Using Score Differences for Search Result Diversification

Sadegh Kharazmi Mark Sanderson Falk Scholer
RMIT University & NICTA, Melbourne, Australia
{sadegh.kharazmi, mark.sanderson, falk.scholer}@rmit.edu.au

David Vallet
NICTA, Sydney, Australia
david.vallet@nicta.com.au

ABSTRACT

We investigate the application of a light-weight approach to result list clustering for the purposes of diversifying search results. We introduce a novel post-retrieval approach, which is independent of external information or even the full-text content of retrieved documents; only the retrieval score of a document is used. Our experiments show that this novel approach is beneficial to effectiveness, albeit only on certain baseline systems. The fact that the method works indicates that the retrieval score is potentially exploitable in diversity.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval–*retrieval models, search process*

General Terms

Algorithm, Theory, Experimentation

Keywords

Diversity, Score Difference, Clustering

1. INTRODUCTION

User queries submitted to an Information Retrieval (IR) system are often ambiguous at different levels [7]. To address such ambiguity, IR systems attempt to *diversify* search results, so that they cover a wide range of possible interpretations (*aspects*, *intents* or *subtopics*) of a query. Consequently, the number of redundant items in a search result list should be decreased, while the likelihood that a user will be satisfied with any of the displayed results should become higher.

In traditional IR, the estimated relevance of a document, which is used to determine the ranking of search results, depends primarily on query-document similarity. In diversified retrieval, search result rankings are based not only on query-document similarity, but also on the other documents that have been retrieved prior to the current document under consideration (i.e., document-document similarity).

Many of the proposed diversification techniques take a greedy approach, comparing a document to all previously retrieved documents, or the subtopics of a query. Also, they may use additional information, such as past user interactions, to identify which of the possible subtopics of a query are more likely to be interesting to the user. Most effective diversification approaches in the literature use techniques that focus on *coverage*, favoring documents that cover as many novel subtopics of a query as possible. This is in contrast to earlier techniques that focus on *novelty*, estimating the newness of a document with respect to those already retrieved. Novelty-based techniques usually exploit implicit information, such as differences in document content.

One source of implicit information derived from search results that appears to have never been investigated are differences in retrieval scores: *score differences*. Retrieval systems will usually, in response to a query, return a list of documents sorted by a relevance score, indicating the degree to which a query and document match. When analyzing a retrieved document list, the differences between the scores of adjacent retrieved documents differ, and this variation might be exploitable. Two documents that receive similar relevance scores are likely to share similar features; they might therefore address the same subtopics of a user's query. Conversely, two adjacent documents that have a large score difference are likely to have fewer features in common, which suggests that the documents might cover different query subtopics.

Our research question is to ask if we can exploit score differences to help with search result diversification.

We develop a simple non-greedy diversification approach that uses differences between the scores of the initially retrieved documents. The approach is experimentally investigated using the TREC framework, comparing to baselines and state-of-the-art diversification approaches.

2. RELATED WORK

There are two key approaches to diversifying search results, based on explicit or implicit evidence [9]. The explicit approaches [1, 10] match retrieved documents to the subtopics of a query, which are "pre-derived" from external sources such as a query log or taxonomies. Implicit approaches attempt to diversify based on a representation of already-retrieved documents.

Maximal Marginal Relevance (MMR) [2] is perhaps the most widely-studied implicit approach. Here, a diverse set of

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

results (S) is built incrementally from an initial retrieved list (R). The results are picked from R using a greedy approach where, in each iteration, the document that is most novel is selected. Novelty in this case is defined as the mean content-based dissimilarity between the candidate document and the already selected documents in S. A tuning parameter λ defines the trade-off between relevance and diversity.

Inspired by Modern Portfolio Theory (MPT) in finance, Wang et al. introduced a new implicit approach that analyses the expected mean and variance of the return of a portfolio [11]. Facility Location Analysis (FLA) was introduced by Zuccon et.al [12] to improve the MPT approach.

