A Context-Dependent Relevance Model

E.K.F. Dang[1*], R.W.P. Luk[1], J. Allan[2]

* Corresponding author

1. Department of Computing,
The Hong Kong Polytechnic University,
Hung Hom, Hong Kong.
Email addresses: {cskfdang, csrluk }@comp.polyu.edu.hk

2. Center for Intelligent Information Retrieval,
School of Computer Science,
140 Governors Drive,
University of Massachusetts,
Amherst, MA 01003-9264.
Email address: allan@cs.umass.edu

Abstract

Numerous past studies have demonstrated the effectiveness of the *relevance model* (RM) for information retrieval (IR). This approach enables relevance or pseudo-relevance feedback to be incorporated within the language modeling framework of IR. In the traditional RM, the feedback information is used to improve the estimate of the *query* language model. In this article, we introduce an extension to RM in the setting of relevance feedback. Our method provides an additional way to incorporate feedback via the improvement of the *document* language models. Specifically, we make use of the context information of known relevant and non-relevant documents to obtain weighted counts of query terms for estimating the document language models. The context information is based on the words (unigrams or bigrams) appearing within a text window centered on query terms. Experiments on several TREC collections show that our context-dependent relevance model can improve retrieval performance over the baseline RM. Together with previous studies within the BM25 framework, our current study demonstrates that the effectiveness of our method for using context information in IR is quite general not limited to any specific retrieval model.

Introduction

Since the proposal of the language modeling approach in information retrieval (IR) by Ponte & Croft (1998), a large amount of subsequent work has demonstrated the effectiveness of this approach. The general language modeling approach regards both the query and documents to be generated according to some probability distribution, i.e. a statistical language model. In the original work of Ponte & Croft (1998), a language model is inferred for each document and the ranking of documents is based on the likelihood of generating the query by random sampling from the corresponding document language models. With no explicit notion of relevance, this approach differs from the traditional probabilistic IR methods, which mostly rank documents according to their probability of relevance (Robertson & Sparck-Jones, 1976).

It is well established that relevance feedback (RF) can help a retrieval system to return relevant documents (Rocchio, 1971; Salton & Buckley, 1990; Harman, 1992; Buckley et al, 1994). Even without user-interaction, 'blind feedback' or pseudo-relevance feedback (PRF) can also be useful (Buckley, Allan, Salton & Singhal, 1995). However, as the query-likelihood language modeling approach does not model relevance explicitly, at first it was not clear how any relevance feedback information may be used directly (e.g. Lafferty & Zhai, 2003). There has been much research in this regard and various feedback methods have been developed. Ponte (2000) extracted terms from the feedback documents to expand the initial query as a direct extension to the query-likelihood model. Another direction is to obtain better estimates of the *query* language model (Lafferty & Zhai, 2001) by using feedback information. Works taking this approach include the relevance model of Lavrenko & Croft (2001), the mixture model and divergence minimization model of Zhai & Lafferty (2001), and the regularized mixture model of Tao & Zhai (2006). Lv & Zhai (2009) performed a comparative study of these feedback methods and found the RM approach to be the most robust.

A reason why relevance feedback is useful in IR is that it may reveal the actual desired context of the information need. This is important because a well-known source of difficulty in IR is the ambiguity of query terms, with a query term possessing multiple meanings. For example, the query 'blackberry' on its own may refer to either a fruit or a mobile device. Various methods have also been studied in the past to deal with the query term ambiguity issue by determining the context of query terms, such as the latent word context model of Brosseau-Villeneuve, Nie & Kando (2014), and the Local Context Analysis of Xu & Croft (2000). Wu et al. (2007, 2008)

proposed a probabilistic retrieval model which mimics a human making a series of local relevance judgments on the 'document-contexts', or simply 'contexts', in a document. They defined a 'context' as a block of text centered on a query term, with the assumption that all relevant information is contained within such query contexts in a document. Dang et al. (2010) studied the use of context-dependent term weights for relevance feedback retrieval. These term weights are BM25 weights (e.g. Robertson & Walker, 1994; Robertson, 2004) adjusted according to the local probability of relevance of each query term occurrence in a document. They found their retrieval results to be comparable with a state-of-the-art adaption of the Markov Random Field (MRF) approach of Metzler & Croft (2005) approach for relevance feedback retrieval (Lease 2008).

In this article, we study an extension of the relevance model (RM) of Lavrenko & Croft (2001) with the use of context information. The novelty of our method is that it incorporates feedback via the improvement of *document* language models. Specifically, we adopt the Boost & Discount (B&D) procedure of Dang et al. (2010) in the setting of relevance feedback to obtain weighted counts of query terms for estimating the document language models. Thus, this new feedback technique differs from the traditional RM, which estimates an improved *query* language model. Another notable feature of the method is that information taken from known non-relevant documents is incorporated into the RM. This is unlike other studies which only utilize relevant documents in the feedback process (e.g. Diaz & Metzler, 2006; Lease, 2008). Experiments on several TREC collections show that our method outperforms the RM baseline. Hence, the significance of our work is the demonstration of the usefulness of context consideration in the language model framework. It provides a new way to go beyond the bag-of-words representation in this framework (Dang et al, 2014). Furthermore, together with the previous studies within the BM25 framework (Dang et al., 2010), our current results show that the effectiveness of our method for using context information in IR is quite general and not limited to any specific retrieval model.

While our method requires relevance feedback, in practice relevance information may be obtained by implicit feedback apart from demanding actual user interaction. For example, some studies have shown that clickthrough data in Web-search applications can be utilized for this purpose. Each 'click' on a page can be regarded as positive feedback (Joachims, Granka, Pan,

Hembroke, & Gay, 2005), while negative inferences may be drawn from pages that are bypassed (Das Sarma, Gollapudi, & Ieong, 2008).

The remainder of this article is organized as follows. In the next section, we present a review of related work, in particular past studies on extending the relevance model (RM) and the use of context information in IR. In the Model Formulations section, we first describe in detail the mathematical formulation of the baseline RM and the context-dependent BM25 term-weights. We then describe the incorporation of context dependence in RM, as well as the adaptation of the MRF approach for relevance feedback. In the Experiments section, we present a comparison of the empirical results of our method with the RM baseline as well as the adapted MRF method. Last, we provide a conclusion and a brief discussion of our future research direction.

Related Work

Extension to the Relevance Model (RM)

In a comparative study of several pseudo-relevance feedback methods in the language modeling framework (Lavrenko & Croft, 2001; Zhai & Lafferty, 2001; Tao & Zhao, 2006), Lv & Zhai (2009) found the RM approach of Lavrenko & Croft (2001) to be the most robust. There have been a number of past studies of extensions to RM, either involving pseudo-relevance feedback as in the original formulation or the feedback of true relevance information. Positive results have been reported by many of these works. Based on past research which showed the advantage of passage retrieval over document retrieval, particularly in the case where documents are long or contain multiple topics, Liu & Croft (2002) studied a passage-based implementation of RM. Each document was segmented into overlapping passages. A relevance model is constructed for the query and the passages instead of documents. Improvement over document-based RM was demonstrated. Diaz & Metzler (2006) used large external corpora, including the web corpus, in the relevance model, to provide feedback information in addition to the usual top ranked documents in the test collection. Lee, Croft & Allan (2008) introduced a cluster-based resampling method to obtain better documents than simply the top-ranked documents for pseudo-relevance feedback. They applied the documents belonging to the top-ranked clusters to the relevance model for query expansion. An extension to RM based on the proximity to query terms was studied by Lv & Zhai (2010). With the notion that words closer to query terms are more likely to be related to the query topic, their positional relevance model assigns weightings to

words in the feedback documents according to the distance to query terms. In the setting of relevance feedback, Lease (2008) employed an adaption of the Markov Random Field (MRF) method of Metzler & Croft (2005). Lease's method in effect incorporated a RM3 variant of the relevance model into MRF and attained better retrieval performance than other peer systems participating at the TREC2008 relevance feedback track. The RM3 variant will be further described in the Model Formulations section below.

