Inf Retrieval (2009) 12:437–460
DOI 10.1007/s10791-008-9065-9

Re-ranking search results using language models of query-specific clusters

Oren Kurland

Received: 13 October 2007 / Accepted: 16 June 2008 / Published online: 10 July 2008

Abstract To obtain high precision at top ranks by a search performed in response to a query, researchers have proposed a cluster-based re-ranking paradigm: clustering an initial list of documents that are the most highly ranked by some initial search, and using information induced from these (often called) *query-specific* clusters for re-ranking the list. However, results concerning the effectiveness of various *automatic* cluster-based re-ranking methods have been inconclusive. We show that using query-specific clusters for automatic re-ranking of top-retrieved documents is effective with several methods in which clusters play different roles, among which is the *smoothing* of *document language models*. We do so by adapting previously-proposed cluster-based retrieval approaches, which are based on (static) query-independent clusters for ranking all documents in a corpus, to the re-ranking setting wherein clusters are query-specific. The best performing method that we develop outperforms both the initial document-based ranking and some previously proposed cluster-based re-ranking approaches; furthermore, this algorithm consistently outperforms a state-of-the-art pseudo-feedback-based approach. In further exploration we study the performance of cluster-based smoothing methods for re-ranking with various (soft and hard) clustering algorithms, and demonstrate the importance of clusters in providing context from the initial list through a comparison to using single documents to this end.

Keywords Query-specific clusters · Cluster-based language models · Cluster-based re-ranking · Cluster-based smoothing

1 Introduction

Users of search engines expect to see the documents most pertaining to their queries at the top ranks of the retrieved results (Croft 1995). A paradigm suggested by several

Initial verison of this paper appeared as Chapter 5 of Kurland (2006).

O. Kurland (✉)
Faculty of Industrial Engineering and Management, Technion—Israel Institute of Technology, Technion City, Haifa 32000, Israel
e-mail: kurland@ie.technion.ac.il

researchers for achieving this goal is to perform an initial search over the entire corpus in response to a query, and then to automatically *re-rank* the most highly ranked documents, so as to improve the precision at the very top ranks of the resultant list. (See, for example, Preece (1973); Willett (1985); Kleinberg (1998); Liu and Croft (2004); Diaz (2005); Kurland and Lee (2005, 2006); Liu and Croft (2006a).) The motivating idea behind the *re-ranking* paradigm is that the ratio of relevant to non-relevant documents in the *initial list* to be re-ranked, that is, the most highly ranked documents from the initial search, tends to be much higher than that in the entire corpus. However, since documents in the list were retrieved in response to a query, it is a challenging task to differentiate the relevant documents from the non-relevant ones.

To approach this challenge of (automatic) re-ranking, several researchers (Preece 1973; Willett 1985; Liu and Croft 2004; Kurland and Lee 2006; Liu and Croft 2006a, b) proposed to cluster the documents in the initial list and utilize information induced from the clusters; those are often termed *query-specific* clusters since the documents upon which clustering is performed were retrieved in response to a query. A potential advantage in using query-specific clusters for re-ranking that researchers (Hearst and Pedersen 1996; Tombros et al. 2002; Liu and Croft 2004; Kurland and Lee 2006) have pointed out is that relevant documents in the initial list might be clustered together—a manifestation of van Rijsbergen's *cluster hypothesis* (van Rijsbergen 1979) in the re-ranking setting. Indeed, there is some empirical evidence that (under different clustering algorithms) there are often some query-specific clusters that contain a high percentage of relevant documents (Hearst and Pedersen 1996; Tombros et al. 2002; Kurland 2006). However, automatically finding these clusters is a very hard challenge (Willett 1985; Liu and Croft 2004). On the other hand, it was shown that users of interactive search systems can use query-specific clusters for quickly detecting relevant documents that they contain (Hearst and Pedersen 1996; Leuski 2001).

A different way by which clusters can be utilized has recently been proposed in the language modeling framework to information retrieval (Ponte and Croft 1998; Croft and Lafferty 2003). Researchers suggested to use information induced from document clusters to *smooth* document language models so as to "enrich" the document representation with corpus-related information (Azzopardi et al. 2004; Kurland and Lee 2004; Liu and Croft 2004; Tao et al. 2006; Wei and Croft 2006). Indeed, cluster-based smoothing has shown promise for ranking an entire corpus when *static query-independent clusters*, which are created offline, were used (Azzopardi et al. 2004; Kurland and Lee 2004; Liu and Croft 2004; Tao et al. 2006; Wei and Croft 2006). However, the results regarding the effectiveness of cluster-based smoothing for the re-ranking setting (using query-specific clusters) have remained inconclusive (Liu and Croft 2004).

