On the Additivity and Weak Baselines for Search Result Diversification Research

Mehmet Akcay
Middle East Technical University & ASELSAN
Ankara, Turkey
meakcay@aselsan.com.tr

Ismail Sengor Altingovde
Middle East Technical University
Ankara, Turkey
altingovde@ceng.metu.edu.tr

Craig Macdonald
University of Glasgow
Glasgow, Scotland, UK
craig.macdonald@glasgow.ac.uk

Iadh Ounis
University of Glasgow
Glasgow, Scotland, UK
iadh.ounis@glasgow.ac.uk

ABSTRACT

A recent study on the topic of additivity addresses the task of search result diversification and concludes that while weaker baselines are almost always significantly improved by the evaluated diversification methods, for stronger baselines, just the opposite happens, i.e., no significant improvement can be observed. Due to the importance of the issue in shaping future research directions and evaluation strategies in search results diversification, in this work, we first aim to reproduce the findings reported in the previous study, and then investigate its possible limitations. Our extensive experiments first reveal that under the same experimental setting with that previous study, we can reach similar results. Next, we hypothesize that for stronger baselines, tuning the parameters of some methods (i.e., the trade-off parameter between the relevance and diversity of the results in this particular scenario) should be done in a more fine-grained manner. With trade-off parameters that are specifically determined for each baseline run, we show that the percentage of significant improvements even over the strong baselines can be doubled. As a further issue, we discuss the possible impact of using the same strong baseline retrieval function for the diversity computations of the methods. Our takeaway message is that in the case of a strong baseline, it is more crucial to tune the parameters of the diversification methods to be evaluated; but once this is done, additivity is achievable.

KEYWORDS

Additivity; result diversification; statistical significance

1 INTRODUCTION

Search result diversification in Information Retrieval (IR) is the process of (re-) ranking the retrieved documents for a query so that the top-ranked results would satisfy the users who all issue the same query but with diverse intents [22]. In the literature, search result diversification methods are broadly categorized as explicit and implicit [24]. In a nutshell, implicit methods only rely on the initially retrieved document list (the so-called candidate documents) to infer different subtopics (a.k.a., aspects or intents) of the query and re-rank the list. In contrast, explicit methods assume that query subtopics are made available (i.e., via using a topical taxonomy [1] or mining query logs [24]) and aim to use these subtopics to re-rank the candidate result list to surface results corresponding to different interpretations higher up the result list. In the last decade, several diversification methods have been investigated for and applied to adhoc text retrieval (e.g., web search [3, 14, 24], tweet search [12, 15]) but also in many other contexts, such as image search [16, 26], database and data stream querying [4, 10], and even recommender systems [25, 27, 28].

In a recent study, Kharazmi et al. [7] investigated the *additivity* of the findings with respect to different types of baselines for various IR tasks. First coined by Armstrong et al. [2], the additivity of a method refers to its capability to improve a strong baseline given an improvement over a weak one. Besides several other very useful and inspiring analyses and discussions, Kharazmi et al. also focussed on diversification, by employing three implicit and three explicit methods, and several baseline runs (i.e., adhoc runs submitted to TREC between 2009 and 2011 without any diversification effort) to investigate the additivity of the possible improvements made by these methods over weak and strong baselines. We emphasize that in this context, the term baseline refers to an adhoc retrieval method/system that returns a candidate result list (i.e., a run in TREC terminology) to be diversified, and a weak baseline run is such a list with a relatively low initial diversity performance (with respect to well-known evaluation metrics such as α-nDCG or ERR-IA). Their findings are quite striking: even when the diversification methods are found to consistently and significantly improve the weak baselines (and this only holds for the explicit diversification methods using the TREC official subtopics), these methods rarely improve the stronger ones; i.e., additivity does not occur.

The implications of the above conclusion are important. It says that in the future, researchers should use stronger baselines even for the initial retrieval stage to demonstrate the power of their diversification method, i.e., a simple adhoc run produced by a typical system (say, Lucene, Terrier or other research prototypes) or method (say, BM25) cannot be considered adequate. Given that at least some

ICTIR'17, October 1–4, 2017, Amsterdam, The Netherlands
© 2017 ACM. 978-1-4503-4490-6/17/10…$15.00
DOI: https://doi.org/10.1145/3121050.3121059

of these stronger baseline runs may involve several additional features that are extracted from external resources, such as proprietary datasets or even public ones that are no longer available (e.g., a modified web site or taxonomy), the necessity of such baselines may slow down the pace of experimentation in this subject area. Therefore, we believe that it is mandatory to repeat the procedure described by Kharazmi et al., and investigate their findings in a timely manner.

Our goal in this paper is to reproduce the major findings of the aforementioned previous work regarding the result diversification task, and question the validity of the resulting claims on additivity via additional experiments and analysis. In the previous works on explicit result diversification, it is widely reported that the official TREC subtopics yield much higher effectiveness than using subtopic definitions from other resources, such as web search engine suggestions (e.g., see [3, 14, 24]). It is also shown that explicit diversification, not surprisingly, outperforms the implicit approaches, especially when the official TREC subtopics are employed (e.g., see [11]). These observations are also verified by Kharazmi et al. in that significant improvements are either rare or even non-existent for the implicit methods and for the explicit methods with ODP-based subtopics even on the weaker baselines (see Fig. 3 in [7]). Thus, in this paper, we essentially focus on repeating the experiments employing explicit diversification methods and the official TREC subtopics, to investigate the new additivity claims of Kharazmi et al.