Context based methods

Various methods using query contexts have shown promising results in past research. Generally such methods are based on term relations with the query terms. These relations may be discovered either globally or locally. Global term relations are obtained with statistics involving the whole corpus, while local relations are extracted from top-ranked documents retrieved for a given query. An example of a global method is the term context model of Pickens & MacFarlane (2006), who found a set of supporting terms for each query term based on global usage patterns.

Several local methods are described in the rest of this section. Among these, Xu & Croft (2000) introduced a Local Context Analysis, which is a query expansion technique based on co-occurrence with query terms in the top-ranked retrieved documents. The latent word context model of Brosseau-Villeneuve, Nie & Kando (2014) tackles query ambiguity via local Dirichlet allocation (Blei et al., 2003), applied to match the contexts of a document and the query. Bai et al. (2005) introduced an Information Flow (IF) method to specify term relations. Combined with pseudo-relevance feedback, query-centered IF relationships are extracted and used to expand the query language model. Wu et al. (2007) introduced a probabilistic retrieval model which mimics a user making a series of local relevance decisions on each document. They assumed that all relevance information is contained within a text window, called a 'context', centered on each query term occurrence in the document. Adopting this definition of a context, Dang et al. (2010) studied a context-dependent term weight in the setting of relevance feedback. Their procedure adjusts the BM25 weight of a query term according to the local probability of relevance of the context at each occurrence of the term. In this article, we study a novel extension to the relevance model by introducing context-dependence using the method of Dang et al. (2010).

Model Formulations

Baseline relevance model

In the original language modeling approach of Ponte & Croft (1998) for information retrieval, a unigram language model is estimated for each document. Documents are ranked according to the likelihood of the query being generated by the respective document language models. Lafferty & Zhai (2001) showed that this ranking can be derived in a more general risk minimization framework, with documents being ranked based on minimal Kullback-Leibler divergence between the query language model θ_Q and the document language model θ_D. Thus, the ranking score $S(Q,D)$ of a document D for the query Q is:

$$S(Q,D) = -D_{KL}(\theta_Q \parallel \theta_D) = -\sum_{w \in V} p(w \mid \theta_Q) \log \frac{p(w \mid \theta_Q)}{p(w \mid \theta_D)}, \quad (1)$$

where the sum is over all words w in the vocabulary V. Equation 1 may be rewritten as:

$$S(Q,D) = \sum_{w \in V} p(w \mid \theta_Q) \log p(w \mid \theta_D) + C_Q, \quad (2)$$

where C_Q is a factor independent of the document D and hence may be ignored for ranking purposes. With the above ranking formula, the retrieval problem becomes that of an estimation of the query and document language models, $\hat{\theta}_Q$ and $\hat{\theta}_D$.

For the estimate $p(w \mid \hat{\theta}_D)$, we adopt the commonly used maximum likelihood (ML) estimator with Dirichlet smoothing (Zhai & Lafferty, 2004):

$$p(w \mid \hat{\theta}_D) = \frac{tf(w,D) + \mu\, p(w \mid C)}{|D| + \mu} \quad (3)$$

where $tf(w,D)$ is the count (i.e. term frequency) of w in D, $|D|$ is the total number of words in D, $p(w|C)$ represents the collection language model, and μ is the constant Dirichlet prior. Further applying a ML estimator for the collection language model, i.e. $p(w|C) = tf(w,C)/|C|$, Equation 3 may be rewritten as:

$$p(w \mid \hat{\theta}_D) = \lambda \frac{tf(w,D)}{|D|} + (1-\lambda) \frac{tf(w,C)}{|C|} \quad (4)$$

where $tf(w,C)$ is the number of occurrences of w in the collection C, $|C|$ is the total number of words in C, and λ is a constant with a value between 0 and 1, given by $\lambda = |D| / (|D| + \mu)$.

Estimating the query model is conceivably harder than estimating the document model because queries are usually short and consequently there is a lack of training data. Several methods have been proposed in the past to obtain a better query language model than the maximum likelihood

(Lavrenko & Croft, 2001; Zhai & Lafferty, 2001; Tao & Zhao, 2006). These methods generally make use of pseudo-relevance feedback. Lv & Zhai (2009) reported a comparative study of these methods. They found the RM3 variant of the relevance model of Lavrenko & Croft (2001), as described below, to be superior.

Unlike the query-likelihood formulation of Ponte & Croft (1998), which considers the query as a random sample from a document language model, Lavrenko & Croft (2001) assumed both the query and the relevant documents to be samples from an unknown relevance model. Thus, assuming the words w in a relevant document and the terms of the query $Q = \{q_1, q_2,\ldots q_k\}$ to be sampled identically and independently from a unigram distribution θ_M, Lavrenko & Croft (2001) derived the probability

$$p(w \mid \hat{\theta}_Q) \propto \sum_{\theta_M \in \Theta} p(\theta_M)\, p(w \mid \theta_M) \prod_{i=1}^{k} p(q_i \mid \theta_M) \tag{5}$$

where Θ represents some finite universe of unigram distributions from which θ_M may be sampled. In practice for ad-hoc retrieval, in the spirit of pseudo-relevant feedback, Lavrenko & Croft (2001) restricted Θ to the models θ_D corresponding to the N_{PRF} top ranked documents obtained in an initial retrieval for the query $\{q_1, q_2,\ldots q_k\}$.

It was found that retrieval performance could be enhanced by an interpolation of the estimate of Equation 5 with the maximum likelihood (ML) query model (Abdul-Jaleel et al., 2004):

$$p(w \mid \theta_Q) = \alpha\, p(w \mid \widetilde{\theta}_Q) + (1-\alpha)\, p(w \mid \hat{\theta}_Q)\,, \tag{6}$$

where α is the mixing factor, $p(w \mid \hat{\theta}_Q)$ is given by Equation 5, and $p(w \mid \widetilde{\theta}_Q)$ represents the ML query model:

$$p(w \mid \widetilde{\theta}_Q) = \frac{tf(w,Q)}{|Q|}\,. \tag{7}$$

In Equation 7, $tf(w,Q)$ is the number of occurrences of the word w in the query Q, and $|Q|$ is the total number of words in the query. Equation 6 may be interpreted as an expansion of the original query Q and is generally referred as RM3 (e.g. Abdul-Jaleel et al., 2004; Diaz & Metzler, 2006; and Lv & Zhai, 2009 & 2010).