We show that using query-specific clusters for automatic re-ranking is in fact effective whether clusters are used for *selecting* documents—specifically, detecting relevant documents by the patterns of membership of documents in clusters—or for *smoothing* document language models. We do so by adapting recently proposed cluster-based retrieval algorithms (Kurland and Lee 2004), which utilize information induced from static query-independent clusters for ranking all documents in a corpus, to the re-ranking setting wherein clusters are query-specific.

We empirically show that the most effective (cluster-based smoothing) re-ranking algorithm that we present not only significantly outperforms the initial document-based ranking over all tested TREC corpora, but also consistently outperforms a state-of-the-art pseudo-feedback-based approach, namely *the relevance model* (Lavrenko and Croft 2001). Moreover, the algorithm also outperforms some previously-proposed cluster-based approaches for re-ranking that utilize information induced from query-specific clusters. In further exploration we study the performance of cluster-based smoothing methods for re-ranking with various clustering algorithms, and demonstrate the importance of clusters in providing context from the initial list through a comparison to using single documents to this end.

The rest of the paper is organized as follows. In Sect. 2 we present the different re-ranking algorithms that we explore. Section 3 describes the connection of our approach to previously-suggested models for re-ranking and to previous approaches for utilizing cluster-based information. We then present an empirical evaluation of the performance of our algorithms in Sect. 4 and conclude in Sect. 5.

2 Retrieval framework

Since we are focused on the re-ranking setting, the algorithms we present are applied not to the entire corpus $\mathcal{C}$, but to a subset $\mathcal{D}_{\text{init}}^{N,q}$ (henceforth $\mathcal{D}_{\text{init}}$), defined as the top N documents retrieved in response to the query q by a given initial retrieval engine. The algorithms also take into account a set $Cl(\mathcal{D}_{\text{init}})$ of (query-specific) clusters of the documents in $\mathcal{D}_{\text{init}}$. We assume that documents in $\mathcal{D}_{\text{init}}$ and clusters in $Cl(\mathcal{D}_{\text{init}})$ are assigned with unique IDs.

The algorithms we present utilize statistical language models (Ponte and Croft 1998; Croft and Lafferty 2003). We use $p_x(y)$ to denote the language-model-based similarity between x (a document or a cluster) and y (a document, a cluster, or a query).[1] Our language-model-induction methods are described in Sect. 2.2.

Clustering Previous work on utilizing query-specific clustering has focused on hard-clustering techniques (e.g., Willett 1985; Hearst and Pedersen 1996; Leuski 2001; Tombros et al. 2002; Liu and Croft 2004). In contrast, here we focus on using overlapping nearest-neighbor clusters that were shown to be effective when utilized in a query-independent fashion (Griffiths et al. 1986; Kurland and Lee 2004; Kurland et al. 2005; Kurland 2006), and which were recently used in the re-ranking setting (Kurland and Lee 2006; Liu and Croft 2006a, b).

Formally, for each document $d \in \mathcal{D}_{\text{init}}$ we define a cluster that contains d and the $k-1$ documents d_i ($d_i \neq d$) from $\mathcal{D}_{\text{init}}$ that yield the highest language-model similarity $p_{d_i}(d)$ (we break ties by document IDs); k is a free parameter. Thus, the set $Cl(\mathcal{D}_{\text{init}})$ contains N (overlapping) clusters. We study the relative merits of this nearest-neighbor clustering approach with respect to hard-clustering techniques in Sect. 4.6.

2.1 Re-ranking algorithms

In what follows we adapt cluster-based retrieval algorithms that were originally designed by Kurland and Lee (2004) for use with *query-independent* (static) clusters, and were shown to be effective for ranking the entire corpus, to the re-ranking setting wherein the clusters are *query-specific*.

The original versions of the algorithms that we consider (Kurland and Lee 2004) operate on clusters that are most similar to the query—i.e., *top-retrieved clusters*—for anchoring the query-independent clustering information to the query at retrieval time. The variants that we present here, on the other hand, consider *all* clusters in $Cl(\mathcal{D}_{\text{init}})$ as these are constructed from $\mathcal{D}_{\text{init}}$—documents that are the most highly ranked by some initial search.

In the algorithms that we present clusters play two different roles, namely *selection* of documents and *smoothing* of documents' language models (Kurland and Lee 2004).

[1] Some other work uses these language-model-based estimates for forming links between textual items and utilizing them with graph-based methods (Kurland and Lee 2005, 2006). We discuss the relation of our methods to these approaches in Sects. 3 and 4.

2.1.1 Cluster-based document selection

Most cluster-based document-selection algorithms aim to identify a subset of clusters that potentially contain a large number of relevant documents (Croft 1980; Willett 1985; Kurland and Lee 2004; Liu and Croft 2004). However, finding query-specific clusters that contain a high percentage of relevant documents is known to be a very hard task (Hearst and Pedersen 1996; Tombros et al. 2002; Liu and Croft 2004). One of the reasons is that query-specific clusters contain documents that are similar to the query to begin with.

