Journal of Universal Computer Science, vol. 30, no. 11 (2024), 1511-1528
submitted: 14/9/2023, accepted: 11/4/2024, appeared: 28/10/2024

A Study of Word Bigrams for Pseudo-relevance Feedback in Information Retrieval

A contribution to the Forum for Negative Results

Edward Kai Fung Dang
(Department of Computing, The Hong Kong Polytechnic University, Hong Kong
https://orcid.org/0000-0002-4917-7216, ekfdang@gmail.com)

Robert Wing Pong Luk
(Department of Computing, The Hong Kong Polytechnic University, Hong Kong
https://orcid.org/0000-0002-9310-8867, csrluk@gmail.com)

Qing Li
(Department of Computing, The Hong Kong Polytechnic University, Hong Kong
https://orcid.org/0000-0003-3370-471X, qing-prof.li@polyu.edu.hk)

Abstract: Traditional information retrieval models mostly adopt a term independence assumption and are based on single terms or unigrams. Past efforts have attempted to go beyond this assumption, such as by using contiguous terms (i.e. word n-grams) or terms appearing in proximity. One such approach employs pseudo-relevance feedback (PRF) in an extended BM25 model, with an expanded query containing bigrams and proximity word pairs besides unigrams. However, the benefit of this approach over the traditional unigram PRF remains inconclusive. We speculate the uncertain effectiveness of bigram PRF in this past work is due to: (1) The new bigrams obtained from the expanded query may be formed by pairing unigrams drawn from different documents. These are potentially noise instead of relevant concepts; (2) The collection statistics of n-grams needed to calculate the document ranking functions, such as their document frequency, is not available in retrieval. Only estimates of these quantities are used instead. We suggest that these issues may be overcome by extracting word n-grams as single units in query expansion, and employing a document index that contains both unigrams and word n-grams. We demonstrate the approach for the case of bigram PRF in an extended BM25. Retrieval experiments are conducted on a range of standard test collections. For the majority of tested collections, the difference between values of the evaluation metrics (Mean Average Precision and the precision-oriented NDCG@20) obtained by our bigram PRF and the unigram PRF baseline is not statistically significant. Thus, our bigram PRF fails to improve over unigram PRF robustly across collections. An analysis of our results reveals 'query drift' due to bigram query expansion terms that represent too-broad topics as a cause for the failure of our approach.

Keywords: bigrams, pseudo-relevance feedback, information retrieval, evaluation
Categories: H.3.1, H.3.3
DOI: 10.3897/jucs.112725

1 Introduction

One of the continuing research directions in information retrieval (IR) is the development of new methods to improve retrieval effectiveness. Many of the traditional information

retrieval models [Dang et al. 2022] adopt the term independence assumption, whereby the occurrences of different terms are statistically independent, both in documents and queries. Correspondingly, these retrieval models are based on individual terms or unigrams, using a 'bag-of-words' representation of documents and queries. However, the term independence assumption is counter-intuitive for textual data. For example, the meaning of word pairs like 'black hole' or 'Middle East' would be lost if the words in the pairs are taken separately. Thus, there has been much research effort to go beyond 'bag-of-words' in IR.

In this regard, some works utilize contiguous terms, such as word n-grams [Song and Croft 1999] or biterms [Srikanth and Srihari 2002], within the language model (LM) framework [Ponte and Croft 1998], achieving some improvement over unigram models. However these early works only used small test collections (size of 2GB or less) in retrieval experiments, so that the effectiveness of these approaches on large collections (such as a few hundred GB or above) needs to be confirmed. Methods utilizing non-adjacent terms have also shown some success. These methods make use of query terms in proximity, i.e. occurring within a given distance. For example, proximity terms may be utilized within the vector space model [Mishne and de Rijke 2005]. Some works introduce query term proximity heuristics [Rasolofo and Savoy 2003, Büttcher et al. 2006, He et al. 2011] to enhance established models like the classical probabilistic BM25 [Robertson et al. 1995]. Significant improvement has been obtained by the Markov Random Field (MRF) approach, which provides a model of various degrees of dependency between query terms [Metzler and Croft 2005].

We are motivated to examine the use of word n-grams in the setting of pseudo-relevance feedback (PRF), an established technique involving query expansion (QE), whereby after an initial retrieval using the given query, a number of terms are automatically extracted from the top ranked retrieved documents and added to the query for a second retrieval (e.g. [Buckley et al. 1995]). The reason for our motivation is that while only a few word n-grams can be made up from the initial query, which often consists of a small number of terms as are typical in web queries [Spink et al. 2002], the QE process in PRF allows more word n-grams to be added to the query. The use of term proximity methods in PRF has been studied by [He et al. 2011], in a proximity-based BM25 model, called BM25P. In their approach, an initial retrieval is performed to extract single terms (unigrams) to make up the expanded query. A bigram model is then applied in the PRF stage, whereby in the second retrieval, term pairs may be formed either between all possible pairs of terms in the expanded query, or only between terms in the original query. In either case, improvements over the unigram model with statistical significance is observed only on some of the smaller tested collections. Thus, the usefulness of n-grams or proximity terms for PRF with basis retrieval models like BM25 remains inconclusive.

We speculate that the limited effectiveness of bigram PRF in [He et al. 2011] is due to: (1) The individual terms extracted to make up the expanded query may come from different top ranked documents. Hence, the new bigrams may be formed by pairing unigrams drawn from different documents. These are potentially noise instead of relevant concepts; (2) Because of the use of a conventional index of single terms, which is prevalent in most IR systems [Perry and Willett 1983], the collection statistics of n-grams needed to calculate the document ranking functions, such as their document frequency (*df*, i.e. the number of documents in the collection in which an n-gram occurs), is not available during retrieval run-time. Only estimates of these quantities are used instead [He et al. 2011, Mitra et al. 1997]. In this work, we tackle these issues with the aim to determine conclusively whether n-grams can enhance the effectiveness of PRF retrieval.

Our approach consists of the following strategies: (1) We enable both single terms

and n-grams to be selected from each top ranked document to form the expanded query in PRF, unlike the procedure of [He et al. 2011]; (2) We employ a document index that contains n-grams in additional to unigrams, so that collection statistics of n-grams are available during retrieval run-time. The inverted index is used, as is common in IR systems [Perry and Willett 1983, Dang et al. 2015]. The inverted index maps each term found in the corpus to a list of postings corresponding to the documents in which the term occurs. Each posting typically contains a document identifier, together with the term's number of occurrences (tf or term frequency) in the document. An auxiliary file may store a dictionary of all the terms, with statistics such as df and total number of occurrences in the corpus (i.e. collection frequency, CF). Therefore, exact n-gram df values are readily available during retrieval, avoiding the need of estimation.

