and Kendall's rank correlation (τ) between the performance of two methods S_{m_1} and S_{m_2}. A high correlation implies that queries with a relatively high (low) score using m_1 also receive a relatively high (low) score using m_2.

For RQ2, we identify two groups of queries. Let $t_q(m, b)$ be the performance difference between an initial baseline result b and a diversification result using method m for a query q as evaluated by a diversification measure. The first group consists of "easy" queries that are improved by at least one method and not hurt by others: $E = \{q | \sum_m t_q(m,b) > 0 \textit{ and } \forall m, t_q(m,b) \geq 0\}$. The second group consists of "difficult" queries that are hurt by at least one method and not improved by others: $D = \{q | \sum_m t_q(m,b) < 0 \textit{ and } \forall m, t_q(m,b) \leq 0\}$, where all diversification methods use the same baseline b. We investigate whether the two groups show different patterns characterized by properties associated with the initial performance of the queries.

Let G_q^K be the top K documents retrieved in response to query q and evaluated by a diversity measure, F_q be the set of facets of q and R_q^N be the N judged relevant documents of q in collection C. The properties we examine are as follows. (i) The performance of the initial ranked list G_q^K in terms of a diversity measure $eval@K$. (ii) The number of relevant documents and facets covered in the top of a ranked list: $R@K = |G_q^K \cap R_q|$ and $F@K = |\{f | f \in G_q^K\} \cap F_q|$. Here, we decompose $eval@K$ into two factors: relevance and diversity, in order to see whether these two factors have a different impact on the diversification performance. (iii) The percentage of relevant documents (facets) covered in the top of a ranked list compared to the total number of relevant documents (facets) for a query in the collection: $R@K\% = R@K/|R|$ and $F@K\% = F@K/|F|$. This takes into account the collection factor, i.e., diversification will not work in a collection without diverse content for a query.

3 Experiments and Results

We conduct our experiments using the ClueWeb category B dataset and the 100 test queries from TREC'09 and '10 Web track diversity task. For evaluation, we take the α-NDCG (@5, 10 and 20) [4], used as official measure at the TREC'09 and '10 diversity track, with α set to 0.5. We use the Markov Random Field model (MRF) [6] with default parameter settings to generate the initial baseline results b. We diversify with the top

100 documents in b using the three diversification methods described in Section 2.[3] We only include the results of a method with its optimal parameter settings found in a preliminary experiment. For MMR, λ is found to be 0.9. Following [5], we use LDA [2] to model the underlying facets of a query and of a document for both IA-select and RR, where the optimal number of facets are 50 and 10 respectively.

Table 1 shows the correlation between the performance of different diversification methods. All methods show significant positive correlation in terms of both ρ and τ. In particular, MMR and IA-select show a remarkably strong correlation, while both methods show a weaker correlation with RR, suggesting that RR behaves somewhat differently. The overall significant correlation indicates agreement between methods on the relative performance of queries, i.e., some queries consistently perform worse when subjected to diversification, or are more difficult to achieve good diversification results on, than others, regardless of the method applied.

Table 1. Performance correlation between diversification methods. All correlations are significant (p-value < 0.01).

Eval. measure	α-NDCG@5		α-NDCG@10		α-NDCG@20	
Corr. coef.	ρ	τ	ρ	τ	ρ	τ
MMR vs. IA-sel	0.896	0.830	0.917	0.828	0.925	0.818
MMR vs. RR	0.471	0.336	0.650	0.502	0.689	0.502
IA-sel vs. RR	0.495	0.376	0.669	0.533	0.675	0.515

We identify D and E from the 100 queries based on α-NDCG@5, 10 and 20 and list in Table 2 statistics of the query properties for these groups as discussed above.[4] (i) In terms of α-NDCG, queries in D have significantly higher scores than those in E, suggesting that queries with relatively good initial performance (set D), tend to be "difficult" for diversification. (ii) Queries in D cover significantly more relevant documents (facets), i.e., $R@K$ ($F@K$), compared to those in E except in the case where $K = 20$ for $F@K$. We see that both relevance and diversity of b has an impact on diversification performance. (iii) In terms of $R@K\%$ ($F@K\%$), queries in D have significantly higher scores than those in E, i.e., a larger percentage of all relevant documents (facets) in the collection is covered in the top of b for queries in D than in E.

Table 2. Contrasting properties of initial ranked lists (D vs. E). $^{\blacktriangle}$ ($^{\triangle}$) indicates a significant difference; p-value $<$.01 (.05) using Wilcoxon rank sum test.

