A comparison between term-independence retrieval models for ad hoc retrieval

EDWARD KAI FUNG DANG and ROBERT WING PONG LUK, The Hong Kong Polytechnic University
JAMES ALLAN, University of Massachusetts, Amherst, USA

In Information Retrieval, numerous retrieval models or document ranking functions have been developed in the quest for better retrieval effectiveness. Apart from some formal retrieval models formulated on a theoretical basis, various recent works have applied heuristic constraints to guide the derivation of document ranking functions. While many recent methods are shown to improve over established and successful models, comparison among these new methods under a common environment is often missing. To address this issue, we perform an extensive and up-to-date comparison of leading term-independence retrieval models implemented in our own retrieval system. Our study focuses on the following questions: (RQ1) Is there a retrieval model that consistently outperforms all other models across multiple collections; (RQ2) What are the important features of an effective document ranking function? Our retrieval experiments performed on several TREC test collections of a wide range of sizes (up to the terabyte-sized Clueweb09 Category B) enable us to answer these research questions. This work also serves as a reproducibility study for leading retrieval models. While our experiments show that no single retrieval model outperforms all others across all tested collections, some recent retrieval models, such as MATF and MVD, consistently perform better than the common baselines.

CCS Concepts: • **Information systems** → **Retrieval models and ranking**; **Retrieval effectiveness**.

Additional Key Words and Phrases: Information retrieval, retrieval model, comparison, multiple hypotheses testing, evaluation

1 INTRODUCTION

In Information Retrieval (IR), there is an ongoing quest to develop new retrieval models or document ranking functions to obtain better retrieval effectiveness. In the Vector Space Model, document ranking is based on a summation of (normalized) weights of matching query terms that make up the representation of document vectors. Generally, the effectiveness of different forms of term weights needs to be tested empirically in order to find the best term weight [77]. Some formal retrieval models are formulated on a theoretical basis and their derivation guided by specific principles. For example, the Probability Ranking Principle (PRP) [74] 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. The well-established BM25 model [72] follows the PRP as its ranking function is derived based on the estimated probability of relevance of the documents. Another highly successful approach is the probabilistic language model (LM) framework [35; 67]. Even though there is no explicit notion of relevance in the original formulation of LM, it has been shown that LM also conforms to the PRP [4; 48; 55; 56]. On the other hand, some other successful probabilistic models, such as those of the Divergence from randomness (DFR) framework [2], are not derived on the basis of the PRP. While some models are not regarded as following the PRP in this article, it is possible that in the future the ranking function of some of these models is shown to be rank equivalent to the probability of relevance and thus conform to the PRP.

Inspired by the functional form and components of some successful retrieval models, Fang et al. [23] introduced a list of heuristic constraints on document ranking functions for good performance. Various recent works have applied these heuristic constraints in designing document ranking functions. For example, Clinchant and Gaussier [11] invoke

Authors' addresses: Edward Kai Fung Dang, cskfdang@comp.polyu.edu.hk; Robert Wing Pong Luk, Department of Computing, The Hong Kong Polytechnic University, Hung Hom, Hong Kong, csrluk@comp.polyu.edu.hk; James Allan, College of Information and Computer Sciences, University of Massachusetts, Amherst, MA, 01003-9264, USA, allan@cs.umass.edu.

the heuristic constraints to derive information-based models, while Goswami et al. [28] investigate all functions that are systematically generated within a certain function space and pruned by the heuristic constraints. The effectiveness of such methods as demonstrated in retrieval experiments [11; 28] supports the heuristic constraints of [23] as good criteria for effective ranking functions.

In practice, as there is no definite theoretical prediction of the actual retrieval effectiveness of a ranking function, the performance of retrieval methods must be evaluated empirically. The standard practice is to compare new retrieval methods with one or more established retrieval models as baselines to confirm the effectiveness of the new methods. Yet, the comparison between some latest promising methods is often lacking. There have been some empirical comparisons or reproducibility studies of retrieval methods in the past (e.g. [6; 17; 40; 90; 91]). However, there are various shortcomings. Document collections of small sizes were used in some of the early studies, e.g. 3204 documents in [17]. On the other hand, while larger document collections were used in more recent works such as [90], some recent promising retrieval models were not covered (e.g. SPUD [16], MVD [65], and the machine generated models of Goswami et al. [28]).

