Utilizing Focused Relevance Feedback

Elinor Brondwine elinor@tx.technion.ac.il
Anna Shtok annabel@tx.technion.ac.il
Oren Kurland kurland@ie.technion.ac.il
Faculty of Industrial Engineering and Management
Technion, Israel

ABSTRACT

We present a novel study of ad hoc retrieval methods utilizing document-level relevance feedback and/or *focused relevance feedback*; namely, passages marked as (non-)relevant. The first method uses a novel mixture model that integrates relevant and non-relevant information at the language model level. The second method fuses retrieval scores produced by using relevant and non-relevant information separately. Empirical exploration attests to the merits of our methods, and sheds light on the effectiveness of using and integrating relevance feedback for textual units of varying granularities.

Keywords: focused relevance feedback

1. INTRODUCTION

Most previous work on using relevance feedback for ad hoc (query-based) document retrieval has focused on utilizing feedback provided at the document level. Utilizing information induced from relevant documents can significantly improve retrieval effectiveness [12, 13]. The effective utilization of non-relevant documents, on the other hand, has been demonstrated mainly for very difficult queries [16, 7].

Relevant documents can also contain non-relevant information. Thus, utilizing *focused relevance feedback*, that is, feedback for passages in relevant documents, can be of merit. For example, using information induced from relevant passages can improve retrieval effectiveness [14].

We present a study of methods that utilize positive and/or negative relevance feedback for documents and/or for passages. Our first method uses a novel mixture model that integrates, at the language model level, information induced from relevant and non-relevant units (documents and/or passages). Our second method fuses retrieval scores attained by using, separately, relevant and non-relevant units.

Empirical evaluation sheds light on the effectiveness of using information induced from relevant and non-relevant units of different granularities and their integration. For example, the best performance of our methods was attained using information induced from relevant passages and non-relevant documents.

Our main novel contribution is the development of methods that (i) use relevance feedback at both document and passage levels, and/or (ii) utilize non-relevant passages.

SIGIR '16, July 17-21, 2016, Pisa, Italy

DOI: http://dx.doi.org/10.1145/2911451.2914695

2. RELATED WORK

Methods using information induced from both relevant and non-relevant documents emphasize terms that appear in the former and downplay the importance of those appearing in the latter (e.g., [6, 12, 13, 16]). A similar principle, although implemented using different techniques, is applied in our methods that utilize also (non-)relevant passages.

Findings about the merits of using information induced from non-relevant documents have largely been inconclusive [6, 5, 13]. Notable exceptions are work on addressing very difficult queries [16, 7] and on the document routing task [15]. In contrast to our work, information induced from (non-)relevant passages in relevant documents was not used.

Shen and Zhai [14] showed the merits of using relevant (but not non-relevant) passages in the mixture model [17]. We extend the mixture model by using both relevant and non-relevant feedback units (documents and/or passages). Automatically identifying (non-)effective passages in (non-)relevant documents and using these for retrieval has shown no merit [11, 9]. In contrast, our methods utilize true relevance feedback for passages and documents.

A mixture model, different than ours, was used to induce a query model from *pseudo* relevant documents using non-relevant documents [10]. In contrast to our work, relevant documents and passage-level feedback were not used.

3. RETRIEVAL FRAMEWORK

Let D_{init} denote an initial list of documents retrieved from corpus C in response to query q by some retrieval method. Suppose that relevance feedback is provided for documents in D_F ($\subset D_{init}$); specifically, R_d and NR_d are the sets of relevant and non-relevant documents in D_F, respectively.

Relevant documents can also contain non relevant information. Accordingly, we further assume that *focused relevance feedback* is provided for relevant documents. Namely, non-overlapping variable length passages of documents in R_d are marked as relevant to q; unmarked passages are considered non-relevant. We concatenate all relevant and non-relevant passages in each relevant document into a single *relevant pseudo passage* and a single *non-relevant pseudo passage*, respectively; the order of concatenation has no effect since we use unigram language models that assume term

independence. R_p and NR_p are the sets of pseudo relevant and non-relevant passages in documents in R_d, henceforth simply referred to as relevant and non-relevant passages, respectively. Herein, R refers to the set of either relevant documents R_d or relevant passages R_p; NR is the set of non-relevant documents (NR_d) or non-relevant passages (NR_p).

