Utilizing Minimal Relevance Feedback for Ad Hoc Retrieval

Eyal Krikon
krikon@tx.technion.ac.il

Oren Kurland
kurland@ie.technion.ac.il

Faculty of Industrial Engineering and Management
Technion — Israel Institute of Technology
Haifa 32000, Israel

ABSTRACT

Using relevance feedback can significantly improve (ad hoc) retrieval effectiveness. Yet, if little feedback is available, effectively exploiting it is a challenge. To that end, we present a novel approach that utilizes document *passages*. Empirical evaluation demonstrates the merits of the approach.

Categories and Subject Descriptors: H.3.3 [Information Search and Retrieval]: Retrieval models; Relevance feedback

General Terms: Algorithms, Experimentation

Keywords: language models, passages, relevance feedback

1. INTRODUCTION

The effectiveness of ad hoc (query-based) retrieval can significantly improve if relevance feedback is available and exploited. Often, terms that are common in the relevant documents — but not very common in the corpus — are used for query expansion [7]. Since a document can be deemed relevant even if only a small part of it contains query-pertaining information, utilizing the commonalities between relevant documents is important. Indeed, using very few relevant documents, which reflects practical settings wherein relevance judgments are scarce, can sometimes fall short [8].

We present a novel retrieval method that addresses the challenge of utilizing very little relevance feedback; specifically, a single relevant document. This is a *query-by-example* task performed in a query-dependent context. We tackle the uncertainty about what makes the document relevant, and more specifically, which terms can represent the underlying information need, by using document *passages* in two capacities. First, rather than treat the relevant document as one unit, its passages are used so as to better focus on the document parts that presumably contain query-pertaining information. Second, to enrich the basis for performing query expansion, passages from documents in an initially retrieved list that are similar to the relevant document (and to its passages) are used as pseudo relevant units. Empirical evaluation demonstrates the merits of our method with respect to various reference comparisons.

2. RETRIEVAL FRAMEWORK

Suppose that some search algorithm is employed in response to query q so as to rank documents d in corpus $\mathcal{D}$ by their presumed relevance to information need I expressed by q. Let $\mathcal{D}_{\text{init}}$ denote the list of documents that are the highest ranked by the induced initial ranking (**initial** in short), and d_{rel} denote the highest ranked relevant document in $\mathcal{D}_{\text{init}}$.

Naturally, we can employ a relevance-feedback-based (**RF**) method using q and d_{rel} to create an expanded query form [7]. Since only a single relevant document is available, we opt to enrich the basis for creating this form; e.g., by employing a pseudo-feedback-based (**PRF**) approach that regards all documents in $\mathcal{D}_{\text{init}}$ as pseudo relevant. Indeed, recent work [6, 10] has demonstrated the potential merits of integrating true and pseudo relevance feedback (**RF+PRF**).

A potential drawback of such approaches is that they often treat a document as one unit. For example, it could be that all the information in the relevant document d_{rel} that pertains to I is confined to a single short *passage* g of d_{rel}. Moreover, the terms in g need not be the most dominant (e.g., frequent) in d_{rel}, and hence, might not be assigned with high enough weight by the relevance-feedback-based method. The same observations hold for documents in $\mathcal{D}_{\text{init}} \setminus \{d_{rel}\}$ for which no (positive) feedback is available. Thus, we use the passages in documents in $\mathcal{D}_{\text{init}}$ (the set of which is denoted G) as pseudo relevant units for performing query expansion. To that end, we estimate $p(g|I)$, the probability that passage g ($\in G$) contains information pertaining to I.

As q and d_{rel} are the only "signals" about I, we estimate $p(g|I)$ by $\lambda p(g|d_{rel}) + (1-\lambda)p(g|q)$; λ is a free parameter. Assuming a uniform prior for passages, we can score g by $\lambda \frac{p(d_{rel}|g)}{\sum_{g' \in G} p(d_{rel}|g')} + (1-\lambda)\frac{p(q|g)}{\sum_{g' \in G} p(q|g')}$. In addition, $\sum_{g'' \in d_{rel}} p(q|g'')p(g''|g)$ is used as an estimate for $p(q|g)$. The goal of using this estimate is addressing the potential vocabulary mismatch between passage g (which could be part of d_{rel}) that pertains to I and q by using d_{rel}'s passages (g'') as proxies for g; the impact of g'' is based on its "direct match" with q, which potentially reflects to some extent the likelihood that g'' contains information pertaining to I as it is part of d_{rel}.[1] Using the estimates just described results in our **PsgF** method that scores g ($\in G$) by:

$$S(g) \overset{def}{=} \frac{\lambda p(d_{rel}|g)}{\sum_{g' \in G} p(d_{rel}|g')} + \frac{(1-\lambda)\sum_{g'' \in d_{rel}} p(q|g'')p(g''|g)}{\sum_{g' \in G}\sum_{g'' \in d_{rel}} p(q|g'')p(g''|g')}. \tag{1}$$

The highest scoring passages are used for query expansion.