Similar to [He et al. 2011], we use BM25 as the basis model. Retrieval experiments are performed on a wide range of standard Text REtrieval Conference (TREC) test collections. We find PRF using a mixture of unigrams and bigrams in QE to have some improvement in Mean Average Precision (MAP), which is an evaluation metric that measures both precision and recall, over unigram-based PRF in the web collections. However, in the majority of cases, the differences are not statistically significant. Also, compared with unigram-based PRF, there is either minimal difference or a degradation in the precision-oriented NDCG@20 (Normalized Discounted Cumulative Gain at the top 20 retrieved documents), with the differences being not statistically significant for all the collections. Thus overall in terms of both evaluation metrics, our approach does not improve over unigram PRF robustly across collections, in contrary to our expectation. An analysis of the retrieval results reveals some inadequacies of our approach. While our approach helps to reduce non-relevant bigrams in the expanded query, some of the added bigrams are not sufficiently specific but represent broad topics that do not focus only on the given query. These act as an added source of 'query drift' resulting in a degradation in retrieval effectiveness.

In summary, the contributions of this article are: (1) We propose the approach of PRF retrieval based on a document index that includes both unigrams and bigrams. Also, new bigrams in the expanded query are extracted as single units from top-ranked documents; (2) Retrieval experiments are performed on test collections covering a wide range of sizes to verify the effectiveness of the approach; (3) Our results show that for the majority of tested collections, the difference between values of both MAP and the precision-oriented NDCG@20 obtained by our bigram PRF and the unigram PRF baseline is not statistically significant; (4) An analysis of the retrieval results provide an explanation of the failing of our approach. Both noise terms that are non-relevant to the query and bigram QE terms that represent too-broad topics lead to 'query drift' that harms retrieval effectiveness.

2 Background

2.1 Term dependency methods in retrieval

There has been much effort in IR to incorporate term dependency in retrieval models. While the LM [Ponte and Croft 1998] assumes all words occur independently, the generalized language model (GLM) [Song and Croft 1999] is an extension that models the probability of sequences of n words, or word n-grams. For example, the bigrams modeled by GLM consist of ordered word pairs. A further extension is explored by [Srikanth and Srihari 2002] with a relaxation of the order of words. That is, unordered word pairs, or biterms, are used in the model. The works of [Song and Croft 1999]

and [Srikanth and Srihari 2002] show that bigram or biterm models can bring some improvement in retrieval effectiveness over the unigram LM. However, these early works only use small test collections with a size of 2GB or less. Hence, the effectiveness of their methods on modern applications with large data sets (such as in the terabyte regime) is unknown.

The use of n-grams and proximity terms has been studied within the vector space model (VSM) [Mishne and de Rijke 2005]. In the VSM, queries and documents are represented as vectors. The document ranking function is a summation of the weights of matching query terms that make up the representation of document vectors. [Mishne and de Rijke 2005] employ a fixed basic ranking formula in which a term may be a single word, a phrase (ordered contiguous words) or words in proximity. They find the use of phrase and proximity terms to be effective for web retrieval.

Some works take into account proximity terms by adding proximity scoring heuristics to the BM25 document ranking function. For example, [Rasolofo and Savoy 2003] consider the set of all possible pairs of query terms. They introduce a proximity score which is a function of the inverse square distance between the words in each pair. They find some improvement in the precision of top-ranked documents, but only marginal improvement in MAP, which is based on both precision and recall. Using a similar method as [Rasolofo and Savoy 2003], the work of [Büttcher et al. 2006] suggests that the enhancement in retrieval effectiveness with the use of proximity terms is larger as the size of the text collection increases.

2.2 Related Work

Here we discuss in more detail the work of [He et al. 2011], which is related to this article. A BM25P model is introduced by [He et al. 2011] to improve the classical BM25 model by utilizing term proximity evidence. The approach of BM25P is similar to that of [Mishne and de Rijke 2005], except BM25P extends the BM25 document ranking function. [He et al. 2011] find that variants of the BM25P model based on bigrams alone do not perform as good as the unigram-based BM25. However, linear combinations of the unigram and bigram variants of BM25P outperform the unigram BM25.

The use of PRF in BM25P is also studied by [He et al. 2011]. Our proposed approach addresses some issues of PRF in BM25P: (1) The individual terms extracted to make up the expanded query may come from different top ranked documents. Hence, the new bigrams may be formed by pairing unigrams drawn from different documents. These are potentially noise instead of relevant concepts; (2) Because of the use of a conventional index of single terms, which is prevalent in most IR systems [Perry and Willett 1983], the collection statistics of n-grams needed to calculate the document ranking functions, such as their document frequency (*df*, i.e. the number of documents in the collection in which an n-gram occurs), is not available during retrieval run-time. Only estimates of these quantities are used instead [He et al. 2011, Mitra et al. 1997].

These issues are illustrated by a sample TREC title query and typical QE terms extracted by various PRF methods (Table 1). Unigram PRF returns single terms extracted from top-ranked documents of an initial retrieval. Table 1 shows the top six unigram QE terms in the example. In conventional bigram PRF, such as in the work of [He et al. 2011], bigarms that are added to the expanded query are obtained by merging the QE terms of unigram PRF. Samples of such composite terms are shown in Table 1. Generally such composite terms are not actual phrases that appear in documents (e.g. ‘academy actress’). If the two component terms occur in separate sentences or different documents, the composite bigram is potentially noise rather than a relevant concept (e.g. ‘actress

Query (Q685)	Oscar winner selection
QE terms (unigram PRF)	academy, film, award, actress, wheeler, best
QE terms (bigram PRF [He et al. 2011])	'academy film', 'academy award', 'academy actress', 'academy wheeler', 'film award', 'actress wheeler'
QE terms (bigram PRF, our approach)	'academy award', 'best picture', 'motion picture', 'best actress', 'best director', 'rain man'

Table 1: Example of a TREC title query and samples of query expansion terms

wheeler'). As for our approach, all the bigrams in the expanded query are necessarily actual phrases extracted from top-ranked documents. We expect these QE bigrams less likely to be noise.

Last, [He et al. 2011] only evaluate their methods by MAP but not by any precision-oriented metric such as NDCG, which is important for applications like web search, where users are mostly interested in a small number of top-ranked documents. In this regard, we report our results in both MAP and NDCG@20.

3 Our approach

In Section 3.1 we discuss two ways of creating word bigrams during indexing of the corpus, depending on the condition whether the pair of words making up a bigram may or may not be separated by stopwords. Section 3.2 describes how PRF is extended to utilize word bigrams in addition to unigrams.

3.1 Indexing of the corpus

Our retrieval system employs an inverted index [Perry and Willett 1983] that contains unigrams and n-grams. While the approach may be applied to all $n \geq 2$, in this article we restrict to $n = 2$, that is, up to bigrams are included. The indexed terms are extracted from the documents after stemming based on Porter's algorithm [Porter 1980] and stopwords removal following the Indri stopword list[1]. Bigrams are defined as contiguous pairs of words that are not separated by 'breaks'. Such breaks include punctuations (but excluding the hyphen) and new paragraphs. Furthermore, we consider two cases — (1) SNB: stopwords not treated as breaks, and (2) SB: stopwords treated as breaks. The set of SNB bigrams extracted from a document is a superset of the SB bigrams — the set of SNB bigrams consists of SB bigrams plus pairs of words separated by stopwords only. Thus SNB bigrams include ordered pairs of words in proximity, though the allowed distance between words is not governed by a specific window size, but variant depending on the text. Table 2 illustrates the indexed terms for a sample text. In this example, stopwords are removed, though stemming is not applied for clarity.