In this article we present an extensive and up-to-date comparison of leading retrieval models. The goal of the study is not just to investigate whether there exists a consistent top performing retrieval model. In addition, by comparing the retrieval models we seek to provide some insight regarding the factors for good retrieval performance. Thus the focus of our study is represented by the following research questions:

RQ 1. *Is there a retrieval model that consistently outperforms all other models across multiple collections?*

RQ 2. *What are the important features of an effective document ranking function?*

To answer the above questions, we have performed retrieval experiments with an extensive list of term-independence (or bag-of-words) retrieval models, i.e. models that assume independence in the occurrence of distinct terms. We focus on term-independence models because many leading retrieval models adopt the term-independence assumption, including some common baselines in IR studies, such as BM25, LM and PL2. Furthermore, an examination of the effectiveness of term-independence retrieval models is a prerequisite for an extension to include term-dependence in the base models, or utilization of these models in multi-stage ranking architectures [91]. Thus, we have excluded comparison with methods using advanced techniques such as proximity matching (e.g. Markov random field [59]). Besides, we have not employed pseudo-relevance feedback (PRF) or other term-expansion techniques, which are well known to enhance retrieval effectiveness (e.g. [57]), as we focus on the features of the base models. All the retrieval models are implemented in our own retrieval system, enabling a fair comparison under a common experimental environment. The empirical comparison is performed in terms of two common metrics — Mean Average Precision (MAP) and normalized discounted cumulative gain at the top 20 documents (NDCG@20). Our retrieval experiments are performed using TREC collections covering different domains (news/federal register documents and webpages) and a wide range of sizes (from the 3GB TREC-6 to the terabyte-sized Clueweb09 Category B).

The novelty of this work, which distinguishes it from other comparative studies of retrieval models [6; 17; 40; 90; 91], includes the following aspects: (a) in addition to an empirical comparison of the retrieval models, we include an analytical comparison by noting the features employed in the various document ranking functions and identifying the important features for effective retrieval; (b) our extensive comparison includes some novel retrieval models not featured in previous studies, such as the machine generated functions of [28], SPUD [16], MATF [64] and MVD models [65]; (c) we have included a review of the retrieval models, using a uniform notation to specify the ranking functions, for the ease of other researchers to implement these functions; (d) we select reliable training sets to calibrate the retrieval models instead of, say, simply selecting the first TREC track on each dataset (Section 4.1); (e) in addition to MAP, we include results of the NDCG metric, which was not used in some previous comparative studies; (f) we employ a hypothesis testing in which retrieval models are compared against the model found to be best performing, instead of against a specific baseline; (g) we apply the Benjamini-Horchberg procedure in statistical significance testing to tackle the multiple comparison problem [8].

In summary, the contributions of our current study are the following findings:

- With respect to RQ1, no single retrieval model is found to consistently outperform all other tested methods with statistical significance and across all collections used.
- Several retrieval models, namely BM25, LM and PL2, which are commonly used as baselines in IR studies, consistently do not rank in the top three in either MAP or NDCG@20 across all collections.
- On the other hand, some recent retrieval models, such as MATF [64] and MVD [65], consistently yield higher MAP and NDCG@20 than the common baselines. Thus, the results suggest that in studies of new retrieval techniques, new good performing retrieval models like these may be used as baselines and the methods should be tested on multiple collections. It is noted that while MATF and MVD perform well, they are typically not supported in common open-source systems (e.g. Indri[1], Terrier[2], Lucene[3] or Anserini[4] [91]).
- To answer RQ2, we identify some common features of the top performing models as important features for an effective retrieval ranking function: (1) a logarithmic function of the term frequency; (2) normalization of the term frequency by the average term frequency; and (3) normalization either based on the number of discrete terms in a document or by the ‘normalization 2’ of the DFR framework [2].
- Our work also serves as a reproducibility study for leading established and recent retrieval models. The retrieval results of the models implemented in our system match those reported in the literature, obtained on systems such as Indri [90], Terrier [91] or Anserini [91].

The rest of the article is organized as follows. Section 2 provides the background of the current work, including a review of some past comparisons of retrieval models (Section 2.1) and a review of the common heuristics applied in past works (Section 2.2). The various retrieval models and document ranking methods included in our own study are then described in Section 3. The details of our experimental study are presented in Section 4. A discussion of the experimental results including an analytical evaluation of the retrieval methods is presented in Section 5. Finally, a conclusion of the study is provided.

