Utilizing inter-passage and inter-document similarities for re-ranking search results

Eyal Krikon[1] and Oren Kurland[1] and Michael Bendersky[2]
(1) Faculty of Industrial Engineering and Management, Technion — Israel Institute of Technology
(2) Center of Intelligent Information Retrieval, Dept. of Computer Science , University of Massachusetts, Amherst MA 01003
krikon@tx.technion.ac.il,kurland@ie.technion.ac.il,bemike@cs.umass.edu

ABSTRACT

We present a novel language-model-based approach to *re-ranking* an initially retrieved list so as to improve precision at top ranks. Our model integrates whole-document information with that induced from *passages*. Specifically, inter-passage, inter-document, and query-based similarities are integrated in our model. Empirical evaluation demonstrates the effectiveness of our approach.

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

General Terms: Algorithms, Experimentation

Keywords: re-ranking, passages, language models, centrality

1. INTRODUCTION

To obtain high precision at the very top ranks of a document-list returned in response to a query, researchers have proposed various *re-ranking* techniques that re-order documents in an initially retrieved list (e.g., [27, 11, 19, 8, 14, 15, 28]). An information source often utilized by these re-ranking methods is *inter-document* similarities. For example, documents that are similar to many other documents in the list, and hence, are considered as *central*, have higher chances of being relevant by the virtue of the way the list was created [14]; that is, in response to the query at hand.

An issue often not accounted for in these re-ranking approaches is that long and/or heterogeneous relevant documents could contain many parts (*passages*) that have no query-related information. While this issue is addressed by methods that use passage-based information for ranking *all* documents in the corpus [24, 6, 26, 10, 18], these methods do not exploit similarity relationships between documents, nor between their passages — a potentially rich source of information for re-ranking as noted above.

We present a novel language-model-based approach to re-ranking that leverages the strengths of the two research directions just described: utilizing inter-item similarities and exploiting passage-based information. Specifically, our model integrates query-similarity information induced from documents and passages with passage and document centrality information; the latter are induced from inter-passage and inter-document similarities, respectively.

Empirical evaluation demonstrates the effectiveness of our model; specifically, with respect to (i) a standard passage-based document retrieval method that does not utilize inter-passage and inter-document similarities, (ii) a re-ranking method that utilizes inter-document similarities but does not utilize passage-based information, and (iii) a state-of-the-art pseudo-feedback-based query expansion approach.

CIKM'09, November 2–6, 2009, Hong Kong, China.

2. RETRIEVAL FRAMEWORK

Let q, d and $\mathcal{D}$ denote a query, a document, and a corpus of documents, respectively. We use g to denote a passage, and write $g \in d$ if g is part of d. The model we present is not committed to any specific type of passages. To measure similarity between texts x and y, we use a language-model-based estimate $p_x(y)$ described in Section 4.

Our goal is to re-rank an initial list $\mathcal{D}_{\text{init}}$ ($\mathcal{D}_{\text{init}} \subset \mathcal{D}$) of documents, which was retrieved in response to q by some search algorithm, so as to improve precision at top ranks.

2.1 Passage-based document re-ranking

We take a probabilistic approach and rank document d in $\mathcal{D}_{\text{init}}$ by the probability $p(d|q)$ of its being relevant to the information need underlying q. Since the query is fixed, we can use the rank equivalence

$$p(d|q) \stackrel{rank}{=} p(q|d)p(d). \tag{1}$$

We interpret $p(q|d)$ as the probability that q can serve as a summary of d's content, or in other words, the probability of "generating" the terms in q from a model induced from d (cf., the language modeling approach to retrieval [23, 7])[1]; $p(d)$ is the prior probability of d being relevant.

Since passages are shorter than documents, and hence, are often considered as more focused (coherent) units, they can potentially aid in generating summaries that are more "informative" than those generated from whole documents. Indeed, it has long been acknowledged that passages can serve as effective proxies for estimating the document-query match ($p(q|d)$ in our case), especially for long and/or heterogeneous documents [24, 6, 26, 10]. Following this observation, and inspired by recent work on ranking *document*

[1]While it is convenient to use the term "generate" in reference to work on utilizing language models for IR, we do not assume an underlying generative theory as opposed to Lavrenko and Croft [17] and Lavrenko [16], inter alia.

clusters by using documents as proxies for clusters [13], we develop our passage-based document re-ranking model.