Explicit approaches are focused on query subtopics, which can be derived from a pre-defined taxonomy such as the Open Directory Project[1] (ODP), internal document features, query logs, or online resources [1, 8, 9]. The two most effective explicit diversification approaches are *xQuAD* [10] and *IASelect* [1].

All of these approaches use an iterative greedy selection approach to rank the most diverse documents.

3. SCORE DIFFERENCES

We hypothesize that documents with similar features (e.g., content, aspects covered, length) will be allocated (by ranking functions) similarity scores that are close together. Conversely, documents with different features are likely to be allocated similarity scores that are further apart.

Let $D_1, D_2, \ldots, D_N$ be an initial ranking of documents which are ordered by a ranking function $s(D)$, and θ be a difference threshold parameter. Then if

$$\frac{|s(D_i) - s(D_{i+1})|}{s(D_{i+1})} < \theta$$

we assume that the documents cover the same query subtopic. If the value is $\geq \theta$, it indicates that the two documents belong to different subtopics.

To test our hypothesis, we set up an experiment to measure the score differences between pairs of adjacent ranked documents that either covered the same or different subtopics of a query. We used the documents, queries, and diversity relevance judgments from the TREC 2009–2011 Web Tracks [3, 4]. Documents were ranked using the Dirichlet-smoothed language model from the Indri IR system[2] with default parameter settings. The query subtopics covered by individual answer documents are defined in the TREC relevance judgments. All non-relevant documents were assigned to an extra "non-relevant" subtopic. In total, 148 topics (TREC Web Track 2009, 2010 and 2011 queries) were used and for each topic, and pairs of documents in the top 100 positions in a ranked list were examined.

We tested the *Language Modeling* (LM) and *Okapi BM25* ranking functions, which are widely used in retrieval research, and have been shown to be effective ranking functions [6]. First, using language modelling as the ranking function to score documents, our analysis shows that for pairs of documents where there is no change in subtopic, the mean measured score difference is 0.065. Conversely, when the subtopic changes, the mean difference in scores is 0.073. Although the differences are small, a pairwise permutation test indicates that they are statistically significant ($p < 0.01$). This analysis suggests that score differences have the potential to be used as a technique to help with result diversification.

We repeated the same experiment using BM25. The mean measured score difference when there is no change in subtopic is 0.069, versus 0.071 when the subtopic changes. A pairwise permutation test indicates that these differences are not statistically significant. We therefore hypothesise that score differences are less likely to work well when the BM25 ranking function is used as a base run.

4. DIVERSIFYING RESULTS USING SCORE DIFFERENCES

To apply the score differences technique for result diversification, first, the score difference between each pair of documents, starting at rank position 1, was calculated. The top 100 documents were then re-ranked by decreasing size of the score difference between each document and the document above it. The documents with the biggest difference between the paired documents would now be top ranked, and they should be documents covering different subtopics.

After some experimentation with this simple approach, we found that it was better to re-rank documents based on a linear combination of the rank positions in the initial ranking and score differences, as shown in Algorithm 1. This approach to diversification, *RankScoreDiff*, does not use any information apart from the similarity scores from an initial retrieval run. It is therefore an implicit diversification approach.

Algorithm 1 RankScoreDiff(L)

$L' \leftarrow$ ScoreDiff(L)
for $1 < i \leq |L|$ **do**
  $Score(L[i]) \leftarrow \frac{1}{Rank(L[i])} + \frac{1}{Rank(L'[i])}$
end for
Sort L on $Score(L[i])$

5. EXPERIMENTAL SETUP

We investigated the effectiveness of *RankScoreDiff* as a diversification approach using the diversity task framework of the TREC Web Track from 2009–2011, which comprises 148 queries. Results are reported over the Clueweb category B collection. The Clueweb Online Services[3] were used to retrieve the top 100 documents for each query, using the same ranking functions and version of Indri described in Section 3. The top 100 documents were diversified using the methods under test.