Below we present two retrieval methods, based on unigram language models, that utilize the relevance feedback. The maximum likelihood estimate (MLE) of term w with respect to a set S of texts is $p_S^{MLE}(w) \stackrel{def}{=} \frac{\sum_{x \in S} c(w,x)}{\sum_{x \in S} \sum_{w' \in x} c(w',x)}$; $c(w,x)$ is the count of w in text x. We use $p_x^{Dir}(w)$ to denote the probability assigned to term w by a Dirichlet smoothed unigram language model induced from text x [17]. The similarity between two language models, $p_y(\cdot)$ and $p_z(\cdot)$, is measured using cross entropy, where higher values correspond to decreased similarity:

$$CE(p_y(\cdot) \,||\, p_z(\cdot)) = -\sum_w p_y(w) \log p_z(w). \quad (1)$$

3.1 Distillation model

The goal of our first method is to "distill" the aspects most relevant to the information need from the feedback. For example, the premise of the mixture model [17] is that terms in relevant documents are generated either by a relevance topic language model or by the corpus language model. We generalize the mixture model to utilize both relevant and non-relevant feedback units (documents and/or passages).

We assume that terms in units in R (i.e., all relevant documents or all relevant passages) are generated by a mixture of (i) the relevance topic model, $p_{rel}(\cdot)$, which we want to estimate; (ii) the corpus language model, $p_C^{MLE}(\cdot)$, which is assumed to represent a *general* non-relevant document; and, (iii) a *query-specific* irrelevance topic model, $p_{NR}^{MLE}(\cdot)$, induced from the non-relevant units (NR). Following work on using negative relevance feedback, if w is a query term we set $p_{NR}^{MLE}(w) \stackrel{def}{=} 0$ and re-normalize the probabilities [16].

We estimate $p_{rel}(\cdot)$ by using the EM algorithm to maximize the log likelihood of units in R:

$$\sum_{x \in R} \sum_w c(w,x) \log \big((1-\lambda_1-\lambda_2) p_{rel}(w) + \lambda_1 p_{NR}^{MLE}(w) + \lambda_2 p_C^{MLE}(w)\big); \quad (2)$$

λ_1 and λ_2 are free parameters. As is common in work on query language models [17, 1], we interpolate $p_{rel}(\cdot)$ with the original query model:

$$p_{distill}(w) \stackrel{def}{=} \lambda_q p_q^{MLE}(w) + (1-\lambda_q) p_{rel}(w); \quad (3)$$

λ_q is a free parameter. We then rank the documents in the corpus using $-CE(p_{distill}(\cdot)||p_d^{Dir}(\cdot))$.

We instantiate Equation 2 using R ($\in \{R_d, R_p\}$) and NR ($\in \{NR_d, NR_p\}$). The resultant four models attained from Equation 3 are denoted **Distill(R,NR)**.

3.2 Score-based fusion

The second retrieval model is based on the principle that documents similar to the relevant units and dissimilar from the non-relevant units should be rewarded. Specifically, we apply a two-step approach inspired by work on using *only* the query and non-relevant documents [16]. First, a relevance topic model, $p_r(\cdot)$, is induced from the relevant units in R ($\in \{R_d, R_p\}$) using *some approach*. Then, the document corpus is ranked using $-CE(p_r(\cdot)||p_d^{Dir}(\cdot))$. Second, the top n documents are *re-ranked* by the similarity of their language models with $p_r(\cdot)$ and dissimilarity from the language models induced from non-relevant units in NR. Formally, documents d are ranked in descending order of the following score-based fusion:

$$-\alpha CE(p_r(\cdot)||p_d^{Dir}(\cdot)) + (1-\alpha) \min_{x \in NR} CE(p_x^{MLE}(\cdot)||p_d^{Dir}(\cdot)); \quad (4)$$

α is a free parameter. As in the distillation model, for query term w and non-relevant unit x ($\in NR$): $p_x^{MLE}(w) \stackrel{def}{=} 0$ and the probabilities are re-normalized[1].