3. EVALUATION

[1] Inter-passage relations were also utilized, for example, for text summarization [3], finding documents similar to a given document [9], and re-ranking search results [4].

SIGIR'11, July 24–28, 2011, Beijing, China.
ACM 978-1-4503-0757-4/11/07.

Experiments were conducted with TREC corpora (disks, queries): AP (1-3, 51-150), TREC8 (4-5, 401-450), WSJ (1-2, 151-200), WT10G (WT10G, 451-550). Titles of TREC topics served for queries. Tokenization, Porter stemming and language model induction were performed using the Lemur toolkit (www.lemurproject.org). Half overlapping 150-terms windows in documents serve for passages [4].

We use $\exp\left(-D\left(p_x^{Dir[0]}(\cdot)\,\middle\|\,p_y^{Dir[\mu]}(\cdot)\right)\right)$ [4] for $p(x|y)$ in Equation 1; D is the KL divergence; $p_z^{Dir[\mu]}(\cdot)$ is the Dirichlet-smoothed unigram language model induced from z with smoothing parameter μ.

The initial list, $\mathcal{D}_{\text{init}}$, is set to the 50 highest ranked documents d in the corpus by $p(q|d)$ with μ set to optimize MAP (at 1000) so as to have an initial list of reasonable quality.

We use relevance model number 3 (RM3) [1] as a query expansion method that assigns the probability $\beta p_q^{JM[0]}(t) + (1-\beta)\sum_{x \in X} p_x^{JM[\alpha]}(t)\,W(x)$ to term t; X is a set of passages or documents; $p_x^{JM[\alpha]}(\cdot)$ is the Jelinek-Mercer smoothed language model induced from x with smoothing parameter α; β is a free parameter; and, $W(x)$ is x's weight ($\sum_{x \in X} W(x) = 1$). Methods that use d_{rel} assign it with either a weight 1 (RF) or ρ (RF+PRF), which is a free parameter; in the latter case, the weight $(1-\rho)$ is distributed among the pseudo-relevant documents. For any method, except PsgF, that uses pseudo-relevant units f ($\in F \subseteq X$), $\frac{p(q|f)}{\sum_{f' \in F} p(q|f')}$ serves as f's relative weight; for PsgF the (normalized) scores assigned in Equation 1 are used. The set X in RM3 is set to (i) d_{rel} for RF, (ii) k highest ranked documents in $\mathcal{D}_{\text{init}}$ for PRF, (iii) d_{rel} and the $k-1$ highest ranked documents d in $\mathcal{D}_{\text{init}}$ ($d \neq d_{rel}$) for RF+PRF, and (iv) the k highest ranked passages by Equation 1 for PsgF.

We use two additional passage-based baselines that utilize RM3. **TopPDrel** sets $X \overset{def}{=} \{g^*\}$ ($W(g^*) \overset{def}{=} 1$); g^* is d_{rel}'s passage with the highest $p(q|g)$ [2]. **AllPDrel** uses all d_{rel}'s passages ($X \overset{def}{=} \{g | g \in d_{rel}\}$) as pseudo relevant units.

For evaluation, d_{rel} is not considered (i.e., the residual corpus approach is employed); MAP(@1000) and p@5 are reported. The two-tailed paired t-test (95% confidence level) serves for determining statistically significant differences.

The following free-parameter-values' ranges are used: $\mu = 2000$ except in $p_d^{Dir[\mu]}(q)$ where we use the value used to create $\mathcal{D}_{\text{init}}$ to maintain consistency with the initial ranking; $k \in \{5, 10, 20, 30, 40, 50\}$; $\lambda \in \{0, 0.1, \ldots, 1\}$; $\rho \in \{0.2, 0.4, 0.6, 0.8\}$; the number of terms used by RM3 is set to $\{25, 50, 75, 100, 500, 1000, 5000, ALL\}$ ("ALL": all terms in the vocabulary); $\alpha \in \{0, 0.1, 0.3, 0.5, 0.7, 0.9\}$; $\beta \in \{0, 0.1, \ldots, 0.9\}$. To give the "best chance" to the reference comparisons RF, PRF, TopPDrel and AllPDrel, their free-parameter values are set to optimize MAP over all queries per corpus. For RF+PRF and PsgF, RM3's parameters are set to the values used by RF and AllPDrel, respectively; the other free parameters are set using leave-one-out cross validation performed over queries (MAP is the optimization criterion).