Associated with the index of our retrieval system, apart from the inverted file containing the postings of all terms in the corpus, an auxiliary file stores some statistics such as df and CF(collection frequency) of each term [Dang et al. 2015]. Therefore, the exact tf and df values of bigrams are available for use in the document ranking formula of some retrieval models (e.g. Equation (6) for BM25), and need not be estimated during

[1] http://www.lemurproject.org/stopwords/stoplist.dft

Document	The universe's age is the most controversial issue addressed so far with the Hubble telescope.
stopwords	the is so with
unigrams only	universe age most controversial issue addressed far Hubble telescope
bigrams (SNB)	age_most most_controversial controversial_issue issue_addressed addressed_far far_Hubble Hubble_telescope
bigrams (SB)	most_controversial controversial_issue issue_addressed Hubble_telescope

Table 2: Example of indexed unigrams and bigrams after removal of stopwords

run-time (e.g. [Mishne and de Rijke 2005]). In the current study, we do not deal with proximity terms. Therefore, term positions need not be stored in the index.

In order to keep the total number of bigrams in the index manageable, for bigrams that occur in each document, we only add to the index those that appear at least 2 times, i.e. $tf \geq 2$. Table 3 shows an example of bigram statistics for the TREC-6 (Disks 4&5 collection, Table 4). The data shows that setting the minimum tf to 2 reduces the number of bigrams significantly. For the example collection, the size of the inverted index with SB type bigrams is about 22% larger than that of an index with only unigrams. This indicates that including bigrams in the index is manageable with regard to storage cost.

	unigrams	unigram + bigrams			
		bigram $tf \geq 1$		bigram $tf \geq 2$	
		SNB	**SB**	**SNB**	**SB**
number of terms (M)	2.2	24.2	12.0	4.1	3.3
size of compressed inverted index (MB)	185.8	1001	550	254	226
increase in size over unigram index	-	439%	196%	37%	22%

Table 3: Example of some unigram and bigram statistics for the TREC-6 (Disks 4&5) collection

3.2 Pseudo-relevance feedback retrieval utilizing unigrams and bigrams

We investigate the use of bigrams in PRF retrieval. The first stage of PRF involves an initial retrieval with the given query q. A query expansion (QE) is then performed, whereby a number of terms are automatically extracted from the top ranked (typically about 20) retrieved documents, which are assumed to be relevant. The selected QE terms are then added to the original query for a second retrieval. Generally, scores are assigned to the terms appearing in the top-ranked documents, and the top scoring N_{QE} terms are selected for QE. In our experiments, we use the following QE term scoring function [Dang et al. 2021]:

$$score_{QE}(t,d) = f(t,d) \times \log_{10}(N/df(t)), \tag{1}$$

where $f(t,d)$ is the number of occurrences of term t in d (term frequency), df is the number of documents in the collection in which t occurs (document frequency), and

N is the number of documents in the collection. We employ the same scoring function $score_{QE}(t, d)$ for both unigrams and bigrams. The validity or effectiveness of using same scoring function for unigrams and bigrams is a first attempt and needs to be verified by retrieval experiments (Section 4.2). A more thorough investigation of appropriate formulae for unigram and bigram selection is a topic for future study.

In the notation of the Vector Space Model (VSM), the expanded query vector $\vec{q}_{PRF}$ is obtained by mixing the normalized original query vector $\vec{q}$ and the normalized vector of QE terms $\vec{q}_{QE}$:

$$\vec{q}_{PRF} = \alpha_m \frac{\vec{q}}{|\vec{q}|} + (1 - \alpha_m) \frac{\vec{q}_{QE}}{|\vec{q}_{QE}|}, \tag{2}$$

where $|\cdot|$ is the city-block length (i.e. the number of terms) and α_m is a mixing factor with a value between 0 and 1. A second retrieval is performed by calculating new ranking scores for all documents in the collection with the expanded query. In VSM, the document ranking function is a summation of the weights of matching query terms that make up the representation of document vectors. The contribution of each matched query term to the ranking function is made up of two factors — (1) $w(t)$, which is the magnitude of the corresponding component in the expanded query vector $\vec{q}_{PRF}$ and (2) a term weight $S(t, d)$ of t in a document d that is retrieval model specific. The factor $w(t)$ is given by:

$$w(t) = \vec{t} \cdot \vec{q}_{PRF}, \tag{3}$$

where $\vec{t}$ is the unit vector of term t. We consider t in Equation (3) being either a unigram (u_1) or bigram (u_1u_2) and $\vec{t}$ is the corresponding unit vector. Past research finds that retrieval based on bigrams alone does not perform as good as the corresponding unigram model, such as the BM25 [He et al. 2011]. However, linear combinations of the unigram and bigram models outperform the unigram BM25 [He et al. 2011]. Therefore, we employ a document ranking function $S(q_{PRF}, d)$ that is a linear mixture of unigram and bigram components, denoted by $S_1(q_{PRF}, d)$ and $S_2(q_{PRF}, d)$, respectively, which are given by the following weighted sums:

$$S_1(q_{PRF}, d) = \sum_{u_1 \in q_{PRF}} w(u_1) S(u_1, d), \tag{4a}$$

$$S_2(q_{PRF}, d) = \sum_{u_1 u_2 \in q_{PRF}} w(u_1 u_2) S(u_1 u_2, d), \tag{4b}$$

where the sums are over matched unigrams and bigrams, respectively, in the expanded query.

The overall document ranking score $S(q_{PRF}, d)$ is a linear mixing of scores based on unigrams and bigrams:

$$S(q_{PRF}, d) = (1 - \beta) S_1(q_{PRF}, d) + \beta S_2(q_{PRF}, d), \tag{5}$$

where β is a mixing constant. In this article, we imply by 'bigram PRF' a linear mixing of unigram and bigram scores in the PRF ranking function as in Equation (5).

In the current study, we employ the BM25 [Robertson et al. 1995] as the basis retrieval model. This classical probabilistic model is chosen because of its high retrieval effectiveness among term-independent models and it is commonly used as a baseline for comparison with other retrieval methods [Dang et al. 2022]. In the BM25 approach, term

frequencies within a document are modeled by a mixture of two Poisson distributions [Robertson and Walker 1994, Robertson et al. 1995]. The BM25 weight of each term t appearing in the document d is:

$$S_{BM}(t,d) = \frac{(k+1)f(t,d)}{f(t,d) + k\left(1 - b + b\frac{|d|}{\Delta}\right)} \cdot \log\left(\frac{N - df(t) + 0.5}{df(t) + 0.5}\right), \qquad (6)$$

where k and b are constants, $|d|$ is the total number of single terms (unigrams) in d (document length), Δ is the average document length, $f(t,d)$ is the number of occurrences of t in d (term frequency), $df(t)$ is the number of documents in the collection containing t (document frequency), and N is the number of documents in the collection. We apply the BM25 weight of Equation (6) to t being either a unigram or bigram, i.e., in Equations (4a) and (4b), $S(t,d) = S_{BM}(t,d)$ for $t \in \{u_1, u_1u_2\}$. Adopting the BM25 term weight to bigrams amounts to assuming a mixture of Poisson distributions for bigram term frequencies, as in [He et al. 2011].

