A Retrieval Model Family Based on the Probability Ranking Principle for Ad Hoc Retrieval

Edward Kai Fung Dang[1*], Robert Wing Pong Luk[1] and James Allan[2]

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

[2]College of Information and Computer Sciences, University of Massachusetts, Amherst, MA 01003-9264, USA, email: allan@cs.umass.edu

[*]Corresponding author

Abstract

Many successful retrieval models are derived based on or conform to the Probability Ranking Principle (PRP). We present a new derivation of a document ranking function given by the probability of relevance of a document, conforming to the PRP. Our formulation yields a family of retrieval models, called Probabilistic Binary Relevance (PBR) models, with various instantiations obtained by different probability estimations. By extensive experiments on a range of TREC collections, improvement of the PBR models over some established baselines with statistical significance is observed, especially in the large Clueweb09 Cat-B collection.

1 Introduction

In Information Retrieval (IR), there has been much research effort to develop new retrieval models for better retrieval effectiveness. The probability ranking principle (PRP) (S. E. Robertson, 1977) has provided a basis of many retrieval models. The PRP asserts that if the probability of relevance of a document to a query is estimated as accurately as possible based on the information available, then ranking documents in decreasing probability of relevance will yield the best retrieval effectiveness (S. E. Robertson, 1977). An early retrieval model based on the PRP is the Binary Independence Model (BIM) (Yu & Salton, 1976; S. E. Robertson & Spärck Jones, 1976), which provides theoretical justification for the inverse document frequency (IDF) factor. However, BIM lacks a term frequency (TF) factor. An improvement over the BIM is the BM25 model with TF-IDF term weightings (S. Robertson, Walker, Jones, Hancock-Beaulieu, & Gatford, 1995; S. Robertson & Zaragoza, 2009). The language model (LM) framework (Ponte & Croft, 1998; Hiemstra, 1998) is a different approach that is also highly successful (Zhai & Lafferty, 2004; Cummins, Paik, & Lv, 2015). Although there is no explicit notion of relevance in original formulation of LM, it has been shown to conform to the PRP (Lafferty & Zhai, 2003; Azzopardi & Roelleke, 2007; Luk, 2008, 2022).

Table 1: Summary of notations

Symbol	Description
N	Number of documents in the collection
$\dot{d}$	binary document vector representation of document d
$\|\dot{d}\|$	length of $\dot{d}$ (number of distinct terms in d)
$\|x\|$	total number of terms in x ($x \in \{$document d, query q, collection $\mathbb{C}\}$)
$\|x\|_p$	p-norm length of x ($x \in \{d, \tilde{d}\}$)
Δ_p	average p-norm document length
$f(t,x)$	TF (number of occurrences) of term t in x ($x \in \{d, q, \mathbb{C}\}$)
$\tilde{f}(t,d)$	normalized TF of term t in document d
$\tilde{d}$	a set-of-tuple representation of document d with normalized TF
$\tilde{d}_t$	element of $\tilde{d}$, given by the tuple $(t, \tilde{f}(t,d))$ for term t
$df(t)$	document frequency of term t (number of documents containing t)
$df_{\mathbb{C}}$	sum of document frequencies over all distinct terms in the collection
$\propto$	rank equivalent relation
$V(\cdot)$	function that returns the set of terms (vocabulary) of the argument

Because of the success of the PRP-based retrieval models, a research direction of interest is the derivation of new retrieval models based on the PRP with the aim to improve retrieval effectiveness over the established models. Previously (Wu, Luk, Wong, & Kwok, 2008) showed that some existing TF-IDF term weightings including the BM25 can be derived from the probability of relevance, but no new ranking function was introduced. In this regard, we present the formal derivation of a new retrieval model family based on the PRP, called Probabilistic Binary Relevance (PBR) models.

The retrieval effectiveness of the new PBR models is evaluated by empirical comparison with several successful models, on a range of TREC test collections. We include as baselines the BM25, LM and SPUD (Cummins et al., 2015) models as they conform to the PRP, and also the PL2 model of the Divergence From Randomness (DFR) framework (Amati & van Rijsbergen, 2002) as it is a highly effective model not derived from the PRP. Improvement of the new PBR models over some of these strong baselines with statistical significance is observed, especially in the terabyte-sized Clueweb09 Cat-B collection. While we have demonstrated the retrieval effectiveness of several instantiations of the PBR models, we have not exhausted the possibilities. New models in the family may be obtained by appropriate estimation of probabilities and interpolation, with potential further enhancement of retrieval effectiveness.

The rest of the paper is organized as follows. Section 2 provides the background on some successful retrieval models, particularly those used as baselines in our study, and discusses some promising approaches. Section 3 describes our approach, including a detailed derivation of the PBR models and several instantiations. Experimental results are presented in Section 4, followed by a conclusion. Table 1 summarizes the notations used in the paper.

2 Background

2.1 Related work

The document ranking function of the BM25 model (S. Robertson et al., 1995) is:

$$S_{BM}(q,d) = \sum_{t \in V(q)} \frac{(k+1)f(t,d)}{f(t,d) + k\left(1 - b + b\frac{|d|}{\Delta}\right)} \cdot \log\left(\frac{N - df(t) + 0.5}{df(t) + 0.5}\right) \cdot \frac{(k'+1)f(t,q)}{k' + f(t,q)}, \tag{1}$$

where k, k' and b are constants and Δ is the average document length. The TF component of the BM25 ranking function incorporates document length normalization, which ensures long documents are not excessively favored over short documents in retrieval (Singhal, Buckley, & Mitra, 1996; Na, 2015). Instead of a simple normalization by the document length $|d|$, the normalization in BM25 takes into account that the length of a document may depend on the document's verbosity and scope (S. E. Robertson & Walker, 1994). This functional form was subsequently formalized as pivoted normalization (Singhal et al., 1996), as a means to avoid over-penalizing long documents.

Another successful retrieval framework is the LM approach (Ponte & Croft, 1998; Hiemstra, 1998). In one formulation (Lafferty & Zhai, 2001; Zhai & Lafferty, 2004), the LM document ranking is given by the Kullback-Leiber (KL) divergence between the query LM and document LM:

$$\begin{aligned} S_{LM}(q,d) &= -\sum_{t \in V(q)} p(t|q) \log \frac{p(t|q)}{p(t|d)}, \\ &\propto \sum_{t \in V(q)} p(t|q) \log p(t|d). \end{aligned} \tag{2}$$

The probability $p(t|d)$ is given by:

$$p(t|d) = \lambda(d)\frac{f(t,d)}{|d|} + (1 - \lambda(d))\frac{f(t,\mathbb{C})}{|\mathbb{C}|}, \tag{3}$$

which includes a smoothing component that assigns a non-zero probability to unseen words, with $\lambda(d)$ being in general a document-dependent parameter with value between 0 and 1. For Dirichlet smoothing, which is found to be effective for keyword queries (Zhai & Lafferty, 2004), $\lambda(d) = |d|/(|d| + \mu)$, with the Dirichlet prior μ being a positive number. Applying Dirichlet smoothing together with an estimate of the query LM, Eq. (2) becomes:

$$S_{LM}(q,d) \propto \sum_{t \in V(q)} \frac{f(t,q)}{|q|} \log\left(\frac{f(t,d) + \mu\, f(t,\mathbb{C})/|\mathbb{C}|}{|d| + \mu}\right). \tag{4}$$

While there is no explicit IDF component in LM, smoothing with a collection-based distribution plays a role similar to the IDF (e.g. (Zhai & Lafferty, 2004)).

