Iterative Search using Query Aspects

Manmeet Singh, W. Bruce Croft
Center for Intelligent Information Retrieval
College of Information and Computer Sciences
University of Massachusetts,
Amherst, MA
(msingh, croft)@cs.umass.edu

ABSTRACT

Pseudo-relevance feedback (PRF) via query expansion has proven to be effective in many information retrieval tasks. In most existing work, the top-ranked documents from an initial search are assumed to be relevant and used for feedback. There are some drawbacks to this approach. One limitation is that there might be other relevant documents which were not retrieved or considered for the the feedback process. Another issue is one or more of the top retrieved documents may be non-relevant, which can introduce noise into the feedback mechanism. Term-level diversification, on the other hand, uses an effective technique for identifying terms associated with query aspects or subtopics. We propose a new iterative feedback method that combines PRF with aspect generation to improve feedback effectiveness. In our experiments, we discovered a new property of convergence of feedback terms that was incorporated into the PRF process. We show that the resulting method significantly outperforms the baseline relevance model.

CCS Concepts

•**H.3.3 Information Search and Retrieval** → Retrieval Models, Relevance Feedback;

Keywords

pseudo-relevance feedback, search diversification, iterative search

1. INTRODUCTION

Local feedback techniques such as pseudo-relevance feedback (PRF) have been used to improve retrieval performance without user interaction[18, 12, 15, 6, 13, 14]. They have been shown to work better than global context analysis[17]. PRF uses the top retrieved documents to extract expansion terms and weights those terms according to some model, such as tf.idf. Although this approach has generally proven to be effective, there are limitations. Non-relevant documents may be retrieved, introducing noise into the term selection and weighting process. Other relevant documents containing important terms may not be retrieved at high ranks. One way to address these limitations is to improve the existing pseudo-relevance feedback techniques [8, 4, 11]. Another is to change the assumptions behind PRF. Term-level diversification uses an effective technique to discover terms that describe different aspects or subtopics related to a query [1]. By using this technique to select terms for PRF, we change the assumption that the top-ranked documents are relevant and instead assume the top-ranked documents describe the important aspects. Another assumption we change is that a single iteration of feedback will be the most effective. In the orginal implementations of relevance feedback, multiple iterations were used and not found to be effective. These experiments were, however, done on small collections and using manual relevance assessments [15]. We propose an automatic iterative method for PRF that incorporates the aspect generation technique from term-level diversification. Specifically, we use the DSPApprox method to discover aspects and compare this to a simple tf.idf term selection method. In our experiments, we show that feedback terms do in fact converge over multiple iterations and that the aspects generated produce significant improvements compared to the well-known RM3 baseline and a single iteration of PRF.

The rest of the paper is organized as follows. Sec. 2 details the related work. In Sec. 3 we explain the methodology and then Sec. 4 details the experimental setup. Sec. 5 shows the results and in Sec. 6 we discuss and analyze the findings. Finally, Sec. 7 concludes the paper.

2. RELATED WORK

A variety of approaches to automatic query expansion for improving the performance of retrieval queries have been previously studied. Pseudo-relevance feedback has proven to be a successful technique for query expansion [9, 6]. It has also been shown to be effective for query classification [18], query translation, and spelling corrections.

Lavrenko and Croft's relevance model (RM3) algorithm [6] is a pseudo-relevance feedback method developed for the language modeling framework. The standard formulation of this method involves submitting an original query (LM), using the resulting ranked list to perform weighted query expansion, and performing a second round of retrieval. We compare our results to this method. Lv and Zhai used a boosting approach-FeedbackBoost [8] that iteratively selects

CIKM '16, October 24–28, 2016, Indianapolis, IN, USA.