Various methods can be used to induce $p_r(\cdot)$ from a set of relevant units, R ($\in \{R_d, R_p\}$). We use the standard mixture model [17] which is a special case of our distillation model from Equation 3 when setting $\lambda_1 = 0$ in Equation 2.[2] Equation 4 is then instantiated using a choice of NR ($\in \{NR_d, NR_p\}$). The four resultant score-based fusion methods are denoted **SF(R,NR)**.

4. EXPERIMENTAL SETUP

For experiments we used the INEX corpus[3] which contains 2,666,190 Wikipedia articles. We used the 120 queries from the ad hoc tracks of 2009 and 2010 for which binary document-level and (arbitrary-length) passage-level relevance judgments are available; unmarked text in relevant documents is considered non-relevant [2]. The average number of relevant documents per query is 86. The average percentage of relevant text in a relevant document is 41.5%; i.e., most text in relevant documents does not pertain to the query. We re-visit this important point below.

Krovetz stemming was applied to documents and queries and stopwords on the INQUERY list were removed. Indri 5.3 (http://www.lemurproject.org/indri) was used for experiments. The initial ranking from which D_{init} is derived is induced using standard language-model-based retrieval [17]: document d is scored by $-CE(p_q^{MLE}(\cdot)||p_d^{Dir}(\cdot))$ (see Equation 1). The Dirichlet smoothing parameter in document language models, μ, was set to 1000 in all methods [18].

The document feedback set, D_F, contains $2k$ documents: the k highest ranked relevant documents (R_d) and the k highest ranked non-relevant documents (NR_d) in D_{init}; $k \in \{1, 2, \ldots, 5\}$; each value entails an experimental setting. The goal was to ameliorate across-query effects that are due to varying numbers of (non-)relevant documents at top ranks.

Mean average precision at cutoff 1000 (MAP) serves as the retrieval evaluation measure. Two evaluation paradigms were employed: **standard** (regular) and **residual** collection. In the residual paradigm [3], all documents in D_F

[1]Equation 4 is conceptually reminiscent of the MultiNeg method from [16] that utilizes the query and non-relevant documents for re-ranking. Experiments — numbers are omitted due to space considerations — reveal the following. The approach in Equation 4 yields better performance in our setting when using min rather than average. The approach is also superior to using a single model, $p_{NR}^{MLE}(\cdot)$, induced from the non-relevant units; cf., the SingleNeg method [16].

[2]We found that using relevance model #3 (RM3) [8, 1] results in similar conclusions to those we present below. Actual results are omitted due to space considerations.

[3]http://www.mpi-inf.mpg.de/departments/databases-and-information-systems/software/inex

were removed from result lists and relevance judgments files; MAP is measured on result lists of 1000 documents. Statistically significant performance differences are determined using the paired two-tailed t-test with $p < 0.05$.

Free-parameter values are set using leave-one-out cross-validation performed over queries per experimental setting; MAP is the optimization measure. The value ranges are as follows: λ_q (Equation 3) is in $\{0.2, 0.5, 0.8\}$; λ_1 and λ_2 (Equation 2) are in $\{0, 0.1, 0.5, 0.9\}$; $\lambda_1 + \lambda_2 < 1$; α (Equation 4) is in $\{0, 0.2, \ldots, 1\}$; the number of documents re-ranked in the score-based fusion method, n, is set to 1000. As is common [17, 1], language models induced using relevance feedback are clipped to ν ($\in \{10, 25, 50\}$) terms.

As noted, the standard mixture model [17], henceforth MM, is a special instance of our distillation model when setting $\lambda_1 = 0$ in Equation 2 (i.e., non-relevant units are not used) and applying Equation 3. MM is also used to induce the relevance topic model, $p_r(\cdot)$, used in Equation 4. Hence, we use **MM(R_d)** which utilizes the relevant documents (R_d) and **MM(R_p)** (also used in [14]) which utilizes relevant passages (R_p) as reference comparisons. The EM algorithm used in the mixture and distillation models converged in 13-14 iterations.