Numerous techniques have been developed in the LM framework, such as the Smoothed Pólya Urn Document (SPUD) model that captures word burstiness, i.e. the tendency of a term to repeat itself in a

document. The document ranking function of a SPUD model that applies Dirichlet smoothing is

$$S_{SPUD}(q,d) = \sum_{t \in V(q)} \frac{f(t,q)}{|q|} \log\left(\frac{1}{\mu|\dot{d}|+1}\left[\mu|\dot{d}|\frac{f(t,d)}{|d|} + \frac{df(t)}{df_{\mathbb{C}}}\right]\right), \tag{5}$$

where μ is a positive constant. A difference between SPUD and the traditional LM (Eq. (4)) is that the normalization in SPUD is based on $|\dot{d}|$, which is the number of distinct terms in a document, instead of the document length $|d|$. We include SPUD as a baseline in our experiments because it was shown to out-perform the traditional LM (Cummins et al., 2015).

The DFR framework of probabilistic IR models (Amati & van Rijsbergen, 2002) are highly effective retrieval models not derived from the PRP. This approach incorporates two main concepts — first, the information gain of a term as measured by the divergence of the term's occurrence distribution in a document from its distribution in the whole collection, which is assumed to be random, and second, normalization of the TF. We use one of the DFR models, the PL2, as a baseline because it has been shown to be effective (He & Ounis, 2003, 2007; P. Yang & Fang, 2016). The PL2 ranking score is given by:

$$S_{PL2}(q,d) = \sum_{t \in V(q)} Inf_1(t,d) \cdot Inf_2(t,d) \tag{6a}$$

where

$$Inf_1(t,d) = tfn \cdot \log_2 \frac{tfn}{\lambda} + \left(\lambda + \frac{1}{12 \cdot tfn} - tfn\right) \cdot \log_2 e + 0.5 \cdot \log_2(2\pi \cdot tfn), \tag{6b}$$

$$Inf_2(t,d) = \frac{1}{tfn+1}, \tag{6c}$$

with $\lambda = f(t,\mathbb{C})/N$, and tfn is a normalization of the TF according to:

$$tfn = f(t,d) \cdot \log_2\left(1 + c\frac{\Delta}{|d|}\right) \tag{6d}$$

where c is a positive constant.

2.2 Other methods

The retrieval models described in Section 2.1 are common IR baselines, as used in our experiments. These are bag-of-words models in which the association between distinct terms is ignored. Apart from these models, some approaches that go beyond bag-of-words have been shown to be effective. These include the cross term approach that models term association (Zhao, Huang, & Ye, 2014), term proximity matching methods such as Markov Random Field (MRF) (Metzler & Croft, 2005), and also the use of term position features (Hammache & Boughanem, 2021). As the objective of this paper is to demonstrate the effectiveness of the new bag-of-words PBR models in a pilot study, we restrict to bag-of-words baselines for a fair comparison. Another highly promising direction is neural retrieval models (Guo et al., 2020). These models typically require extensive training, which is not needed with the PBR models. Pre-training of word representations, such as in BERT-based neural models (W. Yang, Zhang, & Lin, 2019), requires

external data. For non-neural models, an external corpus may be used to improve retrieval effectiveness by query expansion. However in this paper, we focus on base models without query expansion. Therefore, for a comparison with approaches that do not require external data, neural retrieval models are not included in our experiments.

3 Our approach: Probabilistic Binary Relevance Models

This section presents the derivation of the new PBR model family, with a description of several instantiations of the PBR models in Section 3.6.

3.1 Theoretical foundation

Generally a retrieval request is specified by a query q that consists of a string of query terms. A document d may be represented as a vector with each dimension and its value corresponding to a term t in d and the TF, respectively. The use of TF normalized by the document length in ranking functions can enhance retrieval effectiveness (Singhal et al., 1996; Na, 2015). We denote the normalized TF by $\tilde{f}(t,d)$ and represent the document with normalized TF, $\tilde{d}$, by a set of tuples, i.e. $\tilde{d} = \{\tilde{d}_t | t \in V(d)\}$, where $\tilde{d}_t \equiv (t, \tilde{f}(t,d))$.

On the basis of the PRP, we seek to derive the probability of relevance of a document to serve as a retrieval ranking function. We first make a 'bag-of-words' assumption by stating the probability of relevance in terms of the (normalized) document vector, which ignores the word ordering in a document:

Assumption 1 *The probability of relevance of a document d given query q is $p(r|q,\tilde{d})$.*

The use of $\tilde{d}$ in the probability highlights the use of normalized TF in our derivation. Assumption 1 excludes approaches that utilize term association. The assumption is also not invoked in approaches that do not calculate the probability of relevance, such as neural retrieval models (Guo et al., 2020).

3.2 Decomposing the query

By Bayes' rule, we can write $p(r|q,\tilde{d})$ as follows:

$$p(r|q,\tilde{d}) = p(r|\tilde{d})\frac{p(q|r,\tilde{d})}{p(q|\tilde{d})}. \tag{7}$$

In the unigram LM (Ponte & Croft, 1998), by the assumption that query terms occur independently given a particular language model, the query likelihood is decomposed into a product of the probabilities of query terms. Similarly, we decompose the probabilities of Eq. (7) as follows:

Assumption 2 $\frac{p(q|r,\tilde{d})}{p(q|\tilde{d})} = \frac{\prod_{t\in V(q)} p(t|r,\tilde{d})^{f(t,q)}}{\prod_{t\in V(q)} p(t|\tilde{d})^{f(t,q)}}$.

This assumption is a type of linked dependence (Cooper, 1995), as it is equivalent to a combination of the conditions $p(q|r,\tilde{d}) = K\prod_{t\in V(q)} p(t|r,\tilde{d})^{f(t,q)}$ and $p(q|\tilde{d}) = K\prod_{t\in V(q)} p(t|\tilde{d})^{f(t,q)}$, where K is a constant. While $K=1$ corresponds to independence of the terms in the query, a value of K that differs from 1 may be viewed as a crude indicator of departure from independence of the query terms. Thus Assumption 2 is weaker, and more general, than the assumption used in LM.

By Assumption 2, Eq. (7) becomes:

$$p(r|q,\tilde{d}) = \frac{p(r|\tilde{d}) \prod_{t \in V(q)} p(t|r,\tilde{d})^{f(t,q)}}{\prod_{t \in V(q)} p(t|\tilde{d})^{f(t,q)}}. \tag{8}$$

By the conditional probabilities $p(r|\tilde{d}) = p(r,\tilde{d})/p(\tilde{d})$ and $p(t|\tilde{d}) = p(t,\tilde{d})/p(\tilde{d})$, the above becomes:

$$p(r|q,\tilde{d}) = p(r,\tilde{d})p(\tilde{d})^{|q|-1} \prod_{t \in V(q)} \frac{p(t|r,\tilde{d})^{f(t,q)}}{p(t,\tilde{d})^{f(t,q)}}, \tag{9}$$

where $|q|$ is the total number of query terms including repetition counts. Furthermore, by the conditional probability $p(t|r,\tilde{d}) = p(t,r,\tilde{d})/p(r,\tilde{d})$, we have:

$$p(r|q,\tilde{d}) = p(r,\tilde{d})p(\tilde{d})^{|q|-1} \prod_{t \in V(q)} \frac{p(r,t,\tilde{d})^{f(t,q)}}{p(r,\tilde{d})^{f(t,q)} p(t,\tilde{d})^{f(t,q)}}. \tag{10}$$

Using $p(r,t,\tilde{d})/p(t,\tilde{d}) = p(r|t,\tilde{d})$, the above becomes:

$$\begin{aligned} p(r|q,\tilde{d}) &= p(r,\tilde{d})p(\tilde{d})^{|q|-1} \prod_{t \in V(q)} \frac{p(r|t,\tilde{d})^{f(t,q)}}{p(r,\tilde{d})^{f(t,q)}} \\ &= \frac{p(\tilde{d})^{|q|-1}}{p(r,\tilde{d})^{|q|-1}} \prod_{t \in V(q)} p(r|t,\tilde{d})^{f(t,q)} \\ &= \frac{1}{p(r|\tilde{d})^{|q|-1}} \prod_{t \in V(q)} p(r|t,\tilde{d})^{f(t,q)}. \end{aligned} \tag{11}$$

The factor $p(r|\tilde{d})$ in Eq. (11) is query-independent because it is a marginal probability summed over all queries: $p(r|\tilde{d}) = \sum_{q'} p(r,q'|\tilde{d})$. We assume that query-independent document features do not affect the probability of document relevance given a specific query. Hence we make the following assumption:

Assumption 3 *The factor $p(r|\tilde{d})$ does not affect document ranking.*

The above assumption means that query-independent document features are ignored in deriving the document rankings. An example of such features is recency, which has been shown to correlate with relevance in microblog retrieval (Efron, 2012). However, temporal features may not be beneficial for all queries in ad hoc retrieval (Li & Croft, 2003). Another example in web retrieval is the authority of a webpage, which may be indicated by the PageRank score based on links. In document retrieval, while variants of the PageRank may be obtained with automatically-induced links (Kurland & Lee, 2010), such methods are beyond the scope of this paper. In future studies that take into account query-independent features, the document ranking would be scaled by an extra factor as in Eq. (11).