2 BACKGROUND

2.1 Literature review

There have been several past comparative studies of retrieval models in terms of their retrieval effectiveness, covering exact-match, vector space, probabilistic models, etc. Such studies generally compare the performance of the retrieval models empirically, as we briefly review in Section 2.1.1, though there are also some studies that provide an analytical comparison of the retrieval models, as discussed in Section 2.1.2. A summary of the studies is shown in Table 1.

2.1.1 Empirical comparison.

[1] http://www.lemurproject.org/indri.php
[2] http://terrier.org
[3] https://lucene.apache.org/core/
[4] http://anserini.io/

Table 1. Past comparative studies of term-independence retrieval models

Paper	Retrieval models	initial/ PRF	Test collections	Language	Evaluation metric
(a) Empirical comparison					
Dai et al. [17]	VSM, BM25, KL-Dir (LM), KL-JM, KL-Abs [a]	initial	CACM (3204 docs, 64 queries)	English	MAP
Bennett et al. [6]	BM25, KL-Dir (LM), KL-JM, KL-Abs, KL-TS [a]	initial	TREC-8, WT10g, GOV	English	MAP
Lv&Zhai[57]	MLE (LM), RM, DMM, SMM, RMM [b]	PRF	AP88-89, Disks4&5, WT2g	English	MAP, P@10, recall
Yang&Fang[90]	BM25, BM25+ and variants, pivoted, PIV+ and variants, LM, F3LOG & variants, PL2, PL3, LGD, SPL, MATF	initial	Disks1&2, Disks4&5, WT2g, GOV2, Clueweb09, Clueweb12	English	MAP, ERR@20
Yang et al. [91]	BM25, LM, PL2, F2EXP, SPL	initial	Disks4&2, Disks4&5, WT10g, AQUAINT, GOV2, Cluewb09, Clueweb12	English	MAP, NDCG@20
Luk et al. [54]	VSM, 2-Poisson (BM11), Logistic regression, Pircs	initial & PRF	TREC-5, -6, -9, NTCIR II	Chinese	MAP
Amati et al. [1]	BM25, LM, DFR [c]	initial & PRF	CLEF 2003	French, Spanish, Italian	MAP, P@5, P@10
Savoy [78]	VSM, pivoted, Okapi (BM) Prosit (DFR)	initial & PRF	NTCIR-4	Chinese, Japanese, Korean, English	MAP
Khankasikam [41]	Boolean, VSM, BIM	initial	Thai database (5000 docs, 100 queries)	Thai	MAP
(b) Analytical comparison					
Turtle&Croft[84]	Boolean, VSM, Inference network	-	-	-	-
Losee [53]	Boolean, term independence	-	-	-	-
(c) Empirical and Analytical comparison					
Fang et al. [24]	pivoted, BM25, LM, PL2	initial	Disks4&5, Robust04, Robust05	English	MAP
This article	Ltw1, BM25, BM25+, LM, SPUD, F3LOG, PL2, PL3 LGD, SPL, IRRAc, Pivoted, PIV+, Machine generated, MATF, MVD	initial	Disks4&5, wt10g, GOV2, Clueweb09 Cat B	English	MAP, NDCG@20

Note: a. KL-Dir, KL-JM, KL-Abs, KL-TS denote KL divergence with Dirichlet, Jelinek-Mercer, Absolute discounting, and a Two-stage combination of Dirichlet and Jelinek-Mercer smoothing, respectively; b. MLE, RM, DMM, SMM, RMM denote Maximum Likelihood Estimate language model, Relevance model, Divergence minimization model, Simple mixture model and Regularized mixture model, respectively; c. DFR denote Divergence from randomness.

Dai et al. [17]. This work compares the Vector space model (VSM), BM25 model, the language model (LM) and the Indri inference network model. For LM, the KL-divergence formulation [46] was used, with several smoothing methods applied [94]. Retrieval experiments were performed on the CACM dataset of 3204 documents and a set of 64 queries. It was found that Indri model, which combines the language model and an inference network, gives the best performance in terms of MAP. However, a question is whether their conclusions are applicable to the newer and much larger collections (e.g. GOV2 or Clueweb09) that better reflect current and future application requirements.