In doing so, our contributions are three-fold: (1) our experimental findings under exactly the same setup verify the results in [7]; (2) Moreover, our additional experiments where we set the λ trade-off parameter of some diversification methods (i.e. the parameter that balances the relevance to the main query and the diversity with respect to the query subtopics) for each baseline separately show that these methods can actually still significantly improve a non-trivial percentage of strong baselines, too; (3) We discuss the possible impact of using the same strong baseline retrieval function inside the diversification methods, i.e., to compute the relevance of a document to a subtopic, and provide some indirect evidence. Overall, our additional experiments and discussions show the potential of additivity of these diversification methods on strong baselines; and pinpoint the subtle issues (such as parameter tuning and relevance computation of documents to subtopics) that should be carefully handled while applying a diversification method to such baselines.

This paper is organized as follows. In Section 2, we briefly review the explicit diversification methods implemented for this work. In Section 3, we describe our experimental setup following the blueprint in [7]. Section 4 provides results of the repeated and additional experiments. In Section 5, we discuss the impact of modeling the document-subtopic relevance in this context. Finally, in Section 6, we conclude and summarize the main lessons learnt from this work.

2 EXPLICIT RESULT DIVERSIFICATION

In a typical result diversification scenario, for a query q, the adhoc retrieval results (i.e., a candidate ranking that typically includes from 50 up to 1000 documents) are given. The goal is to create a final ranking S of top-k documents (in practice, k is usually at most 20) that both maximizes the relevance to query q and its subtopics q_i, and minimizes the redundancy with respect to these subtopics [6]. As discussed above, both the earlier studies and Kharazmi et al. reported that the best diversification performance is obtained by the methods that utilize an explicitly modeled set of the query's subtopics, T_q, provided beforehand. Therefore, in this work, we implement three explicit diversification approaches, namely, IA-Select [1] and xQuAD [19] as employed in [7], as well as CombSum, as a recently proposed method that is shown to be comparable to or better than both of the former methods [14] and PM2 [3], which is another state-of-the-art approach. Note that, in [7], another variant of xQuAD (referred to as xQuADRel [29]) has also been considered, but since their experiments revealed that it is always inferior to IA-Select and xQuAD in yielding significant improvements over the baseline runs, we use CombSum instead of xQuADRel. In the following, we briefly review these methods as implemented in our setup:

IA-Select This is a best-first greedy method [1] that scores the documents in each iteration and selects the one that is most likely to cover all query subtopics that are not yet covered by the documents that have already been selected for the final top-k results, S, in the previous iterations. While the original definition of the IA-Select's scoring function employs a slightly different notation, following the practice in [14, 19], we present it as follows:

$$S(q,d) = \sum_{q_i \epsilon T_q} P(q_i|q)P(d|q_i) \prod_{d_j \epsilon S} (1 - P(d_j, q_i)) \quad (1)$$

In IA-Select, $P(q_i|q)$ is the likelihood (or, importance) of subtopic q_i for the query q. The probability $P(d|q_i)$ represents the likelihood of observing document d for the subtopic q_i and is usually modeled based on the relevance score $rel(d, q_i)$ (normalized to the [0, 1] range) of the retrieval system that generates the candidate ranking (see Sections 3 and 5 for further discussions).

xQuAD Again operating in iterations, eXplicit Query Aspect Diversification (xQuAD) [19] is based on a probabilistic mixture framework that takes into account the relevance to the main query q as well as the relevance and diversity with respect to the query's subtopics. Its scoring function is as follows:

$$S(q,d) = (1-\lambda)P(d|q) + \lambda \sum_{q_i \epsilon T_q} \left[P(q_i|q)P(d|q_i) \prod_{d_j \epsilon S} (1 - P(d_j|q_i)) \right] \quad (2)$$

where $P(d|q)$ is typically modeled as $rel(d, q)$, i.e., the (normalized) relevance score of d for q as generated by a retrieval system, while the other probabilities are defined as in the case of IA-Select. Note that there is a trade-off parameter λ to balance the relevance and diversity of the results in the final ranking. For $\lambda = 0$, the final ranking is exactly the same as the candidate ranking, and for $\lambda = 1$, the $P(d|q)$ component is totally discarded, as in IA-Select.

CombSum This method is an adaptation of the score-based ranking aggregation technique CombSum [5, 9] to the diversification problem [14]. Instead of running in iterations, CombSum first ranks the documents for each subtopic by computing $P(d|q_i)$, and then combines these rankings and the ranking for the main query using the following function, where the probabilities are defined as above:

$$S(q,d) = (1-\lambda)P(d|q) + \lambda \sum_{q_i \epsilon T_q} P(q_i|q)P(d|q_i) \quad (3)$$

In summary, each of the diversification methods contains a notion of document-query relevance estimation (e.g. $P(d|q)$), which we denote as $rel(d, q)$, as well as a document-subtopic relevance (e.g. $P(d|q_i)$), which we denote as $rel(d, q_i)$. Furthermore, both xQuAD

Table 1: Score intervals for categorizing baseline runs.

Baseline level	ERR-IA@20	α-nDCG@20	P-IA@20
Weak	$\leqslant 0.18$	$\leqslant 0.23$	$\leqslant 0.10$
Medium	>0.18 & $\leqslant 0.33$	>0.23 & $\leqslant 0.41$	>0.10 & $\leqslant 0.19$
Strong	>0.33	>0.41	>0.19

Table 2: Number of baseline runs in each year and level w.r.t. each metric.