By Assumption 3, Eq. (11) reduces to the following:

$$p(r|q,\tilde{d}) \propto \prod_{t \in V(q)} p(r|t,\tilde{d})^{f(t,q)}. \tag{12}$$

3.3 Decomposing the document

The conditional probabilities of Eq. (12) may be written as follows:

$$p(r|q,\tilde{d}) \propto \prod_{t \in V(q)} \frac{p(\tilde{d},r,t)^{f(t,q)}}{p(\tilde{d},t)^{f(t,q)}} = \prod_{t \in V(q)} \frac{p(\tilde{d}|r,t)^{f(t,q)} p(r,t)^{f(t,q)}}{p(\tilde{d}|t)^{f(t,q)} p(t)^{f(t,q)}}. \quad (13)$$

By an assumption of linked dependence (Cooper, 1995) similar to that applied to the query (Assumption 2), the probabilities of $\tilde{d}$ in Eq. (13) may be written as products of the probabilities of the constituent tuples, $\tilde{d}_u \equiv (u, \tilde{f}(u,d))$, where $u \in V(d)$:

Assumption 4 $\frac{p(\tilde{d}|r,t)}{p(\tilde{d}|t)} = \frac{\prod_{u \in V(d)} p(\tilde{d}_u|r,t)}{\prod_{u \in V(d)} p(\tilde{d}_u|t)}$.

By Assumption 4, Eq. (13) becomes:

$$p(r|q,\tilde{d}) \propto \prod_{t \in V(q)} \left\{ \frac{p(r,t)^{f(t,q)}}{p(t)^{f(t,q)}} \times \left[\prod_{u \in V(d)} \frac{p(\tilde{d}_u|r,t)}{p(\tilde{d}_u|t)} \right]^{f(t,q)} \right\}$$

$$= \prod_{t \in V(q)} \left\{ p(r|t)^{f(t,q)} \times \left[\prod_{\substack{u \in V(d) \\ \wedge u \in V(q)}} \frac{p(\tilde{d}_u|r,t)}{p(\tilde{d}_u|t)} \times \prod_{\substack{v \in V(d) \\ \wedge v \notin V(q)}} \frac{p(\tilde{d}_v|r,t)}{p(\tilde{d}_v|t)} \right]^{f(t,q)} \right\}, \quad (14)$$

where the component in the square bracket is partitioned into terms that appear in the query and those not found in the query.

3.4 Simplification for matching the document and query terms

A bag-of-words representation of the document $\tilde{d}$ (as in Assumption 1) indicates that the occurrences of distinct terms are assumed to be independent. Such term independence further implies the following assumption for a term v that is not the query term t:

Assumption 5 $p(\tilde{d}_v|r,t) \approx p(\tilde{d}_v|t)$ *for* $t \in V(q)$ *and* $v \neq t$.

The above assumption means that given a query term t, the probability of occurrence of a different term v in a retrieved document d does not depend on whether d is relevant or not. Assumption 5 may not be true for example if the term v is a synonym or a statistically associated term (Stiles, 1961; Salton, 1963) of the query term t. It may be problematic also if v is another query term so that relevance is related to the query term v as well as t. However in Assumption 2, it is also assumed that relevance does not affect the decomposition of the query into individual query terms. In this case, Assumption 5 is reasonable.

Using Assumption 5, Eq. (14) can be simplified as follow:

$$
\begin{aligned}
& p(r|q,\tilde{d}) \\
& \propto \prod_{t \in V(q)} \left\{ p(r|t)^{f(t,q)} \times \left[\prod_{\substack{u \in V(d) \\ \wedge u \in V(q)}} \frac{p(\tilde{d}_u|r,t)}{p(\tilde{d}_u|t)} \times \prod_{\substack{v \in V(d) \\ \wedge v \notin V(q)}} \frac{p(\tilde{d}_v|t)}{p(\tilde{d}_v|t)} \right]^{f(t,q)} \right\} \\
& = \prod_{t \in V(q)} \left\{ p(r|t)^{f(t,q)} \times \left[\prod_{\substack{u \in V(d) \\ \wedge u \in V(q)}} \frac{p(\tilde{d}_u|r,t)}{p(\tilde{d}_u|t)} \right]^{f(t,q)} \right\}. \qquad (15)
\end{aligned}
$$

The above may be further simplified by partitioning the terms $u \in V(q)$ into the term that equates to a given query term t and other query terms v that are different from t (i.e., $v \neq t$) , as well as applying Assumption 5 to these other query terms v, as follows:

$$
\begin{aligned}
p(r|q,\tilde{d}) & \propto \prod_{t \in V(q)} p(r|t)^{f(t,q)} \left[\prod_{\substack{u \in V(d) \\ \wedge u = t}} \frac{p((\tilde{d}_u|r,t)}{p((\tilde{d}_u|t)} \times \prod_{\substack{v \in V(d) \\ \wedge v \in V(q) \\ \wedge v \neq t}} \frac{p(\tilde{d}_v|r,t)}{p(\tilde{d}_v|t)} \right]^{f(t,q)} \\
& \approx \prod_{t \in V(q)} p(r|t)^{f(t,q)} \left[\prod_{\substack{u \in V(d) \\ \wedge u = t}} \frac{p(\tilde{d}_u|r,t)}{p(\tilde{d}_u|t)} \times \prod_{\substack{v \in V(d) \\ \wedge v \in V(q) \\ \wedge v \neq t}} \frac{p(\tilde{d}_v|t)}{p(\tilde{d}_v|t)} \right]^{f(t,q)} \\
& \propto \prod_{t \in V(q)} p(r|t)^{f(t,q)} \left[\frac{p(\tilde{d}_t|r,t)}{p(\tilde{d}_t|t)} \right]^{f(t,q)} \\
& = \prod_{t \in V(q)} p(r|t)^{f(t,q)} \left[\frac{p(r|\tilde{d}_t,t)}{p(r|t)} \right]^{f(t,q)} \\
& = \prod_{t \in V(q)} \left[p(r|t,\tilde{d}_t) \right]^{f(t,q)}. \qquad (16)
\end{aligned}
$$

3.5 Estimation of the probabilities

In order to estimate the probability $p(r|t,\tilde{d}_t)$ in Eq. (16) for a query term t that appears in document d, we first consider its complement, $p(\bar{r}|t,\tilde{d}_t)$. We follow the approach of (Wu et al., 2008), whereby retrieval is modeled as a user making a series of local judgment of relevance for each document. By the Query-Centric Assumption (Wu et al., 2008), any information relevant to a given query is found in the texts centered on query term occurrences. Thus, each query term occurrence is judged as either locally relevant (r_{loc}) or locally nonrelevant ($\bar{r}_{loc}$) by examining the text around the term. The overall judgment of relevance of the document is determined by combining the local judgments according to the disjunctive relevance decision (DRD) principle (Kong, Luk, Lam, Ho, & Chung, 2004), which states that a document is judged relevant to a query if any part of the document is relevant to the query. Conversely an irrelevant

document means that all parts of it are not relevant to the query. The DRD principle conforms to the TREC evaluation policy[1] (Dang, Luk, & Allan, 2021).