Bennett et al. [6]. The empirical study of [6] compares the retrieval performance of the BM25 model and the LM using various smoothing methods [94], as well as a two-stage smoothing scheme [93] in which the document language model is first smoothed by a Dirichlet prior, followed by a further interpolation with a query background model using Jelinek-Mercer smoothing. It was found that for ad-hoc retrieval with short title queries, LM with Dirichlet smoothing outperformed the other models on both news and web collections. However in [6], only fixed sets of standard parameters are used for the various retrieval models, without training the parameters. Thus the study leaves the question whether the relative performance of the models changes when the models are appropriately trained on different test collections.

Yang and Fang [90]. Recently, [90] presented a comprehensive reproducibility study of retrieval models, by performing retrieval experiments with a large number of retrieval models that have been shown to be effective in the past, together with variants of these models. The retrieval models studied by [90] cover several major classes — BM25, Pivoted normalization [79], LM, DFR models [2] and Information-based models [11]. Furthermore, these retrieval models were compared on a wide range of standard TREC test collections, including the news and congressional records collection of Disks 4&5, the small webpage collection of WT2g, the GOV2 webpage collection for terabyte track, and the terabyte-sized Clueweb09 and Clueweb12 Web track collections.

Our comparative study presented in this article covers some established basic retrieval models as well as their best variants as found by [90]. With the view of current and future applications with large corpus in the terabyte regime, we select the retrieval model variants found to attain the best performance for Clueweb09 in [90]. Specifically, these include the BM25+ variant, the PIV+ variant of pivoted normalization, the F3LOG variant of LM and the PL3 variant of DFR. The information-based Log-logistic distribution (LGD) and and Smoothed power-law distribution (SPL) models [11] are also included in our study, as in [90].

Yang et al. [91]. The primary objective of [91] is to answer to the need of reproducible baselines in IR research. They demonstrated that the Anserini toolkit, built on Lucene, can fulfill this objective. In particular, they presented a reproducibility study of several retrieval models, including the BM25, LM, PL2 of the DFR framework, the F2EXP of the axiomatic approach [25] and the information-theoretical SPL model [11], by performing retrieval on several test collections. The reproduced results were shown to be comparable to those reported in the comparison research papers.

Lv and Zhai [57]. As pseudo-relevance feedback (PRF) is a well known approach to enhance retrieval effectiveness, various works have introduced PRF techniques in the language modeling framework. Most of these techniques attempt to improve the estimation of the query language models. Lv and Zhai [57] performed a comparative study of several PRF methods for estimating query language models. These include the Relevance model (RM) [49], Divergence minimization model (DMM) [92], Simple mixture model (SMM) [92] and Regularized mixture model (RMM) [82]. It was found that the RM3 variant of the Relevance model and the SMM performed well, with RM3 being more robust to the setting of feedback parameters.

While the established retrieval models were mostly developed with English documents, some past works have investigated the application of these retrieval models to non-English test collections and compared their performance. Overall, these works indicate that the effectiveness of the retrieval models and their ranking functions generally is not limited to any specific language of the queries and documents. We review some of these works as follows.

Luk and Kwok [54]. A comprehensive comparison of retrieval models for Chinese documents was provide by [54]. Retrieval experiments were performed on three Chinese test collections, using the vector space model (VSM) with various TF-IDF weights, the (2-Poisson) BM11 model, Logistic regression [13] and Pircs [45] models. As spaces do not delimit words in the Chinese language, various indexing strategies were also investigated in [54], such as character, word, bigram and Pircs indexing. The BM11 model and Pircs model were found to yield good retrieval effectiveness, while the tested variants of VSM were generally not effective.

Amati et al. [1]. This work studied the performance of several retrieval models on three CLEF 2003 monolingual test collections, namely the French, Spanish and Italian collections. The tested retrieval models are highly successful for English language retrieval — the BM25 model, language model (LM), and DFR model. Both an initial retrieval with the given queries and PRF retrieval with expanded queries were tested. With evaluation based on MAP, P@5 and P@10, the experiments showed that for all three collections, DFR yielded better performance over BM25 and LM. It was also shown that PRF was effective in enhancing the performance for BM25 and DFR, but detrimental to LM. The retrieval results supported the hypothesis that the effectiveness of a retrieval model is only moderately affected by the language of the queries and document collection.

Savoy [78]. The work of [78] compared a range of retrieval methods on the NTCIR-4 test collection, comprised of news documents in several languages — English, Chinese, Japanese and Korean. The tested retrieval methods include nine different term weights in the Vector Space model, as well as Okapi (BM) and DFR [2] models. For retrieval on all the different language documents, the experiments indicate that the common aspects in the effective retrieval models are document-length normalization of the term frequency and a term discriminative factor (such as the inverse document frequency).