2.1.1 Model derivation

We use $p(g_i|d)$ to denote the probability that some passage g_i in the corpus is chosen as d's proxy for "query generation" ($\sum_{g_i} p(g_i|d) = 1$); $p(g_i)$ will be used to denote the prior probability of choosing g_i as a query-generator for any document. Using probability algebra, Equation 1 becomes

$$p(d|q) \stackrel{rank}{=} p(d) \sum_{g_i} p(q|d, g_i)p(g_i|d). \tag{2}$$

To estimate $p(q|d, g_i)$, the probability of generating q from d and g_i, we use a simple mixture of the probabilities of generating q from each [3]: $\lambda p(q|d)+(1-\lambda)p(q|g_i)$; λ is a free parameter. Using this estimate in Equation 2 and applying probability algebra yields the following ranking principle:

$$Score(d) \stackrel{def}{=} \lambda p(d)p(q|d) + (1-\lambda) \sum_{g_i} p(q|g_i)p(d|g_i)p(g_i). \tag{3}$$

Equation 3 scores d using a two-component mixture model. The first component is based on the probability of generating q directly from d and on the prior probability of d being relevant. The second is based on the probability of generating q from passages. The relative impact of passage g_i depends on its (i) prior probability of being a query-generator ($p(g_i)$), (ii) "association" with d ($p(d|g_i)$), and (iii) probability of generating q ($p(q|g_i)$). Indeed, if g_i is strongly associated with (e.g., textually similar to) d, and it is a-priori a good candidate for query generation, then it can serve as d's "faithful" proxy; yet, g_i can potentially be more effective than d for query generation by the virtue of being more focused.

To alleviate the computational cost of estimating Equation 3, we make the assumption that d's most effective proxies are its passages — i.e., that $p(d|g_i)$ is much larger for passages in d than for passages not in d. Consequently, we truncate the summation in Equation 3 to yield[2]:

$$Score(d) \stackrel{def}{=} \lambda p(d)p(q|d) + (1-\lambda) \sum_{g_i \in d} p(q|g_i)p(d|g_i)p(g_i). \tag{4}$$

Algorithm. The scoring function in Equation 4 is not committed to any specific paradigm of estimating probabilities. Following common practice in work on language models for IR [18, 7], we use the language-model-based estimate $p_x(y)$ for $p(y|x)$.

The remaining task for deriving a specific algorithm from Equation 4 is estimating the document and passage priors, $p(d)$ and $p(g)$, respectively. Note that insofar we have not used the fact that the documents we are ranking are those in the initially retrieved list $\mathcal{D}_{\text{init}}$. In fact, Equation 4 can be used to rank all documents in the corpus; if we were to do so, then the natural choice for the priors would be a uniform distribution, as is standard practice in the language modeling framework [7].

However, some previous work on re-ranking search results has demonstrated the merits of using a measure of the *centrality* of a document with respect to $\mathcal{D}_{\text{init}}$ as a *document-bias* [14]. Specifically, a document is considered central if it is similar to many other (central) documents in $\mathcal{D}_{\text{init}}$. The idea is that central documents reflect the context of $\mathcal{D}_{\text{init}}$, which was retrieved in response to the query, and, hence, have higher chances of being relevant [14, 15]. Following the same line of reasoning we hypothesize that *passage centrality* with respect to $\mathcal{D}_{\text{init}}$ is indicator for passage relevance, that is, for the passage "ability" to serve as an effective query generator. Specifically, we consider a passage of a document in $\mathcal{D}_{\text{init}}$ to be central to the extent it is similar to many other (central) passages of documents in $\mathcal{D}_{\text{init}}$.