In this study, we focus on the effectiveness of using bigrams in the PRF second retrieval with an expanded query that includes bigrams as well as unigrams, in comparison with QE of only unigram terms. Therefore for both cases, we use the unigram BM25 for the initial retrieval, so that QE terms are extracted from the same set of top-ranked documents.

4 Experiments

Section 4.1 describes in detail our experimental methodology and setup. Our retrieval results are reported in Section 4.2, followed by an analysis of the results in Section 4.2.3.

4.1 Methodology and setup

We have performed retrieval experiments on a range of TREC test collections (Table 4) to evaluate our method, which is implemented on our own retrieval system [Dang et al. 2022]. For diversity, we have tested on both news and webpage collections. The size of these collections spans from about 3GB (Robust04) to the terabyte regime (Clueweb09 Category B subset of English-language pages). The Clueweb09 collection contains spam data. Using the spam scores distributed for this collection [Cormack et al. 2011], we filter Cluseweb09 Cat-B to the set of documents with Fusion spam scores in the 60^{th} percentile, as in the setting of some past works (e.g. [Huston and Croft 2014]). The performance of the retrieval models is evaluated in terms of widely used metrics — Mean Average Precision (MAP, e.g. [Ayter et al. 2015]), which takes into account both precision and recall (based on the top 1000 retrieved documents), and the precision-oriented NDCG@20, which reflects user preference (e.g. [Ferro and Silvello 2018]).

Short title queries, each averaging about 2 to 3 query terms are used for the initial retrieval because such query lengths are typical in web searches [Spink et al. 2002]. The free parameters in the retrieval model and PRF are calibrated to maximize the MAP by training on a selected set of 50 title queries. The calibrated parameters are applied to other sets of queries for testing. Training is performed separately for the various test collections because of their different document nature (news and webpages) and the very different sizes. For the news collection, the TREC-6 queries are used for training because the dataset of TREC-6 is a superset of the dataset used for the other news tracks. As earlier

Type	news/federal register; congressional records (CR)		news/federal register	
Dataset	Disks 4&5		Disks 4&5 - Congressional Record (Robust04)	
N	556,075		528,153	
Size (GB)	3.27		3	
TREC	6		7, 8, Robust 2003-2004	
queries	training		testing	
	50 queries		200 queries	

Type	webpages					
Dataset	WT10g		GOV2		Clueweb09 Cat-B	
N	1,692,096		25,205,179		50,189,002 (not spam-filtered) 33,378,688 (spam-filtered)†	
Size (GB)	10		426		≈ 1500	
TREC	9	10	Terabyte		Web track	
			2004-2005	2006	2009-2011	2012
queries	testing	training	testing	training	testing	training
	50 queries	50 queries	100 queries	50 queries	150 queries	50 queries

† Filtered documents with Waterloo Fusion spam scores up to the 60^{th} percentile.

Table 4: Summary of TREC test collections

tracks also used news collections, we expect search engines participating in TREC-6 to be well calibrated (e.g., with calibration based on TREC-1 to -3) so there was less chance for them to miss relevant documents. For a similar reason, for the webpage collections we select the last track on each dataset (i.e., TREC-10 of WT10g, Terabyte-2006 of GOV2 and web track 2012 of Clueweb09 Cat-B) to be the training set because the systems participating in the later tracks of TREC are expected to be better trained, so that there is less chance for them to miss relevance documents in the last track.

The methodology of using distinct sets of queries for training and testing has been used in the literature (e.g., [Lease2009, Roy et al. 2019, Trotman et al. 2014]) as an alternative to cross-validation. We employ this calibration methodology instead of cross-validation for the following reasons: (1) This methodology uses less training data than cross-validation for collections with large sets of testing queries, so that it should be a stronger test of the robustness of a retrieval method. For a typical 5-fold cross validation on a set of 50 queries, 10 queries of these are first selected as test queries and the remaining 40 queries are used as training queries. The training/testing process is repeated four more times, using a different subset of 10 queries for testing and the remaining 40 queries for training. In this case, the number of training queries is in effect four times that of the test queries. In comparison, our methodology uses a fixed set of 50 queries for training, and the calibrated model is tested on up to 200 queries (Table 4); (2) This approach reflects retrieval systems in practice, as these are generally calibrated beforehand with a fixed number of training queries and utilized for retrieval with thousands or millions of unseen queries. Thus, cross-validation is not very realistic because it uses much more training data than test data; (3) This methodology allows the retrieval performance for both the training set and testing sets of queries to be examined, as in Table 5; (4) This methodology enables checking whether good performance of a trained model can generalize to good results in testing using fixed parameters, as an indication of the robustness of the retrieval model.

4.2 Retrieval results

Table 5 summarizes the retrieval results. The primary objective of this study is to examine the effectiveness of our bigram PRF approach against the conventional unigram PRF. Table 5 presents the results of these methods applied to a BM25 base model. The results for BM25 without QE (‘initial retrieval’) are also provided as a common point of reference.

The two right-most columns of Table 5 show the MAP and NDCG@20 values for the test queries on various collections. Statistical analysis based on the randomization test [Smucker et al. 2007] is performed to compare the results of the different methods on the test queries to determine whether any observed difference is statistically significant (p-value less than 0.05).

4.2.1 Bigram types

For bigram PRF, we first compare the use of SNB (stopwords not treated as breaks) and SB (stopwords treated as breaks) bigrams in our approach. PRF retrieval using these two types of bigrams is performed on the Robust04, WT10g and GOV2 collections. The results for the test queries presented in Table 5 indicate quite minimal differences between SNB and SB bigrams, in both MAP and NDCG@20. The relative difference between the two methods is small, generally within about 1%. Moreover, we have confirmed all the differences to be not statistically significant.

Numerically, SB bigram PRF yields marginally higher MAP and NGCG@20 values than SNB bigram PRF on the tested collections, despite the lack of statistical significance in the differences. Because SB bigrams incur a smaller storage cost than SNB bigrams (Table 3), our results suggest that SB bigrams are the better choice for PRF retrieval. Correspondingly, we only perform PRF retrieval with SB bigrams on the Cluebweb09 Cat-B collection because of the long time needed for PRF training on this terabyte-sized collection.

