Cluster-Based Query Expansion

Inna Gelfer Kalmanovich and Oren Kurland
Faculty of Industrial Engineering and Management
Technion — Israel Institute of Technology
Haifa 32000, Israel
innagel@tx.technion.ac.il,kurland@ie.technion.ac.il

ABSTRACT

We demonstrate the merits of using *document clusters* that are created offline to improve the overall effectiveness and performance robustness of a state-of-the-art pseudo-feedback-based query expansion method — the *relevance model*.

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

General Terms: Algorithms, Experimentation

Keywords: query expansion, clusters, relevance model

1. INTRODUCTION

Pseudo-feedback-based (PF) query expansion methods augment the query with terms from documents in an initially retrieved list. The list is often composed of documents most highly ranked by using document-query similarities [9, 6]. However, some (or even many) of the documents in the initial list may not be relevant. Furthermore, not all query-related aspects might be manifested in the list [3]. Hence, the expanded query might exhibit *query drift* [9], that is, represent an information need different than that underlying the original query. Indeed, there are many queries for which state-of-the-art PF expansion methods yield retrieval performance that is substantially inferior to that of using the original query with no expansion — the *performance robustness* problem [2, 7].

The potentially degraded "quality" of the initial list is often caused by the virtue of the way it is created, that is, using surface-level document-query similarities. Thus, we propose to perform query-expansion based on *clusters* of similar documents that are created offline. These clusters, which will be selected based on their query-similarities, can potentially be considered as better reflecting the corpus-context of the query than individual documents are [4, 5, 8]. A case in point, the clusters can contain relevant documents that do not exhibit high surface-level query similarity, but which are similar to documents that do.

We show that utilizing clusters created offline can improve the overall effectiveness and performance robustness of a state-of-the-art PF-based query expansion method — the *relevance model* [6]. The performance also transcends that of a recently-proposed approach that utilizes *query-specific* clusters to enhance the relevance model performance [7].

SIGIR'09, July 19–23, 2009, Boston, Massachusetts, USA.
ACM 978-1-60558-483-6/09/07.

2. METHODS

Let q, d, and $\mathcal{D}$ denote a query, a document, and a corpus respectively. We assume that the corpus is clustered offline into the set of clusters $Cl(\mathcal{D})$. Our algorithms operate in the language modeling (LM) framework. Specifically, we use $p_x(\cdot)$ to denote the (smoothed) unigram language model induced from text x (a document or a cluster).

Pseudo-feedback-based query expansion methods in the LM framework construct a language model that represents an expanded-query form from language models of documents in an initially retrieved list $\mathcal{D}_{\text{init}}$ [6, 13]. Often, $\mathcal{D}_{\text{init}}$ contains the k documents d in the corpus that yield the highest *query-likelihood* [10], $\prod_{q_i} p_d(q_i)$; $\{q_i\}$ is the set of query terms. We use **DocQE** to denote such a retrieval model that utilizes only documents to perform expansion.

The expansion methods just mentioned are not "aware" of the textual items from which the language models they utilize are induced. Thus, following the observations made in Section 1 we can use language models of clusters rather than documents to construct an expanded query form. Specifically, we use the language models of the k clusters c in $Cl(\mathcal{D})$ that yield the highest query-likelihood $\prod_{q_i} p_c(q_i)$; **ClustQE** denotes the resultant retrieval method.

Since the clusters are created in a query-independent fashion, the k selected ones might contain documents that are not query-related. These documents can belong to the selected clusters due to non-query-related inter-document similarities. To potentially ameliorate this problem, we consider the **DocClustQE** approach that uses both documents and clusters that are similar to the query to perform the expansion. Specifically, the k elements x in $\mathcal{D} \cup Cl(\mathcal{D})$ that yield the highest query-likelihood $\prod_{q_i} p_x(q_i)$ are used.

Related work. Previously-suggested query-expansion models that utilize clusters created offline (or topic models) [8, 11, 12] rank cluster-based smoothed document LMs with the expanded form. In contrast, we utilize cluster-based information *only* for creating the expanded-query form; standard corpus-based smoothed document LMs are used to rank documents. Furthermore, integrating documents and clusters for query-expansion as in DocClustQE, which has performance merits (see Section 3), was not explored [8, 11, 12].

3. EVALUATION

We conducted experiments on the TREC corpora specified in Table 1. We applied tokenization, Porter stemming, and stopword removal (using the INQUERY list) via the Lemur toolkit (www.lemurproject.org), which was also used

	AP				ROBUST				WSJ				SJMN			
	queries:51-150 disks: 1-3				queries: 301-450, 601-700 disks: 4-5 (-CR)				queries: 151-200 disks: 1-2				queries: 51-150 disk: 3			
	MAP	% >	p@5	% >	MAP	% >	p@5	% >	MAP	% >	p@5	% >	MAP	% >	p@5	% >
LM	22.3	–	44.7	–	24.6	–	**49.6**	–	32.7	–	55.6	–	19.3	–	33.2	–
DocQE	29.0^{l}	52.5	49.3	45.0	30.4^{l}	**54.2**	49.0	48.6	39.9^{l}	**60.0**	58.8^{l}	58.0	24.6^{l}	45.0	38.4^{l}	48.0
SamplingQE	28.8^{l}	49.5	47.9	45.4	$\mathbf{30.7}^{l}$	53.6	49.2	**51.2**	39.2^{l}	56.0	60.4	58.0	24.1^{l}	44.0	36.2	44.0
ClustQE	$\mathbf{30.1}^{l}_{s}$	**56.6**	50.5^{l}	48.0	25.6^{d}_{s}	49.4	48.6	47.0	$\mathbf{40.9}^{l}$	**60.0**	60.8^{l}	**66.0**	$\mathbf{26.1}^{l}$	**47.0**	42.0^{ld}_{s}	**53.0**
DocClustQE	$\mathbf{30.1}^{ld}_{s}$	55.6	$\mathbf{52.1}^{l}$	**50.0**	30.4^{l}	53.8	49.1	49.0	$\mathbf{40.9}^{l}$	**60.0**	**62.0**	**66.0**	26.0^{l}	46.0	$\mathbf{43.0}^{ld}_{s}$	**53.0**