DOI: http://dx.doi.org/10.1145/2983323.2983903

and combines basis feedback methods. In each iteration, a basis feedback method is selected to improve those queries on which the already selected basis feedback methods perform poorly in terms of both effectiveness and robustness. It uses a linear combination of these basis feedback methods as its final feedback model. Kurland and Lee proposed an iterative pseudo-feedback approach to ad hoc information retrieval using cluster-based language models [5]. In this method, an initial set of documents are retrieved based on vector space model. Then the retrieved documents are analyzed based on the context of all terms in a document and query using clustering. Li and He used a pseudo relevance feedback method [3] named iterative probabilistic one-class SVMs to re-rank the retrieved images; but their intended application is not a direct scoring of potential retrieval candidates.

Aspect based-retrieval approach has been studied in retrieval and diversification tasks [1, 16]. An aspect representation determines how the information needs underlying a query are represented as multiple aspects of the query. Generating these aspects automatically is not as well understood. A topic term extraction algorithm named DSPApprox was proposed by Lawrie and Croft [6, 7] for hierarchical multi-document summarization. This method was also incorporated in term level diversification by Dang and Croft using the vocabulary of the retrieved documents for automatic aspect or topic generation [1]. Harman and Buckley recognized that weak aspect coverage [2] in the resulting feedback model or final expansion model is something algorithms must detect and remedy to avoid query drift. Our technique also addresses this issue by covering various aspects and then using the most important ones for the feedback process.

3. METHODOLOGY

Our method identifies the possible aspects iteratively and uses the most important ones for PRF. To generate these aspects, we compare the DSPApprox method and a simple tf.idf term selection method. Let S be the set of possible aspects that could be generated by a query Q. Let K be the number of aspects generated per query in every pass. Let N be the number of documents we use to generate these aspects. Here K, N are free parameters whose values need to be determined. Let the feedback terms generated using the top N documents on query Q in the intermediate step by aspect generation method be $t_1, t_2, t_3..., t_K$. Now we have generated K new queries- $Q \cup t_1, Q \cup t_2, ..., Q \cup t_K$ and our set S contains maximum K aspects. We repeat the process with each of the generated K queries until no new terms are found. We discard duplicate terms generated in any of the iterations. For preprocessing, we removed stop-words, used stemming and case folding before applying the algorithm.

The problem formulated above can be represented as a k-ary tree data-structure, where root node represents the base query. The K possible child-nodes, represent the intermediate queries $Q \cup t_1, Q \cup t_2, ..., Q \cup t_K$. Each of these child nodes can further have K child nodes making maximum of K^2 nodes (possible queries) at level two and the process goes on the same way for every level. There was considerable overlap in the aspects generated, and as duplicate aspects are discarded, there were never K^2 nodes at level 2 or higher. Let $S_0, S_1, S_2, ...$ be the aspect sets generated at each level. Our final aspect set S is $\{S_0 \cup S_1 \cup S_2, ...\}$. Algorithmically, this problem could be solved by breadth first search using a queue data-structure. This approach is what we use in our iterative search method. Though we had put thresholds on the maximum number of feedback terms to be considered and on the number of levels the algorithm should search iteratively, we never reach any of the threshold conditions as terms converged before reaching these thresholds.

Our research shows a new property of convergence in the application of multi-level PRF (i.e., there are no new terms after a certain level). This convergence was achieved on all the queries of the dataset with the DSPApprox method as well as tf.idf weighting. This happened due to the overlap of aspects with every feedback query. From the converged set, we then selected the best terms for PRF.

One might argue that the above algorithm may be slow and hence not good for an online setting even if it produces good results. This concern is addressed by the fact that the above algorithm could easily be parallelized by distributing the feedback queries on multiple servers while maintaining a shared set for the aspects generated across systems. Our analysis section also shows that the algorithm runs fast even without parallelism because the convergence is achieved quite quickly.

3.1 Aspect Generation

Aspects denote the multiple possible intents, interpretations, or subtopics associated with a given query. It is an explicit idea expressed within a query. For example, in the query "painting on wood" there is a "painting" aspect and a "wood" aspect. Logically, documents covering both of these aspects would be considered relevant. More generally, a document will have a higher degree of relevance the more query aspects it includes.