Suppose there are n occurrences of the term t in document d. In the case of no TF normalization, $\tilde{f}(t,d) = f(t,d) = n$. Then $p(\bar{r}|t, \tilde{d}_t = (t,n))$ is the probability that the d is nonrelevant, given the query term t and the information that there are n occurrences of t in d. We estimate this probability by applying the DRD principle, which conforms to TREC evaluation:

Assumption 6 *$p(\bar{r}|t, \tilde{d}_t = (t,n))$ is approximated as the probability that all n occurrences of the query term t in the document d are locally nonrelevant to the query.*

Let r_{t_i} denote the binary local relevance of the i^{th} occurrence of t. By Assumption 6 we have

$$p(\bar{r}|t, \tilde{d}_t = (t,n)) \approx p(r_{t_1} = \bar{r}_{loc}, r_{t_2} = \bar{r}_{loc}, \cdots, r_{t_n} = \bar{r}_{loc}). \tag{17}$$

We assume that in the case of all occurrences of t being nonrelevant, the probability for the j^{th} occurrence, $p(r_{t_j} = \bar{r}_{loc} | r_{t_1} = \bar{r}_{loc}, r_{t_2} = \bar{r}_{loc}, \cdots, r_{t_{j-1}} = \bar{r}_{loc})$ can be estimated by Laplace's law of succession (e.g., (Feller, 1968) and (Amati & van Rijsbergen, 2002)):

$$p(r_{t_j} = \bar{r}_{loc} | r_{t_1} = \bar{r}_{loc}, r_{t_2} = \bar{r}_{loc}, \cdots, r_{t_{j-1}} = \bar{r}_{loc}) \approx \frac{j}{j+1}. \tag{18}$$

The value of the probability of Eq. (18) increases with j and tends to 1 if j is large. An interpretation of the Laplace's law of succession estimate in Eq. (18) is that the local relevance judgments of each occurrence of t are not taken independently. If the first $j-1$ occurrences of t are all judged to be nonrelevant, for example indicating the usage of term t in the document is different from that in the given query, then the j^{th} occurrence is more likely to be nonrelevant.

The joint probability $p(r_{t_1} = \bar{r}_{loc}, r_{t_2} = \bar{r}_{loc}, \cdots, r_{t_n} = \bar{r}_{loc})$ that all n occurrences of t are nonrelevant may be calculated by the chain rule:

$$p(r_{t_1}, r_{t_2}, \cdots, r_{t_n}) = p(r_{t_1}) \cdot p(r_{t_2}|r_{t_1}) \cdot p(r_{t_3}|r_{t_1}, r_{t_2}) \cdots p(r_{t_n}|r_{t_1}, r_{t_2}, \cdots, r_{t_{n-1}}). \tag{19}$$

Applying Eq. (18) and Eq. (19) to Eq. (17) leads to

$$\begin{aligned} p(\bar{r}|t, \tilde{d}_t = (t,n)) &\approx p(r_{t_1} = \bar{r}_{loc}, r_{t_2} = \bar{r}_{loc}, \cdots, r_{t_n} = \bar{r}_{loc}) \\ &\approx \frac{1}{2} \cdot \frac{2}{3} \cdots \frac{n-1}{n} \cdot \frac{n}{n+1} = \frac{1}{n+1}. \end{aligned} \tag{20}$$

The above derivation of the probability is based on an integral number of trials. In general with TF normalization, in order to calculate the probability $p(\bar{r}|t, \tilde{d}_t = (t, \tilde{f}(t,d)))$ with a non-integral value of $\tilde{f}(t,d)$, we make an assumption that it is obtained by a simple linear interpolation:

Assumption 7 *$p(\bar{r}|t, \tilde{d}_t = (t, \tilde{f}(t,d)))$ is given by a linear interpolation of the probabilities for the closest integral occurrence frequencies below and above $\tilde{f}(t,d)$.*

The closest integers below and above $\tilde{f}(t,d)$ are $n_{floor} = \lfloor \tilde{f}(t,d) \rfloor$ and $n_{ceiling} = \lceil \tilde{f}(t,d) \rceil$, respectively. The linear interpolation in Assumption 7 is given by $p(\bar{r}|t, \tilde{d} = (t, \tilde{f}(t,d))) \approx (n_{ceiling} - \tilde{f}(t,d)) \times p(\bar{r}|t, \tilde{d}_t =$

[1] https://trec.nist.gov/data/reljudge_eng.html

$(t, n_{floor}))$ + $(\tilde{f}(t,d) - n_{floor}) \times p(\bar{r}|t, \tilde{d}_t = (t, n_{ceiling}))$ for $\tilde{f}(t,d) \neq$ integer.

With the probabilities $p(\bar{r}|t, \tilde{d}_t = (t, n_{floor}))$ and $p(\bar{r}|t, \tilde{d}_t = (t, n_{ceiling}))$ being specified based on Laplace's law of succession (Eq. (20)), Assumption 7 leads to:

$$\begin{aligned} p(\bar{r}|t, \tilde{d}_t = (t, \tilde{f}(t,d))) &\approx \frac{(n_{ceiling} - \tilde{f}(t,d))}{(n_{floor}+1)} + \frac{(\tilde{f}(t,d) - n_{floor})}{(n_{ceiling}+1)} \\ &\approx \frac{1}{\tilde{f}(t,d)+1}. \end{aligned} \tag{21}$$

In Eq. (21), the interpolation is approximated by the same functional form as obtained for integral number of occurrences of query term t (Eq. (20)), but with n of Eq. (20) replaced by the normalized frequency $\tilde{f}(t,d)$. With this approximation, we obtain $p(\bar{r}|t, \tilde{d}_t) \approx \frac{1}{\tilde{f}(t,d)+1}$ for all values of $\tilde{f}(t,d)$.

For binary relevance and using Eq. (21),

$$p(r|t, \tilde{d}_t) = 1 - p(\bar{r}|t, \tilde{d}_t) \approx \frac{\tilde{f}(t,d)}{\tilde{f}(t,d)+1}. \tag{22}$$

Now we discuss document length normalization in more detail. In particular, we need to determine the weight $\tilde{f}(t,d)$ in the normalized document $\tilde{d}$, in terms of the observed occurrence frequency of term t in the document, $f(t,d)$. The general form of the normalized weight is $\tilde{f}(t,d) = a(d)f(t,d)$, where $a(d)$ is a document-dependent normalization factor. The p-norm lengths of the document vector and the normalized document vector are, respectively, $|d|_p = \sqrt[p]{\sum_t f(t,d)^p}$ and $|\tilde{d}|_p = \sqrt[p]{\sum_t \tilde{f}(t,d)^p} = \sqrt[p]{\sum_t a(d)^p f(t,d)^p} = a(d)|d|_p$. Therefore, $\tilde{f}(t,d)/|\tilde{d}|_p = f(t,d)/|d|_p$ and thus

$$\tilde{f}(t,d) = \frac{f(t,d)}{|d|_p/|\tilde{d}|_p}. \tag{23}$$

While it is common in the literature (e.g. (Spärck Jones, Walker, & Robertson, 2000)) to apply normalization with the L1 norm (or city-block) document length (i.e. $|d|_p$ with $p = 1$), Eq. (23) is a generalization to p-norm document length. In (Spärck Jones et al., 2000), a scaling is applied to the document length so that the TF normalization is $f(t,d)/NF$, with the normalization factor NF being $NF = |d|/\Delta$, where Δ is the average document length. This scaling ensures that a document of average length will get the same ranking score after document length normalization as without normalization. Thus we set $|\tilde{d}|_p$ in Eq. (23) to $|\tilde{d}|_p = \Delta_p/\kappa$, where Δ_p is the average p-norm document length and κ is a scaling factor.