Effectiveness was measured with α−nDCG, a widely-used metric that incorporates both relevance and diversity into a single score. The parameter α, which sets the relative importance of these two evaluation considerations, was set to 0.5, as recommended by the creators of the measure [5]. In the subsequent presentation of results, two-tailed paired t-tests are used to evaluate statistical significance.

5.1 Implementation and Tuning

For diversity approaches that require explicitly defined subtopics as an input (*xQuAD* and *IASelect*), two sources of

[1] http://www.dmoz.org/

[2] Version 5.2, http://lemurproject.org/indri.php

[3] http://boston.lti.cs.cmu.edu/Services/

Runs	ODP Subtopics			TREC Subtopics		
	$\alpha-$nDCG@5	$\alpha-$nDCG@10	$\alpha-$nDCG@20	$\alpha-$nDCG@5	$\alpha-$nDCG@10	$\alpha-$nDCG@20
Initial Run (LM)	0.235	0.276	0.315	0.235	0.276	0.315
MMR	0.233	0.274	0.317	0.233	0.274	0.317
MPT	0.235	0.277	0.316	0.232	0.277	0.319
FLA-MPT	0.240	0.280	0.320	0.240	0.280	0.320
RankScoreDiff	0.246 $*^{\uparrow}$	0.274	0.324 *	0.246 $*^{\uparrow}$	0.274	0.324 *
xQuAD	0.246 $^{\uparrow}$	0.286 $^{\Uparrow}$	0.326 $^{\Uparrow}$	0.318 ** $^{\uparrow}$	0.357 ** $^{\Uparrow}$	0.396 **
RankScoreDiff + xQuAD	0.258 * $^{\Uparrow}$	0.291 $^{\Uparrow}$	0.332	0.318 ** $^{\Uparrow}$	0.358 ** $^{\Uparrow}$	0.397 ** $^{\Uparrow}$
IASelect	0.266	0.298	0.337	0.321 ** $^{\Uparrow}$	0.365 ** $^{\Uparrow}$	0.400 **
RankScoreDiff + IASelect	0.283 * ‡ $^{\Uparrow}$	0.315 * ‡ $^{\uparrow}$	0.347 *	0.323 ** $^{\Uparrow}$	0.367 ** $^{\Uparrow}$	0.405 **

Table 2: Effectiveness of diversification approach using language modeling as baseline. For approaches that need explicit representation of subtopics, TREC official subtopics and ODP subtopics were used.

	Method	Spearman's ρ
Implicit	MMR	0.92
	MPT	0.86
	$FLA-MPT$	0.83
Explicit	$xQuAD_{ODP}$	0.78
	$IASelect_{ODP}$	0.75
	$xQuAD_{Trec}$	0.67
	$IASelect_{Trec}$	0.71

Table 1: The Spearman correlation of *RankScoreDiff* with other diversification approaches in terms of effectiveness measured by $\alpha-$nDCG@20.

subtopic definitions were used: first, the TREC Web Track official subtopics; and second, subtopics derived from the ODP using TextWise[4] services, with three levels of categorization to generate subtopics. These subtopics represent an upper-bound on effectiveness (perfect knowledge from the relevance judgments), and a reasonable but imperfect approach, respectively.

All approaches used in the experiments were trained to provide the best possible uniform diversification, to ensure that comparisons between the methods are fair. For approaches that required the tuning of parameters, this was carried out using 10-fold cross-validation to determine the best value on each collection. Parameters were tuned at increments of 0.1, and the best λ value obtained as 0.8 for the ODP subtopics and 0.9 for the official TREC subtopics.

5.2 Experimental Results

We investigated the impact on effectiveness of the proposed approach as a diversification feature and compared it to existing approaches in the literature [1, 2, 10, 11, 12].

Table 1 shows the Spearman correlation between our proposed approach and other diversification approaches. The results show that, in general, *RankScoreDiff* is more strongly correlated with implicit diversification approaches than with explicit approaches; the difference with the latter becomes more pronounced when the *TREC* (perfect) subtopics are available.