	$K=5$		$K=10$		$K=20$	
Query set	E	D	E	D	E	D
# queries	41	18	31	17	27	16
α-NCDG@K	0.076	$0.284^{\blacktriangle}$	0.145	$0.281^{\triangle}$	0.191	$0.339^{\blacktriangle}$
$R@K$	0.65	$2.00^{\blacktriangle}$	2.10	$3.71^{\blacktriangle}$	5.33	$7.69^{\triangle}$
$R@K\%$	0.03	$0.18^{\blacktriangle}$	0.09	$0.24^{\blacktriangle}$	0.24	$0.39^{\triangle}$
$F@K$	0.46	$1.17^{\blacktriangle}$	0.84	$1.29^{\triangle}$	1.18	1.63
$F@K\%$	0.16	$0.51^{\blacktriangle}$	0.46	$0.72^{\triangle}$	0.33	$0.59^{\blacktriangle}$

The phenomena listed under (ii) and (iii) can be explained as follows. Given that all diversification methods do not generate perfect results, during re-ranking, diversification can hurt a result list by replacing a top ranked relevant and "novel" document by a non-relevant document or a relevant but "non-novel" document, where a "novel" document covers the facet of a query that is not (adequately) covered by the documents ranked before it. Intuitively, such replacement would have a higher chance to occur if an initial result list whose top K documents cover a large number of relevant docu-

[3] We did not remove spam. The performance of the three methods are between the median and the best of systems taking part in the diversity task at the TREC 2009 Web track.

[4] Since we only re-rank the top 100 documents, $|F_q|$ and $|R_q|$ are the relevant documents (facets) of a query covered by the top 100 documents in the initial ranked list.

ments or diverse facets, especially if most of the documents ranked below top K are non-relevant or non-novel, e.g., as indicated by a high $R@K\%(F@K\%)$. Also, a high $R@K\%(F@K\%)$ implies that there is little room for improvement. E.g., in the case of $K = 10$, on average 72% of the facets are covered by the initial top 10 documents for queries in D, the potential improvement through diversification lies in finding the other 28% of the facets, while the potential for E is 54%, as only 46% of the facets are covered by the initial top 10 documents.

4 Discussion and Conclusion

We investigated the performance of queries in result diversification with three conceptually different diversification methods. Across methods, some queries are more difficult than others for diversification. Further, queries with relatively good initial performance in terms of relevance or diversity tend to deteriorate through diversification.

The contribution of our analysis is two-fold. (i) We provide empirical evidence which confirms that some queries stand to gain more from diversification than others, independent of the diversification method used. (ii) Our analysis provides insights in the properties that should be focused on when identifying such queries.

We plan to look into predictors for the properties analyzed in this study, i.e, properties confirmed to have a high correlation with diversification performance.

Acknowledgements. This research was partially supported by the European Union's ICT Policy Support Programme as part of the Competitiveness and Innovation Framework Programme, CIP ICT-PSP under grant agreement no. 250430, the Fish4Knowledge project and the PROMISE Network of Excellence, funded and co-funded by the 7th Framework Programme of the European Commission, grant agreement no. 257024 and no. 258191, the DuOMAn project carried out within the STEVIN programme funded by the Dutch and Flemish Governments under project nr STE-09-12, the Netherlands Organisation for Scientific Research under project nrs 612.061.814, 612.061.815, 640.-004.802, 380-70-011, the Center for Creation, Content and Technology, the Hyperlocal Service Platform project funded by the Service Innovation & ICT program, the WAHSP project funded by the CLARIN-nl program, and under COMMIT project Infiniti.

5 References

[1] R. Agrawal, S. Gollapudi, A. Halverson and S. Ieong. Diversifying Search Results. In *WSDM'09*, 2009.

[2] D. Blei, A. Ng, M. Jordan and J. Lafferty. Latent Dirichlet Allocation. In *JMLR*, 2003.

[3] J. Carbonell and J. Goldstein. The Use of MMR, Diversity-based Reranking for Reordering Documents and Producing Summaries. In *SIGIR'98*, 1998.

[4] C. Clarke, M. Kolla, G. Cormack, O. Vechtomova, A. Ashkan, S. Büttcher, and I. Mackinnon. Novelty and Diversity in Information Retrieval Evaluation. In *SIGIR'08*, 2008.

[5] J. He, E. Meij and M. de Rijke. Result Diversification Based on Query-specific Cluster Ranking. In *JASIST*, 62(3), 2011.

[6] D. Metzler and W. B. Croft. A Markov Random Field Model for Term Dependencies. In *SIGIR'05*, 2005.

[7] R. Santos, C. Macdonald and I. Ounis. Selectively Diversifying Web Search Results. In *CIKM'10*, 2010.