Furthermore in (Spärck Jones et al., 2000), instead of applying the simple TF normalization factor $NF = |d|/\Delta$, a mixed normalization factor is introduced with a tuning constant b between 0 and 1, such that $NF = (1-b) + b|d|/\Delta$. With $b < 1$, the mixed normalization factor will give a document ranking score larger than the full normalization ($b = 1$). This has the effect of the pivoted normalization of (Singhal et al., 1996) which avoids over-penalizing long documents. Applying the mixed normalization, our weight in the normalized document then becomes:

$$\tilde{f}(t,d) = \frac{f(t,d)}{\kappa\left[(1-\beta) + \beta\frac{|d|_p}{\Delta_p}\right]}, \tag{24}$$

where β is a mixture parameter between 0 and 1. Substituting the above expression for $\tilde{f}(t,d)$ into

Eq. (22), we obtain

$$p(r|t,\tilde{d}_t) \approx \frac{f(t,d)}{f(t,d)+\kappa\left[(1-\beta)+\beta\frac{|d|_p}{\Delta_p}\right]} = TF_{BMp}(t,d). \tag{25}$$

The factor $TF_{BMp}(t,d)$ as defined in Eq. (25) is a p-norm version of the BM25 TF factor (Eq. (1)). In this paper, we focus on $p=2$, i.e. L2 norm (or Euclidean) document length. The retrieval effectiveness of variations with a general $|d|_p$ document length is left for future investigation.

3.6 Instantiations of the ranking formula

A document ranking formula may be obtained by substituting the probability of Eq. (25) in Eq. (16) to calculate $p(r|q,\tilde{d})$. However if any query term t is missing from a document, i.e. $f(t,d)=0$, this would lead to an overall $p(r|q,\tilde{d})=0$. To avoid this zero probability problem, we interpolate the probabilities of Eq. (16) with a background probability of relevance based on the collection $p(r|t,\mathbb{C})$, similar to the LM approach, as follows:

$$\begin{aligned} p(r|q,\tilde{d}) &\propto \prod_{t\in V(q)} \left[\lambda_d p(r|t,\tilde{d}_t) + (1-\lambda_d)p(r|t,\mathbb{C})\right]^{f(t,q)}, \\ &\propto \sum_{t\in V(q)} \log\left(\lambda_d p(r|t,\tilde{d}_t) + (1-\lambda_d)p(r|t,\mathbb{C})\right)^{f(t,q)}, \\ &= \sum_{t\in V(q)} f(t,q)\times\log\left(\lambda_d p(r|t,\tilde{d}_t) + (1-\lambda_d)p(r|t,\mathbb{C})\right), \end{aligned} \tag{26}$$

where λ_d is mixture parameter. In general, the parameter λ_d may be document-dependent. Variations of document ranking may be obtained by appropriate choices of λ_d and the form of $p(r|t,\mathbb{C})$ in Eq. (26).

The first ranking formula is obtained by setting λ_d as used in the SPUD model, i.e. $\lambda_d = \mu|\dot{d}|/(\mu|\dot{d}|+1)$, where μ is a positive constant (see Eq. (5)). To estimate $p(r|t,\mathbb{C})$, we make the following assumption:

Assumption 8 *An occurrence of the query term t in a document is an indication that the document is relevant.*

This assumption is consistent with most retrieval models that assign a higher ranking score to documents with a higher occurrence frequency of a query term, as formalized by the heuristics of (Fang, Tao, & Zhai, 2004). Assumption 8 is only applied to estimate the background probability $p(r|t,\mathbb{C})$ and not to $p(r|t,\tilde{d}_t)$. By this assumption, a document that contains a query term is relevant. So, the number of relevant documents based on the query term t is $df(t)$ (without information about the other query terms) and the probability of selecting a relevant document from the collection is $df(t)/N$. Using this probability of document relevance as an estimate of $p(r|t,\mathbb{C})$, the ranking formula of this instantiation, which will be called PBRn, becomes:

$$\begin{aligned} S_{PBRn}(q,d) &= \sum_{t\in V(q)} f(t,q)\times\log\left(\frac{1}{\mu|\dot{d}|+1}\left[\mu|\dot{d}|\times TF_{BMp}(t,d)+\frac{df(t)}{N}\right]\right) \\ &\propto \sum_{t\in V(q)} f(t,q)\times\log\left(\frac{1}{\mu|\dot{d}|+1}\left[\mu|\dot{d}|\times TF_{BMp}(t,d)\frac{N}{df(t)}+1\right]\right). \end{aligned} \tag{27}$$

The factor $N/df(t)$ in Eq. (27) serves as a term discrimination factor which does not favor common terms, like the traditional IDF.

We consider a second variation similar to PBRn, but with a different estimate of $p(r|t,\mathbb{C})$. By Assumption 8 the occurrence of a query term, and hence the presence of a non-zero dimension of the query term in the binary document vector $\dot{d'}$, indicates that the document d' is relevant. Thus, again based on Assumption 8, $p(r|t,\mathbb{C})$ is estimated by the probability that a document vector $\vec{d'}$ has a non-zero dimension of query term t. The number of distinct terms, i.e. number of non-zero dimensions, in a document d' is $|\dot{d'}|$. Considering all the documents in the whole collection, the number of instances of t appearing as a non-zero dimension is $df(t)$, while the total number of non-zero dimensions is $\sum_{d'} |\dot{d'}|$. Therefore, the probability of t being a non-zero dimension in a document vector is $df(t)/\sum_{d'} |\dot{d'}|$. Moreover, $\sum_{d'} |\dot{d'}|$ is equal to the sum of document frequency of all distinct terms over the collection, i.e. $\sum_{d'} |\dot{d'}| = \sum_{t'} df(t')$ (Cummins et al., 2015). Therefore, $p(r|t,\mathbb{C})$ is estimated by the probability of finding a non-zero dimension of the query term in a document vector, i.e. $p(r|t,\mathbb{C}) \approx df(t)/\sum_{t'} df(t') = df(t)/df_{\mathbb{C}}$. This estimate of $p(r|t,\mathbb{C})$ has the same form as the collection smoothing employed in SPUD (Eq. (5)). This variation will be called PBRs, with the ranking formula being:

$$\begin{aligned} S_{PBRs}(q,d) &= \sum_{t \in V(q)} f(t,q) \times \log\left(\frac{1}{\mu|\dot{d}|+1}\left[\mu|\dot{d}| \times TF_{BMp}(t,d) + \frac{df(t)}{df_{\mathbb{C}}}\right]\right) \\ &\propto \sum_{t \in V(q)} f(t,q) \times \log\left(\frac{1}{\mu|\dot{d}|+1}\left[\mu|\dot{d}| \times TF_{BMp}(t,d)\frac{df_{\mathbb{C}}}{df(t)} + 1\right]\right). \end{aligned} \tag{28}$$

As a third variation, considering the whole collection, there are $f(t,\mathbb{C})$ occurrences of term t out of a total of $\sum_{t'} f(t',\mathbb{C}) = |\mathbb{C}|$ terms, so that the probability of the occurrence of t is $f(t,\mathbb{C})/|\mathbb{C}|$. By Assumption 8, an occurrence of t in a document indicates that it is relevant. Therefore, $p(r|t,\mathbb{C})$ is estimated by the probability of occurrence of t, so that $p(r|t,\mathbb{C}) \approx f(t,\mathbb{C})/|\mathbb{C}|$. This is an approximate upper bound estimate as some occurrences of t are relevant and some are not. For λ_d, we apply the same form as the Dirichlet smoothing of LM (Zhai & Lafferty, 2004) (Section 2). Together with Eq. (25), the ranking formula of this instantiation, which will be called the PBRc, is then:

$$S_{PBRc}(q,d) = \sum_{t \in V(q)} f(t,q) \times \log\left(\lambda_d \times TF_{BMp}(t,d) + (1-\lambda_d)\frac{f(t,\mathbb{C})}{|\mathbb{C}|}\right), \tag{29}$$

where $\lambda_d = |d|/(|d|+\mu)$, with μ being a positive constant.