The results of our effectiveness experiments are shown in Tables 2 and 3, for the LM and BM25 initial retrieval runs, respectively. To carry out a detailed analysis, different baselines were considered. For this reason the following comparisons were made:

- A significant difference between the measured technique and the initial run is shown using * ($p < 0.05$) and ** ($p < 0.01$).
- A significant difference between the measured technique and implicit approaches (MMR, MPT, FLA-MPT and *RankScoreDiff*), which are independent of external knowledge such as subtopics, is shown using $^{\uparrow}$ ($p < 0.05$) and $^{\Uparrow}$ ($p < 0.01$). (The symbol indicates that a technique is significantly better than all four of the implicit approaches at the specified level.)
- A significant difference between a state-of-the-art explicit diversification method (*xQuAD* or *IASelect*), compared to *RankScoreDiff* combined with that method, is shown using † ($p < 0.05$) and ‡ ($p < 0.01$).

Language model as a baseline

Table 2 shows the results when using LM as an initial retrieval run. It can be seen that *RankScoreDiff* (row 5) significantly improves over the base run (row 1) for $\alpha-$nDCG@5 and $\alpha-$nDCG@20. Although there is some marginal improvement in comparison with other implicit approaches (rows 2-4), this improvement is only significant for $\alpha-$nDCG@5. The results suggest that *RankScoreDiff* is competitive in comparison with implicit approaches, but it is not as good as the explicit approaches (rows 6 and 8). However, *RankScoreDiff* can also be used in combination with the explicit approaches (rows 7 and 9 of the table).

Using ODP subtopics, the combination of *RankScoreDiff* with an explicit approach improves over using the explicit approach on its own in most cases. The improvement is significant for $\alpha-$nDCG@5 and $\alpha-$nDCG@10 when *IASelect* and *RankScoreDiff* are combined. Using the TREC (perfect) subtopics, marginal improvements are obtained when combining *RankScoreDiff* with the explicit approaches, however the combined approach is not significantly different compared with the original explicit approach. In addition, the combined approaches are always significantly better than the base run, and are usually significantly better than the implicit approaches.

OKAPI BM25 as a baseline

Table 3 shows results when using BM25 as a baseline run. The improvements in effectiveness over the base run are marginal for all implicit approaches, including *RankScoreDiff*.

Furthermore, Table 3 shows that for the ODP subtopics, even the explicit approaches do not lead to significant improvements over the base run. Similarly, combining *RankScoreDiff* with *xQuAD* and *IASelect* for the ODP subtopics does not improve significantly on the baseline, and in some cases reduces effectiveness. When using TREC (perfect) subtopics, all explicit approaches (on their own, or combined

[4]http://www.textwise.com/

Runs	ODP Subtopics			TREC Subtopics		
	$\alpha-$nDCG@5	$\alpha-$nDCG@10	$\alpha-$nDCG@20	$\alpha-$nDCG@5	$\alpha-$nDCG@10	$\alpha-$nDCG@20
Initial Run (BM25)	0.268	0.300	0.336	0.268	0.300	0.336
MMR	0.266	0.300	0.337	0.266	0.300	0.337
MPT	0.270	0.301	0.336	0.272	0.302	0.337
FLA-MPT	0.275	0.305	0.340	0.275	0.305	0.340
RankScoreDiff	0.270	0.298	0.335	0.270	0.298	0.335
xQuAD	0.273	0.309	0.344	0.335 ** ⇑	0.377 ** ⇑	0.407 **
RankScoreDiff + xQuAD	0.275	0.309	0.339	0.341 ** ⇑	0.378 ** ⇑	0.409 **
IASelect	0.263	0.294	0.331	0.348 **⇑	0.389 ** ⇑	0.420 **
RankScoreDiff + IASelect	0.256	0.288	0.323	0.343 ** ⇑	0.384 ** ⇑	0.413 **

Table 3: Effectiveness of diversification approach using Okapi (BM25) as baseline. For approaches that need explicit representation of subtopics, TREC official subtopics and ODP subtopics were used.

with *RankScoreDiff*) improve significantly over the base run and over the implicit approaches. The combination of *RankScoreDiff* and *xQuAD* could marginally improve over *xQuAD* on its own for $\alpha-$nDCG@5, while this is not the case for *IASelect*.