Table 1: Performance numbers. The best result in a column is boldfaced. Statistically significant differences with LM, DocQE, and SamplingQE are marked with 'l', 'd', and 's', respectively.

for LM induction. Topic titles served as queries. Unless otherwise specified, we use Dirichlet-smoothed unigram language models with the smoothing parameter set to 1000.

We use MAP and p@5 for performance evaluation measures. Statistically-significant differences of performance are determined using the two-sided Wilcoxon test at a 95% confidence level. We also report the performance *robustness* of the methods (denoted '% >'); that is, the percentage of queries for which a method posts performance (MAP/p@5) that is superior to that of the LM query-likelihood model [10] (denoted **LM**) that does not perform query expansion.

We use *relevance model* number 3 (RM3), which is a state-of-the-art pseudo-feedback-based query-expansion method [6, 1], for experiments. We set the Jelinek-Mercer smoothing parameter of the LMs from which RM3 is constructed to 0.9 following previous recommendations [6]. The other free parameters of RM3 are set in the tested methods to values that optimize MAP. Specifically, the number of elements, k, used for RM3 construction is chosen from $\{10, 50, 100, 500\}$; the number of terms is set to values in $\{25, 50, 100, 500, ALL\}$, where '$ALL$' stands for the number of unique terms in the corpus; and, the parameter that governs the interpolation with the original query model is set to values in $\{0, 0.1, \ldots, 0.9\}$.

Following previous work on cluster-based retrieval we use (overlapping) nearest-neighbors clusters of 5 documents that are created prior to retrieval time [4, 5, 11]. A cluster is represented by the concatenation of its constituent documents. The order of concatenation has no effect since we use unigram language models that assume term independence.

As a reference comparison we use a recently-proposed cluster-based sampling method (denoted **SamplingQE**) for constructing RM3 [7]. Specifically, we take the 50 documents in the corpus that yield the highest query-likelihood and cluster them into 50 *query-specific* (overlapping) nearest-neighbors clusters of 5 documents. Then, we use the constituent documents (with repetitions) of the m query-specific clusters that yield the highest query-likelihood to construct RM3. The free parameters' values are set to optimize MAP; specifically, m is set to values in $\{5, 10, 20, 30, 40, 50\}$, and the other free parameters are set to values in the ranges specified above.

As can be seen in Table 1 all query-expansion-based methods post in most cases better performance than that of the non-expansion-based LM approach. We can also see that our proposed methods, ClustQE and DocClustQE, post performance that is in a majority of the relevant comparisons (corpus × evaluation measure) superior to, and more robust than, that of the DocQE and SamplingQE[1] methods. In addition, note that the performance of DocClustQE is at least as good, and robust, as that of ClustQE in a majority of the relevant comparisons.

Thus, we conclude that there is merit in using information induced from clusters created offline for query-expansion, especially when integrated with information induced from individual documents as in DocClustQE.

Acknowledgments We thank the reviewers for their comments, and Lillian Lee for discussions of ideas presented in this paper. The paper is based upon work supported in part by IBM's and Google's faculty research awards. Any opinions, findings and conclusions or recommendations expressed in this material 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] G. Amati, C. Carpineto, and G. Romano. Query difficulty, robustness, and selective application of query expansion. In *Proceedings of ECIR*, pages 127–137, 2004.

[3] C. Buckley. Why current IR engines fail. In *Proceedings of SIGIR*, pages 584–585, 2004. Poster.

[4] O. Kurland and L. Lee. Corpus structure, language models, and ad hoc information retrieval. In *Proceedings of SIGIR*, pages 194–201, 2004.

[5] O. Kurland, L. Lee, and C. Domshlak. Better than the real thing? Iterative pseudo-query processing using cluster-based language models. In *Proceedings of SIGIR*, pages 19–26, 2005.

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

[7] K.-S. Lee, W. B. Croft, and J. Allan. A cluster-based resampling method for pseudo-relevance feedback. In *Proceedings of SIGIR*, pages 235–242, 2008.

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

[9] M. Mitra, A. Singhal, and C. Buckley. Improving automatic query expansion. In *Proceedings of SIGIR*, pages 206–214, 1998.

[10] F. Song and W. B. Croft. A general language model for information retrieval (poster abstract). In *Proceedings of SIGIR*, pages 279–280, 1999.

[11] T. Tao, X. Wang, Q. Mei, and C. Zhai. Language model information retrieval with document expansion. In *Proceedings of HLT/NAACL*, pages 407–414, 2006.

[12] X. Wei and W. B. Croft. LDA-based document models for ad-hoc retrieval. In *Proceedings of SIGIR*, pages 178–185, 2006.

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

[1]The performance patterns of SamplingQE are in accordance with those originally reported [7]. That is, the performance for heterogeneous corpora (ROBUST) is better than that for homogeneous corpora (AP, WSJ, SJMN).