In our current work reported in this article, we focus on relevance feedback (RF) retrieval. In this case relevance judgment is made on N_{RF} feedback documents, giving a set R of known relevant documents and a set I of known non-relevant documents. Given the known relevant

documents, we may replace the $p(w \mid \hat{\theta}_Q)$ of Equation 5 by the 'true relevance model' (Diaz & Metzler, 2006):

$$p(w \mid \hat{\theta}_Q) = \frac{1}{|R|} \sum_{D \in R} p(w \mid \theta_D) . \tag{8}$$

Performing retrieval with RM3, Abdul-Jaleel et al. (2004) did not expand the query with all words contained in the top ranked feedback documents. Instead, the model was truncated to include 200 words of the highest probability. Similarly, in our RF experiments, we retain the top N_{QE} words ranked according to the probability of Equation 8, using a maximum likelihood estimate:

$$p(w \mid \hat{\theta}_Q) \propto \frac{1}{N_{RF}} \sum_{D \in R} \frac{tf(w,D)}{|D|} . \tag{9}$$

Summarizing the above, our final ranking function of a document D for a query Q in RF retrieval based on RM3 is obtained by substituting the various estimates in Equation 2, yielding:

$$S(Q,D) \propto \sum_{w \in Q_{QE}} s(w) \log\left(p(w \mid \hat{\theta}_D)\right) \tag{10}$$

where the sum is over all words in the expanded query Q_{QE},

$$s(w) = \alpha \frac{tf(w,Q)}{|Q|} + (1-\alpha) \frac{1}{N_{RF}} \sum_{D \in R} \frac{tf(w,D)}{|D|}, \tag{11}$$

and $p(w \mid \hat{\theta}_D)$ is given by Equation 4. We will use Equation 10 as the baseline retrieval model.

Context-dependence in the BM25 framework

Dang et al. (2010) demonstrated the effectiveness of incorporating context-dependence in the BM25 framework by adjusting the term frequency of the initial query terms according to the local probability of relevance of query 'contexts'. A context is defined as a block of text centered on a query term. In the Boost and Discount (B&D) procedure of Dang et al. (2010), the count of a specific query term within a document is either boosted or discounted depending on whether there is a stronger overall evidence of relevance or non-relevance for the contexts of the query term. Specifically, the unigram B&D procedure calculates an additive factor for the term frequency of the query term q_i in document D:

$$tf_{BD}^{(u)}(q_i, D) = tf(q_i, D) + \Delta tf_{BD}^{(u)}(q_i, D), \tag{12}$$

where

$$\Delta tf_{BD}{}^{(u)}(q_i, D) = M^{(u)} \sum_{k \in Loc(q_i, D)} [P^{(u)}(R \mid c(D, k, C_m^{(u)})) - 0.5]. \quad (13)$$

In Equation 13, $Loc(q_i,D)$ denotes the set of all locations in D where the query term q_i occurs, and $c(D, k, C_m^{(u)})$ is the context of size $C_m^{(u)}$ centered on the term at location k (i.e. with ($C_m^{(u)}$-1)/2 words on each side of k). As signified by the superscript (u), $P^{(u)}(R \mid c(D, k, c_m^{(u)})$ is the probability of relevance of the context according to the evidence provided by unigrams in the context. The multiplicative constant $M^{(u)}$ may be interpreted as a document length which converts the probability $P^{(u)}(R \mid c(D, k, C_m^{(u)}))$ to a frequency count. $P^{(u)}\left(R \mid c(D, k, C_m^{(u)})\right)$ has a value between 0 and 1, with $P(R) = 0$ and $P(R) = 1$ corresponding respectively to a firm belief that the context $c(D,k, C_m^{(u)})$ is non-relevant and a belief that it is relevant. The value $P(R)$=0.5 means there is total uncertainty regarding the relevance of the context, in which case Equation 13 indicates that the context gives zero contribution to $\Delta tf_{BD}^{(u)}(q_i, d)$. When there is evidence to suggest that the context is either relevant or non-relevant ($P(R)$>0.5 or $P(R)$<0.5 respectively), the term frequency will be either 'boosted' or 'discounted'.

The B&D procedure relies on some given relevance information to estimate the probability $P^{(u)}$ in Equation 13. In the relevance feedback (RF) setting, this information is provided by the known relevant and non-relevant documents. From these given documents, B&D extracts terms from the contexts of query terms to obtain sets of 'boost terms' $S_B^{(u)}(q_i)$, and 'discount terms' $S_D^{(u)}(q_i)$, respectively. As these terms will serve to provide the evidence of relevance or non-relevance respectively of a context, terms that are common to both sets are ambiguous and thus removed from the sets. The likelihood of a context in an unseen document being relevant or non-relevant may be deduced by noting its similarity with the sets $S_B^{(u)}(q_i)$ and $S_D^{(u)}(q_i)$. The B&D procedure thus estimates the probability $P^{(u)}$ according to a logistic regression model (e.g. Kleinbaum, 2002):

$$P^{(u)}(R \mid c(D, k, C_m^{(u)})) = f(\gamma_B^{(u)} X_B^{(u)}(D, k, C_m^{(u)}) - \gamma_D^{(u)} X_D^{(u)}(D, k, C_m^{(u)})), \quad (14)$$

where $f(\cdot)$ is the logistic function, i.e. $f(z) = 1/(1+e^{-z})$, $\gamma_B^{(u)}$ and $\gamma_D^{(u)}$ are logistic coefficients, and $X_B^{(u)}$ and $X_D^{(u)}$ represent positive and negative evidence, respectively, for the context $c(D, k, C_m^{(u)})$ being relevant. $X_B^{(u)}$ and $X_D^{(u)}$ are calculated by a sum of weighted counts of words in $c(D,$

k, $C_m^{(u)}$) matching those in the boost and discount sets, $S_B^{(u)}(q_i)$ and $S_D^{(u)}(q_i)$. Dang et al. (2010) found it effective to use an inverse document frequency (*idf*) weighting of any matched word t:

$$w(t,q_i) = idf(t)/idf_0, \tag{15}$$

where $idf(t)=\log_{10}((N-df(t)+0.5)/(df(t)+0.5))$ with N being the total number of documents in the collection, and $df(t)$ is the document frequency of term t in the collection. This expression of *idf* follows the BM25 form. The factor $idf_0 = \log_{10}((N+0.5)/0.5)$ normalizes the weight w to between 0 and 1. The *idf* gives reduced weighting to terms that are too common in the collection and hence would not serve as good discriminators of a document. While there is no q_i dependency in the idf weighting defined by Equation 15, such dependency may be included in more complex weightings.

In the BM25 formulation, both queries and documents are represented as vectors, with dimensions corresponding to words in the corpus. The vector elements are given by the actual term frequency counts, $tf(w,q)$ and $tf(w,D)$, of a word w in the query and document respectively. Having obtained the adjusted term frequency $tf_{BD}(q_i,D)$ of Equation 12, B&D computes a modified BM25 term weight:

$$BM25_{BD}(q_i,D) = f_{BM}^{(u)}(q_i,D) \times \log\left(\frac{N - df(q_i) + 0.5}{df(q_i) + 0.5}\right) \tag{16a}$$

$$f_{BM}^{(u)}(q_i,D) = \frac{tf_{BD}^{(u)}(q_i,D)}{abs(tf_{BD}^{(u)}(q_i,D)) + k\left(1 - b + b\frac{|D|_2}{\Delta}\right)}, \tag{16b}$$

where k and b are constant parameters, $|D|_2$ is the Euclidean length of the document vector D, and Δ is the average Euclidean length of all documents in the collection. As explained by Dang et al. (2010), the modified form $f_{BM}^{(u)}(q_i,D)$ of Equation 16b deals with cases where discount is sufficiently strong such that $tf_{BD}^{(u)}(q_i,D)$ (Equation 12) becomes negative. In particular, by including the $abs(\cdot)$ function, Equation 16b allows $f_{BM}^{(u)}(q_i,D)$ to become negative when $tf_{BD}^{(u)}(q_i,D)$ is negative, as an indication that the document D is likely to be non-relevant based on the evidence of the contexts.