Automatic construction of these query aspects has been difficult and the techniques have generally relied on manually created descriptions of the aspects. Term-level diversification is also an explicit approach which uses terms without grouping. We use a greedy algorithm named DSPApprox to generate these aspects. The algorithm [10] iteratively selects terms from the candidate topic term set T. The utility of each term is the product of its topicality and predictive ability. At each step, the algorithm selects the topic term $t^* \in T$ with maximum utility. Then, it decreases the predictive ability of other topic terms that predict the same vocabulary. This ensures that topic terms from the uncovered part of the vocabulary will emerge for selection in the next iteration. The algorithm stops once the utility of all candidate topic terms reaches zero, indicating that all vocabulary has been covered and we get the list of top feedback terms. In the tf.idf based selection method top terms with the highest weights are used and added to the aspect set. In the end, top terms with highest tf.idf weighing are selected from the final set (after conver- gence) as aspects. As the DSPApprox approach covers more diverse topics [1] the set of terms generated after convergence are different than the tf.idf term selection method.

4. EXPERIMENTAL SETUP

4.1 Dataset

We experimented with the Robust Track of TREC 2004 which has 528,155 news articles. We use all the 250 queries in this set for our experiments. The Robust Track data is

a standard ad hoc retrieval testbed with an emphasis on the overall reliability of IR systems and it contains difficult queries for a heterogeneous data set. This data set is called "Robust04" in the following discussion.

4.2 Baselines

In order to evaluate the effectiveness of our technique, we used two methods as baselines for comparison. The first baseline (RM3) is one of the most effective and robust pseudo feedback methods under language modeling framework. It uses the query likelihood score of a feedback document as the weight and estimates a query language model (for feedback) based on weighted aggregation of term counts in the feedback documents. The second baseline uses the terms generated by single iteration of PRF. This baseline helps us evaluate whether the top terms we chose from the converged set are better than the terms chosen from the set obtained after level one, highlighting the importance of the convergence property.

4.3 Parameter Settings

Our technique needs to choose optimal values for the following hyperparameters in order to get maximum prediction accuracy:

K: # of aspects requested per query

N: # of top documents considered to generate aspects

r: # of feedback terms added to original query

$fborigWt$: weight(importance) to base query terms vs feedback terms

We used grid search method to determine the best values for these parameters. The best combination we obtained was K = 50, N = 10, r = 10, fborigWt = 0.8. The tuning of other free paramters- maximum # of aspects to be considered and maximum # of levels to explore was not required as we quickly obtained convergence for all 250 queries.

5. RESULTS

The results were evaluated on various standard metrics. The first and second columns in the table shown below are the baseline results obtained using the RM3 and one-level of PRF respectively. The next two columns presents the results obtained by our iterative search method which uses query aspects generated via tf.idf weighting and the DSPApprox technique respectively.

Table 1: Results on Robust04. † indicates significantly better than baseline for 0.05 as threshold for p value

Metrics vs Method	RM3	Baseline-2	Using TF-IDF	Using DSPApprox
MAP	0.250	0.256	0.267 †	0.277 †
P@10	0.425	0.427	0.442 †	0.445 †
P@20	0.362	0.366	0.376	0.375
NDCG@10	0.425	0.417	0.432	0.436 †
NDCG@20	0.412	0.409	0.421 †	0.424 †
R@10	0.139	0.140	0.149 †	0.147 †
R@20	0.210	0.217	0.221 †	0.221 †

The mean average precision (MAP) improved by 6.8% and 10.8% respectively compared to the relevance model (RM3). Our method also does a better job on ranking the top 10 and 20 documents. The results confirm that our method significantly outperforms the baseline methods.

Figure 1. shows the number of aspects which were generated for each query after the convergence was achieved. The average # of aspects generated by tf-idf weighing and DSPApprox method were 188 and 148 respectively. Hence, on an average iterative search using DSPApprox method converges faster.