5. EXPERIMENTAL RESULTS

Figure 1 depicts the performance results. We see that the more feedback documents are used (i.e., higher k), the more effective the retrieval. Specifically, for the residual evaluation paradigm (Figures 1(c) and 1(d)), where MAP values decrease with increasing k due to removing the given relevant documents from *all* rankings [4], the relative performance improvements of the feedback-based methods with respect to the initial ranking increase as a function of k.

Relevant units. Figure 1 shows that in *all* cases, using relevant passages ($R = R_p$) yields better performance than using relevant documents ($R = R_d$): compare a solid curve with white markers ($R = R_p$) to a dotted curve with gray markers of the same type ($R = R_d$). This finding can be attributed to the fact that relevant documents contain much non-relevant information as mentioned in Section 4.

The distillation model. Figures 1(a) and 1(c) show that in comparison to MM(R), which does not utilize non-relevant units, and regardless of the choice of R, using also non-relevant documents in our distillation model improves retrieval effectiveness in a vast majority of cases. For the standard evaluation, the distillation method yields improvements in 9 out of 10 cases (Distill(R_d,NR_d) vs. MM(R_d) and Distill(R_p,NR_d) vs. MM(R_p) over 5 values of k); in 7 cases the improvements are statistically significant. For the residual evaluation, effectiveness is improved in 7 out of 10 cases with 4 improvements being statistically significant.

The effectiveness of using $NR = NR_p$ depends on the relevant units used. Distill(R_d,NR_p) outperforms MM(R_d) in all cases (often statistically significantly) for both evaluation paradigms. Moreover, Distill(R_d,NR_p) outperforms Distill(R_d,NR_d) in all cases for the residual evaluation, although few improvements are statistically significant. The merits of using non-relevant passages ($NR = NR_p$) to distill a relevance topic model from relevant documents can be attributed to the fact that relevant documents contain much non-query-pertaining text (see Section 4). However, using the relevant passages alone, MM(R_p), is more effective than using relevant documents and non-relevant passages, Distill(R_d,NR_p), and is as effective as using non-relevant passages in addition to relevant passages, Distill(R_p,NR_p). That is, using non-relevant passages to distill a relevance topic model from either relevant documents or relevant passages has no merit over using only the relevant passages.

Distill(R_p,NR_d) is the most effective distillation model in a vast majority of cases; most improvements over other distillation models and the mixture models are statistically significant for both evaluation paradigms. Thus, in contrast to non-relevant passages in relevant documents, non-relevant documents can be effectively used to distill a relevance topic model from relevant passages[4].

The score-based fusion model. The performance of the score-based fusion model (Equation 4) is presented in Figures 1(b) and 1(d). The curves of SF(R,NR_p) (almost) coincide with the curves of MM(R) regardless of the choice of R. This means that using the similarity of a document to non-relevant passages has little merit. In contrast, SF(R,NR_d) outperforms MM(R), regardless of the choice of R. The improvements are statistically significant for all 10 cases (5 values of k $\times$ 2 choices of R) for the standard evaluation, and in 4 out of 10 cases for the residual evaluation.

Overall, the most effective score-based fusion model for both evaluation paradigms is SF(R_p,NR_d). The improvements it posts over the other methods (specifically, MM) are statistically significant in a vast majority of the cases[5].

6. SUMMARY

Our distillation and score-based fusion methods use relevance feedback for documents and passages in different ways. The distillation model utilizes both relevant and non-relevant units in a mixture model to rank the entire corpus. The score-based fusion model re-ranks a list retrieved using information induced only from relevant units by using, in addition, dissimilarities with non-relevant units.

Despite these differences, the conclusions regarding the merits of using the different types of (non-)relevant units are similar in most cases. That is, using relevant passages is superior to using relevant documents regardless of the non-relevant units used. Yet, using non-relevant documents in addition to relevant passages is of much merit and results in the best performance for both methods. A noticeable difference between the two methods is the effectiveness of using non-relevant passages in addition to relevant documents in the distillation model. No such merits were observed for the score-based fusion method.