Subsequently, Dang, Luk & Allan (2014) extended the work to include n-grams in the B&D procedure. In particular, they showed that including bigrams (n=2) could improve retrieval performance over unigram B&D, while larger values of n did not lead to further improvement. In

Dang et al. (2014), a bigram is defined as an ordered pair of words in a document after stop-word removal. An additional requirement is that the adjacent members (t_i, t_{i+1}) should not be separated by any punctuation, excluding hyphens, in the document. A hyphenated word is considered as a single word. Similar to unigram B&D, sets of boost and discount bigrams, $S_B^{(b)}(q_i)$ and $S_D^{(b)}(q_i)$, are extracted from the query term contexts of the known relevant and non-relevant documents. Bigrams common to both sets are removed from the sets. The local probability of relevance of each query term context in an unseen document is estimated according to the evidence of bigrams:

$$P^{(b)}(R \mid c(D,k,C_m^{(b)})) = f(\gamma_B^{(b)} X_B^{(b)}(D,k,C_m^{(b)}) - \gamma_D^{(b)} X_D^{(b)}(D,k,C_m^{(b)})), \quad (17)$$

where $X_B^{(b)}$ and $X_D^{(b)}$ are calculated by a sum of idf-weighted counts of boost and discount bigrams found in $c(D, k, C_m^{(b)})$. Term frequencies adjusted by bigram B&D, $tf_{BD}^{(b)}(q_i, D)$, as well as the BM25 factor $f_{BM}^{(b)}(q_i, D)$, may then be obtained in the same way as the unigram analogues, Equations 12, 13 and 16b. Treating unigrams and bigrams as independent term features, a combined BM25 factor is given by a linear mixture of the unigram and bigram components:

$$f_{BM}(q_i, D) = \alpha_m f_{BM}^{(u)}(q_i, D) + (1 - \alpha_m) f_{BM}^{(b)}(q_i, D) \quad (18)$$

where α_m is a mixing constant having a value between 0 and 1. Using Equation 18, the final adjusted BM25 term weight is:

$$BM25_{BD}(q_i, D) = f_{BM}(q_i, D) \times \log\left(\frac{N - df(q_i) + 0.5}{df(q_i) + 0.5}\right). \quad (19)$$

Dang et al. (2014) identified two ingredients in their bigram B&D method that were needed to attain robust performance improvement over using unigrams only. First, they found that in calculating the *idf* weighting for bigrams, it was necessary to use a 'local' document frequency. As opposed to a 'global' *df*, which counts the number of documents in the whole collection that contain the bigram, the local *df* limits the count to those documents that also contain one or more of the query terms. The local bigram *df* is thus query-dependent. Second, it was necessary to reduce the amount of noise by filtering out bigrams with large *df* values.

Context-dependent relevance model (CDRM)

Inspired by the promising results of the B&D method of Dang et al. (2010, 2014), we wish to investigate an adaptation of the method to introduce context-dependence in the language modeling framework. In the setting of relevance feedback, we apply the B&D procedure to the RM3 relevance model. B&D calculates a weighted frequency count of the initial query terms $\{q_1, \ldots q_k\}$ based on the available relevance information. Thus for these terms, in the smoothed ML document model of Equation 4, we modify the ML component by replacing the simple count of words $tf(q_i,D)$ with a B&D weighted count $tf_{BD}(q_i, D)$:

$$p_{BD}(q_i \mid \hat{\theta}_D) = \lambda \frac{tf_{BD}(q_i, D)}{|D|} + (1-\lambda)\frac{tf(q_i, C)}{|C|}. \tag{20}$$

Following Dang et al. (2014), we treat unigrams and bigrams as independent features. Then the weighed term frequency in Equation 20 is given by:

$$tf_{BD}(q_i, D) = tf(q_i, D) + \Delta tf_{BD}{}^{(u)} + \Delta tf_{BD}{}^{(b)}, \tag{21}$$

with

$$\Delta tf_{BD}{}^{(x)} = M^{(x)} \sum_{k \in loc(q_i, d)} [P^{(x)}(R \mid c(D, k, C_m^{(x)})) - 0.5], \tag{22}$$

where $x \in \{u,b\}$ is a label indicating either unigram (u) or bigram (b). In Equation 22, the probabilities $P^{(x)}$ are the same as in the BM25 formulation, given by Equations 14 and 17. It is possible that the amount of term frequency discounting is so large that the overall value of $p_{BD}(q_i \mid \hat{\theta}_D)$ in Equation 20 becomes negative. To avoid divergence of the score defined in Equation10, we modify it to the following:

$$S_{CDRM}(Q, D) \propto \sum_{w \in Q} s(w) \log\Big(\max\Big(p_{BD}(w \mid \hat{\theta}_D), \varepsilon\Big)\Big) + \sum_{w \in Q_{QE} \setminus Q} s(w) \log\Big(p(w \mid \hat{\theta}_D)\Big), \tag{23}$$

where ε is a small positive number, and the second sum is over the RM3 expanded query excluding the initial query terms.

In summary, the CDRM approach calculates a document ranking score $S_{CDRM}(Q, D)$ in RF retrieval by making use of the feedback information in two ways. First, as in the traditional RM approach, the known relevant documents are used to improve the estimation of the query model in the form of a query expansion (Equation 11). Second, CDRM additionally uses feedback information to estimate the document model of the initial query terms, $p_{BD}(q_i \mid \hat{\theta}_D)$ (Equation

20). The idea is that using B&D weighted counts, $tf_{BD}(q_i, D)$, which are based on the query term contexts, can yield a better estimate of the document model than using the simple ML.

Table 1 lists the various parameters involved in the CDRM.

TABLE 1. Summary of parameters

Symbol	Description
N_{RF}	Number of relevant judgments made in RF
N_{QE}	Number of query expansion terms selected from judged relevant documents
α	Mixing factor in the RM3 query expansion [Equations 6,11]
μ	constant Dirichlet prior which specifies the mixing factor λ in Equation 4
$C_B^{(x)}$	Size of context in known relevant documents for extracting Boost terms
$C_D^{(x)}$	Size of context in known irrelevant documents for extracting Discount terms
$C_m^{(x)}$	Size of context in unjudged documents for estimation of the local probability of relevance [Equations 14,17]
$\gamma_B^{(x)}$	Logistic coefficient for Boost terms [Equations 14,17]
$\gamma_D^{(x)}$	Logistic coefficient for Discount terms [Equations 14,17]
$M^{(x)}$	Multiplicative factor controlling the strength of B&D [Equation 13]
df_B	Document frequency threshold for Boost bigram pruning
df_D	Document frequency threshold for Discount bigram pruning

Adaption of the Markov Random Field approach in relevance feedback

Lease (2008) adapted the Markov Random Field (MRF) approach of Metzler & Croft (2005) for relevant feedback (RF) retrieval. This approach, called MRF-RF in the rest of this article, was found to be extremely effective, achieving top performance at the TREC 2008 Relevance Feedback Track (Buckley & Robertson, 2008). We will include this method in our experimental comparison of the performance of various techniques.