To derive specific document and passage centrality measures, we use a previously-proposed method for document-centrality induction [14]. We construct a document-only graph from all documents in $\mathcal{D}_{\text{init}}$ and a passage-only graph from all passages of documents in $\mathcal{D}_{\text{init}}$. Edge-weights in these graphs represent inter-item similarities. We then use the PageRank [4] score of item x with respect to its ambient graph as its centrality value $Cent(x)$. For documents not in $\mathcal{D}_{\text{init}}$ and for passages that are not parts of documents in $\mathcal{D}_{\text{init}}$ we set $Cent(x) \stackrel{def}{=} 0$. Consequently, $Cent(d)$, which serves as the document bias $p(d)$, and $Cent(g)$, which serves as the passage bias $p(g)$, are probability distributions over all documents in the corpus, and over all passages of documents in the corpus, respectively[3].

Using the estimates described above we instantiate Equation 4 to produce our **PsgAidRank** re-ranking algorithm:

$$Score_{PsgAidRank}(d) \stackrel{def}{=} \tag{5}$$
$$\lambda Cent(d)p_d(q) + (1-\lambda) \sum_{g_i \in d} p_{g_i}(q)p_{g_i}(d)Cent(g_i).$$

We note that setting $\lambda = 1$ results in a recently-proposed re-ranking method [14], which utilizes only document centrality and document-query generation information.

3. RELATED WORK

The most commonly used methods for passage-based document retrieval utilize document-query and passage-query similarities (e.g.,[5, 6, 26, 10, 18, 3]), but in contrast to PsgAidRank, passage and document centrality are not utilized. We demonstrate the merits of PsgAidRank over one such common approach in Section 4.2.

There is a large body of work on using graph-based approaches to (re-)rank documents (e.g., see Kurland [12] for a survey, and also [2, 20]). We compare PsgAidRank's performance with that of some of these methods in Section 4.2. Centrality induced over passage-only graphs was also used for text summarization [9, 21] and sentence retrieval for question answering [22]; however, inter-document and document-passage similarities were not utilized.

Information induced from clusters of similar documents in the initial list was also utilized for re-ranking (e.g., [27,

[2]In general, the summation in Equation 4 for long documents contains more passages than for shorter documents. However, note that the relative contribution of passage g_i to the sum is weighted by its document association-strength $p(d|g_i)$. Indeed, normalizing the sum by the number of passages in a document has shown no performance merits.

[3]We use the term "bias" and not "prior" as these are not "true" prior-distributions, because of the virtue by which $\mathcal{D}_{\text{init}}$ was created (in response to the query). Yet, these biases form valid probability distributions by construction.

19, 15, 28]). We compare one such graph-based method [15] with PsgAidRank in Section 4.2.

Inter-passage similarities *within* a document were used to induce passage language models [3], and to devise discriminative passage-based document retrieval models [25]; the latter uses an initial standard passage-based document ranking, and can therefore potentially benefit from using PsgAidRank instead. Also, note that PsgAidRank uses inter-passage similarities both *across* and *within* documents.

Recent work on passage-based document retrieval [3] resembles ours in that it uses passages as proxies for documents. However, only a single passage from each document is used as its proxy, and document and passage centrality are not utilized.

4. EVALUATION

Language-model induction. Let $p_x^{Dir[\mu]}(\cdot)$ denote the unigram, Dirichlet-smoothed, language model induced from text x where μ is the smoothing parameter [29]. We adopt a previously-proposed estimate [14, 15]: $p_y(x) \stackrel{def}{=} \exp\left(-D\left(p_x^{Dir[0]}(\cdot) \,\middle|\middle|\, p_y^{Dir[\mu]}(\cdot)\right)\right)$, where D is the KL divergence. While the estimate does not constitute a probability distribution (as is the case for probabilities assigned by unigram language models to term sequences), normalizing it to produce such a distribution yields no performance merits.