Therefore, we will focus here on a different cluster-based selection approach, which exploits the structure induced by overlapping clusters. Specifically, if we think of clusters as potentially representing aspects manifested in the initial list $\mathcal{D}_{\text{init}}$, one might opt to rank high documents that exhibit as many such aspects as possible, specifically, documents that belong to many clusters. An alternative view for the potential in utilizing the structure induced by clusters might be based on the fact that documents that belong to many of the clusters exhibit (high) similarity to many other documents in the initial list $\mathcal{D}_{\text{init}}$. Thus, such documents could be considered as *central* with respect to the initial list—a notion recently explored via a graph-based framework and which was shown to be connected with relevance (Kurland and Lee 2005, 2006; Kurland 2006).

Utilizing the structure induced by clusters as described above is the idea underlying Kurland and Lee's (2004) best-performing cluster-based selection method—the **bag-select** algorithm. In its original form, the bag-select algorithm ranks high documents from the (entire) corpus that exhibit high similarity to the query and that belong to many *top-retrieved* query-independent clusters. Dropping the notion of "top-retrieved clusters", as we deal with query-specific clusters, we focus on the centrality of a document with respect to the initial list $\mathcal{D}_{\text{init}}$ as measured by its membership in clusters from $\mathcal{C}l(\mathcal{D}_{\text{init}})$.

As noted above, the original version of the bag-select algorithm (Kurland and Lee 2004) also takes into account the document-query similarity information. This is done for coping with the fact that the clusters in this work (Kurland and Lee 2004) are query-independent. Case in point, top-retrieved query-independent clusters might contain documents that do not pertain to the query, but which are similar to documents that are based on information not related to the query. While it might seem at a first glance that using document-query similarity information for re-ranking $\mathcal{D}_{\text{init}}$ is redundant, experimental results show that using this information is actually important. This finding is in line with some recent work on graph-based re-ranking (Kurland and Lee 2005). Indeed, some of the query-specific clusters might exhibit "aspects" not pertaining to the query, or more specifically, contain a high percentage of non-relevant documents. Therefore, using document-query similarity information, and hence considering document-specific characteristics, might help to ameliorate the overgeneralization caused by scoring documents based only on cluster-induced information. Further support to the importance of "query-anchoring" is given in work on *score regularization* for re-ranking (Diaz 2005), which shows that documents from the initial list that are highly similar both to the query and to other documents in the list that are similar to the query tend to be relevant.

Given the observations made at the above, we set the re-ranking version of the bag-select algorithm to score document d by

$$Score_{bag-select}(d) \stackrel{def}{=} p_d(q) \cdot \#(c \in \mathcal{C}l(\mathcal{D}_{\text{init}}) : d \in c).$$

The bag-select algorithm utilizes two sources of information: the number of clusters to which the document belongs and the document-query similarity. In the next section we show how these two sources of information, along with additional ones, can be modeled and integrated using a probabilistic approach.

2.1.2 Cluster-based smoothing

In work on language models for ad hoc retrieval, several researchers have proposed to smooth the document language model with that of the cluster(s) with which it is associated (Azzopardi et al. 2004; Kurland and Lee 2004; Liu and Croft 2004; Wei and Croft 2006). The idea is to enrich the document representation with corpus-context information. Such an approach can help, for example, to deal with the *synonymy* problem, and more generally, with the *sparse data problem*. Applying cluster-based smoothing in the re-ranking setting with query-specific clusters means that the context-information is drawn from the initial list $\mathcal{D}_{\text{init}}$ rather than from the entire corpus. Hence, such an approach can be thought of as query-specific (cluster-based) smoothing: the information used for smoothing is drawn from documents that are (relatively) similar to the document in hand and to the query.

To study whether utilizing context from $\mathcal{D}_{\text{init}}$ to enrich a document representation yields effective re-ranking performance, we adapt Kurland and Lee's (2004) *aspect-based* algorithms, which are named after the aspect models of Hofmann and Puzicha (1998). Aspect models are an approach for modeling a corpus based on the assumption that each document exhibits (or is "generated" by) a mixture of aspects. The algorithm for finding the aspects, in terms of language models, induces clustering of documents as it estimates document-aspect association probabilities, and aspects might be thought of as (soft) clusters.

Kurland and Lee (2004) conceptually adopt the basic formulation underlying the aspect models and use it with static (query-independent) existing clusters for ranking all documents in a corpus. Specifically, the **aspect-t** algorithm (Kurland and Lee 2004) is based on estimating the conditional probability $p(q|d)$—often termed *query likelihood* (Song and Croft 1999). The idea is to estimate the probability that the query terms are generated by a (probabilistic) model induced from a document. Using simple probability rules, this probability can be written as

$$p(q|d) = \sum_{c \in \mathcal{C}l(\mathcal{D}_{\text{init}})} p(q|d,c)p(c|d). \tag{1}$$