The MRF approach is a framework for modeling term dependencies. It assumes a joint distribution over queries and documents, $P_\Lambda(Q, D)$, which is parameterized by Λ and constructed from a graph G. The graph consists of a document node and a node for each of the query terms q_i. A property of the MRF is that a random variable represented at a node is independent of its non-neighbors given the observed values of its neighbors. Term dependencies

are thus specified by the configuration of edges of the graph. Furthermore, $P_\Lambda(Q,D)$ may be factored over cliques (i.e. fully connected subgraphs) of the graph:

$$P_\Lambda(Q,D) = \frac{1}{Z_\Lambda} \prod_{c \in C(G)} \psi(c;\Lambda) \tag{24}$$

where $C(G)$ is the set of cliques in G, $\psi(c;\Lambda)$ is a non-negative potential function, and Z_Λ is a normalization constant. Each potential function may be expressed in the form: $\psi(c;\Lambda)=\exp[\lambda_c f(c)]$, where $f(c)$ is a feature function and λ_c is the weight of the feature. For retrieval, documents are ranked according to the posterior:

$$\begin{aligned} P_\Lambda(D \mid Q) &= P_\Lambda(Q,D) / \mathrm{P}(Q) \\ &\overset{rank}{=} \sum_{c \in C(G)} \log \psi(c;\Lambda) \\ &= \sum_{c \in C(G)} \lambda_c f(c) . \end{aligned} \tag{25}$$

Metzler and Croft (2005) considered three types of query term features: individual terms (f_T), ordered contiguous terms appearing in the query (f_O), and proximity terms (f_U). Proximity terms refer to query terms which appear ordered or unordered within a window of N terms. Then, the ranking function of Equation 25 may be expressed as a mixture of three components:

$$P_\Lambda(D \mid Q) \propto \lambda_T f_T + \lambda_O f_O + \lambda_U f_U . \tag{26}$$

In Equation 26, each feature component is calculated by the maximum likelihood estimate of the feature in the document D, smoothed by its estimate in the collection.

In the case where $\lambda_T = 1$ and $\lambda_O = \lambda_U = 0$, Equation 26 becomes equivalent to the unigram language model. To adapt the MRF approach to relevance feedback, Lease (2008) replaced the individual term component, f_T, of Equation 26 by the RM3:

$$P'_\Lambda(Q,D) \propto \lambda_T S(Q,D) + \lambda_O f_O + \lambda_U f_U \tag{27}$$

where $S(Q,D)$ is given by Equation 10. We call this adaption the MRF-RF method.

Experimental environment and settings

System and relevance feedback settings

The environment of our CDRM experiments follows that of our previous work on context-dependent term weights in the BM25 framework (Dang et al., 2014), to allow for a direct

comparison of the results of the two approaches. The RF experiments were performed on the title queries of the TREC-2005, 6, 7 and 8 collections, with 50 queries for each collection. Some statistics of these collections are given in Table 2. Porter stemming (Porter, 1980) is applied to the collection documents and the queries, and stop-words are removed. Our retrieval system is passage-based, following past studies which found that passage-retrieval could perform better than whole document-retrieval, when documents are long or cover multiple topics (Callan, 1994; Kaszkiel & Zobel, 1997). Liu & Croft (2002) also found passage-retrieval to be effective in the language modeling setting. As in Dang et al. (2014), for our CDRM experiments our system defines passages as non-overlapping blocks in a document, with each passage consisting of 250 words. A language model is constructed for each passage instead of the whole document (Equation 4) and corresponding ranking scores (Equations 10 and 23) are obtained for all the passages. The retrieval output is a list of the original documents ranked according to the scores of their highest ranking constituent passages (Callan, 1994).

TABLE 2. Some statistics for TREC-6, -7, -8, 2005 test collections

	TREC-6	TREC-7	TREC-8	TREC-2005
Average # of title query terms	2.5	2.4	2.4	2.6
Av. # of relevant docs per query	92.22	93.48	94.56	131.22
Number of documents	556,077	528,155	528,155	1,033,461

The conception of RF is that a retrieval system presents to the user a list of N_{RF} documents returned by an initial retrieval. The user then judges these documents as relevant or non-relevant. Information from the judged documents is fed back to the system for a second retrieval. In our experiments, a feedback document is 'judged' as relevant or non-relevant by referring to the pool of judged documents provided by TREC. Any document which does not appear in the judged pool is regarded as non-relevant, following the usual practice in TREC evaluations (Voorhees, 1998).

The performance metric that we use for RF retrieval is the residual MAP (Ruthven & Lalmas, 2003). To calculate this metric, the N_{RF} documents judged by RF are first removed from both the pool of documents assessed by TREC and from our retrieved list. The residual MAP is defined as the MAP calculated based on the remaining (residual) collection. This metric has the advantage of avoiding artificial promotion of any judged relevant document to the top of the

retrieval result. Because the residual MAP depends on the exact list of N_{RF} documents judged by RF, values of the metric can only be compared between different retrieval methods if the same list of feedback documents is used for retrieval by each of these methods.

In this study, we primarily test the effectiveness of incorporating context-dependence in RM. Our context-dependent *tf* of Equations 21 and 22 is inspired by a corresponding formulation in the BM25 framework. Hence, apart from comparing with the standard RM baseline, we also compare the new method with the context-dependent BM25 approaches. In order to make the comparison of residual MAP values possible with our previous BM25 results as reported in Dang et al. (2014), we use the same set of N_{RF} feedback documents in our current experiments as in that previous work. This list is the top N_{RF} documents returned by a pseudo-relevant feedback (PRF) retrieval in a BM25 framework (Dang et al., 2014). RF experiments are performed with both N_{RF} =20 and N_{RF} =10.

In standard RM (Lavrenko & Croft, 2001; Abdul-Jaleel et al., 2004; Lv & Zhai, 2009), having obtained the improved query model based on either PRF or RF, a new ranking score is assigned to all documents according to Equation 10 to yield the final retrieved list. Corresponding in our current study, we assign new ranking scores to the passages of all documents in the collection. This amounts to performing a 're-retrieval' with the expanded query implied by Equation 11. Hence our current procedure differs from that of Dang et al. (2010, 2014), who performed a 're-ranking' of passages retrieved for the initial unexpanded query. Re-ranking of the passages conforms to the Query-centric assumption (Wu et al, 2008) adopted in Dang et al. (2010, 2014) that all relevant information is contained within contexts centered on query terms. The performance of 're-retrieval' and re-ranking of passages for RM will be compared in future studies.

Parameter calibrations

In our experiments, we calibrate the various parameters by maximizing the performance metric, i.e. residual MAP for a training collection (Metzler & Croft, 2005; Dang et al., 2010). Because of the large number of parameters, we do not perform an exhaustive grid search of the globally optimal set of parameters. Rather, we seek the local optimal value of each parameter, by varying the parameter over a range of values while keeping the remaining parameters constant.