Dataset	Baseline level	ERR-IA	α-nDCG	P-IA	Total
TREC2009	Weak	31	16	25	31
	Medium	0	15	6	
	Strong	0	0	0	
TREC2010	Weak	11	3	7	26
	Medium	15	22	18	
	Strong	0	1	1	
TREC2011	Weak	0	0	0	14
	Medium	6	5	3	
	Strong	8	9	11	
Total	Weak	42	19	32	71
	Medium	21	42	27	
	Strong	8	10	12	

and CombSum have a parameter λ, which controls the trade-off between the importance between a document's relevance to the original query, and the coverage of subtopics. In the rest of this paper, we study the setting and instantiation of the methods with respect to $rel(d, q_i)$ and λ.

3 REPRODUCED EXPERIMENTAL SETUP FOR THE ANALYSIS OF ADDITIVITY IN DIVERSIFICATION

Baseline runs and categories. As in [7], we only use the TREC Web Track adhoc retrieval track submissions that are on the ClueWeb09 collection (Part-B) and employs no diversification methods. There were 34, 26 and 16 such runs submitted to TREC 2009, 2010 and 2011, respectively. Following [7], we remove the five lowest scoring (w.r.t. α-nDCG@20) of these 76 runs, to obtain a total of 71 runs.

Kharazmi et al. [7] categorized these runs into three levels, namely, weak, medium and strong baselines, based on the scores of certain evaluation metrics. In particular, for a given evaluation metric, the score range (i.e., the range between the minimum and maximum scores of the baseline runs) is partitioned into three equally sized regions to form these groups. We consider the ranges for α-nDCG and ERR-IA as proposed in [7], and further employ a third metric, P-IA, for additional insights. Table 1 provides the score boundaries, and Table 2 shows the number of runs that fall into each range with respect to each metric. Note that, the number of runs in each level with respect to α-nDCG and ERR-IA metrics (cf. Table 2) exactly match to those values reported in Table 3 of [7].

Diversification methods, query sets and subtopics. As discussed before, we focus on the most-effective diversification scenario, namely explicit diversification with the official TREC subtopics. We pre-process the subtopic descriptions so that they look like real user queries, as has been performed in the literature [21]. In particular, we remove stopwords and generic terms like "find", "look for" and "information". We implement xQuAD, CombSum and IA-Select, as described in Section 2.

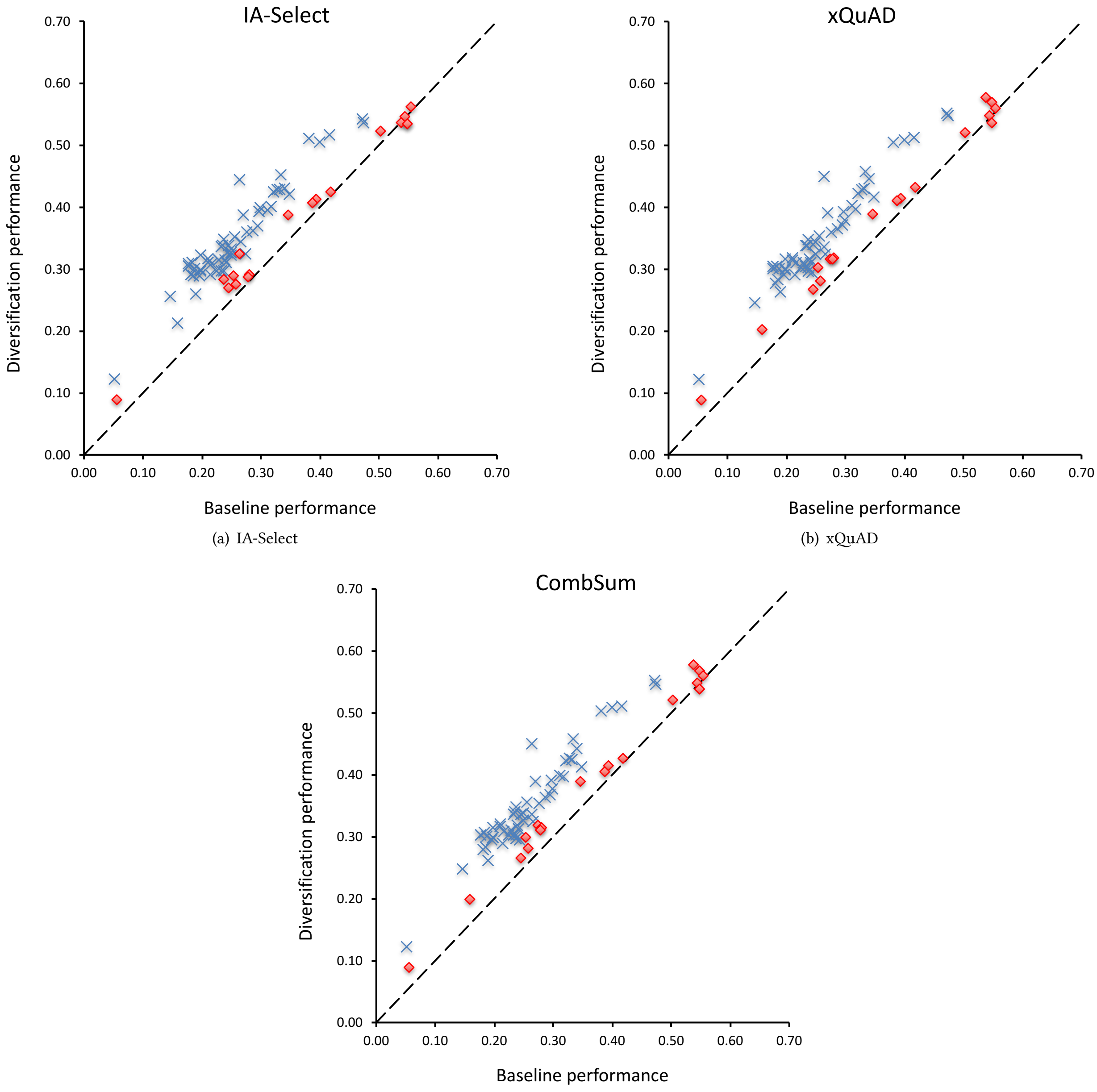
The query sets and their official subtopics from TREC 2009 to 2011 are used to diversify the top-100 candidate documents from the baseline runs of the corresponding year. As discussed in Section 2, all the aforementioned methods need to compute $rel(d, q)$ and $rel(d, q_i)$, i.e., the relevance score of the query and its subtopics to a candidate document, respectively. As all of the baseline runs provide the actual scores along with the candidate document ranking, we use the normalized version of these scores (by the sum of the scores of top-100 documents) for the former component, $rel(d, q)$ (Note that some runs involve negative scores that required further pre-processing before normalization).