4 Experiments

The experimental environment and our training and testing methodology are described in Section 4.1. Our hypothesis testing procedure is discussed in Section 4.2. Section 4.3 presents the results of our retrieval experiments.

4.1 Setup and methodology

Retrieval experiments are performed on our own retrieval system (e.g. (Dang et al., 2021)) using a wide range of standard TREC test collections that span the history of TREC evaluation (Table 2).

Table 2: Summary of TREC collections

Type	News / Congress records			
Dataset	Disks 1&2		Disks 4&5	Disks 4&5 - CR
N	714,857		556,075	528,153
Size (GB)	3.8		3.27	3
TREC	2	3	6	7&8, Robust 2003&2004
Queries	101-150 (training)	151-200 (testing)	301-350 (training)	351-450, 601-700 (testing)

Type	Webpages		
Dataset	WT10g	GOV2	
N	1,692,096	25,205,179	
Size (GB)	10	426	
TREC	9&10	Terabyte 2004&2005	Terabyte 2006
Queries	451-550 (testing)	701-800 (testing)	801-850 (training)

Type	Webpages	
Dataset	Clueweb09 Category B	
N	50,189,002	
Size (GB)	≈ 1500	
TREC	Web track 2009-2011	Web track 2012
Queries	1-150 (testing)	151-200 (training)

For diversity, collections of both newswire articles and webpages are used. To evaluate the scalability of retrieval methods, the collection sizes span from about 3GB (Disks 4&5) to the terabyte regime (Clueweb09 Category B). Stopword removal and stemming based on Porter's algorithm (Porter, 1980) are applied. While the Clueweb09 collection contains some spam data, we do not apply spam filtering as there is at the moment no standard setting of filtering. The performance of the retrieval models is evaluated in terms of two widely used metrics (e.g. (Ferro & Silvello, 2018)), namely the MAP (based on the top 1000 retrieved documents), which takes into account both precision and recall, and the precision-oriented NDCG@20, which is found to reflect user preference. The values of the metrics are obtained based on binary relevance for all tracks.

Short title queries, each averaging about 2 to 3 query terms are used in our retrieval. It is well known that parameter values of retrieval models can affect their performance significantly (e.g. (Zhao et al., 2014)). Thus, it is crucial to calibrate the parameters of both the baselines and the PBR models as best as possible for a meaningful comparison. The free parameters of each retrieval model are calibrated by training on a selected set of 50 title queries (i.e. development topics (Lease, 2009; Roy, Bhatia, & Mitra, 2019)) and applied to other sets of queries for testing. Such training and testing methodology has been used also by other researchers (Lease, 2009; Roy et al., 2019; Trotman, Puurula, & Burgess, 2014). Table 2 shows the training and testing sets used in our experiments. Training is performed separately for the news and webpage collections because of their different document nature. Also, separate training is performed for Clueweb09 Cat-B because it has a much larger size and contains spam while the other two webpage collections do not. For the webpage collections we select the last track (i.e. Terabyte-2006 of GOV2 and Web track 2012 of Clueweb09) to be the training set because the systems participating in the later tracks of TREC should be better trained, so that there is less chance of missing relevance documents on the last track for TREC to form the pool of judged relevant documents. Calibration is performed by a grid search of the parameters that maximize the MAP. Table 3 shows the range over which the various parameters are searched and the optimized values obtained for the various collections. The search range is adjusted such that a peak in MAP is observed within the range. To ensure a strong baseline and confirm the optimized parameters for BM25, we purposely applied a wide search range for

Table 3: Range of retrieval model parameters in grid search and optimized values for various test collections

Retrieval model	Free parameters and grid search range	Optimized value			
		Disks 1&2	Disks 4&5	WT10g/GOV2	Clueweb09
BM25	k:0.1-2000, b:0.1-1.4	1.5, 0.25	0.6, 0.4	0.8, 0.35	3, 0.15
LM	μ:100-3000	1700	400	700	2000
SPUD	μ:0.0001-0.01	0.001	0.003	0.0015	0.0004
PL2	c:1-15	5	10	8	12
PBRn	κ:1-100, β:0.6-1.4, μ:0.001-0.05	16, 0.5, 0.005	11, 1.0, 0.02	20, 0.75, 0.015	2, 0.7, 0.03
PBRs	κ:1-2500, β:0.6-1.4, μ:0.001-0.02	1400, 0.85, 0.003	500, 1.1, 0.008	2000, 0.9, 0.008	5, 0.9, 0.011
PBRc	κ:10-2000, β:0.6-1.4, μ:20-600	800, 1.05, 800	400, 1.2, 300	1500, 1.0, 300	20, 0.8, 300

Table 4: Comparison of MAP between our search engine and the literature

Model	Retrieval system	GOV2 Queries 701-850	Clueweb09 (Category B) Queries 51-200
BM25	Our Search Engine	0.300	0.104
	Yang & Fang (2016)	0.297	0.089
LM	Our Search Engine	0.302	0.100
	Yang & Fang (2016)	0.299	0.090
PL2	Our Search Engine	0.303	0.107
	Yang & Fang (2016)	0.303	0.089

k and b to cover the range as used for the κ and β of the PBR models.

In order to ensure that our implementations of the baseline models perform at a similar level as found in the literature, we compare the MAP performance as shown in Table 4. The performance of our search engine is slightly higher or similar to that of (P. Yang & Fang, 2016) for the various retrieval models on GOV2 and Clueweb09 Cat-B. The MAP performance by (P. Yang & Fang, 2016) is based on optimizing all the queries of the collections, whereas our search engine is only optimized for the training subset of queries. If we optimize our search engine using all the queries of the collections, we expect our search engine to perform even better. Therefore, the baseline retrieval models of our search engine are at least as strong as those reported in the literature.

4.2 Hypothesis testing

We investigate whether the retrieval effectiveness of the PBR models is better than that of the baseline models. For this purpose, we introduce Null Hypothesis Families (NHF) corresponding to each of the evaluation metrics that we use, i.e. MAP and NDCG@20:

> **NHF-MAP**: There is no difference in retrieval effectiveness in terms of MAP between the PBR models and baselines such as BM25, LM, SPUD or PL2.

and

> **NHF-NDCG**: There is no difference in retrieval effectiveness in terms of NDCG@20 between the PBR models and baselines such as BM25, LM, SPUD or PL2.

The Null Hypothesis Families NHF-MAP and NHF-NDCG involve many statistical tests in the comparison between the PBR models and baselines tested in this study, leading to the multiple comparisons problem (Carterette, 2012). As a results, corrections to the significance levels are needed (Dror, Baumer,

Table 5: Retrieval performance on Disks 1&2 collections

	Disks 1&2			
	TREC-2 Q101-150 (training)		TREC-3 Q151-200 (testing)	
	MAP	NDCG@20	MAP	NDCG@20
BM25	.2057	.4539	.2712	.5411
LM	.1988	.4390	.2581	.5164
SPUD	.2011	.4471	**.2723**	.5447
PL2	.1953	.4346	.2653	.5373
PBRn	.1964	.4439	$.2618_s$	.5368
PBRs	.1977	.4483	$.2682^l$	$\mathbf{.5542}^l$
PBRc	.1961	.4431	$.2600_s$	.5367
FDR threshold	-	-	.0125	.004

Note: The superscripts/subscripts b,l,s,p indicate a higher/lower value than BM25, LM, SPUD and PL2, respectively, with statistical significance.

Table 6: Retrieval performance on Disks 4&5 collections

	Disks 4&5 TREC-6 Q301-350 (training)		Disks 4&5 - CR TREC-7&8, Robust 2003&2004 Q351-450,601-700 (testing)	
	MAP	NDCG@20	MAP	NDCG@20
BM25	.2444	.4427	.2573	.4137
LM	.2480	.4477	.2547	.4140
SPUD	.2530	.4566	**.2613**	.4226
PL2	.2494	.4560	.2561	.4204
PBRn	.2541	.4593	.2548	.4171
PBRs	.2592	.4650	.2591	**.4237**
PBRc	.2550	.4609	.2586	.4213
FDR threshold	-	-	.004	.004

Note: For the testing results, none of the PBR models differs from the baselines with statistical significance.