Acknowledgments. We thank the reviewers for their comments. This paper is based upon work supported in part by the German Research Foundation (DFG) via the German-Israeli Project Cooperation (DIP, grant DA 1600/1-1), the Israel Science Foundation under grant no. 433/12, and the Technion-Microsoft Electronic Commerce Research Center.

[4]Extending the distillation model to use both NR_p and NR_d showed merit over using each alone when relevant documents are used but not when relevant passages are used.

[5]We found that a score-based fusion method that extends Equation 4 by using both non-relevant documents and non-relevant passages does not statistically significantly outperform SF(R_d,NR_d) and SF(R_p,NR_d).

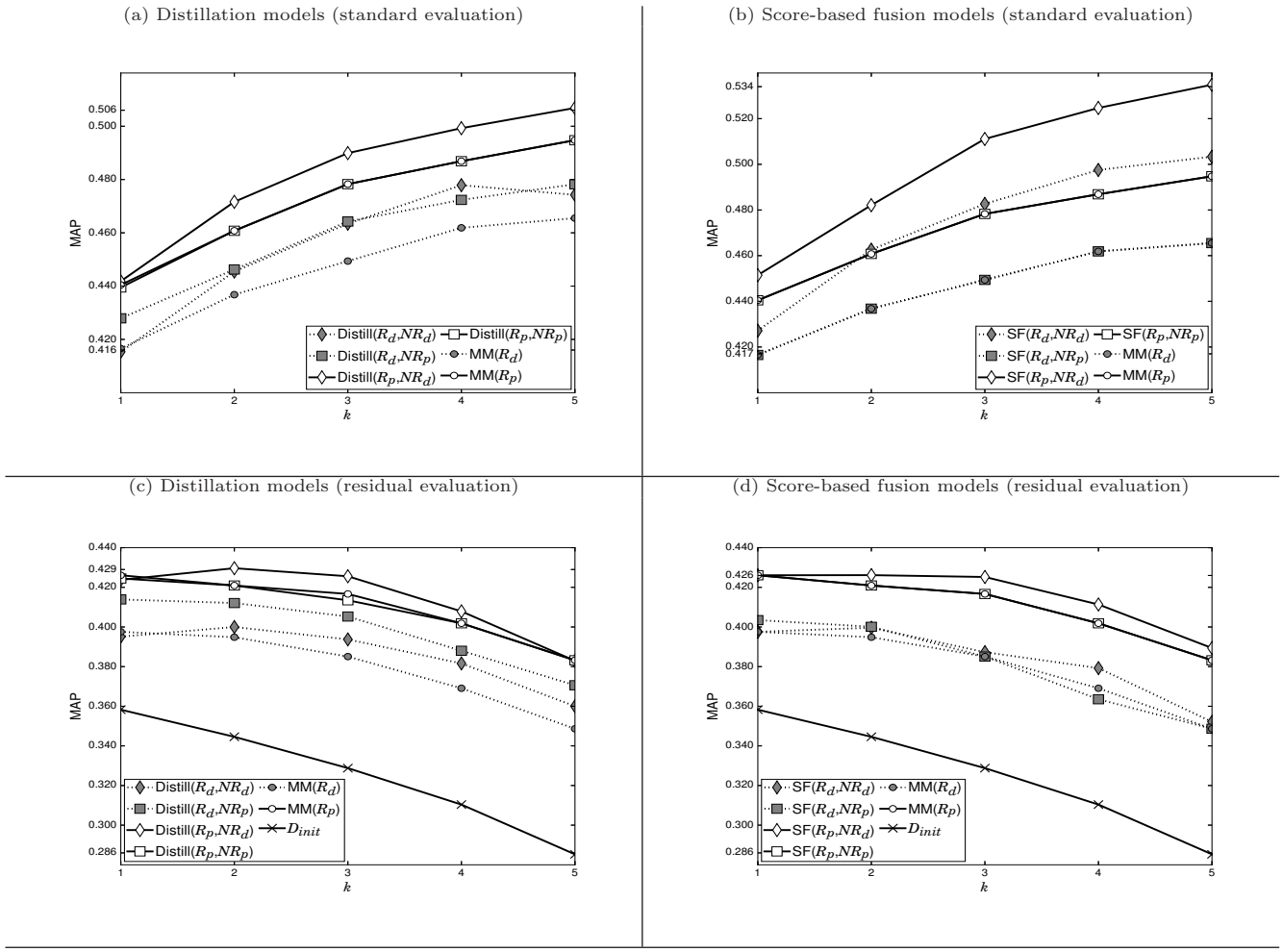

Figure 1: MAP as a function of k. The performance of MM(R) is presented for reference. The MAP of the initial result list (D_{init}) in the standard evaluation paradigm, which does not depend on k, is 0.368; the MAP in the residual evaluation is displayed. The color of the markers (white or gray) and curve style indicate the relevant units used (R_d or R_p). The type of the markers indicates the non-relevant units (NR_d or NR_p).

7. REFERENCES

[1] N. Abdul-Jaleel, J. Allan, W. B. Croft, F. Diaz, L. Larkey, X. Li, M. D. Smucker, and C. Wade. UMASS at TREC 2004 — novelty and hard. In *Proc. of TREC-13*, 2004.

[2] P. Arvola, S. Geva, J. Kamps, R. Schenkel, A. Trotman, and J. Vainio. Overview of the inex 2010 ad hoc track. In *Comparative Evaluation of Focused Retrieval - INEX*, pages 1–32. 2010.

[3] C. Buckley and S. Robertson. Relevance feedback track overview: TREC 2008. In *Proc. of TREC-17*, 2008.

[4] C. Cirillo, Y. K. Chang, and J. Razon. Evaluation of feedback retrieval using modified freezing, residual collection and test and control groups. *The SMART retrieval system-experiments in automatic document processing*, pages 355–370, 1971.