4.2.2 Comparison of unigram PRF and bigram PRF

We compare the performance of bigram PRF with unigram PRF on various test collections. As shown in Table 5, bigram PRF yields higher MAP values than unigram PRF on the tested web collections (WT10g, GOV2 and Cluebweb09 Cat-B). However, statistical significance is only observed in the single case of MAP obtained by SB bigram PRF over unigram PRF on GOV2. Besides, the numerical differences on WT10g and Clueweb09 Cat-B are minimal. On the Robust04 collection, unigram PRF is found to yield a slightly higher MAP than both SNB and SB bigram PRF.

In terms of the precision-oriented NDCG@20, unigram PRF is found to obtain higher values of this metric than either SNB or SB bigram PRF, on the Robust04, WT10g and GOV2 test collections. On Clueweb09 Cat-B, bigram PRF yields a marginally higher NDCG@20 value than unigram PRF, just as in the case of MAP on this collection. We confirm that in all cases, the observed differences in NDCG@20 obtained by unigram and bigram PRF are not statistically significant. The results indicate that the use of a mixture of bigrams and unigrams in PRF is unable to increase the precision of the top ranked retrieved documents (such as in the top 20 as evaluated by NDCG@20) over the use of unigrams alone. Any observed improvement in MAP is thus likely to be due to more relevant documents being pushed into the top 1000 and the consequent increase in recall.

As an additional point of reference, we compare the unigram PRF and bigram PRF performances against the common baseline of unigram-based BM25 without QE. Some observations are: (1) Unigram PRF and bigram PRF (using either SNB or SB bigrams) obtain higher MAP and NDCG@20 values than the BM25 baseline on the majority of tested collections, with the only exception being NDCG@20 on WT10g; (2) For MAP, the better performance of unigram PRF or bigram PRF over the baseline BM25 is mostly

	TREC-6 (training)		Robust04 (testing)	
	MAP	NDCG@20	MAP	NDCG@20
Initial, unigram	0.2444	0.4427	0.2573	0.4137
PRF, unigram	0.2656	0.4329	**0.2964***	**0.4243**
PRF, u+b(SNB)	0.2643	0.4251	0.2927*	0.4209
PRF, u+b(SB)	0.2652	0.4267	0.2957*	0.4231

	WT10g (training)		WT10g (testing)	
	MAP	NDCG@20	MAP	NDCG@20
Initial, unigram	0.2037	0.3292	0.2060	**0.3132**
PRF, unigram	0.2241	0.3487	0.2112	0.3080
PRF, u+b(SNB)	0.2233	0.3301	0.2093	0.2915
PRF, u+b(SB)	0.2236	0.3346	**0.2117**	0.2981

	GOV2 (training)		GOV2 (testing)	
	MAP	NDCG@20	MAP	NDCG@20
Initial, unigram	0.3044	0.4940	0.2971	0.4559
PRF, unigram	0.3180	0.4929	0.3234*	**0.4634**
PRF, u+b(SNB)	0.3408	0.4929	0.3325*	0.4594
PRF, u+b(SB)	0.3421	0.4988	**0.3333***†	0.4609

	Clueweb09 Cat-B (training)		Clueweb09 Cat-B (testing)	
	MAP	NDCG@20	MAP	NDCG@20
Initial, unigram	0.1394	0.2100	0.0841	0.2528
PRF, unigram	0.1624	0.2476	0.1044*	0.2780*
PRF, u+b(SB)	0.1631	0.2471	**0.1047***	**0.2789***

Table 5: Comparison of the unigram-based BM25 (Initial and PRF) with the extended BM25 model (PRF, unigram+bigram) on various test collections. The symbols $*$ and † indicate statistical significance in the improvement (at the 95% confidence level) over the initial (unigram) retrieval and PRF (unigram) retrieval, respectively

statistically significant, except on WT10g. The different behavior of WT10g may be due to the smaller number of test queries (50 queries) compared with the other test collections (100-200 queries); (3) For NDCG@20, the difference in performance of unigram PRF or bigram PRF over the baseline BM25 is mostly not statistically significant, except on Cluebweb09 Cat-B.

In summary, our bigram PRF approach yields higher MAP values than unigram PRF in the web collections, though the performance enhancement is statistically significant only in one of the tested collections (GOV2). On the largest collection tested (Clueweb09 Cat-B), our bigram PRF obtains better MAP and NDCG@20 values than unigram PRF. This suggests the use of bigrams may be effective for collections at least as large as Cluweb09 Cat-B. However, unigram PRF tends to yield better NDCG@20 than bigram PRF in the other smaller tested collections.

4.2.3 Analysis

As our bigram PRF method does not improve over unigram PRF robustly across collections, in term of either the MAP or NDCG@20 metrics, we conduct an analysis in an attempt to understand the reason behind the failing. Table 5 indicates that for the Robust04 collection, the conventional unigram PRF performs better than our bigram PRF in both MAP and NDCG@20. Therefore, our analysis focuses on this collection.

The Robust04 collection contains 200 test queries (Table 4). We compare the AP (Average Precision) values for individual queries, as obtained by our bigram PRF (SB) and conventional unigram PRF. Let dAP denotes the difference of the AP values obtained by the two methods, that is, dAP = AP(our bigram PRF, SB) - AP(unigram PRF). We seek the queries for which our bigram PRF enhances the performance over unigram PRF, as well as the queries for which there is not much difference, and queries for which our bigram PRF have an negative effect. Figure 1 depicts a query-by-query plot of dAP, sorted by the dAP values. The figure shows there are roughly the same number of queries for which dAP is positive or negative.

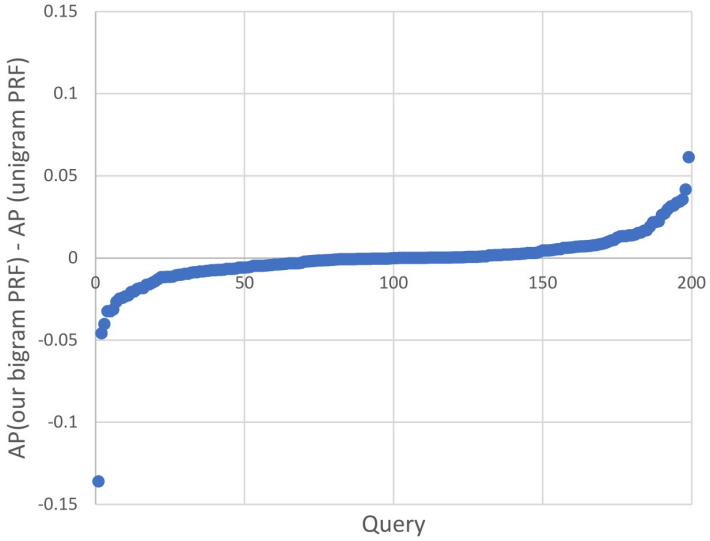

Figure 1: Query-by-query plot of dAP for the Roust04 test queries

We select three queries from each group of queries for which dAP is positive (Group P), roughly zero (Group Z), or negative (Group N), as presented in Table 6. For each query, samples of QE terms with the highest scores according to Equation (1) are shown, including unigram QE terms and bigram QE terms selected by our method. Also shown in the table are the AP values obtained by the initial retrieval, unigram PRF and our bigram PRF.