Bogomolov, & Reichart, 2017; Raiber & Kurland, 2019). Specifically we apply the widely-adopted Benjamini-Hochberg procedure (Benjamini & Hochberg, 1995). Suppose there are m null hypotheses in the family and the corresponding statistical test p-values are sorted in ascending order, the Benjamini-Hochberg threshold is the largest p-value that is smaller than or equal to $k/m \times \alpha_{FDR}$ where k is the rank of the p-value ($k = 1$ for the smallest p-value) and α_{FDR} is the false discovery rate. We set α_{FDR} to 5%, corresponding to a statistical significance confidence level of 95%. In this study, within each Null Hypothesis Family NHF-MAP and NHF-NDCG there are 12 hypotheses (4 baseline models $\times$ 3 PBR models). The threshold p-value of the Benjamini-Hochberg procedure is reported in Tables 5 to 8. Statistical significance means that the p-value is smaller than or equal to the stated FDR p-value threshold. We obtain the p-values by the randomization test (Smucker, Allan, & Carterette, 2007; Parapar, Losada, Presedo-Quindimil, & Barreiro, 2020).

Following the Benjamini-Hochberg procedure in effect entails a more stringent condition for the rejection of the null hypotheses. For example in Table 7 for the NDCG@20 results, with the False Discovery Rate set to $\alpha = 0.05$, the FDR threshold is found to be 0.033, so that a null hypothesis in the NHF-NDCG family is rejected only if the corresponding p-value is less than 0.033 instead of 0.05.

4.3 Results

Tables 5 to 8 summarize the retrieval results on various test collections. Across the collections, the best performing model in MAP or NDCG@20 in the testing data (shown in bold in the tables) is mostly either PBRs or PBRc. The only exception is for MAP on the newswire collections (Tables 5 and 6), for

Table 7: Retrieval performance on WT10g and GOV2 collections

	GOV2 Terabyte 2006 Q801-850 (training)		WT10g TREC-9&10 Q451-550 (testing)		GOV2 Terabyte 2004&5 Q701-800 (testing)	
	MAP	NDCG@20	MAP	NDCG@20	MAP	NDCG@20
BM25	.3044	.4940	.2084	.3238	.2971	.4559
LM	.3083	.4748	.2076	.3214	.2988	.4543
SPUD	.3210	.5025	.2136	.3281	.3123	.4677
PL2	.3072	.4735	.2084	.3268	.3011	.4536
PBRn	.3284	.5302	.2149	.3362	.3065	.4713
PBRs	.3301	.5308	.2161	$.3410^{blsp}$	$\mathbf{.3128}^{l}$	**.4750**
PBRc	.3264	.5257	**.2175**	$\mathbf{.3430}^{blsp}$	.3098	.4741
FDR threshold	-	-	.004	.033	.008	.004

Note: The superscripts *b,l,s,p* indicate a higher value than BM25, LM, SPUD and PL2, respectively, with statistical significance.

Table 8: Retrieval performance on Clueweb09 Cat-B collection

	Clueweb09 Cat-B			
	Web Track 2012 Q151-200 (training)		Web Track 2009-2011 Q1-150 (testing)	
	MAP	NDCG@20	MAP	NDCG@20
BM25	.1170	.1437	.0967	.2023
LM	.1082	.1277	.0939	.1750
SPUD	.1172	.1308	.0938	.1767
PL2	.1108	.1265	.1018	.1845
PBRn	.1381	.2067	.1007	$.2489^{blsp}$
PBRs	.1378	.2112	.0996	$.2449^{blsp}$
PBRc	.1419	.2074	$\mathbf{.1060}^{bls}$	$\mathbf{.2566}^{blsp}$
FDR threshold	-	-	.013	.05

Note: The superscripts *b,l,s,p* indicate a higher value than BM25, LM, SPUD and PL2, respectively, with statistical significance.

which SPUD is the top model, though the difference between SPUD and the PBR models is statistically significant only for PBRn and PBRc on TREC-3 (Table 5), and not in other cases. For the newswire collections, while there are cases where BM25 and PL2 attain higher MAP or NDCG@20 than the PBR models, all of the differences are not statistically significant. On the other hand, for the webpage collections (Tables 7 and 8), the PBRn/PBRs/PBRc categorically perform better than BM25 and LM in both MAP and NDCG@20, with statistical significance in some cases as shown in the tables.

Overall, the better performance of the PBR models over the baselines is more obvious for the larger webpage collections with more documents than the newswire collections. The PBRc model is the top performer in terms of both MAP and NDCG@20 for the WT10g and Clueweb09 Cat-B collections, while the PBRs model yields the highest MAP and NDCG@20 values on the GOV2 collection. The results suggest that the new PBR retrieval models are particularly suited for larger collections in the terabyte regime. This supports the anticipation that the new model, in which the product of probabilities over query terms corresponds to conjoining of the terms, performs better in large collections with more likelihood of finding documents having co-occurrences of different query terms.

The PBR models generally improve over the baselines more in NDCG@20 than in MAP. For the newswire collections (Tables 5 and 6), PBRs attains the best NDCG@20 even though some of the baselines perform better in MAP. On WT10g (Table 7), the PBRs and PBRc models yield better NDCG@20 than all the baseline models with statistical significance, and likewise for all the PBR models on Clueweb09 (Table 8). However, in these cases, the corresponding numerical improvement in MAP may not be statistically significant. This observation suggests that the PBR models are more effective over the tested

baselines in the precision-oriented NDCG@20 than in MAP that also takes recall into account.

Among the PBR models, the tested variants are mostly comparable in retrieval effectiveness based on MAP or NDCG@20 across collections. Overall PBRs and PBRc appear to be the better models as both of them attain the best MAP or NDCG@20 on two or three collections. Across collections, the relative performance of the SPUD-like PBRs and the LM-like PBRc appear to match the comparison between SPUD and LM. For example, on the newswire collections and also GOV2, where SPUD performs better than LM, PBRs also shows better performance than PBRc in MAP and NDCG@20. On the WT10g and Cluweb09 collections, where LM and SPUD show comparable or smaller differences in performance, PBRc performs better than PBRs.

5 Conclusion and future work

We have presented a detailed derivation of the novel PBR retrieval model family based on the PRP. Retrieval performance of several instantiations of the PBR models is evaluated by comparing with strong baselines including the BM25, LM, SPUD and PL2 models on a range of TREC collections. Improvement of the new models over the baselines with statistical significance is observed, especially in the large Clueweb09 Cat-B collection. The effectiveness of the PBR models suggests that they can serve as base models of various retrieval techniques for further retrieval effectiveness improvement, e.g. relevance feedback / pseudo-relevance feedback, or metasearch approaches like reciprocal rank fusion (Cormack, Clarke, & Büttcher, 2009).

One of the limitations of this work is that it relies on the assumptions made in deriving the PBR models, such as the term independence assumption that excludes term association, which may not hold in some retrieval scenarios. However, it is not known whether such assumptions, whether they hold or not, have a (statistically significant) impact on retrieval effectiveness of the PBR models. Therefore, it is necessary to evaluate the effect of going beyond such assumptions by experiments. Hence, possible topics of future study include combining PBR models with term association methods, such as cross terms (Zhao et al., 2014) or MRF techniques (Metzler & Croft, 2005) that incorporate various degrees of term dependencies, and the use of term position features (Hammache & Boughanem, 2021). Furthermore, we have ignored query-independent features (by Assumption 3). While we show that the PBR models attain good performance without consideration of such features, it is of interest to investigate whether such features can further enhance retrieval effectiveness. Last, the effectiveness of new instantiations of the PBR models may also be explored.

References

Amati, G., & van Rijsbergen, C. J. (2002). Probabilistic models of information retrieval based on measuring the divergence from randomness. *ACM Transactions on Information Systems*, *20*(4), 357–389.