The calibrated parameters are then applied to test collections. As in Dang et al. (2010, 2014), we use TREC-2005 as the training collection and the TREC-6, 7 and 8 collections for testing.

In RF retrieval, since a user makes relevance judgments on N_{RF} feedback documents, a piece of information that becomes available to the retrieval system is $N_R@N_{RF}$, the number of relevant documents among these judged documents. The number $N_R@N_{RF}$ generally differs from query to query. Dang et al. (2010) found that the set of parameters which maximized the MAP for queries with small $N_R@N_{RF}$ were quite different from the set of parameters optimal for queries with large $N_R@N_{RF}$. Hence, they suggested a 'Split' calibration scheme, whereby queries of the training collection were divided into two groups − those with $N_R@N_{RF} \leq 3$ and $N_R@N_{RF} > 3$, and separate parameter calibrations were performed for these two groups. For RF retrieval with the test collections, the test queries were likewise divided into the two groups according to $N_R@N_{RF}$ of the N_{RF} feedback documents, and the appropriate set of calibrated parameters is applied. We also adopt the Split calibration scheme in the current study. The calibrated parameters of the relevance model (RM) baseline and context-dependent RM methods are summarized in Table 3, for N_{RF} =20 and N_{RF} =10.

Table 3 supports the use of the Split calibration scheme for CDRM. The table shows that the sets of parameters calibrated for $N_R@N_{RF} \leq 3$ and $N_R@N_{RF} > 3$ queris are somewhat different. For example, the best values of M, the multiplicative factor controlling the strength of B&D, are generally much larger for $N_R@N_{RF} > 3$ than for $N_R@N_{RF} \leq 3$ queries. The table also shows that the known non-relevant documents can contribute to retrieval effectiveness, as indicated by non-zero $\gamma_D^{(x)}$ values for the discount components.

TABLE 3. Summary of the calibrated parameters for RM baseline, RM + B&D unigram, and RM + B&D uni- & bi-gram. Separate sets of parameters are obtained for queries with number of known relevant documents $N_R \leq 3$ and $N_R > 3$.

N_{RF}	Method	N_R	α	N_{QE}	μ	γ_B	γ_D	M	C_B	C_D	C_m	df_B	df_D
20	RM baseline	≤ 3	.12	100	600	-	-	-	-	-	-	-	-
		> 3	.1	240	400	-	-	-	-	-	-	-	-
	RM+ B&D unigram	≤ 3	.12	100	600	.4	.12	4	21	11	51	-	-
		> 3	.1	240	400	.4	.2	8	21	11	61	-	-
	RM + B&D uni- & bi-gram	≤ 3	0.12	100	600	Unigram							
						.4	.12	4	21	11	51	-	-
						Bigram							
						.1	.1	2	121	91	81	120	120
		> 3	.1	240	600	Unigram							
						.4	.2	8	21	11	61	-	-
						Bigram							
						.4	.12	22	121	91	81	160	80

N_{RF}	Method	N_R	α	N_{QE}	μ	γ_B	γ_D	M	C_B	C_D	C_m	df_B	df_D
10	RM baseline	≤ 3	.17	240	600	-	-	-	-	-	-	-	-
		> 3	.12	220	400	-	-	-	-	-	-	-	-
	RM+ B&D unigram	≤ 3	.17	220	1000	.28	.12	2	21	11	71	-	-
		> 3	.1	220	800	.2	.05	16	21	11	51	-	-
	RM + B&D uni- & bi-gram	≤ 3	.17	220	1000	Unigram							
						.28	0	2	21	11	71	-	-
						Bigram							
						.1	.2	6	111	111	101	200	120
		> 3	.1	220	800	Unigram							
						.2	0	16	21	11	51	-	-
						Bigram							
						.1	.2	16	111	111	101	200	120

More details on B&D settings

The B&D procedure calculates idf-weighted counts of the boost and discount n-grams, $X_B^{(x)}$ and $X_D^{(x)}$ of Equations 14 and 17. As in Dang et al. (2014), we include unigrams and bigrams in our B&D procedure for RM. Dang et al. (2014) tested several definitions of the bigram document frequency. They found that in order to yield robust performance improvement with bigram B&D over unigram B&D in the BM25 framework, it was necessary to use a 'local' bigram *df*, as described above in the *Context-dependence in the BM25 framework* section. In that work, which performs a re-ranking of passages in the RF stage instead of a re-retrieval, a local bigram *df* may be suitable. For the current study, since we perform a re-retrieval with the RM3 expanded query, involving passages of the whole collection, it may be more appropriate to use a

global bigram *df*. Hence, in the current experiments, we use the actual collection count of bigram *df*. Future studies of the B&D method for RM may include a further investigation of the use of a local bigram *df*.

Dang et al. (2014) also found that in order for B&D bigrams to be effective, it was necessary to reduce the amount of noise by filtering off the boost and discount bigrams that are too frequent. In the current study of bigram B&D in RM, we apply the *df* thresholds df_B and df_D, i.e., we remove from the sets $S_B^{(b)}(q_i)$ and $S_D^{(b)}(q_i)$ those bigrams with $df > df_B$ and $df > df_D$ respectively.

Experiment results

The results of RF retrieval by various methods for N_{RF} =20 and 10 are summarized in Table 4a. These results are also depicted graphically in Figure 1. We have tested two methods of context-dependent RM. The first method uses B&D with unigrams only (i.e. $M^{(b)} = 0$ in Equation 22) , while the second method include both unigrams and bigrams in B&D. The residual MAP values shown in the table are obtained for 50 title queries of each collection. The percentages are relative differences compared with the RM baseline.

The results show that both context-dependent RM methods can improve retrieval performance compared with the RM baseline. Numerically, both B&D methods yield better MAP values than the baseline for all collections and for both N_{RF} = 20 and N_{RF} = 10. The relative improvement is in the range of 4.4% to 9.1% for N_{RF} = 20, and 2.8% to 7.3% for N_{RF} = 10. As confirmed by the randomization test (Smucker, Allan & Carterette, 2007), the improvement is statistically significant at the 95% confidence level (Table 4b) in almost all cases. The only exception is the *n*-gram (uni- & b-igram) B&D result for N_{RF} = 20 in TREC-8. Comparing the results of the two methods of context-dependent RM, using *n*-gram B&D gives better MAP values than using only unigram B&D in most cases, the exception again being N_{RF} = 20 in TREC-8. A possible study for refining the B&D method for context-dependent RM is a test of using a local bigram *df* (Dang et al., 2014).

TABLE 4a. Summary of residual MAP values obtained by various methods in relevance feedback with 20 or 10 judgments, for four TREC collections. The superscripts *a* and *b* indicate statistically significant improvement (95% confidence level, randomization test) over the RM baseline and RM+B&D unigram methods, respectively. The percentages are relative difference compared with the RM baseline. The best performance for each collection is highlighted in bold.