[5] M. D. Dunlop. The effect of accessing nonmatching documents on relevance feedback. *ACM Transactions on Information Systems*, 15(2):137–153, 1997.

[6] E. Ide. New experiments in relevance feedback. *The SMART retrieval system*, pages 337–354, 1971.

[7] M. Karimzadehgan and C. Zhai. Improving retrieval accuracy of difficult queries through generalizing negative document language models. In *Proc. of SIGIR*, pages 27–36, 2011.

[8] V. Lavrenko and W. B. Croft. Relevance-based language models. In *Proc. of SIGIR*, pages 120–127, 2001.

[9] Y. Li, X. Tao, A. Algarni, and S. Wu. Mining specific and general features in both positive and negative relevance feedback. In *Proc. of TREC*, 2009.

[10] Y. H. Peng Zhang and D. Song. Approximating true relevance distribution from a mixture model based on irrelevance data. In *Proc. of SIGIR*, pages 107–114, 2009.

[11] S. E. Robertson, H. Zaragoza, and M. J. Taylor. Microsoft cambridge at TREC-12: HARD track. In *Proc. of TREC*, pages 418–425, 2003.

[12] J. J. Rocchio. Relevance feedback in information retrieval. In G. Salton, editor, *The SMART Retrieval System: Experiments in Automatic Document Processing*, pages 313–323. Prentice Hall, 1971.

[13] I. Ruthven and M. Lalmas. A survey on the use of relevance feedback for information access systems. *Knowledge Engineering Review*, 18(2):95–145, 2003.

[14] X. Shen and C. Zhai. Active feedback-UIUC TREC-2003 HARD experiments. In *Proc. of TREC*, pages 662–666, 2003.

[15] A. Singhal, M. Mitra, and C. Buckley. Learning routing queries in a query zone. In *Proc. of SIGIR*, pages 25–32, 1997.

[16] X. Wang, H. Fang, and C. Zhai. A study of methods for negative relevance feedback. In *Proc. of SIGIR*, pages 219–226, 2008.

[17] C. Zhai and J. D. Lafferty. Model-based feedback in the language modeling approach to information retrieval. In *Proc. of CIKM*, pages 403–410, 2001.

[18] C. Zhai and J. D. Lafferty. A study of smoothing methods for language models applied to ad hoc information retrieval. In *Proc. of SIGIR*, pages 334–342, 2001.