	AP		TREC8		WSJ		WT10G	
	MAP	p@5	MAP	p@5	MAP	p@5	MAP	p@5
initial	21.8	40.9	23.0	40.8	27.2	43.2	15.6	25.1
RF	32.3	61.1	26.2	54.7	35.3	58.4	23.2	41.3
PRF	30.0	50.8^{r}	26.2	44.9^{r}	33.8	50.4	18.0^{r}	29.2^{r}
RF+PRF	32.9^{p}	$\mathbf{63.4}^{rp}$	26.0	53.5^{p}	35.6	57.2	22.6^{p}	41.5^{p}
TopPDrel	29.3^{r+}	59.4^{p+}	25.2	54.7^{p}	34.8	$\mathbf{58.8}^{p}$	21.9^{p}	42.0^{p}
AllPDrel	32.3_{t}	63.0_{t}^{rp}	26.3_{t}	$\mathbf{55.5}^{p}$	35.5	56.8	$\mathbf{23.8}_{+t}^{rp}$	$\mathbf{43.3}^{p}$
PsgF	$\mathbf{33.4}_{ta}^{rp}$	61.9^{p}	$\mathbf{27.4}_{t}^{r+}$	55.1^{p}	$\mathbf{36.5}_{t}^{r}$	56.8	23.4_{t}^{p}	42.0^{p}

Table 1: Performance numbers; 'r', 'p', '+', 't' and 'a' mark statistically significant differences with RF, PRF, RF+PRF, TopPDrel and AllPDrel respectively. The best result in a column is boldfaced.

Results, conclusions, and future work. Table 1 shows that all methods outperform the initial ranking used to create $\mathcal{D}_{\text{init}}$. (Most of these improvements are statistically significant; hence, they are not marked to avoid cluttering.) The MAP performance of our PsgF method is superior (often, statistically significantly so) for most corpora to that of the reference comparisons. Specifically, PsgF always MAP-outperforms RF (often, statistically significantly), which is the standard approach of using d_{rel} as a whole unit. PsgF also outperforms RF+PRF in most relevant comparisons. As both methods integrate true and pseudo feedback, we see that using passages (by PsgF as opposed to RF+PRF) is of merit. We also see that TopPDrel is often inferior to AllPDrel and PsgF, which suggests that using several (weighted) passages rather than a single one is beneficial. Finally, except for WT10G, PsgF is MAP-superior to AllPDrel and posts more statistically significant MAP improvements over the other methods than AllPDrel does. This implies that using passages not in, as well as in, d_{rel} can be more effective than using just those in d_{rel}. However, in terms of p@5, PsgF is often outperformed by AllPDrel. Thus, integrating these two methods, along with utilizing term-proximity models [6, 5], is an interesting future venue.

Acknowledgments This paper is based upon work supported in part by the ISF under grant no. 557/09, and by IBM's SUR award. Any opinions, findings and conclusions or recommendations expressed here are the authors' and do not necessarily reflect those of the sponsors.

4. 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 *Proceedings of TREC-13*, 2004.

[2] J. Allan. Relevance feedback with too much data. In *Proceedings of SIGIR*, pages 337–343, 1995.

[3] G. Erkan and D. R. Radev. LexPageRank: Prestige in multi-document text summarization. In *Proceedings of EMNLP*, pages 365–371, 2004. Poster.

[4] E. Krikon, O. Kurland, and M. Bendersky. Utilizing inter-passage and inter-document similarities for reranking search results. *ACM Transactions on Infortmation Systems*, 29(1):3, 2010.

[5] H. Lang, D. Metzler, B. Wang, and J.-T. Li. Improved latent concept expansion using hierarchical markov random fields. In *Proceedings of CIKM*, pages 249–258, 2010.

[6] M. Lease. Incorporating relevance and pseudo-relevance feedback in the markov random field model. In *Proceedings of TREC 2008*, 2008.

[7] 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.

[8] E. L. Terra and R. Warren. Poison pills: harmful relevant documents in feedback. In *Proceedgins of CIKM*, pages 319–320, 2005.

[9] X. Wan, J. Yang, and J. Xiao. Towards a unified approach to document similarity search using manifold-ranking of blocks. *Information Processing and Management*, 44(3):1032–1048, 2008.

[10] L. Zhao, C. Liang, and J. Callan. Extending relevance model for relevance feedback. In *Proceedings of TREC 2008*, 2008.