N_{RF}	TREC	RM	RM + B&D unigram	RM + B&D uni- & bi-gram	BM25 + B&D unigram	BM25 + B&D uni- & bi-gram	MRF-RF
20	2005	.3033	.3185[a] +5.01%	**.3310** [ab] +9.13%	.3163 +4.29%	.3249 +7.12%	.2980 -1.75%
	6	.2618	.2747 [a] +4.93%	**.2835** [ab] +8.29%	.2619 +0.04%	.2729 +4.24%	.2364 -9.7%
	7	.2297	.2399 [a] +4.44	**.2463** [ab] +7.23%	.2295 -0.09%	.2363 +2.87%	.2402 +4.57%
	8	.2786	**.2967** [a] +6.5%	.2925 +4.99%	.2776 -0.36%	.2913 +4.56%	.2476 -11.13%
10	2005	.2996	.3081 [a] +2.84%	**.3165** [ab] +5.64%	.3044 +1.6%	.3089 +3.1%	.3083 +2.9%
	6	.2767	.2885 [a] +4.26%	**.2902** [a] +4.88%	.2609 -5.71%	.2677 -3.25%	.2421 -12.5%
	7	.2332	.2425 [a] +3.99%	**.2502** [ab] +7.29%	.2348 +0.69%	.2439 +4.59%	.2506 +7.46%
	8	.2884	.2964 [a] +2.77%	**.2990** [ab] +3.68%	.2715 -5.86%	.2795 -3.09%	.2697 -6.48%

Table 4b. Randomization test *p*-values for the comparison between retrieval methods. The superscripts *a* and *b* indicate statistically significant improvement (95% confidence level) over the RM baseline and RM+B&D unigram methods, respectively.

N_{RF}	TREC	RM + B&D unigram vs. RM	RM + B&D uni- & bi-gram vs. RM	BM25 + B&D unigram vs. RM	BM25 + B&D uni- & bi-gram vs. RM	MRF-RF vs. RM	RM + B&D uni- & bi-gram vs. RM + B&D unigram	RM + B&D unigram vs. MRF-RF	RM + B&D uni- & bi-gram vs. MRF-RF
20	2005	.002 [a]	.002 [a]	.303	.107	.453	.014 [b]	.039	.014
	6	.0004 [a]	.000 [a]	.996	.411	/	.002 [b]	.000	.000
	7	.023 [a]	.007 [a]	.982	.568	.245	.048 [b]	.968	.508
	8	.008 [a]	.546	.957	.475	/	.983	.000	.000
10	2005	.0026 [a]	.000 [a]	.733	.496	.237	.003 [b]	.978	.373
	6	.0022 [a]	.000 [a]	.303	.562	/	.521	0	0
	7	.000 [a]	.000 [a]	.866	.324	.051	.048 [b]	.365	.957
	8	.021 [a]	.008 [a]	.173	.485	/	.044 [b]	.001	.000

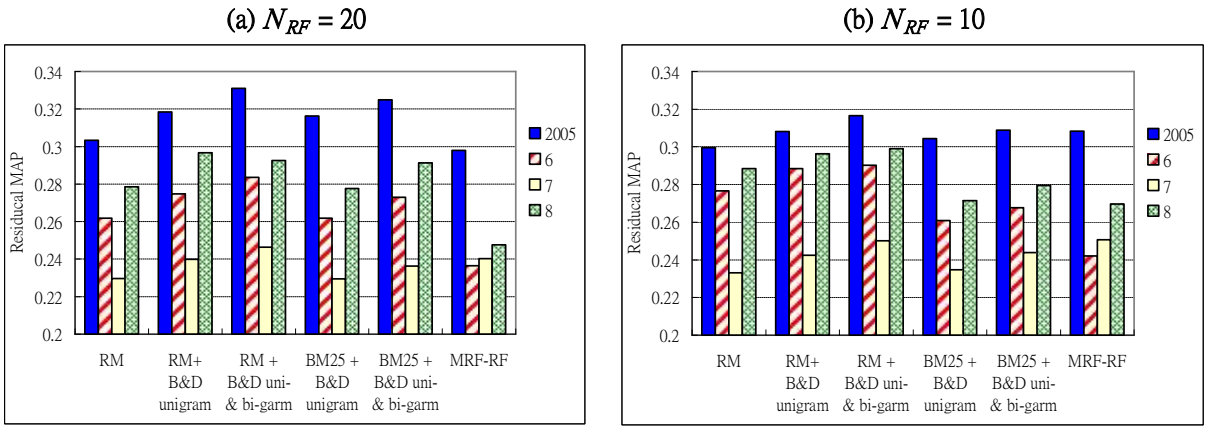

Figure 1. Residual MAP obtained by various relevance feedback methods on the TREC-2005, -6, -7 and -8 collections, for (a) N_{RF} = 20 and (b) N_{RF} = 10.

Table 4a also includes the retrieval results for context-dependent BM25 (either with only unigrams in B&D or with both unigrams and bigrams) and MRF-RF, the adaption of MRF by Lease (2008) for RF retrieval. Calibrations of these methods were performed for TREC-2005, as in the case of the RM methods. These results were previously reported in Dang et al. (2014). Since the first component of the document-ranking formula of the MRF-RF method (Equation 27) is essentially the standard RM, in principle MRF-RF should perform better than the RM baseline. Table 4a shows that this is the case for N_{RF} =10. However for N_{RF} =20, the MRF-RF result (MAP=0.2980) is slightly lower than the RM baseline (MAP=0.3033), though the difference is not statistically significant (p=0.453). One reason for the smaller MRF-RF value may be related to the difference between the document-based Indri system (e.g. Metzler & Croft, 2005) used to obtain this result, and our passage-based system. Overall, while the BM25 and MRF-RF methods can yield better MAP values than the RM baseline in some cases, none of the improvement is statistically significant at the 95% confidence level. This contrasts with our RM method with unigram B&D, which yields statistically significant improvement over the baseline across all collections.

Conclusion

We have studied an extension to the relevance model (RM) of Lavrenko & Croft (2001) in the relevance feedback (RF) setting. While the traditional RM utilizes feedback information to improve the estimate of the query language model, our method additionally incorporates context

information from the feedback documents to adjust the document language models. In traditional RM, the document language model is a smoothed maximum likelihood (ML) estimate, calculated based on actual counts of term occurrences in a document. Our method either boosts or reduces the counts of the query terms according to the evidence of relevance or non-relevance inferred from the local contexts. The context information is based on the unigrams or bigrams appearing within a text window centered on the query terms. Our unigram-based context-dependent RM showed statistically significant performance improvement over the traditional RM across all the tested TREC-6, -7, -8 and -2005 collections, with either 10 or 20 feedback documents. Further improvement could be obtained with bigram-based context-dependent RM, with the improvement being statistically significant in three out of four of the tested collections. Together with the previous studies within the BM25 framework (Dang et al., 2010 & 2014), our current results show that the effectiveness of our method to make use of context information in IR is quite general and not limited to any specific retrieval model.

Future work on our context method may proceed in two directions. The first is a refinement of the current method, such as the weighting of context terms according to the distance from the query term at the context center or other positional considerations. Regarding the bigram-based method, further work may also include studies using a local document frequency for the bigrams (Dang et al., 2014) instead of the collection *df* used in the current work. Another direction is the application of our method to other approaches in IR, such as MRF (Metzler & Croft, 2005). In fact, since context dependence is introduced in our procedure by means of a weighted count of query terms, it may be applied to any IR method which involves a term frequency count. Furthermore, it may be of interest to investigate the incorporation of context-dependence beyond IR, such as in text categorization (e.g. Bekkerman & Allan, 2003; Liu, 2010).

Acknowledgement

Robert Luk thanks the Center for Intelligent Information Retrieval, University of Massachusetts, for facilitating his development of the basic IR system when he was on leave there. This work was partly supported by the Center for Intelligent Information Retrieval.