4.1 Experimental setup

We conducted experiments on TREC collections that were used in some previous work on re-ranking [19, 14, 15]: (i) AP (disks 1-3, queries: 51-64, 66-150), (ii) TREC8 (disks 4,5 (-CR), queries: 401-450), and (iii) WSJ (disks 1-2, queries: 151-200). We use titles of TREC topics for queries. We applied tokenization and Porter stemming via the Lemur toolkit[4], which was also used for language-model induction.

As in some previous work on re-ranking [14, 15], we set $\mathcal{D}_{\text{init}}$, the list upon which re-ranking is performed, to be the 50 highest-ranked documents by an *initial ranking* induced over the corpus using $p_d(q)$ — i.e., a standard language model approach; the document language model smoothing parameter (μ) is set to a value optimizing MAP (at 1000) so as to yield an initial ranking of a reasonable quality.

The goal of re-ranking methods is to improve precision at top ranks. Therefore, we use the precision of the top 5 and 10 documents (p@5, p@10) for evaluation metrics. Statistically significant differences of performance are determined using the two-tailed Wilcoxon test at a 95% confidence level.

As in some prior work on graph-based re-ranking [14, 15, 2], we set the values of the free parameters of PsgAidRank so as to optimize p@5;[5] such performance optimization is employed for all methods that we consider, unless otherwise specified. We set λ, the interpolation parameter of PsgAidRank, to a value in $\{0, 0.1, \ldots, 1\}$. The centrality induction method that we use incorporates two free parameters [14]: the graph out-degree is set in *both* the document and passage graphs to α percent of the number of nodes in the graph, where $\alpha \in \{4, 8, 18, 38, 58, 78, 98\}$; PageRank's damping factor, δ, is set for both graphs to a value in {0.05,0.1,0.2,. . . ,0.9,0.95}. The document and passage language models smoothing parameter, μ, is set to 2000 [29] in all the methods we consider, except for the estimate $p_d(q)$ where we use the value chosen to create $\mathcal{D}_{\text{init}}$ so as to maintain consistency with the initial ranking.

For passages we use half-overlapping fixed windows of 150 terms that are marked prior to retrieval time, as these were shown to be effective for document retrieval [6, 25], specifically, in the language-model framework [18, 3].

[4]www.lemurproject.org

[5]If two parameter settings yield the same p@5, we choose the one *minimizing* p@10 so as to provide conservative estimates of performance.

4.2 Experimental results

Recall that the initial ranking used to create $\mathcal{D}_{\text{init}}$ is based on a language model approach ($p_d(q)$) wherein the smoothing parameter (μ) is set to optimize MAP. We therefore also compare PsgAidRank with *optimized baselines*, which use $p_d(q)$ to rank *all* documents in the corpus, and in which μ is set to optimize p@5 and p@10, independently. As can be seen in Table 1, PsgAidRank substantially outperforms both the initial ranking and the optimized baselines — often to a statistically significant degree.

We also compare the performance of PsgAidRank with that of a commonly-used passage-based document-retrieval approach (denoted *InterPsgDoc*), which in language model terms scores d by $\lambda p_d(q) + (1-\lambda)\max_{g_i \in d} p_{g_i}(q)$ [5, 6, 26, 3]. In contrast to PsgAidRank, InterPsgDoc does not utilize inter-document and inter-passage similarities. We use InterPsgDoc to re-rank the documents in $\mathcal{D}_{\text{init}}$. The parameter λ is set to a value in $\{0, 0.1, \ldots, 1\}$ so as to optimize p@5 performance. As can be seen in Table 1, the performance of PsgAidRank is substantially better (and often to a statistically significant degree) than that of InterPsgDoc.

Another reference comparison that we report in Table 1 is relevance model number 3 (RM3) (a state-of-the-art pseudo-feedback-based query-expansion method) [17, 1], which is used to rank *all* documents in the corpus. (We set the values of the free parameters of RM3 as in some prior work [13] so as to optimize p@5.) As can be seen, PsgAidRank's performance is superior in all cases to that of RM3. While the performance differences are not statistically significant, PsgAidRank posts more statistically-significant improvements over the initial ranking, optimized baselines, and InterPsgDoc than RM3 does.