The basic conceptual assumption underlying aspect models is that a query is independent of a document given a cluster (Hofmann and Puzicha 1998). That is, the query terms could be viewed as being generated directly from models of clusters (aspects) that generate the terms in the document. Using this assumption we get that the above probability is

$$\sum_{c \in \mathcal{C}l(\mathcal{D}_{\text{init}})} p(q|c)p(c|d). \tag{2}$$

Following recent work on cluster-based smoothing (Kurland and Lee 2004; Liu and Croft 2004; Tao et al. 2006), we post the constraint, which we will later relax, that a document can be "represented" (i.e., smoothed) *only* by the clusters to which it belongs.[2]

[2] Such a constraint can potentially alleviate the computational cost of estimating the document-cluster association strength for all available clusters and documents; this cost is significant when using, for example, static overlapping clusters (Kurland and Lee 2004). An implicit assumption underlying this constraint is that the best clusters to use for representing a document are those that contain it. We return to this point later on.

Thus, we truncate the summation from Eq. 2 (hence the suffix "-t" for "truncated")[3] and in addition use a language-model-based similarity measure for conditional probabilities to derive the aspect-t algorithm:

$$Score_{aspect-t}(d) \stackrel{def}{=} \sum_{c \in \mathcal{C}l(\mathcal{D}_{\text{init}}):d \in c} p_c(q) p_d(c).$$

Is is important to note that the original scoring function of the aspect-t algorithm (Kurland and Lee 2004) is slightly different than the one presented here, and not only due to the shift from using (top-retrieved) query-independent clusters to using *all* available query-specific clusters from $\mathcal{C}l(\mathcal{D}_{\text{init}})$. Directly adapting Kurland and Lee's model to the re-ranking setting, by using query-specific instead of query-independent clusters, yields the scoring function $\sum_{c:d \in c} p_c(q) p_c(d)$. This model is a result of using Bayes rule upon Eq. 2 and assuming uniform prior distributions for documents and clusters. Our formulation here, on the other hand, is not dependent on these assumptions, and, as it turns out, yields much better re-ranking performance than that of the originally suggested model (Kurland and Lee 2004).

The assumption that a query is independent of a document given a cluster can cause overgeneralization. That is, representing a document *only* via the clusters to which it belongs ignores potentially important information with regard to the document-specific characteristics. Such information can help to estimate the document-query "match". Hence, we drop this independence assumption, and in addition (i) use the estimate $\lambda p(q|d) + (1 - \lambda)\, p(q|c)$ for $p(q|d, c)$ where λ is a free parameter, (ii) apply some probability algebra, and (iii) use a language-model-based similarity measure for conditional probabilities in Eq. 1, to derive Kurland and Lee's best-performing model, **interpolation-t**:[4]

$$Score_{interpolation-t}(d) \stackrel{def}{=} \lambda p_d(q) + (1 - \lambda) \sum_{c \in \mathcal{C}l(\mathcal{D}_{\text{init}}):d \in c} p_c(q) p_d(c).$$

Note that the interpolation-t algorithm anchors the cluster-based ranking of the aspect-t algorithm to the query by interpolation with the query-similarity score $p_d(q)$. This anchoring makes sense when query-independent clustering is used as in the original proposal of interpolation-t (Kurland and Lee 2004). However, as is the case for the bag-select algorithm from the above, and as we will be shown in Sect. 4, this anchoring has the potential to improve re-ranking effectiveness, even though the clusters are query-specific. This further demonstrates the importance in utilizing document-specific characteristics for ameliorating the overgeneralization caused by the use of clusters as proxies for ranking documents.

We also note that the interpolation-t algorithm can be conceptually viewed as a generalized version of the models of Liu and Croft (2004) and Wei and Croft (2006) that use cluster-based smoothing of document language models. (The former uses k-means clusters and the latter uses LDA clusters (Blei et al. 2003);[5] note that the interpolation-t algorithm does not require the clusters to be overlapping.)

Document-cluster relationship Both the aspect-t and interpolation-t algorithms use clusters as "representatives" ("proxies") of their constituent documents. However, if the different clusters are thought of as potentially representing different aspects manifested in

[3] The aspect-based models were originally termed "aspect-x" (Kurland and Lee 2004).

[4] The original name of this algorithm was *interpolation* (Kurland and Lee 2004).

[5] We hasten to point out that the models in Liu and Croft (2004) and Wei and Croft (2006) operate at the term-level, that is, interpolation is performed upon estimates of term probabilities. In contrast, interpolation-t operates at the score level by fusion of language-model-based similarity scores.

the initial list $\mathcal{D}_{\text{init}}$, then a document can be associated (with varying degrees of strength) with different aspects (clusters) regardless of which clusters it belongs to. Thus, we consider the alternative of smoothing a document language model with the language models of *all* clusters in $\mathcal{C}l(\mathcal{D}_{\text{init}})$ to a degree controlled by the document-cluster language-model-based similarity. Doing so results in a formulation that is more "faithful" to the original probabilistic formulation in Eqs. 1 and 2 than those of the aspect-t and interpolation-t algorithms. (Recall that the latter two use truncation of the summation in Eqs. 1 and 2.) We thereby define the algorithms **aspect-f** and **interpolation-f** using the scoring functions: ("-f" stands for using the full summation in Eqs. 1 and 2)

$$Score_{aspect-f}(d) = \sum_{c \in \mathcal{C}l(\mathcal{D}_{\text{init}})} p_c(q) p_d(c),$$

and

$$Score_{interpolation-f}(d) = \lambda p_d(q) + (1-\lambda) \sum_{c \in \mathcal{C}l(\mathcal{D}_{\text{init}})} p_c(q) p_d(c),$$

respectively.