Group P. For two out of the three Group P sample queries in Table 6 (Q685 and Q440), unigram PRF yields poorer AP values than the initial retrieval using the original title queries. This is despite unigram PRF out-performing the initial retrieval in MAP by about 15% with statistical significance when averaged over 200 test queries on the

Group P. Positive *dAP*	
Query (Q685)	Oscar winner selection
QE terms (unigram PRF)	academy, film, award, actress, wheeler, best
QE terms (bigram PRF)	'academy award', 'best picture', 'motion picture', 'best actress', 'best director', 'rain man'
AP(init, uni-PRF, bi-PRF)	0.2980, 0.1997, 0.2610
Query (Q440)	child labour
QE terms (unigram PRF)	FLSA, employ, hour, carpet, law, work
QE terms (bigram PRF)	'child labour', 'labour law', 'labour regulation', 'child benefit', 'child care', 'new law'
AP(init, uni-PRF, bi-PRF)	0.1999, 0.1691, 0.2035
Query (Q656)	lead poisoning children
QE terms (unigram PRF)	paint, iron, blood, lead-based, FDA, hazard
QE terms (bigram PRF)	'lead poisoning', 'lead paint', 'young child', 'lead level', 'lead hazard', 'learning disability'
AP(init, uni-PRF, bi-PRF)	0.2171, 0.2346, 0.2643
Group Z. *dAP* ≈ 0	
Query (Q361)	clothing sweatshops
QE terms (unigram PRF)	garment, sew, labour, worker, wage, immigrant
QE terms (bigram PRF)	'garment industry', 'minimum wage', 'sewing machine', 'labor law', 'child labour', 'shop owner'
AP(init, uni-PRF, bi-PRF)	0.4221, 0.6270, 0.6270
Query (Q623)	toxic chemical weapon
QE terms (unigram PRF)	destruction, convention, substance, binary, agent, military-chemical
QE terms (bigram PRF)	'chemical weapon', 'toxic substance', 'toxic agent', 'biological weapon', 'mustard gas', 'chemical warfare'
AP(init, uni-PRF, bi-PRF)	0.2629, 0.3504, 0.3504
Query (Q679)	opening adoption records
QE terms (unigram PRF)	electronic, signature, FDA, adopted, parent, birth
QE terms (bigram PRF)	'electronic record', 'adopted parents', 'birth parents', 'paper record', 'birth certificate', 'open system'
AP(init, uni-PRF, bi-PRF)	0.7861, 0.6829, 0.6829
Group N. Negative *dAP*	
Query (Q601)	Turkey Iraq water
QE terms (unigram PRF)	Syria, Euphrates, Israel, Iran, Tigris, Ankara
QE terms (bigram PRF)	'middle east', 'Turkey official', 'Turkey foreign', 'northern Iraq', 'foreign minster', 'Turkey government'
AP(init, uni-PRF, bi-PRF)	0.5270, 0.5226, 0.3866
Query (Q682)	adult immigrant English
QE terms (unigram PRF)	school, eligible, status, class, literacy, noncitizen
QE terms (bigram PRF)	'eligible immigrant', 'immigration status', 'school district', 'Los Angeles', 'consent form', 'adult education'
AP(init, uni-PRF, bi-PRF)	0.2243, 0.2238, 0.1836
Query (Q663)	Agent Orange exposure
QE terms (unigram PRF)	veteran, Vietnam, herbicide, cancer, disease, child
QE terms (bigram PRF)	'agent orange', 'Vietnam veteran', 'herbicide exposure', 'veteran affair', 'assistance program', 'Vietnam War'
AP(init, uni-PRF, bi-PRF)	0.4228, 0.7751, 0.7427

Table 6: Analysis of Robust04 test queries

Robust04 collection (Table 5). It appears some of the individual QE terms returned by unigram PRF for these queries correspond to fairly general concepts (e.g. {film, award, best} for Q685 or {employ, hour, carpet} for Q440). These added individual terms may

act as noise and are detrimental to the retrieval performance. On the other hand, the bigram QE terms mostly appear to be sensible concepts related to the original queries. Moreover, many of the bigram QE terms represent relevant concepts not explicitly covered by the original query terms or top-scored unigram QE terms (e.g. 'best picture', 'best director' of Q685). In these cases, the added bigrams are beneficial to retrieval and improve the performance over unigram PRF, as indicated by positive *dAP* values.

Group Z. Similar to the Group P queries, many of the bigram QE terms in Group Z appear to be relevant concepts. Many of these are formed by combining either the original query terms or top-ranked unigram QE terms (e.g. 'chemical weapon', 'toxic substance', 'toxic agent' of Q623). Thus, the concepts represented by these QE bigrams are already covered by the terms in the expanded query of unigram PRF. In this case, bigram PRF may not bring much more improvement over unigram PRF.

Group N. For the Group N sample queries of Table 6, unigram PRF improves the AP values over the initial retrieval (Q663) or has little overall effect (Q601 and Q682). The effectiveness of unigram PRF depends on whether the helpful terms out-weigh the noise terms in the expanded query. As for the bigram QE terms, there is also a mixture of relevant (e.g. 'Vietnam veteran', 'veteran affair', 'Vietnam War' in Q663) and noise terms (e.g. 'herbicide exposure' in Q663). Furthermore it appears that for some queries, many of the bigram QE terms represent fairly broad concepts that do not specifically focus on the given query (e.g. 'Middle East', 'Turkey official', 'Northern Iraq', 'foreign minister', 'Turkey government' in Q601). While it is conceivable that some of these bigrams may be found in relevant documents, their generality may cause non-relevant documents containing them to be pushed up in ranking among the retrieved documents. These problems correspond to 'query drift', a well-known issue of PRF [Ruthven and Lalmas 2003]. In this case, both the noise and non-specific QE bigrams contribute to a drop in retrieval effectiveness compared with unigram PRF.

4.3 Discussion

Because of the small probability of two words (that are not stopwords) co-occurring consecutively by random, word bigrams that appear either multiple times in a document or in multiple documents are likely to represent true intended concepts in these documents. For short queries that are often used in web retrieval [Spink et al. 2002], only a small number of word bigrams can be made up from the query terms. The QE process of PRF allows more word bigrams to be added to the query. Thus, the effect of word bigrams is expected to be greater in PRF than retrieval without QE. Also due to the small occurrence probability of specific word bigrams, large collections may contain more documents in which they appear. Thus, the effectiveness of bigram PRF may be more obvious with large collection sizes. Therefore PRF retrieval on large collections is expected to be a good testing ground of the usefulness of word bigrams in IR.