For the latter component, $rel(d, q_i)$, the ideal case would be to obtain the document-subtopic scores using the exact retrieval system that yielded the candidate documents in each run. However, given the number and complexity of the methods in the baseline runs, this is practically unattainable. Thus, to compute document-subtopic scores, we employ a variant of the well-known Okapi BM25 weighting model [17], setting its parameters as follows $k_1 = 1.2$ and $b = 0.50$ (again, the actual BM25 scores are sum-normalized over the top-100 documents[1]). To do so, we use an index of ClueWeb09 Part-B collection using the open source Zettair retrieval system [30]. Although Kharazmi et al. [7] do not specify how exactly these computations are made, we verified through a personal communication [8] that they employed Indri with Okapi BM25 for computing $rel(d, q_i)$ in their work. Having said that, we further discuss the impact of this choice in Section 5.

In all our experiments, following [19], the subtopic probabilities $P(q_i|q)$ are computed uniformly as $1/|T_q|$, where T_q is the set of subtopics for a given query q. For xQuAD and CombSum, we set the trade-off parameter λ in three ways: First, we use the fixed value of $\lambda = 0.9$ reported in [7], which is said to be obtained via a 5-fold cross-validation process and by testing all values in the [0,1] range with a step size of 0.1 over the training sets. As confirmed by Kharazmi et al. [8], in their work the cross-validation has been applied over the runs, i.e., the best-performing λ parameter is determined over the training runs and then applied for the test runs in each fold. Second, we applied a similar procedure (i.e., a 5-fold cross-validation and scanning the [0,1] range) to determine the λ value that maximizes the α-nDCG@20 specifically for each run, i.e., in a *localized* fashion. In this case, for each run, we determine the best-performing λ over the training queries and then apply to the test queries—this mimics typical deployment of a run in a production environment, as well as in various research papers such as [19, 20]. Our λ parameter is more likely to be adjusted to the particular characteristics of a given run, rather than set for all runs in a training fold, as performed by [7]. Thirdly, we also report the performance using the *best* λ, which is again obtained per run but without using cross validation, as an upper-bound.

Evaluation metrics. To evaluate the diversification effectiveness, we compute three common metrics, namely, α-nDCG (with the default $\alpha = 0.5$), ERR-IA and Precision-IA, at the cut-off value of 20, using the `ndeval` software. We provide the evaluation results for our experiments at github.com/altingovde/ICTIR2017-DivAdditivity.

[1] This normalization is used in [14]; Santos [18] uses a slightly different normalization.



(a) IA-Select

(b) xQuAD

(c) CombSum

Figure 1: Scatter plots showing effectiveness of xQuAD and CombSum with trade-off parameter λ = 0.9. X- and y-axis show α-nDCG@20 scores for the baseline run and its diversified version, respectively. Points plotted as blue crosses are statistically significant improvements over the baseline, while red diamonds indicate no significiant difference.

4 EXPERIMENTS AND EVALUATION

As our first goal is reproducing the findings of Kharazmi et al., Section 4.1 presents our experiments conducted in the same setup as theirs (to the greatest extent possible) and employing the reported Global λ value of 0.9 for xQuAD and CombSum methods. In the additional experiments given in Section 4.2, keeping all the other setup details the same, we demonstrate the impact of using the Local and Best λ values during diversification.

4.1 Reproduced Results using Global Trade-off Parameter

In this section, we report our findings for all three diversification methods using the global λ value of 0.9. Figure 1 presents the performances of the diversification methods applied over the baseline (non-diversified) runs in terms of the α-nDCG metric. From the figure, we observe that, as in [7], while weaker baselines (closer to y-axis in the plots) are almost always significantly improved

Table 3: Ratio of runs significantly improved for each baseline level using xQuAD and CombSum. Note that [7] did not report the CombSum method or P-IA metric.

Level	Method	Results from [7]			Our results (Global $\lambda = 0.9$)		
		ERR-IA	α-nDCG	P-IA	ERR-IA	α-nDCG	P-IA
Weak	xQuAD	34/42	18/19	N/A	30/42	17/19	21/32
	CombSum	N/A	N/A	N/A	30/42	17/19	22/32
Medium	xQuAD	3/21	25/42	N/A	13/21	33/42	14/27
	CombSum	N/A	N/A	N/A	13/21	33/42	12/27
Strong	xQuAD	0/8	1/10	N/A	1/8	3/10	7/12
	CombSum	N/A	N/A	N/A	1/8	3/10	7/12

Table 4: Ratio of runs significantly improved for each baseline level using IA-Select.