2.2 Language-model-based similarity induction

In this section we present our estimate for the language-model similarity $p_x(y)$. For language model induction we treat documents and queries as term sequences.

While there are various approaches for representing clusters (Liu and Croft 2006b, 2008), our focus here is on the merits (or lack thereof) of our re-ranking methods. Therefore, we take the standard approach, which was shown to be effective in several applications of cluster-based retrieval (Kurland and Lee 2004, 2006; Liu and Croft 2004), and represent a cluster by the term sequence that results from concatenating its constituent documents; the order of concatenation has no effect since we are only going to define unigram language models that assume term independence.

We use $tf(w \in x)$ to denote the number of times that term w occurs in the text (or text collection) x. The *maximum likelihood estimate* (MLE) of w with respect to x is defined as

$$\widetilde{p}_x^{MLE}(w) \stackrel{def}{=} \frac{tf(w \in x)}{\sum_{w'} tf(w' \in x)}.$$

To cope with the *zero probability problem*, namely, the assignment of zero probability to unseen terms, we adopt the widely used Dirichlet-smoothed estimate (Zhai and Lafferty 2001; Croft and Lafferty 2003):

$$\widetilde{p}_x^{[\mu]}(w) \stackrel{def}{=} \frac{tf(w \in x) + \mu \cdot \widetilde{p}_{\mathcal{C}}^{MLE}(w)}{\sum_{w'} tf(w' \in x) + \mu};$$

μ is a free parameter that controls the amount of reliance on corpus statistics. We extend the estimate just described to a term sequence $\mathbf{w} = w_1 w_2 \cdots w_n$ using the term-independence assumption:

$$p_x^{[\mu]}(\mathbf{w}) \stackrel{def}{=} \prod_{j=1}^{n} \widetilde{p}_x^{[\mu]}(w_j). \tag{3}$$

Using the estimate from Eq. 3 for estimating the similarity $p_x(y)$ will result in longer texts y being assigned lower similarity values than shorter texts are. Also, for very long texts (as is the case for clusters, for example), the estimate might cause underflow problems. Therefore, we use a previously proposed estimate (Lavrenko et al. 2002; Kurland and Lee 2004, 2005), which is based on the Kullback Leibler divergence $D(\cdot||\cdot)$ (Cover and Thomas 1991)

$$p_x^{KL,\mu}(\mathbf{w}) \stackrel{def}{=} \exp\left(-D(\widetilde{p}_{\mathbf{w}}^{MLE}(\cdot)||\widetilde{P}_x^{[\mu]}(\cdot))\right). \tag{4}$$

Using some probability algebra (see, for example, Lafferty and Zhai 2001), it can be shown that the estimate from Eq. 4 is equivalent to

$$p_x^{KL,\mu}(\mathbf{w}) = H(\mathbf{w}) \cdot p_x^{[\mu]}(\mathbf{w})^{\frac{1}{|\mathbf{w}|}}, \tag{5}$$

where H is the entropy function.

Thus, the estimate $p_x^{KL,\mu}(\mathbf{w})$ avoids the length-bias caused by the unigram language model through length normalization. Furthermore, the entropy of a document (language model) was shown to be connected with relevance in the re-ranking setting (Kurland and Lee 2005); hence, our similarity estimate "favors" documents that have a higher "prior" probability of being relevant to the query.

Also, it is important to point out that while the estimates $p_x^{KL,\mu}(\mathbf{w})$ and $p_x^{[\mu]}(\mathbf{w})$ are equivalent for the purpose of ranking documents in response to a fixed query (Lafferty and Zhai 2001), in the re-ranking setting this equivalence does not hold since we estimate similarities between different pairs of text items.

Finally, we note that while the estimate $p_x^{KL,\mu}(\mathbf{w})$ does not form a valid probability distribution, normalizing it for cases wherein one might be needed (e.g., for the distribution of clusters over a document in the aspect models) results in degraded re-ranking performance and therefore we use the estimate as is.

3 Related work

Preece (1973) was perhaps the first to suggest the use of query-specific clusters, although he did not present specific retrieval models for utilizing them.