In our experiments, our bigram PRF approach yields higher MAP values than unigram PRF with statistical significant on the large GOV2 collection, but not on the smaller Robust04 and WT10g collections. This observation is consistent with our expectation. However, the improvement of MAP on the largest Clueweb09 Cat-B is small and not statistically significant. Other works in the literature have also observed relatively small performance improvements with term dependency models on the Clueweb09 Cat-B collection, compared with other collections like Robust04 and GOV2 (e.g. [Bendersky et al. 2012, Huston and Croft 2014]). A possible factor causing the different behavior in Cluweb09 Cat-B may be the presence of spam.

The motivation of our approach to select word bigrams as single units from top-ranked documents in PRF is to reduce the likelihood of them being noise, as illustrated by the example of Table 1. In reality, our experiments show that our bigram PRF approach can outperform unigram-based PRF in terms of MAP in some collections, but generally not for the precision-oriented NDCG@20. The poorer NDCG@20 suggests there is noise in the bigram QE terms that affects the top-ranked documents. This is supported by the analysis in Section 4.2.3. Table 6 shows that the added noise may come from bigram QE terms that represent non-relevant concepts (e.g. 'herbicide exposure' in Q663). Furthermore we find that in some cases the bigram QE terms represent broad and non-specific concepts, which act as an added source of 'query-drift' harming the retrieval effectiveness.

Overall, our approach may perform as well as or surpass existing methods, though not robustly across all test collections. Thus, our results are negative as they do not meet our aim to demonstrate conclusively the effectiveness of word n-grams in PRF retrieval. Based on the problem of induction, [Luk 2019] argues that negative results are worth publishing, as has been suggested by [Prechelt 1997]. In general, new methods whose performance is not statistically different from the state-of-the-art are worth publishing because future modifications to these methods may achieve statistically significant improvements over the state-of-the-art [Luk 2019]. In our case, because query drift is found to be an issue in bigram PRF, modifications to our method will have to include a more stringent criterion of word bigram selection in the expanded query. We suggest a few ways this may be done. As we identify non-specific bigrams to be a source of query drift, the scoring function of Equation (1) may be modified to remedy this issue. For example, a possible choice of bigram QE scoring function is:

$$score_{QE,bigram}(t,d) = \log(f(t,d)) \times \log_{10}(N/df(t)). \tag{7}$$

The above function reduces the contribution of the term frequency component, hence putting more emphasis on the discriminatory inverse document frequency component. Alternatively, bigrams QE terms that have large document frequency values above a threshold may be removed. Another possibility is to only count bigrams that occur within a fixed-size window (known as a 'document context') around the original query terms in the top-ranked documents, as these bigrams are more likely to be related to the original query [Wu et al. 2008]. It may also be possible to reduce query drift due to the QE terms by performing a *re-ranking* instead of re-retrieval in the PRF step. That is, new ranking scores based on the expanded query are calculated for the retrieved documents of the initial retrieval instead of performing a new retrieval. This would avoid new non-relevant documents being retrieved. Last, another method is to directly control the number of both unigrams and word bigrams added in QE instead of letting them be determined according to the scores.

A feature of our approach is to include word bigrams in the indexing, so that collection statistics such as exact values of the df of word bigrams are available for calculating ranking scores. Our results corroborate with previous work such as that of [He et al. 2011], in showing statistically significant improvement in MAP on GOV2. However similar to [He et al. 2011], we do not demonstrate a robust performance enhancement across all tested collections. This suggests that the method of estimating df values of word bigrams as used by [He et al. 2011] and others are as good as the use of exact values. Thus our results support the use of estimated word bigram df values in practice.

There are two restrictions of word n-grams: (1) the constituent terms must be adjacent; (2) the order of the individual terms is fixed. Both of these restrictions may explain the

lack of significant improvement to retrieval performance across collections with the use of word bigrams in our experiments. This conjecture is supported by the success of the Markov Random Field (MRF) [Metzler and Croft 2005, Metzler and Croft 2007] approach, which makes use of both ordered and un-ordered query terms in proximity.

A possible direction for future studies is a combination of our bigram PRF approach with other techniques, such as MRF. For example in the LM framework, [Lease2008] introduces an approach that adapts MRF to relevance feedback with known relevant documents. This method was found to be extremely effective, achieving top performance at the TREC 2008 Relevance Feedback Track [Buckley and Robertson 2008]. We may explore the use of bigrams in the method of [Lease2008] in a PRF setting, whereby the query is expanded with both unigrams and bigrams (extracted as single units from top-ranked documents) instead of just unigrams. This allows our bigram PRF approach to be used in conjunction with proximity term techniques. Last, more tests with the Clueweb09 Cat-B with a higher level of spam filtering, or larger collections without spam, are needed to verify that our method can improve both MAP and NDCG@20, as our bigram PRF seems to work better for large collections.

5 Conclusion

To tackle the issues of past research on using proximity terms in PRF with an extended BM25 model [He et al. 2011], we propose the approach of employing a document index that consists of word n-grams in addition to unigrams, and extracting word n-grams as single units in QE. We have tested the approach for the case of PRF based on a mixture of unigrams and bigrams. Retrieval experiments are performed on a range of TREC test collections. While our bigram PRF approach shows some improvement in MAP over unigram PRF in the web collections, in the majority of cases the differences are not statistically significant. Also, compared with unigram-based PRF, there is either minimal difference or a degradation in NDCG@20, with the differences being not statistically significant for all the collections. Thus overall, in terms of both evaluation metrics, our approach does not improve over unigram PRF robustly across collections. An analysis of the retrieval results reveals that both noise terms that are non-relevant to the query and bigram QE terms that represent too-broad topics contribute to ‘query drift’ that harms retrieval effectiveness.

Although in general our bigram PRF approach performs similarly as unigram PRF, our approach obtains higher numerical values of MAP and NDCG@20 on the terabyte-sized Clueweb09 Cat-B test collection. This shows that our bigram PRF approach potentially may outperform unigram PRF on large collections of size similar to or larger than Clueweb09 Cat-B. This is encouraging in that it is possible some modification of our bigram PRF method or combination with other techniques may lead to more robust improvement in retrieval effectiveness, particularly for large collections. As mentioned by [Luk 2019] on the basis of the induction problem, methods matching state-of-the-art performance are worth publishing as modifications of them may outperform the state-of-the-art. Therefore, our bigram PRF approach is worth further exploration for more robust improvement on large collections.

Acknowledgements

This work is supported by the Hong Kong PolyU project P0030932.

References

[Ayter et al. 2015] Ayter, J., Chifu, A.-G., Déjean, S., Desclaux, C., and Mothe, J.: "Statistical analysis to establish the importance of information retrieval parameters"; J. Uni. Comp. Sc., 21, 13 (Dec 2015) 1767-1789.

[Bendersky et al. 2012] Bendersky, M., and Croft, W. B.: "Modeling higher-order term dependencies in information retrieval using query hypergraphs"; Proc. 35 th Ann. Intl. ACM SIGIR Conf. (Aug 2012) 941-950.

[Buckley et al. 1995] Buckley, C., Salton, G., Allan, J., and Singhal, A.: "Automatic query expansion using SMART: TREC-3"; NIST special publication sp (Apr 1995) 69-80, https: //trec.nist.gov/pubs/trec3/t3_proceedings.html.