Khankasikam [41]. A comparative study of retrieval models for Thai documents was performed by [41]. In this work, the traditional Boolean model, VSM, and the basic probabilistic Binary independence model (BIM) are used on a test collection of 5000 Thai documents and 100 queries. The retrieval experiments show that the traditional retrieval models are successfully applicable to Thai information retrieval. Moreover, the results are consistent with studies of English text retrieval in that the probabilistic model and vector space model perform better than the Boolean model.

2.1.2 Analytical comparison. Apart from empirical comparison of retrieval models, some works in the literature present theoretical comparisons of the models with an analysis of their formal properties or characteristics, in order to understand the differences in their retrieval performance.

Turtle and Croft [84]. An early analytic comparison of retrieval models was given by [84]. They compared three classes of retrieval models, namely the Boolean model, Vector space model and the probabilistic Inference network model. It was shown that the Boolean and VSM could be represented within the inference network model. Thus, the differences between the models can be viewed as differences in the way probabilities are estimated and combined in a probabilistic model.

Losee [53]. An analytic approach to the evaluation of the performance of retrieval models was developed by [53]. In this approach, retrieval models are compared in terms of the average search length (ASL), i.e. the position of the average relevant document in the ranked list of documents. Specifically, for a query that consists of two query terms, the ASL is calculated as a function of p, the probability that a particular term is present in a relevant document, and c, which represents the expected proportion of documents containing both terms. For example, the approach leads to the conclusion that models that take into account term dependence are generally superior to Boolean models. However, it is not clear whether the approach may be generalized to other retrieval needs, such as where high precision at the top n documents, or high recall are important.

2.2 Retrieval heuristics

While retrieval models may be formulated on diverse theoretical bases, most of the existing models share several common elements — first, the count of query term occurrences in documents (i.e. term frequency); second, a factor that gives larger weights to terms that are more discriminative, such as rare terms in the corpus; and third, document-length normalization, which avoids over-preference of long documents. Drawing insights from established retrieval models, Fang et al. [23] presented a list of six heuristic constraints on the functional form of an effective retrieval scoring function. These constraints mostly concern the above mentioned common elements of retrieval models. Subsequent works have restated these heuristics in analytic forms (e.g. [11; 28]) or introduced slight variations (e.g. [16]). Let S denotes a document ranking function. Following the nomenclature of [23] and using the notations of Table 2, the heuristics [23] and their analytic forms [28] are briefly described as follows:

TFC1. A larger count of query term in a document should correspond to a higher ranking score, thus $\partial S/\partial f(t,d) > 0$.

TFC2. The document score S should be a concave function of $f(t,d)$, i.e. $\partial^2 S/\partial f(t,d)^2 < 0$.

TDC. This constraint gives higher weight to more discriminative terms. Fang et al. [23] associate lower document frequency $df(t)$ with higher term discriminatory power. In this case, the TDC may be stated as $\partial S/\partial df(t) < 0$.

LNC1. The first length normalization constraint corresponds to assigning a lower score to a document that is longer solely due to having more non-query terms. This is consistent with the analytic form: $\partial S/\partial |d| < 0$.

LNC2. The second length normalization constraint adjusts the scoring function to avoid over-penalizing long documents. It states that if d_1 and d_2 are two documents such that $|d_1| = k \times |d_2|$ for some $k > 1$, and $f(t,d_1) = k \times f(t,d_2)$ for all terms t, then the score of d_1 should be at least as large as that of d_2, i.e. $S(q,d_1) \geq S(q,d_2)$. For example, LNC2 applies to the scenario where a document d_1 is formed by concatenating a document d_2 k times.

TF-LNC. Given a query q with a single word w, if d_1 and d_2 are two documents such that $f(w,d_1) > f(w,d_2)$ and $|d_1| = |d_2| + f(w,d_1) - f(w,d_2)$, then $S(q,d_1) > S(q,d_2)$. This constraint also avoids over-penalizing long documents, such as those generated by adding more query terms.

It is apparent that two assumptions underlie the above heuristics. First, term-independence is assumed, as the heuristics only deal with each distinct term independently from other terms. Second, query terms appearing in a document are taken as an indication of relevance, while terms not contained in the query are taken to be non-relevant. The second assumption, as reflected in the LCN1 and TF-LNC constraints, is not always true as it fails to deal with word mismatch (e.g. synonyms of query terms) or words with multiple meanings.