References

Abdul-Jaleel, N., Allan, J., Croft, W. B., Diaz, F., Larkey, L., Li, X., Metzler, D., Smucker, M. D., Strohman, T., Turtle, H., & Wade, C. (2004). UMass at TREC 2004: Novelty and hard. In TREC ’04.

Bekkerman, R., & Allan, J. (2003). Using bigrams in text categorization. CIIR Technical Report, IR-408, University of Massachusetts, 2003.

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

Brosseau-Villeneuve, B., Nie, J.-Y., & Kando, N. (2014). Latent word context model for information retrieval. Information Retrieval, 17, 21-51.

Buckley, C., Salton, G., & Allan, J. (1994). The effect of adding relevance information in a relevance feedback environment. In Proceedings of ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 292-300).

Buckley, C., Allan, J., Salton, G., & Singhal, A. (1995). Automatic query expansion using Smart: Trec3. In Proceedings of the Third Text Retrieval Conference (TREC-3) (pp. 69-80).

Buckley, C., & Robertson, S. (2008). Relevance feedback track overview: TREC2008. In Proceedings of the Seventeenth Text Retrieval Conference (TREC).

Callan, J. P. Passage-based evidence in document retrieval. (1994). In Proceedings of ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 302-310).

Dang, E. K. F., Luk, R. W. P., Allan, J., Ho, K. S., Chan, S. C. F., Chung, K. F. L., & Lee, D. L. (2010). A new context-dependent term weight computed by Boost and Discount using relevance information. Journal of the American Society of Information Science and Technology, 61(12), p. 2514-2530.

Dang, E. K. F., Luk, R. W. P., & Allan, J. (2014). Beyond bag-of-words: Bigram-enhanced context-dependent term weights. Journal of the American Society of Information Science and Technology, 65(6), p.1134-1148.

Das Sarma, A., Gollapudi, S., & Ieong, S. (2008). Bypass rates: reducing query abandonment using negative inferences. In Proceedings of ACM SIGKDD Conference on Knowledge Discovery and Data Mining (pp. 177-185).

Diaz, F., & Metzler, D. (2006). Improving the estimation of relevance models using large external corpora. In Proceedings of ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 154-161).

Harman, D. (1992). Relevance feedback revisited. In Proceedings of ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 1-10).

Joachims, T., Granka, L., Pan, B., Hembroke, H., & Gay, G. (2005). Accurately Interpreting Clickthrough Data as Implicit Feedback. In Proceedings of ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 154-161).

Kaszkiel M., & Zobel, J. (1997). Passage retrieval revisited. In Proceedings of ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 178-185).

Kleinbaum, D.G. (2002). Logistic regression: A self-learning text. New York: Springer.

Lafferty, J. D., & Zhai, C. X. (2001). Document language models, query models, and risk minimization for information retrieval. In Proceedings of ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 111-119).

Lafferty, J. D., & Zhai, C. X. (2003). Probabilistic relevance models based on document and query generation. In W.B. Croft & J. Lafferty (Eds.) Language modeling for information retrieval, Kluwer Academic Publishers (pp. 1-10).

Lavrenko, V., & Croft, W. B. (2001). Relevance-based language model. In Proceedings of ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 120-127).

Lease, M. (2008). Incorporating relevance and pseudo-relevance feedback in the Markov random field model. In Proceedings of the 17th Text Retrieval Conference (TREC).

Lee, K. S., Croft, W. B., & Allan, J. (2008). A cluster-based resampling method for pseudo-relevance feedback. In Proceedings of ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 235-242).

Liu, X., & Croft, W. B. (2002). Passage retrieval based on language models. In Proceedings of CIKM'02 (pp.375-382).

Liu, R.-L. (2010). Context-based term frequency assessment for text classification. Journal of the American Society of Information Science and Technology, 61(2), p.300-309.

Lv, Y., & Zhai, C. X. (2009). A comparative study of methods for estimating query language models with pseudo feedback. In Proceedings of CIKM'09 (pp. 1895-1898).

Lv, Y., & Zhai, C. X. (2010). Positional relevance model for pseudo-relevance feedback. In Proceedings of ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 178-185).

Metzler, D., & Croft, W. B. (2005). A Markov random field model for term dependencies. In Proceedings of ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 472-479).

Pickens, J. & MacFarlane, A. (2006) Term context models for information retrieval. In Proceedings of ACM Conference on Information and Knowledge Management (pp. 559-566).

Ponte, J. M., & Croft, W. B. (1998). A language modeling approach to information retrieval. In Proceedings of ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 275-281).

Ponte, J. M. (2000). Language models for relevance feedback. In W.B. Croft (Ed.) Advances in information retrieval: Recent research from the Center for Intelligent Information Retrieval, Kluwer Academic Publishers. (pp. 73-95).

Porter, M. (1980). An algorithm for suffix stripping. Program, 14(3), p. 130-137.

Rocchio, J. J. (1971). Relevance feedback in information retrieval. In G. Salton (Ed.) The SMART retrieval system: Experiments in automatic document processing (pp. 313-323).

Robertson, S. E., & Spärck-Jones, K. (1976). Relevance weighting of search terms. Journal of the American Society for Information Science, 27(3), pp. 129-146.

Robertson, S. E., & Walker, S. (1994). Some simple approximations to the 2-Poisson model for probabilistic weighted retrieval. In Proceedings of ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 232-241)

Robertson, S. E. (2004). Understanding inverse document frequency: On theoretical arguments for IDF. Journal of Documentation, 60(5), pp. 503-520.

Ruthven, I., & Lalmas, M. (2003). A survey on the use of relevance feedback for information access systems. The Knowledge Engineering Review, 8(2), pp. 95-145.

Salton, G., & Buckley, C. (1990). Improving retrieval performance by relevance feedback. Journal of the American Society for Information Science, 41(4), pp. 288-297.

Smucker, M. D., Allan, J., & Carterette, B. (2007). A comparison of statistical significance tests for information retrieval evaluation. In Proceedings of ACM Conference on Information and Knowledge Management (pp. 623-632).

Tao, T., & Zhai, C. (2006). Regularized estimation of mixture models for robust pseudo-relevance feedback. In Proceedings of ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 162-169).

Voorhees, E. M. (1998). Variations in relevance judgments and the measurement of retrieval effectiveness. In Proceedings of ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 315-323).

Wu, H. C., Luk, R. W. P., Wong, K. F., & Kwok, K. L. (2007). A retrospective study of a hybrid document-context based retrieval model. Information Processing and Management, 43, p. 1308-1331.

Wu, H. C., Luk, R. W. P., Wong, K. F., & Kwok, K. L. (2008). Interpreting TF-IDF weights as making relevance decisions, ACM Transactions on Information Systems, 26(3), Article 13.

Xu, J., & Croft, W. B. (2000). Improving the effectiveness of information retrieval with local context analysis. ACM Transactions on Information Systems, 18(1), p. 79-112.

Zhai, C., & Laferty, J. (2001). Model-based feedback in the language modeling approach to information retrieval. In Proceedings of CIKM'01, (pp.403-410).

Zhai, C., & Laferty, J. (2004). A study of smoothing methods for language models applied to information retrieval. ACM Transactions on Information Systems, 22(2), p. 179-214.