[Buckley and Robertson 2008] Buckley, C., and Robertson, S.: "Relevance feedback track overview: TREC-2008"; Proc. 17th Text REtrieval Conference (TREC-2008) (Nov 2008), MD:NIST Special Publication 8.

[Büttcher et al. 2006] Büttcher, S., Clarke, C. L. A., and Lushman, B.: "Term proximity scoring for ad-hoc retrieval on very large text collections"; Proc. 29 th Ann. Intl. ACM SIGIR Conf. (Aug 2006) 621-622.

[Cormack et al. 2011] Cormack, G. V., Smucker, M. D., and Clarke C. L. A.: "Efficient and effective spam filtering and re-ranking for large web datasets"; Inf. Retrieval, 14, 5 (2011) 441-465.

[Dang et al. 2015] Dang, E. K. F., Luk, R. W. P. and Allan, A.: "Fast forward index methods for pseudo-relevance feedback retrieval"; ACM Trans. Inf. Sys., 33, 4, Article 19 (May 2015) 1-33.

[Dang et al. 2021] Dang, E. K. F., Luk, R. W. P. and Allan, A.: "A principled approach using fuzzy set theory for passage-based document retrieval"; IEEE Trans. Fuzzy Sys., 29, 7 (2021) 1967-1977.

[Dang et al. 2022] Dang, E. K. F., Luk, R. W. P. and Allan, A.: "A comparison between term-independence retrieval models for ad hoc retrieval"; ACM Trans. Inf. Sys., 40, 3, Article 62 (Jul 2022) 1-37.

[Ferro and Silvello 2018] Ferro, N., and Silvello, G.: "Toward an anatomy of IR system component performances"; J. Ass. Inf. Sc. and Tech., 69, 2 (Nov 2018) 187-200.

[He et al. 2011] He, B., Huang, J. X., and Zhou, X.: "Modeling term proximity for probabilistic information retrieval models"; Info. Sciences, 181 (Jul 2011) 3017-3031.

[Huston and Croft 2014] Huston, S., and Croft, W. B.: "A comparison of retrieval models using term dependencies"; Proc. 23rd ACM CIKM (Nov 2014) 111-120.

[Lease2008] Lease, M.: "Incorporating relevance and pseudo-relevance feedback in the Markov Random Field model"; Proc. 17th Text Retrieval Conference TREC2008 (Nov 2008). Retrieved from http://trec.nist.gov/pubs/trec17/papers/brownu.rf.rev.pdf

[Lease2009] Lease, M.: "An improved Markov random field model for supporting verbose queries"; Proc. 32 nd Ann. Intl. ACM SIGIR Conf. (Aug 2009) 476-483.

[Luk 2019] Luk, R. W. P.: "Why is Bayesian confirmation theory rarely practiced?"; Science and Philosophy, 7, 1 (Jun 2019) 3-20.

[Metzler and Croft 2005] Metzler, D., and Croft, W. B.: "A markov random field model for term dependencies"; Proc. 28 th Ann. Intl. ACM SIGIR Conf. (Aug 2005) 472-479.

[Metzler and Croft 2007] Metzler, D., and Croft, W. B.: "Latent concept expansion using markov random fields"; Proc. 30 th Ann. Intl. ACM SIGIR Conf. (Aug 2007) 311-318.

[Mishne and de Rijke 2005] Mishne, G., and de Rijke, M.: "Boosting web retrieval through query operations"; Adv. Information Retrieval: 27^{th} ECIR (Mar 2005) 502-516.

[Mitra et al. 1997] Mitra, M., Buckley, C., Singhal, A., and Cardie, C.: "An analysis of statistical and syntactic phrases"; Proc. RIAO-97 (Jun 1997) 200-214.

[Perry and Willett 1983] Perry, S. A., and Willet, P.: "A review of the use of inverted files for best match searching in information retrieval systems"; J. Info. Science, 6, 2-3 (Feb 1983) 59-66.

[Ponte and Croft 1998] Ponte, J. M., and Croft, W. B.: "A language modeling approach in information retrieval"; Proc. 21^{st} Intl. ACM SIGIR Conf. (Aug 1998) 275-281.

[Porter 1980] Porter, M. F.: "An algorithm for suffix stripping"; Program 14, 3 (Mar 1980) 130-137.

[Prechelt 1997] Prechelt L.:"Why we need an explicit forum for negative results"; J. Uni. Comp. Sc., 3, 9 (Sep 1997) 1074-1083.

[Rasolofo and Savoy 2003] Rasolofo, Y., and Savoy, J.: "Term proximity scoring for keyword-based retrieval systems"; Proc. 25^{th} Euro. Conf. Info. Ret. (Apr 2003) 207-218.

[Robertson and Walker 1994] Robertson, S. E., and Walker, S.: "Some simple effective approximations to the 2-Poisson model for probabilistic weighted retrieval"; Proc. 17^{th} Intl. ACM SIGIR Conf. (1994) 232-241.

[Robertson et al. 1995] Robertson, S. E., Walker, S., Jones, S., Hancock-Beauleiu, M. M., and Gatford, M.: "Okapi at TREC-3."; NIST special publication sp (Apr 1995) 109.

[Roy et al. 2019] Roy, D., Bhatia, S., and Mitra, M.:"Selecting discriminative terms for relevance model"; Proc. 42 nd Ann. Intl. ACM SIGIR Conf. (Jul 2019) 1253-1256.

[Ruthven and Lalmas 2003] Ruthven, I., and Lalmas, M.:"A survey on the use of relevance feedback for information access systems"; The Knowldege Engineering Review, 18:2 (Jun 2003) 95-145.

[Smucker et al. 2007] Smucker, M. D., Allan, J., and Carterette, B.: "A comparison of statistical significance tests for information retrieval evaluation"; Proc. 16^{th} ACM CIKM (Nov 2007) 623-632.

[Song and Croft 1999] Song, F., and Croft, W. B.: "A general language model for information retrieval"; Proc. 8^{th} ACM CIKM (Nov 1999) 316-321.

[Spink et al. 2002] Spink, A., Jansen, B.J., Wolfram, D., and Saracevic, T.:"From e-sex to e-commerce: Web search changes"; IEEE Computer 3, 35 (Mar 2002) 107-109.

[Srikanth and Srihari 2002] Srikanth, M., and Srihari, R.: "Biterm language models for document retrieval"; Proc. 25^{th} Intl. ACM SIGIR Conf. (Aug 2002) 425-426.

[Trotman et al. 2014] Trotman, A., Puurula, A., and Burgess, B.:"Improvements to BM25 and language models examined"; Proc. 19^{th} Australasian Document Computing Symposium (Nov 2014) 58-65.

[Wu et al. 2008] Wu, H.C., Luk, R.W.P., Wong, K.F., and Kwok, K.L.:"Interpreting TF-IDF weights as making relevance decisions"; ACM Trans. Inf. Sys., 26, 3, Article 13 (Jun 2008) 1-37.