Various past works (e.g. [11; 28]) support the heuristic constraints of [23] to be good criteria for effective retrieval models. Fang and Zhai [25] introduced an axiomatic approach, in which established retrieval models, e.g. the BM25 or PL2, are modified such that the resulting ranking functions better satisfy certain heuristic constraints.

Table 2. Summary of notations

Symbol	Description
N	Number of documents in the collection
d	document d (typically regarded as a sequence of terms)
$\|d\|$	length of document d (total number of terms in d)
$\dot{d}$	binary document vector of the document d
$\|\dot{d}\|$	length of binary document vector $\dot{d}$ (number of distinct terms in d)
Φ	average number of distinct terms in a document
$\|\mathbb{C}\|$	total number of terms in the collection $\mathbb{C}$
q	query q (typically regarded as a sequence of terms)
$\|q\|$	total number of terms in the query q
t	query term t
$f(t,d)$	term frequency (number of occurrences) of term t in document d
$f(t,q)$	number of occurrences of term t in query q
$f(t,\mathbb{C})$	number of occurrences of term t in the collection $\mathbb{C}$
$f_{avg}(d)$	average term frequency in document d
$df(t)$	document frequency of term t (number of documents containing t)
$df_{\mathbb{C}}$	sum of document frequencies over all distinct terms in the collection
Δ	average document length
$\propto$	rank equivalent relation
$V(\cdot)$	function that returns the set of terms or the vocabulary of the argument

3 REVIEW OF RETRIEVAL MODELS COMPARED IN THIS STUDY

In order to provide a comprehensive comparison of term-independence retrieval models, we have implemented established and successful models covering various theoretical frameworks, as well as recent promising models shown to outperform the established models. The models tested in our comparison cover several categories: Vector Space (Section 3.1), BM25-based (Section 3.2), language modeling (LM) framework (Section 3.3), Divergence From Randomness (DFR) (Section 3.4), Information-based (Section 3.5), Divergence From Independence (DRI) (Section 3.6), pivoted normalization (Section 3.7), machine-generated (Section 3.8) and the promising MATF-like models (Section 3.9).

Table 3 summarizes some past record of the retrieval models in the literature. Some of the models attained the top performance in TREC, demonstrating the effectiveness of these models. As the BM25, LM and PL2 models are widely used as baselines in empirical studies in the literature, Table 3 indicates the models which have been shown to outperform these baselines, with some example references in the literature.

3.1 Vector space model with conventional TF-IDF term weights

The Vector Space Model (VSM) is one of the earliest retrieval models. In VSM, documents and queries are represented as vectors, with dimensions corresponding to distinct terms. The ranking function is given by the dot product of the document and query vectors. In each document vector, the numerical value of each dimension is typically specified by a term weight based on occurrence statistics of the corresponding term in the document or collection. In particular, term weights of the term frequency × inverse document frequency (TF-IDF) form were found to be more effective than a simple Boolean model that only considers the presence or absence of terms.

3.1.1 Ltw1. Several variants of the TF-IDF term weight were tested by Luk and Kwok [54] in VSM for Chinese retrieval. In particular, the variant called Ltw1 was found to obtain the best MAP values for some collections [54]. The Ltw1

Table 3. Past record of the retrieval models in the literature

Retrieval method		top TREC in MAP	used as	outperform (MAP/NDCG)			tested in collection	
Category	Model		baseline	BM25	LM	PL2	GOV2	Clueweb09
	Ltw1							
BM25-based	BM25	TREC-6, -7^{b}	✓				✓	✓
	BM25+			$\checkmark^{*c}$	$\checkmark^{c}$	$\checkmark^{c}$	✓	✓
Language model	LM		✓				✓	✓
	SPUD				$\checkmark^{*e}$		✓	
	F3LOG			$\checkmark^{f}$	$\checkmark^{*f}$	$\checkmark^{f}$	✓	✓
DFR	PL2		✓				✓	✓
	PL3			$\checkmark^{f}$	$\checkmark^{f}$	$\checkmark^{f}$	✓	✓
Information-based	LGD			$\checkmark^{f}$	$\checkmark^{f}$	$\checkmark^{f}$	✓	✓
	SPL			$\checkmark^{f}$	$\checkmark^{f}$	$\checkmark^{f}$	✓	✓
DFI	IRRAc	Web track 2012^{g}						✓
pivoted	pivoted		✓	$\checkmark^{f}$	$\checkmark^{f}$	$\checkmark^{f}$	✓	✓
	PIV+			$\checkmark^{f}$	$\checkmark^{f}$	$\checkmark^{f}$	✓	✓
Machine generated	Gos1			$\checkmark^{*h}$	$\checkmark^{*h}$		✓	
	Gos3				$\checkmark^{h}$		✓	
MATF-like	MATF		✓	$\checkmark^{*i}$	$\checkmark^{*i}$	$\checkmark^{*i}$	✓	✓
	MVD			$\checkmark^{*j}$	$\checkmark^{*j}$	$\checkmark^{*j}$		✓