Azzopardi, L., & Roelleke, T. (2007). Explicitly considering relevance within the language modeling framework. In *Proceedings of the 1st International Conference on Theory of Information Retrieval* (pp. 125–134).

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, *57*(1), 289–300.

Carterette, B. A. (2012). Multiple testing in statistical analysis of systems-based information retrieval experiments. *ACM Transactions on Information Systems*, *30*(1, Article 4), 1–34.

Cooper, W. S. (1995). Some inconsistencies and misidentified modeling assumptions in probabilistic information retrieval. *ACM Transactions on Information Systems*, *13*(1), 100–111.

Cormack, G. V., Clarke, C. L. A., & Büttcher, S. (2009). Reciprocal rank fusion outperforms Condorcet and individual rank learning methods. In *Proceedings of the 32nd Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval* (pp. 758–759).

Cummins, R., Paik, J. H., & Lv, Y. (2015). A Pólya Urn Document language model for improved information retrieval. *ACM Transactions on Information Systems*, *33*(4, Article 21), 1–34.

Dang, E. K. F., Luk, R. W. P., & Allan, J. (2021). A principled approach using fuzzy set theory for passage-based document retrieval. *IEEE Transactions on Fuzzy Systems*, *29*(7), 1967–1977.

Dror, R., Baumer, G., Bogomolov, M., & Reichart, R. (2017). Replicability analysis for natural language processing: Testing significance with multiple datasets. *Transactions of the Association for Computational Linguistics*, *5*, 471–486.

Efron, M. (2012). Query-specific recency ranking: Survival analysis for improved microblog retrieval. In *Proceedings of the TAIA-12 Workshop associated to SIGIR-12.*

Fang, H., Tao, T., & Zhai, C. (2004). A formal study of information retrieval heuristics. In *Proceedings of the 27th Ann. Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval* (pp. 49–56).

Feller, W. (1968). *An introduction to probability theory and its application, vol. 1, 3rd ed.* Weley, New York.

Ferro, N., & Silvello, G. (2018). Toward an anatomy of IR system component performances. *Journal of the Association for Information Science and Technology*, *69*(2), 187–200.

Guo, J., Fan, Y., Pang, L., Yang, L., Ai, Q., Zamani, H., ... Cheng, X. (2020). A deep look into neural ranking models for information retrieval. *Information Processing & Management*, *57*(6), 102067.

Hammache, A., & Boughanem, M. (2021). Term position-based language model for information retrieval. *Journal of the Association for Information Science and Technology*, *72*(5), 627–642.

He, B., & Ounis, I. (2003). A study of parameter tuning for term frequency normalization. In *Proceedings of the 12th ACM Conf. on Information and Knowledge Management (CIKM'03)* (pp. 10–16).

He, B., & Ounis, I. (2007). Parameter sensitivity in the probabilistic model for ad-hoc retrieval. In *Proceedings of the 16th ACM Conf. on Information and Knowledge Management (CIKM'07)* (pp. 263–272).

Hiemstra, D. (1998). A linguistically motivated probabilistic model of information retrieval. In *Proceedings of European Conference on Digital Libraries* (pp. 569–584).

Kong, Y. K., Luk, R. W. P., Lam, W., Ho, K. S., & Chung, F. L. (2004). Passage-based retrieval using parameterized fuzzy set operators. In *ACM SIGIR Workshop on Mathematical/Formal Methods for Information Retrieval.*

Kurland, O., & Lee, L. (2010). Pagerank without hyperlinks: Structural reranking using links induced by language models. *ACM Transactions on Information Systems*, *28*(4), 1–38.

Lafferty, J., & Zhai, C. X. (2001). Document language models, query models and risk minimization for information retrieval. In *Proceedings of the 24th Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval* (pp. 111–119).

Lafferty, J., & Zhai, C. X. (2003). Probabilistic relevance models based on document and query generation. In *Language modeling for information retrieval* (pp. 1–10). Springer.

Lease, M. (2009). An improved markov random field model for supporting verbose queries. In *Proceedings of the 32nd Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval* (pp. 476–483).

Li, X., & Croft, W. B. (2003). Time-based language models. In *Proceedings of the 12th ACM Conf. on Information and Knowledge Management (CIKM'03)* (pp. 469–475).

Luk, R. W. P. (2008). On event space and rank equivalence between probabilistic retrieval models. *Information Retrieval*, *11*, 539–561.

Luk, R. W. P. (2022). Why is information retrieval a scientific discipline. *Foundations of Science*, To appear.

Metzler, D., & Croft, W. B. (2005). A markov random field model for term dependencies. In *Proceedings of the 28th Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval* (pp. 472–479).

Na, S.-H. (2015). Two-stage document length normalization for information retrieval. *ACM Transactions on Information Systems*, *33*(2, Article 8), 1–40.

Parapar, J., Losada, D. E., Presedo-Quindimil, M. A., & Barreiro, A. (2020). Using score distributions to compare statistical significance tests for information retrieval evaluation. *Journal of the Association for Information Science and Technology*, *71*(1), 98–113.

Ponte, J. M., & Croft, W. B. (1998). A language modeling approach in information retrieval. In *Proceedings of the 21st Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval* (pp. 275–281).

Porter, M. F. (1980). An algorithm for suffix stripping. *Program*, *14*(3), 130–137.

Raiber, F., & Kurland, O. (2019). Relevance feedback: The whole is inferior to the sum of its parts. *ACM Transactions on Information Systems*, *37*(4, Article 44), 1–28.

Robertson, S., Walker, S., Jones, S., Hancock-Beaulieu, M., & Gatford, M. (1995). Okapi at TREC-3. In *Proceedings of the 3rd Text REtrieval conference (TREC-3)* (p. 109). Gaithersburg, MD: NIST Special Publication 500-226.

Robertson, S., & Zaragoza, H. (2009). The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, *3*(4), 333–389.

Robertson, S. E. (1977). The probability ranking principle in ir. *Journal of Documentation*, *33*, 294–304.

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

Robertson, S. E., & Walker, S. (1994). Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *Proceedings of the 17th Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval* (pp. 232–241).

Roy, D., Bhatia, S., & Mitra, M. (2019). Selecting discriminative terms for relevance model. In *Proceedings of the 42nd Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval* (pp. 1253–1256).

Salton, G. (1963). Associative document retrieval techniques using bibilographic information. *Journal of the ACM*, *10*(4), 440–457.

Singhal, A., Buckley, C., & Mitra, M. (1996). Pivoted document length normalization. In *Proceedings of the 19th Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval* (pp. 21–29).

Smucker, M. D., Allan, J., & Carterette, B. (2007). A comparison of statistical significance tests for information retrieval evaluation. In *Proceedings of the 16th ACM Conf. on Information and Knowledge Management (CIKM'07)* (pp. 623–632).

Spärck Jones, K., Walker, S., & Robertson, S. E. (2000). A probabilistic model of information retrieval: development and comparative experiments part 2. *Information Processing and Management*, *36*, 809–840.

Stiles, H. (1961). The association factor in information retrieval. *Journal of the ACM*, *8*(2), 271–279.

Trotman, A., Puurula, A., & Burgess, B. (2014). Improvements to BM25 and language models examined. In *Proceedings of the 2014 Australasian Document Computing Symposium* (p. 58-65).

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), 1–37.

Yang, P., & Fang, H. (2016). A reproducibility study of information retrieval models. In *Proceedings of the 2016 ACM International Conference on the Theory of Information Retrieval* (pp. 77–86).

Yang, W., Zhang, H., & Lin, J. (2019). Simple applications of bert for ad hoc document retrieval. *arXiv preprint arXiv:1903.10972*.

Yu, C., & Salton, G. (1976). Precision weighting - an effective automatic indexing method. *Journal of the ACM*, *23*(1), 76–88.

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

Zhao, J., Huang, J. X., & Ye, Z. (2014). Modeling term associations for probabilistic information retrieval. *ACM Transactions on Information Systems*, *32*(2, Article 7), 1–47.