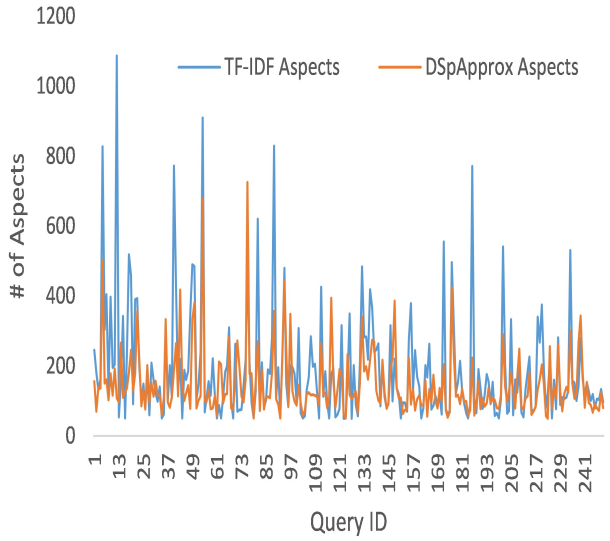

Figure 1: # of aspects generated for each query after convergence

6. DISCUSSION

Our experimental results clearly show that the quality of aspects generated through the iterative process were much better as it led to significant improvement in the accuracy. Our method covered most of the aspects and then used the top ones as feedback terms while the baseline methods were choosing highest tf.idf weighing terms from the top N documents. Iteratively, our algorithm explores relevant documents which in turn gives us more effective feedback terms. To further strengthen our argument, we tried using feedback terms from a much larger pool of documents and the accuracy declined compared to the baseline results. Thus, the iterative process retrieves higher numbers of relevant documents that we couldn't get through RM3 based pseudo-relevance feedback.

There are several other questions that need to be addressed. In first place, why does convergence even happen? Intuitively, generation of similar aspects led to the convergence. Our algorithm converged by level 2 and 3 for K values 50, 25 respectively. Another question which arises from our work is whether the DSPApprox method and tf.idf selection method generated the same set of terms after convergence? The answer to this question is no as neither are the terms generated by both the methods the same nor is any of them subset of the other. DSPApprox tries to identify terms based

on topics, hence it would try not to retrieve a term which is already covered in terms of topicality. This seems to be the most appropriate reason why the average number of aspects retrieved by this algorithm are less compared to the other technique we used. There is a high overlap, however, in the terms generated by both the algorithms.

One might argue that more number of non-relevant documents could be generated as we are doing multiple passes. Even if one or more of the top retrieved documents is non-relevant (which could introduce noise in feedback process), our method doesn't just use these documents as it iteratively explores them and the final set of feedback terms are taken from the converged set. Our experiments with the tf.idf selection method also proved this hypothesis. P@10 tells us how many in the top 10 retrieved documents are relevant and nDCG@10 tells how close we did compared to the ideal ranking if we consider the top 10 documents. We used the top N (= 10) documents for feedback in every pass and as P@10, nDCG@10 have improved, this shows that our method generates fewer non-relevant documents and less noisy feedback terms. Our assumption was that top-ranked documents describe the important aspects instead of being relevant when we used term-level diversification. This hypothesis is also supported by the improvement in the MAP values.

7. CONCLUSIONS

In this paper, we presented an iterative algorithm which does multiple passes of PRF. We experimentally discovered a new property of convergence that helped us to select better feedback terms, thereby improving the retrieval effectiveness. We also conclude that the knowledge of initial set of top documents is not necessarily sufficient to generate high quality feedback terms. The technique also proved effective to avoid noisy terms which is a drawback in PRF. Our approach is efficient and can be easily parallelized to make it more useful for online setting.

There are several possible avenues along which our work could be extended in future. We can cluster the terms from the converged set and then use it for diversification. Our technique may prove effective for query classification tasks. We are also interested in using other families of feedback methods in our iterative search.