Willett (1985) proposed to rank query-specific clusters and then to use the constituent documents of the highest-ranked ones to create a document-based ranking. He noted that the limited effectiveness of the approach could be attributed to the correlation-based ranking that was used to rank the clusters in response to the query. Liu and Croft (2004) took a similar approach for re-ranking, but used a language-model-based estimate for the query-cluster similarity; however, the resultant performance did not transcend that of the initial ranking. We compare the re-ranking performance of this cluster-ranking approach with that of the methods from Sect. 2 in Sect. 4.1.

Several researchers showed that if the documents at the top ranks of an initially retrieved list are clustered, then there is a cluster (a.k.a *the optimal cluster*) that if retrieved in its entirety, yields performance that is better than that of the initial ranking (Hearst and Pedersen 1996; Tombros et al. 2002; Kurland 2006). Moreover, such a cluster exists for different clustering approaches: partitioning (Hearst and Pedersen 1996), hierarchical agglomerative clustering (Tombros et al. 2002) and nearest-neighbor (soft) clustering (Kurland 2006, Chapter 7). While automatically detecting the optimal cluster is a difficult challenge (Willett 1985; Liu and Croft 2004; Kurland 2006, 2008; Liu and Croft 2006a; Kurland and Domshlak 2008), this clustering pattern—which gives support to van

Rijsbergen's cluster hypothesis (van Rijsbergen 1979) in the re-ranking setting—helps users to more quickly detect relevant documents if the results are visualized (and navigated) using cluster-based interfaces (Hearst and Pedersen 1996; Leuski 2001).

In work on cluster-based retrieval in the language modeling framework researchers have proposed to smooth a document language model with those of query-independent clusters so as to utilize corpus-context in representing documents (Azzopardi et al. 2004; Kurland and Lee 2004; Liu and Croft 2004; Tao et al. 2006; Wei and Croft 2006). Liu and Croft (2004) examined this cluster-based smoothing approach for re-ranking, having a document language model smoothed with that of the single query-specific (hard) cluster to which it belongs. As stated in Sect. 2, Liu and Croft's model can be viewed as a specific case of the interpolation-t algorithm when implemented with a hard clustering approach. Similarly, the re-ranking model of Lee et al. (2001), who use hierarchical agglomerative clustering, is also a special case of the interpolation-t algorithm: a document is scored by interpolation of its query-similarity score with the query-cluster similarity score of the single cluster to which it belongs; however, the clusters that are used are static query-independent clusters that are related to the query and not query-specific clusters. In Section 4.6 we present the relative merits of using nearest-neighbor overlapping clusters with respect to hard clusters for the interpolation-t and interpolation-f algorithms.

Query-specific clusters reflect inter-document similarities within the initial list. There has been an increasing use of graph-based techniques for modeling these inter-document similarities for document (re-) ranking (Daniłowicz and Baliński 2000; Kurland and Lee 2005; Zhang et al. 2005; Kurland and Lee 2006). The general idea is to identify documents that are *central* with respect to the initial list—i.e., similar to many (central) documents in the list—using graph-based methods, and use this centrality as criterion for ranking. Kurland and Lee (2006) show that it is more effective in general to incorporate both document-based and cluster-based information in the graphs than to use only the former as is the case in Daniłowicz and Baliński (2000) and Kurland and Lee (2005). Specifically, Kurland and Lee (2006) use the HITS (hubs and authorities) algorithm (Kleinberg 1998) over bipartite graphs of documents on the one side and query-specific clusters on the other side (with edge weights representing cluster-document similarities) to find central documents and clusters. They show that document authoritativeness (as induced by HITS) is connected with relevance and that authoritative query-specific clusters contain a high percentage of relevant documents. We compare the principles underlying their methods, and their performance, to those of ours in Sect. 4.5.

In a related vein, Baliński and Daniłowicz (2005) and Diaz (2005) apply score regularization to ensure that similar documents within an initially retrieved list receive similar scores. Recall that the interpolation-f algorithm assigns high scores to documents that are similar both to the query and to clusters that are similar to the query. Now, replacing clusters with documents (i.e., each document serves as a cluster), we get that a score of a document depends on its similarity to the query and on the similarity to the query of documents to which it is similar—the underlying principle of score regularization (Diaz 2005). We study this algorithm in Sect. 4.7.

Finally, it is important to note that a disadvantage of using query-specific clustering is the computational cost involved in creating the clusters. In contrast to offline clustering, wherein the clusters are created once and then used for all queries, with query-specific clustering each query requires a new clustering to be performed upon the list of retrieved documents. Therefore, several researchers proposed fast algorithms for clustering retrieved results (Cutting et al. 1992; Zamir and Etzioni 1998). The focus of the work in this paper, on the other hand, is on the potential effectiveness in exploiting clustering and not on the efficiency of the clustering method. In fact, as we show in Sect. 4.6, our best performing algorithm (in terms of effectiveness) yields very good precision-at-top-ranks performance with several different clustering methods.