Level	Results from [7]			Our results		
	ERR-IA	α-nDCG	P-IA	ERR-IA	α-nDCG	P-IA
Weak	35/42	18/19	N/A	30/42	18/19	22/32
Medium	9/21	29/42	N/A	13/21	32/42	13/27
Strong	1/8	3/10	N/A	1/8	3/10	6/12

by the application of the diversification method (measured using a paired two-tailed t-test for $p < 0.05$), the improvements for the stronger baselines are not significant in most of the cases for all three methods. The trends across IA-Select, xQuAD and CombSum are similar, and for xQuAD and IA-Select they are consistent with the previously reported findings (see the top row of Fig. 3 in [7]).

As in [7], none of the diversification methods (with official subtopics) yields a significant degradation in the performance compared to their baseline runs; i.e, all the significant changes are improvements. We report the ratio of runs that are statistically significantly improved for each method per baseline category, i.e., weak, medium and strong, with respect to three evaluation metrics in Table 3 (for xQuAD and CombSum, with $\lambda = 0.9$) & Table 4 (for IA-Select[2]). Note that, the denominator of the ratios in the latter results denotes the number of baseline runs at each level for each metric, as provided in Table 2. Both Tables 3 & 4 report the respective results repeated from Table 5 in [7]. By comparing the columns across Tables 3 & 4, we note that our findings are generally consistent with the previous work: almost all weak baselines and the majority of the medium-level baselines are improved by the diversification methods, while the improvements for the strong baselines are rather moderate (i.e., no more than 30% for α-nDCG and ERR-IA). Having said that, for xQuAD, we find a considerably larger number of significant improvements over the medium and strong runs in terms of the ERR-IA metric. Yet another interesting finding is that, when the P-IA metric (which is not reported in [7]) is considered, the percentage of significantly improved strong baselines exceeds 50%, i.e., not really a moderate ratio as for the other two metrics. Overall, we conclude that we can successfully reproduce the main results of [7] for the result diversification task.

4.2 Additional Results using Local Trade-off Parameter

In this section, we investigate the impact of the trade-off parameter λ on the performance of the xQuAD and CombSum methods (recall that IA-Select has no λ parameter). To this end, for each run, we optimize λ (for the α-nDCG@20 metric) using a 5-fold cross validation and scanning the [0, 1] range with a step size of 0.01 . In Figure 2, we present the distribution of these Local λ values over the training folds (i.e., given 71 runs and 5-fold CV, we consider 355 folds in total). The plot clearly justifies our choice of setting the trade-off parameter separately for each run, as the values are quite scattered over the bins, e.g., even the largest bin (for the range [0.9, 1]) yields the best performance during the training for less than one fourth of the total number of folds.

In Table 5, we report the diversification performance using the Local λ values per run, as described above. In comparison to the Global λ column (repeated from Table 3 to facilitate comparison), there is a clear increase in the ratio of significantly improved runs for all baseline levels and methods in terms of all metrics. We concentrate on the strong baselines, as the majority of the other baselines are shown to improve even when the Global λ value is utilized. Table 5 reveals that xQuAD and CombSum yields statistically significant improvements for 70% and 60% (i.e., 7/10 and 6/10) of the strong baselines for the α-nDCG metric, respectively. In terms of the ERR-IA metric, both diversification methods now significantly improve 37.5% (i.e., 3/8) of the strong baselines. Even for P-IA, there is an improvement in the ratio of significantly improved strong runs (i.e., from 7/12 to 8/12). As before, we also plot the performance of diversification methods (with Local λ) applied over the baseline runs in Figure 3, which further reveals that none of the diversification methods yield any drop in the performance, as well as pictorially showing the larger number of significant increases for the two methods.

The Best λ column in Table 5 shows that when the optimal λ (for the α-nDCG@20 metric) for each baseline run is set, the ratio of significantly improved strong runs reaches 62.5% and 80% in terms of the ERR-IA and α-nDCG metrics (i.e., 5/8 and 8/10, respectively). While we essentially provide this setting as an upper-bound, given the small number of queries in the TREC campaigns, one could use a leave-one-out cross validation strategy per run, which would yield a similar performance to the Best λ. Finally note that all experiments in this paper use the same λ value applied uniformly over all the queries in the test fold. In the literature, it has been shown that different queries have different levels of ambiguity, and therefore benefit from different λ values [20] to further improve

[2] Recall that IA-Select has no λ parameter.

Table 5: Ratio of runs that are significantly improved for each baseline level using XQuAD and CombSum with Global, Local and Best λ values.

Level	Diversity method	Global λ = 0.90			Local λ			Best λ		
		ERR-IA	α-nDCG	P-IA	ERR-IA	α-nDCG	P-IA	ERR-IA	α-nDCG	P-IA
Weak	xQuAD	30/42	17/19	21/32	33/42	19/19	24/32	37/42	19/19	28/32
	CombSum	30/42	17/19	22/32	32/42	19/19	24/32	37/42	19/19	29/32
Medium	xQuAD	13/21	33/42	14/27	18/21	36/42	18/27	20/21	40/42	19/27
	CombSum	13/21	33/42	12/27	17/21	36/42	18/27	20/21	40/42	20/27
Strong	xQuAD	1/8	3/10	7/12	3/8	7/10	8/12	5/8	8/10	8/12
	CombSum	1/8	3/10	7/12	3/8	6/10	8/12	5/8	7/10	8/12