Further analysis, along the lines of [13], of the different information types utilized by PsgAidRank, revealed that (i) centrality information is more effective than query-generation information for re-ranking, but using both is superior to using each alone, and (ii) whole-document information is more effective than passage-based information, but using both is superior to using each alone. (Actual numbers are omitted due to space considerations.)

Comparison with graph-based methods for re-ranking. PsgAidRank utilizes document and passage centrality that are induced using a graph-based approach. Hence, in Table 2 we compare its performance with that of previous re-ranking methods that utilize graph-based centrality. Specifically, DocGraph [14] that scores d by its PageRank score as induced over a document-solely graph scaled by $p_d(q)$ (recall from Section 2 that DocGraph is a specific case of PsgAidRank with $\lambda = 1$); ClustDocGraph [15] that scales d's authority score — induced using HITS [11] over a bipartite document-cluster graph — by $p_d(q)$; and, PsgDoc-

	AP		TREC8		WSJ	
	p@5	p@10	p@5	p@10	p@5	p@10
init. rank.	45.7	43.2	50.0	45.6	53.6	48.4
opt. base.	46.5	43.9	51.2	46.4	56.0	49.4
InterPsgDoc	46.1	41.7	50.4	46.0	54.0	48.8
RM3	50.3^{i}	48.6^{io}_{p}	53.6	46.2	58.8^{i}_{p}	51.0
PsgAidRank	$\mathbf{53.7}^{io}_{p}$	$\mathbf{49.3}^{io}_{p}$	**55.2**	**47.4**	$\mathbf{61.2}^{io}_{p}$	$\mathbf{52.8}^{io}$

Table 1: Main result. The best result in a column is boldfaced; 'i', 'o' and 'p' mark statistically significant differences with the the initial ranking, the optimized baselines and InterPsgDoc, respectively.

Graph that scales the authority score of d's most "authoritative" passage — induced (using HITS) over a bipartite document-passage graph — by $p_d(q)$. The free parameters of these methods are set to values as originally reported, and which correspond to the graph-parameters' values that we use for PsgAidRank, so as to optimize p@5.

Table 2 shows that PsgAidRank outperforms DocGraph (although not to a statistically-significant degree) and PsgDocGraph in most cases. In addition, we can see that the performance of ClustDocGraph and PsgAidRank do not dominate each other, and are not statistically distinguishable, across corpora. Thus, an interesting question is how to integrate cluster-based and passage-based information for document re-ranking, which we leave for future work.

	AP		TREC8		WSJ	
	p@5	p@10	p@5	p@10	p@5	p@10
init. rank.	45.7	43.2	50.0	45.6	53.6	48.4
DocGraph	53.3^{i}	49.2^{i}	54.0^{i}	48.0^{i}	57.2^{i}	49.6
ClustDocGraph	$\mathbf{53.7}^{i}$	$\mathbf{49.3}^{i}$	$\mathbf{57.2}^{i}$	**49.0**	57.2	51.0
PsgDocGraph	50.3	47.3	55.6	47.8	53.2	49.2
PsgAidRank	$\mathbf{53.7}^{i}$	$\mathbf{49.3}^{i}$	55.2	47.4	$\mathbf{61.2}^{i}_{p}$	$\mathbf{52.8}^{i}_{p}$

Table 2: Comparison with previous graph-based re-ranking methods. Statistically significant differences between PsgAidRank and the initial ranking and PsgDocGraph are marked with 'i' and 'p', respectively. (There are no statistically significant differences between PsgAidRank and DocGraph, ClustDocGraph.) The best result in a column is boldfaced.

5. CONCLUSION

We presented a novel language-model-based approach to re-ranking an initially retrieved list so as to improve precision at top ranks. Our model integrates inter-passage, inter-document, passage-query, and document-query similarity information. Empirical evaluation demonstrated the merits of our approach.

Acknowledgments We thank the reviewers for their comments. The paper is based upon work supported in part by IBM's and Google's faculty research awards, and upon work supported in part by the Center for Intelligent Information Retrieval. Any opinions, findings and conclusions or recommendations expressed in this material are the authors' and do not necessarily reflect those of the sponsors.