Overall, the results in Tables 2 and 3 show that *RankScoreDiff* is equivalent in effectiveness with other implicit approaches, although with no significant improvement over strong base runs. However, in the absence of perfect subtopics, *RankScoreDiff* can potentially be used in combination with explicit approaches to provide a boost in effectiveness. We note that a particular feature of *RankScoreDiff* is that it is computationally much less intensive than all other diversification approaches, being based only on information that is already available with a base run.

6. CONCLUSIONS

This paper examined a novel approach that uses the differences in original retrieval scores as evidence of diversity, based on the assumption that similar documents will receive similar retrieval scores with respect to a given query, and that similar documents could represent a similar subtopic. We experimentally evaluated the use of a score difference technique to diversify search results. In contrast with existing diversification techniques, which need additional document representations or external subtopics, our proposed approach only needs the relevance score provided by a ranking function. From the results, diversifying using score differences is competitive with other implicit diversification approaches. However, none of these approaches regularly lead to significant improvements over a base run. When perfect subtopic knowledge is not available, the *RankScoreDiff* approach can potentially boost the effectiveness of state-of-the-art explicit diversification techniques.

Our analysis of the distribution of score differences showed that the approach is directly affected by the ranking function that generates the initial retrieval scores.

In future work, we plan to investigate how particular features of ranking functions interact with the score differences approach. For example, a parameterised ranking function such as BM25 allows individual effects such as length normalisation, or the relative emphasis of TF and IDF effects, to be isolated and explored.

7. ACKNOWLEDGMENTS

This work was supported in part by the Australian Research Council (DP130104007), as well as NICTA Victoria which is funded by both the Federal and State governments.

8. REFERENCES

[1] R. Agrawal, S. Gollapudi, A. Halverson, and S. Ieong. Diversifying search results. In *Proc. WSDM*, pages 5–14. ACM, 2009.

[2] J. Carbonell and J. Goldstein. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proc. SIGIR*, pages 335–336. ACM, 1998.

[3] C. Clarke, N. Craswell, and I. Soboroff. Overview of the TREC 2009 web track. In *Proc. TREC*, 2009.

[4] C. Clarke, N. Craswell, I. Soboroff, and G. Cormack. Preliminary overview of the trec 2010 web track. In *Proc. TREC*, volume 10, 2010.

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

[6] W. B. Croft, D. Metzler, and T. Strohman. *Search engines: Information retrieval in practice*. Addison-Wesley Reading, 2010.

[7] F. Radlinski, P. Bennett, B. Carterette, and T. Joachims. Redundancy, diversity and interdependent document relevance. In *Proc. SIGIR*, volume 43, pages 46–52. ACM, 2009.

[8] F. Radlinski and S. Dumais. Improving personalized web search using result diversification. In *Proc. SIGIR*, pages 691–692. ACM, 2006.

[9] R. L. T. Santos, C. Macdonald, and I. Ounis. On the role of novelty for search result diversification. *Information Retrieval*, 2012.

[10] R. L. T. Santos, J. Peng, C. Macdonald, and I. Ounis. Explicit search result diversification through sub-queries. In *Proc. ECIR*, pages 87–99, Milton Keynes, UK, 2010. Springer.

[11] J. Wang and J. Zhu. Portfolio theory of information retrieval. In *Proc. SIGIR*, pages 115–122. ACM, 2009.

[12] G. Zuccon, L. Azzopardi, D. Zhang, and J. Wang. Top-k retrieval using facility location analysis. In *Proc. ECIR*, pages 305–316. Springer, 2012.