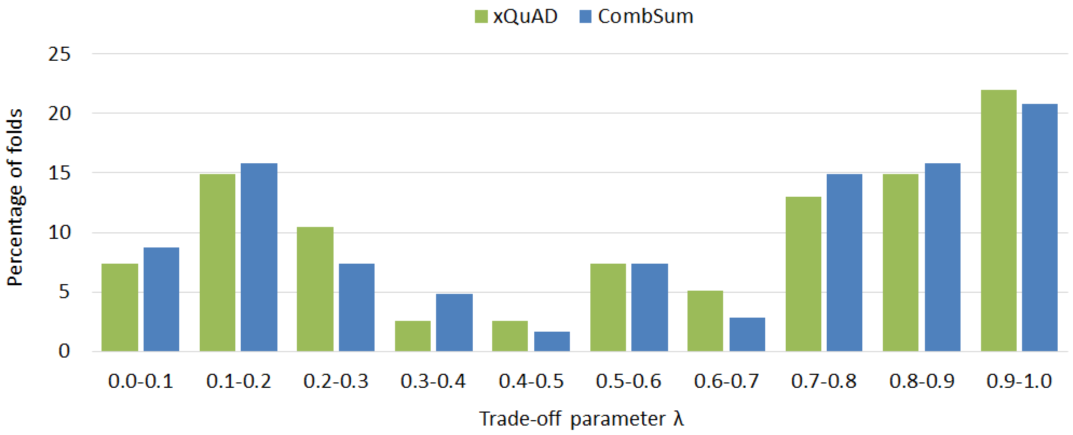


Figure 2: Distribution of trade-off λ parameter values over the 355 training folds.

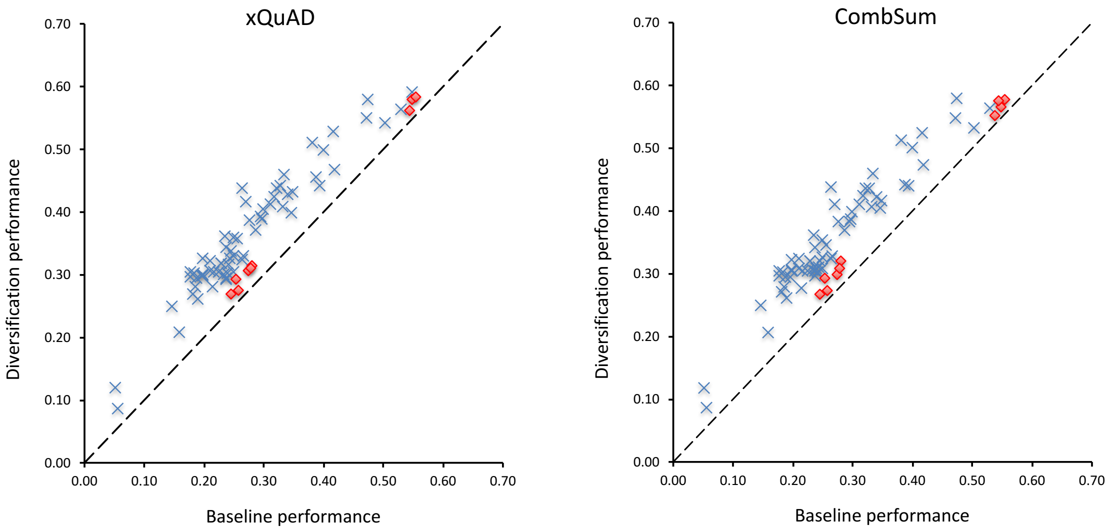


Figure 3: Scatter plots showing effectiveness of xQuAD and CombSum with the Local λ trade-off parameter. X- and y-axis show α-nDCG@20 scores for the baseline run and its diversified version, respectively. Points plotted as blue crosses are statistically significant improvements over the baseline, while red diamonds indicate no significiant difference.

Table 6: α-nDCG@20 scores of the diversified results over ten strong baselines (wrt. α-nDCG) using xQuAD. We compute $rel(d,q)$ either based on the original scores in the run or using BM25, while the $rel(d,q_i)$ scores are always computed by BM25. The bold results are the ones that are higher within the same λ setup. The underlined scores are the highest in that row.

Run id	Original	Global $\lambda = 0.90$		Local λ	
	α-nDCG@20	Original+BM25	BM25+BM25	Original+BM25	BM25 + BM25
2011SiftR1	0.5483	**0.5764**	0.5433	**0.5917**	0.5357
2011SiftR2	0.5372	**0.5793**	0.5442	**0.5615**	0.5362
DFalah11	0.5021	0.5240	**0.5266**	**0.5427**	0.5336
Otago2011cn	0.4160	**0.5170**	0.5141	**0.5286**	0.5178
liaQEWikiGoo	0.4725	**0.5548**	0.5440	**0.5490**	0.5430
srchvrs11b	0.5546	**0.5624**	0.5618	**0.5842**	0.5663
UAmsM705tiLS	0.5298	**0.5648**	0.5342	**0.5635**	0.5527
uogTrB47Vm	0.5691	0.5375	**0.5405**	**0.5795**	0.5376
uwBBadhoc	0.4731	**0.5397**	0.5327	**0.5788**	0.5418
uogTrB67	0.4178	**0.4310**	0.4265	**0.4674**	0.4330

the diversification performance – i.e. they can even outperform a uniform Best λ setting.

Indeed, there are other factors that may affect and potentially improve the diversification performance, such as the normalization of $rel(d,q)$ and $rel(d,q_i)$ scores, and setting the subtopic probabilities, $P(q_i|q)$. For the former component, Santos [18] employs a strategy that is again based on sum-normalization yet yields strict probability values for $P(d|q)$ and $P(d|q_i)$, while Ozdemiray and Altingovde [14] propose an alternative normalization strategy that improves the diversification effectiveness. It is also shown that exploiting the score distribution of candidate documents for each subtopic yields more accurate estimation of subtopic probabilities and subsequently, higher diversification performance [13]. With such optimizations, even more runs in Table 5 could have yielded statistically significant improvements; yet this direction is not explored here and left as a future work.