6. 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*, pages 715–725, 2004.

[2] M. Bendersky and O. Kurland. Re-ranking search results using document-passage graphs. In *Proceedings of SIGIR*, pages 853–854, 2008. poster.

[3] M. Bendersky and O. Kurland. Utilizing passage-based language models for document retrieval. In *Proceedings of ECIR*, pages 162–174, 2008.

[4] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. In *Proceedings of WWW*, pages 107–117, 1998.

[5] C. Buckley, G. Salton, J. Allan, and A. Singhal. Automatic query expansion using SMART: TREC3. In *Proceedings of TREC-3*, pages 69–80, 1994.

[6] J. P. Callan. Passage-level evidence in document retrieval. In *Proceedings of SIGIR*, pages 302–310, 1994.

[7] W. B. Croft and J. Lafferty, editors. *Language Modeling for Information Retrieval*. Number 13 in Information Retrieval Book Series. Kluwer, 2003.

[8] F. Diaz. Regularizing ad hoc retrieval scores. In *Proceedings of CIKM*, pages 672–679, 2005.

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

[10] M. Kaszkiel and J. Zobel. Effective ranking with arbitrary passages. *Journal of the American Society for Information Science*, 52(4):344–364, November 2001.

[11] J. Kleinberg. Authoritative sources in a hyperlinked environment. Technical Report Research Report RJ 10076, IBM, May 1997.

[12] O. Kurland. *Inter-document similarities, language models, and ad hoc retrieval*. PhD thesis, Cornell University, 2006.

[13] O. Kurland. The opposite of smoothing: A language model approach to ranking query-specific document clusters. In *Proceedings of SIGIR*, 2008.

[14] O. Kurland and L. Lee. PageRank without hyperlinks: Structural re-ranking using links induced by language models. In *Proceedings of SIGIR*, pages 306–313, 2005.

[15] O. Kurland and L. Lee. Respect my authority! HITS without hyperlinks utilizing cluster-based language models. In *Proceedings of SIGIR*, pages 83–90, 2006.

[16] V. Lavrenko. *A Generative Theory of Relevance*. PhD thesis, University of Massachusetts Amherst, 2004.

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

[18] X. Liu and W. B. Croft. Passage retrieval based on language models. In *Proceedings of CIKM*, pages 375–382, 2002.

[19] X. Liu and W. B. Croft. Cluster-based retrieval using language models. In *Proceedings of SIGIR*, pages 186–193, 2004.

[20] Q. Mei, D. Zhang, and C. Zhai. A general optimization framework for smoothing language models on graph structures. In *SIGIR*, pages 611–618, 2008.

[21] R. Mihalcea. Graph-based ranking algorithms for sentence extraction, applied to text summarization. In *The Companion Volume to the Proceedings of ACL*, pages 170–173, 2004.

[22] J. Otterbacher, G. Erkan, and D. R. Radev. Using random walks for question-focused sentence retrieval. In *Proceedings of HLT/EMNLP*, pages 915–922, 2005.

[23] J. M. Ponte and W. B. Croft. A language modeling approach to information retrieval. In *Proceedings of SIGIR*, pages 275–281, 1998.

[24] G. Salton, J. Allan, and C. Buckley. Approaches to passage retrieval in full text information systems. In *Proceedings of SIGIR*, pages 49–58, 1993.

[25] M. Wang and L. Si. Discriminative probabilistic models for passage based retrieval. In *Proceedings of SIGIR*, pages 419–426, 2008.

[26] R. Wilkinson. Effective retrieval of structured documents. In *Proceedings of SIGIR*, pages 311–317, 1994.

[27] P. Willett. Query specific automatic document classification. *International Forum on Information and Documentation*, 10(2):28–32, 1985.

[28] L. Yang, D. Ji, G. Zhou, Y. Nie, and G. Xiao. Document re-ranking using cluster validation and label propagation. In *Proceedings of CIKM*, pages 690–697, 2006.

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