4 Conclusions

We showed that algorithms that were originally designed for using static query-independent clusters for ranking an entire corpus in response to a query can be adapted to utilize query-specific clusters for effectively re-ranking documents in an initially retrieved list so as to improve precision at the top ranks. The best-performing algorithm that we developed consistently outperforms both the initial document ranking and a state-of-the-art pseudo feedback method. In further exploration we studied the effect of various—both hard and soft—clustering algorithms on the effectiveness of cluster-based smoothing for the re-ranking task, and showed the importance of clusters in providing context from the initial list by comparison to using single documents to this end.

References

Abdul-Jaleel, N., Allan, J., Croft, W. B., Diaz, F., Larkey, L., Li, X., Smucker, M. D., & Wade, C. (2004). UMASS at TREC 2004—novelty and hard. In *Proceedings of the Thirteenth Text Retrieval Conference (TREC-13)* (pp. 715–725).

Azzopardi, L., Girolami, M., & van Rijsbergen, K. (2004). Topic based language models for ad hoc information retrieval. In *Proceedings of International Conference on Neural Networks and IEEE International Conference on Fuzzy Systems* (pp. 3281–3286).

Baliński, J., & Daniłowicz, C. (2005). Re-ranking method based on inter-document distances. *Information Processing and Management, 41*(4), 759–775.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research, 3*, 993–1022.

Buckley, C., Salton, G., Allan, J., & Singhal, A. (1994). Automatic query expansion using SMART: TREC3. In *Proceedings of the Third Text Retrieval Conference (TREC-3)* (pp. 69–80).

Connell, M., Feng, A., Kumaran, G., Raghavan, H., Shah, C., & Allan, J. (2004). UMass at TDT 2004. TDT2004 System Description.

Cover, T. M., & Thomas, J. A. (1991). Elements of Information Theory. Wiley series in telecommunications. Wiley-Interscience, New York.

Croft, W. B. (1980). A model of cluster searching based on classification. *Information Systems, 5*, 189–195.

Croft, W. B. (1995). What do people want from information retrieval? D-Lib magazine.

Croft, W. B., & Lafferty, J. (Eds.). (2003). Language modeling for information retrieval. In *No. 13 in Information Retrieval Book Series*. Kluwer.

Cronen-Townsend, S., Zhou, Y., & Croft, W. B. (2004). A language modeling framework for selective query expansion. Tech. Rep. IR-338, Center for Intelligent Information Retrieval, University of Massachusetts.

Cutting, D. R., Karger, D. R., Pedersen, J. O., & Tukey, J. W. (June 1992). Scatter/Gather: A cluster-based approach to browsing large document collections. In *15th Annual International SIGIR* (pp. 318–329). Denmark.

Daniłowicz, C., & Baliński, J. (2000). Document ranking based upon Markov chains. *Information Processing and Management, 41*(4), 759–775.

Diaz, F. (2005). Regularizing ad hoc retrieval scores. In *Proceedings of the Fourteenth International Conference on Information and Knowledge Managment (CIKM)* (pp. 672–679).

Diaz, F., & Metzler, D. (2006). Improving the estimation of relevance models using large external corpora. In *Proceedings of SIGIR* (pp. 154–161).

El-Hamdouchi, A., & Willett, P. (1986). Hierarchic document clustering using ward's method. In *Proceedings of SIGIR* (pp. 149–156).

Griffiths, A., Luckhurst, H. C., & Willett, P. (1986). Using interdocument similarity information in document retrieval systems. *Journal of the American Society for Information Science (JASIS), 37*(1), 3–11, reprinted in K. S. Jones & P. Willett (Eds.), *Readings in information retrieval* (pp. 365–373), Morgan Kaufmann, 1997.

Hearst, M. A., & Pedersen, J. O. (1996). Reexamining the cluster hypothesis: Scatter/Gather on retrieval results. In *Proceedings of SIGIR* (pp. 76–84).

Hofmann, T., & Puzicha, J. (1998). Unsupervised learning from dyadic data. Tech. Rep. TR-98-042, International Computer Science Institute (ICSI).

Hu, X., Bandhakavi, S., & Zhai, C. (2003). Error analysis of difficult TREC topics. In *Proceedings of SIGIR* (pp. 407–408), poster.

Kleinberg, J. (1998). Authoritative sources in a hyperlinked environment. In *Proceedings of the 9th ACM-SIAM Symposium on Discrete Algorithms (SODA)* (pp. 668–677), extended version in *Journal of the ACM, 46*, 604–632, 1999.

Kurland, O. (2006). Inter-document similarities, language models, and ad hoc retrieval. Ph.D. thesis, Cornell University.

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

Kurland, O., & Domshlak, C. (2008). A rank-aggregation approach to searching for optimal query-specific clusters. In *Proceedings of SIGIR*.

Kurland, O., & Lee, L. (2004). Corpus structure, language models, and ad hoc information retrieval. In *Proceedings of SIGIR* (pp. 194–201.