8. ACKNOWLEDGMENTS

This work was supported in part by the Center for Intelligent Information Retrieval and in part by NSF IIS-1160894. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the sponsor.

9. REFERENCES

[1] V. Dang and W. B. Croft. Diversity by proportionality: An election-based approach to search result diversification. In *35th International ACM SIGIR Conference on Research and Development in IR*, SIGIR '12, pages 65–74, NY, USA, 2012. ACM.

[2] D. Harman and C. Buckley. The nrrc reliable information access (ria) workshop. In *ACM SIGIR Conference*, '04, pages 528–529, NY, USA, 2004. ACM.

[3] J. He, M. Li, Z. Li, H.-J. Zhang, H. Tong, and C. Zhang. *Advances in Multimedia Information Processing - PCM 2004, Part II*, chapter PRF Based on Iterative Probabilistic One-Class SVMs in Web Image Retrieval, pages 213–220. Springer Berlin Heidelberg, Berlin, Heidelberg, 2005.

[4] G. Kumaran and J. Allan. Simple questions to improve pseudo-relevance feedback results. In *29th Annual International ACM SIGIR Conference on Research and Development in IR*, SIGIR '06, pages 661–662, NY, USA, 2006. ACM.

[5] O. Kurland, L. Lee, and C. Domshlak. Better than the real thing? iterative pseudo-query processing using cluster-based language models. *CoRR*, abs/cs/0601046, 2006.

[6] V. Lavrenko and W. B. Croft. Relevance based language models. In *24th Annual International ACM SIGIR Conference on Research and Development in IR*, SIGIR '01, pages 120–127, NY, USA, 2001. ACM.

[7] D. Lawrie, W. B. Croft, and A. Rosenberg. Finding topic words for hierarchical summarization. In *Proceedings of SIGIR 01*, pages 349–357, 2001.

[8] Y. Lv, C. Zhai, and W. Chen. A boosting approach to improving pseudo-relevance feedback. In *34th International ACM SIGIR Conference on Research and Development in IR*, SIGIR '11, pages 165–174, NY, USA, 2011. ACM.

[9] M. Mitra, A. Singhal, and C. Buckley. Improving automatic query expansion. In *21st Annual International ACM SIGIR Conference on Research and Development in IR*, SIGIR '98, pages 206–214, New York, NY, USA, 1998. ACM.

[10] D. Petkova and W. B. Croft. Hierarchical language models for expert finding in enterprise corpora. In *18th International Conference on Tools with AI*, pages 599–606, 2006.

[11] K. Raman, R. Udupa, P. Bhattacharya, and A. Bhole. On improving pseudo-relevance feedback using pseudo-irrelevant documents. In *ECIR 2010*, 2010.

[12] S. E. Robertson and K. S. Jones. Relevance weighting of search terms. *JASIS*, 27(3):129–146, 1976.

[13] S. E. Robertson, S. Walker, S. Jones, M. Hancock-beaulieu, and M. Gatford. Okapi at TREC3. In *Text REtrieval Conference*, 1994.

[14] G. Salton, J. Allan, and A. Singhal. Automatic text decomposition and structuring. *Information Processing & Management*, 32(2):127–138, 1996.

[15] G. Salton and C. Buckley. Improving retrieval performance by relevance feedback. *Journal of the American Society for Information Science*, 41:288–297, 1990.

[16] R. L. Santos, C. Macdonald, and I. Ounis. Exploiting query reformulations for web search result diversification. In *19th International Conference*, WWW '10, pages 881–890, NY, USA, 2010. ACM.

[17] J. Xu and W. B. Croft. Query expansion using local and global document analysis. In *19th Annual International ACM SIGIR Conference on Research and Development in IR*, SIGIR '96, pages 4–11, NY, USA, 1996. ACM.

[18] C. Zhai and J. Lafferty. Model-based feedback in the language modeling approach to information retrieval. In *Tenth International Conference*, CIKM '01, pages 403–410, NY, USA, 2001. ACM.