Our findings in this section imply that the diversification methods in question may still significantly improve the strong baselines as they do for the weaker ones. However, one might need to be more rigorous and careful tuning of the parameters in the case of strong baselines in comparison to applying them over the weaker baselines. This is contradictory to the claim by Kharazmi et al. in [7], that additivity "almost never" occurs for such diversification methods for strong baselines. Although it is preferable/recommendable to choose the stronger baselines whenever available, significant improvements over reasonable baselines may still be indicative, as well. Note that we still strongly encourage the use of a stronger baseline whenever available, as the actual improvements over the latter, albeit significant or not, would make more sense in real-world applications. We simply show that when such baselines are not available, using a medium level baseline is still viable.

5 DISCUSSION: IMPACT OF DOCUMENT-SUBTOPIC RELEVANCE

While we use BM25 to compute $rel(d,q_i)$ in the experiments of Section 4, this might be an important simplification, and the overall performance of the diversification methods (both metric scores and their statistical significance) could be increased by computing such scores using the exact retrieval function used to compute the document-subtopic relevance score $rel(d,q)$.

Indeed, in a report on their TREC 2010 Web track participation using xQuAD, Santos et al. [23] showed that deploying a supervised learning-to-rank approach for both $rel(d,q)$ and $rel(d,q_i)$ could result in increased diversification effectiveness, as measured by α-nDCG, while outperforming the corresponding learned baseline run by 6%. On the other hand, using learning-to-rank only for $rel(d,q)$ within xQuAD only improved the baseline by 3.4%. As no significance tests were conducted in [23], this admittedly anecdotal evidence suggests that using strong baselines for both $rel(d,q)$ and $rel(d,q_i)$ are important for properly attaining the highest effectiveness, a point not considered by [7].

To further investigate the impact of the $rel(d,q_i)$ component would involve implementing all of the adhoc retrieval methods used in the baseline runs – unfortunately an unfeasible task. Instead, to illustrate the impact, we undertake the reverse, and for the top-100 documents of each run, we also compute the $rel(d,q)$ component using the typical BM25 function (i.e., in effect, once the top 100 candidate documents are obtained from the corresponding TREC adhoc run, we only use BM25 for diversification). Interestingly, Kharazmi et al. also applied this strategy [8] (i.e., computed both components $rel(d,q)$ and $rel(d,q_i)$ using BM25), and hence our analysis here may also help for shedding light on their findings in [7].

We experiment with the trade-off parameter computed as globally and locally, as in Section 4, and focus only on the strong baselines (wrt. α-nDCG), as the others are improved anyway. Table 6 shows that for both ways of setting λ, using the original scores of the run for the $rel(d,q)$ component almost always yields better α-nDCG scores (i.e., 18 out of 20 cases) and furthermore, for 7 (of 10) strong runs the highest scores are obtained by using their original scores together with the Local λ values, confirming the findings in the previous section. More interestingly, even with the Local or Best λ values, the percentage of statistically significant improvements over the non-diversified baseline is very low (i.e., 3/10) when BM25 is employed instead of the original scores (see Table 7). We believe that this latter observation further explains the low number of significant improvements for strong runs in [7]. Given the impact of the $rel(d,q)$ function in this setup, we argue

Table 7: Ratio of strong runs that are significantly improved (wrt. α-nDCG) using XQuAD (with Global, Local and Best λs) and $rel(d, q)$ computed using the original scores or BM25.

Diversity	Global $\lambda = 0.90$		Local λ		Best λ	
method	Original+BM25	BM25+BM25	Original+BM25	BM25+BM25	Original+BM25	BM25+BM25
xQuAD	3/10	2/10	7/10	3/10	8/10	3/10
CombSum	3/10	2/10	6/10	3/10	7/10	3/10

that this provides further evidence that using a $rel(d, q_i)$ score that matches the $rel(d, q)$ score used by the actual run may further improve the diversification performance, and therefore change the conclusions concerning the additivity of these methods.

6 CONCLUSIONS

In this work, we considered the additivity of search result diversification methods in general, and in particular we reproduced the recent study of Kharazmi et al. [7] which applied diversification methods to the TREC baselines runs. Going further than [7], we showed that the setting of the relevance/diversity trade-off parameter λ is key to the overall conclusions of the experiments. Indeed, we found that additivity is more likely to occur when λ is set appropriately for each baseline run, especially for the stronger runs. Furthermore, we showed evidence that the mismatch of the retrieval models used to calculate the relevance of a document to a query (denoted $rel(d, q)$) and to its subtopics (denoted $rel(d, q_i)$) has the effect of underestimating the additivity of the diversification methods. Overall, we have identified and shown two confounding aspects of the experiments reported in [7], which raise questions about the authors' conclusion that the diversification methods "almost never" improve over strong baselines. In fact our study shows that with the appropriate automatic tuning of the parameters of the diversification methods for each of the strong baselines (as might be performed in a deployment setting), additivity is achievable.

ACKNOWLEDGMENTS

We are grateful to the authors of [7] for their timely and detailed responses to our inquiries regarding their work. We also thank to Rodrygo L. T. Santos for fruitful discussions. This work is partially funded by the Royal Society (UK) under the Newton International Exchanges Scheme with grant no. NI140231. I. S. Altingovde is also supported by Turkish Academy of Sciences Distinguished Young Scientist Award (TÜBA-GEBİP 2016).

REFERENCES

[1] Rakesh Agrawal, Sreenivas Gollapudi, Alan Halverson, and Samuel Ieong. 2009. Diversifying Search Results. In *Proceedings of WSDM*. 5–14.