Kurland, O., & Lee, L. (2005). PageRank without hyperlinks: Structural re-ranking using links induced by language models. In *Proceedings of SIGIR* (pp. 306–313).

Kurland, O., & Lee, L. (2006). Respect my authority! HITS without hyperlinks utilizing cluster-based language models. In *Proceedings of SIGIR* (pp. 83–90).

Kurland, O., Lee, L., & Domshlak, C. (2005). Better than the real thing? Iterative pseudo-query processing using cluster-based language models. In *Proceedings of SIGIR* (pp. 19–26).

Lafferty, J. D., & Zhai, C. (2001). Document language models, query models, and risk minimization for information retrieval. In *Proceedings of SIGIR* (pp. 111–119).

Lavrenko, V., Allan, J., DeGuzman, E., LaFlamme, D., Pollard, V., & Thomas, S. (2002). Relevance models for topic detection and tracking. In *Proceedings of the Human Language Technology Conference (HLT)* (pp. 104–110).

Lavrenko, V., & Croft, W. B. (2001). Relevance-based language models. In *Proceedings of SIGIR* (pp. 120–127).

Lavrenko, V., & Croft, W. B. (2003). Relevance models in information retrieval. In W. B. Croft & J. Lafferty (Eds.), *No. 13 in Information Retrieval Book Series* (pp. 11–56).

Lee, K.-S., Park, Y.-C., & Choi, K.-S. (2001). Re-ranking model based on document clusters. *Information Processing and Management, 37*(1), 1–14.

Leuski, A. (2001). Evaluating document clustering for interactive information retrieval. In *Proceedings of the Tenth International Conference on Information and Knowledge Managment (CIKM)* (pp. 33–40).

Liu, X., & Croft, W. B. (2004). Cluster-based retrieval using language models. In *Proceedings of SIGIR* (pp. 186–193).

Liu, X., & Croft, W. B. (2006a). Experiments on retrieval of optimal clusters. Tech. Rep. IR-478, Center for Intelligent Information Retrieval (CIIR), University of Massachusetts.

Liu, X., & Croft, W. B. (2006b). Representing clusters for retrieval. In *Proceedings of SIGIR* (pp. 671–672), poster.

Liu, X., & Croft, W. B. (2008). Evaluating text representations for retrieval of the best group of documents. In *Proceedings of ECIR* (pp. 454–462).

Metzler, D., Diaz, F., Strohman, T., & Croft, W. B. (2005). Using mixtures of relevance models for query expansion. In *Proceedings of the Fourteenth Text Retrieval Conference (TREC).*

Ponte, J. M., & Croft, W. B. (1998). A language modeling approach to information retrieval. In *Proceedings of SIGIR* (pp. 275–281).

Porter, M. F. (1980). An algorithm for suffix stripping. *Program, 14*(3), 130–137, reprinted in K. S. Jones & P. Willett (Eds.), *Readings in information retrieval*, Morgan Kaufmann, 1997.

Preece, S. E. (1973). Clustering as an output option. In *Proceedings of the American Society for Information Science* (pp. 189–190).

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

Shah, C., & Croft, W. B. (2004). Evaluating high accuracy retrieval techniques. In *Proceedings of SIGIR* (pp. 2–9).

Song, F., & Croft, W. B. (1999). A general language model for information retrieval (poster abstract). In *Proceedings of SIGIR* (pp. 279–280).

Tao, T., Wang, X., Mei, Q., & Zhai, C. (2006). Language model information retrieval with document expansion. In *Proceedings of HLT/NAACL* (pp. 407–414).

Tombros, A., Villa, R., & van Rijsbergen, C. (2002). The effectiveness of query-specific hierarchic clustering in information retrieval. *Information Processing and Management, 38*(4), 559–582.

van Rijsbergen, C. J. (1979). *Information retrieval* (2nd ed.). Butterworths.

Voorhees, E. M. (2005). Overview of the TREC 2005 robust retrieval task. In *Proceedings of the Fourteenth Text Retrieval Conference (TREC)*.

Voorhees, E. M., & Harman, D. K. (2005). *TREC: Experiments and evlautaion in information retrieval*. The MIT Press.

Wei, X., & Croft, W. B. (2006). LDA-based document models for ad-hoc retrieval. In *Proceedings of SIGIR* (pp. 178–185).

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

Zamir, O., & Etzioni, O. (1998). Web document clustering: a feasibility demonstration. In *Proceedings of SIGIR* (pp. 46–54).

Zhai, C., & Lafferty, J. D. (2001). A study of smoothing methods for language models applied to ad hoc information retrieval. In *Proceedings of SIGIR* (pp. 334–342).

Zhang, B., Li, H., Liu, Y., Ji, L., Xi, W., Fan, W., Chen, Z., & Ma, W.-Y. (2005). Improving web search results using affinity graph. In *Proceedings of SIGIR* (pp. 504–511).