[2] Timothy Armstrong, Alistair Moffat, William Webber, and Justin Zobel. 2009. Improvements that don't add up: ad-hoc retrieval results since 1998. In *Proceedings of CIKM*. 601–610.

[3] Van Dang and W. Bruce Croft. 2012. Diversity by proportionality: an election-based approach to search result diversification. In *Proceedings of SIGIR*. 65–74.

[4] Elena Demidova, Peter Fankhauser, Xuan Zhou, and Wolfgang Nejdl. 2010. *DivQ*: diversification for keyword search over structured databases. In *Proceedings of SIGIR*. 331–338.

[5] Joseph A. Fox and Edward Shaw. 1994. Combination of multiple sources: The TREC-2 interactive track matrix experiment. In *Proceedings of SIGIR*.

[6] Sreenivas Gollapudi and Aneesh Sharma. 2009. An Axiomatic Approach for Result Diversification. In *Proceedings of WWW*. 381–390.

[7] Sadegh Kharazmi, Falk Scholer, David Vallet, and Mark Sanderson. 2016. Examining Additivity and Weak Baselines. *Transactions on Information Systems* 34, 4 (2016), 23.

[8] Sadegh Kharazmi, Falk Scholer, David Vallet, and Mark Sanderson. 2017. Personal communication. (May 2017).

[9] Joon Ho Lee. 1997. Analyses of multiple evidence combination. In *Proceedings of SIGIR*. 267–276.

[10] Enrico Minack, Wolf Siberski, and Wolfgang Nejdl. 2011. Incremental diversification for very large sets: a streaming-based approach. In *Proceedings of SIGIR*. 585–594.

[11] Kaweh Djafari Naini, Ismail Sengor Altingovde, and Wolf Siberski. 2016. Scalable and Efficient Web Search Result Diversification. *Transactions on the Web* 10, 3 (2016), 15:1–15:30.

[12] Kezban Dilek Onal, Ismail Sengor Altingovde, and Pinar Karagoz. 2015. Utilizing Word Embeddings for Result Diversification in Tweet Search. In *Proceedings of AIRS*. 366–378.

[13] Ahmet Murat Ozdemiray and Ismail Sengor Altingovde. 2014. Query Performance Prediction for Aspect Weighting in Search Result Diversification. In *Proceedings of CIKM*. 1871–1874.

[14] Ahmet Murat Ozdemiray and Ismail Sengor Altingovde. 2015. Explicit search result diversification using score and rank aggregation methods. *JASIST* 66, 6 (2015), 1212–1228.

[15] Makbule Gulcin Ozsoy, Kezban Dilek Onal, and Ismail Sengor Altingovde. 2014. Result Diversification for Tweet Search. In *Proceedings of WISE*. 78–89.

[16] Monica Lestari Paramita, Jiayu Tang, and Mark Sanderson. 2009. Generic and Spatial Approaches to Image Search Results Diversification. In *Proceedings of ECIR*. 603–610.

[17] Stephen Robertson and Hugo Zaragoza. 2009. The Probabilistic Relevance Framework: BM25 and Beyond. *Foundations and Trends in Information Retrieval* 3, 4 (2009), 333–389.

[18] Rodrygo L. T. Santos. 2013. *Explicit web search result diversification*. Ph.D. Dissertation. University of Glasgow.

[19] Rodrygo L. T. Santos, Craig Macdonald, and Iadh Ounis. 2010. Exploiting query reformulations for web search result diversification. In *Proceedings of WWW*. 881–890.

[20] Rodrygo L. T. Santos, Craig Macdonald, and Iadh Ounis. 2010. Selectively diversifying web search results. In *Proceedings of CIKM*. 1179–1188.

[21] Rodrygo L. T. Santos, Craig Macdonald, and Iadh Ounis. 2011. Intent-aware search result diversification. In *Proceedings of SIGIR*. 595–604.

[22] Rodrygo L. T. Santos, Craig Macdonald, and Iadh Ounis. 2015. Search Result Diversification. *Foundations and Trends in Information Retrieval* 9, 1 (2015), 1–90.

[23] Rodrygo L. T. Santos, Richard McCreadie, Craig Macdonald, and Iadh Ounis. 2010. University of Glasgow at TREC 2010: Experiments with Terrier in Blog and Web Tracks. In *Proceedings of TREC*.

[24] Rodrygo L. T. Santos, Jie Peng, Craig Macdonald, and Iadh Ounis. 2010. Explicit Search Result Diversification through Sub-queries. In *Proceedings of ECIR*. 87–99.

[25] David Vallet, Martin Halvey, Joemon M. Jose, and Pablo Castells. 2011. Applying soft links to diversify video recommendations. In *Proceedings of CBMI*. 73–78.

[26] Reinier H. van Leuken, Lluis Garcia Pueyo, Ximena Olivares, and Roelof van Zwol. 2009. Visual diversification of image search results. In *Proceedings of WWW*. 341–350.

[27] Saúl Vargas, Linas Baltrunas, Alexandros Karatzoglou, and Pablo Castells. 2014. Coverage, redundancy and size-awareness in genre diversity for recommender systems. In *Proceedings of RecSys*. 209–216.

[28] Saul Vargas and Pablo Castells. 2011. Rank and relevance in novelty and diversity metrics for recommender systems. In *Proceedings of RecSys*. 109–116.

[29] Saul Vargas, Pablo Castells, and David Vallet. 2012. Explicit relevance models in intent-oriented information retrieval diversification. In *Proceedings of SIGIR*. 75–84.

[30] Zettair. 2016. Zettair open-source search engine. http://www.seg.rmit.edu.au/zettair/. (2016).