Note: a. Luk et al. [54] ; b. Robertson et al. [71]; c. Lv & Zhai [58], Yang & Fang [90]; d. Lavrenko & Croft [49]; e. Cummins et al. [16]; f. Yang & Fang [90]; g. Dincer [20], Clarke et al. [10]; h. Goswami et al. [28] ; i. Paik [64]; j. Paik [65]. The better performance (in terms of MAP/NDCG) over the BM25/LM/PL2 baseline is based on at least one of the tested collections, and the * symbol indicates statistical significance.

scoring function is as follows:

$$S_{Ltw1}(q,d) = \sum_{t \in V(q)} f(t,q) \times \log(f(t,d)+1) \times \log\left(\frac{N}{df(t)}+1\right), \tag{1}$$

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

While better MAP performance could be obtained for some collections using the Ltw1 function with an extra $1/|d|$ weighting factor, we have found in our study that the Ltw1 function without the $1/|d|$ factor obtains better performance for Clueweb09. Some past work (e.g. Lee et al. [51]) showed that cosine similarity (i.e. dividing the dot product score by Euclidean lengths of the document and query) is more effective. However, we also found in our study that the Ltw1 of Eq. (1) performs better than the cosine similarity on Clueweb09. Hence, we only report Ltw1 in the comparison in this article.

3.2 BM25-based

The BM25 is a well established retrieval model that estimates the probability of relevance of documents [72; 73; 75]. Following the probability ranking principle (PRP) of Robertson [74], it was further shown in [19] that if the probability of relevance is estimated as accurately as possible, then the retrieval will yield optimal effectiveness in a range of measures, such as MAP, precision at top-n documents (P@n) and R-precision.

3.2.1 BM25. In this approach term frequencies within a document are modeled by a mixture of two Poisson distributions [72; 73; 75]. For a query q the BM25 weight of a query term $t \in V(q)$ in a document d has a TF-IDF form. The BM25 document ranking is obtained by summing the weights of query terms as follows:

$$S_{BM}(q,d) = \sum_{t \in V(q)} \frac{(k+1)f(t,d)}{f(t,d) + k\left(1 - b + b\frac{|d|}{\Delta}\right)} \cdot \log\left(\frac{N - df(t) + 0.5}{df(t) + 0.5}\right) \cdot \frac{(k'+1)f(t,q)}{k' + f(t,q)}, \tag{2}$$

where k, k' and b are constants, $f(t,d)$ and $f(t,q)$ are the term frequency of t in the document d and query q respectively, $df(t)$ is the document frequency of t, $|d|$ is the L1 norm (or city-block) length of d (i.e. the total number of terms), Δ is the average document length in the collection, and N is the number of documents in the collection. 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 [62; 79]. 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 [75]. This functional form was subsequently formalized as pivoted normalization [79] to improve over a simple $|d|$ normalization which over-penalizes long documents.

3.2.2 BM25+. Numerous variants of the BM25 model have been studied in the past for possible performance improvements [83]. Some of these variants are designed according to an axiomatic approach [25], whereby a base model such as BM25 is modified to better satisfy heuristic constraints. For example, the BM25+ ranking function was introduced by Lv and Zhai [58] to deal with the issue of over penalizing long documents in BM25. Specifically, a constant is added to the TF component to act as a lower-bound contribution for query terms even with a single occurrence and no matter how long the document is. In the work of Yang and Fang [90], BM25+ was found to improve over BM25 and yield the best effectiveness among several BM25 variants tested on Clueweb09. Therefore, this variant is included the current study. The ranking function of BM25+, S_{BM+} is given by:

$$S_{BM+}(q,d) = \sum_{t \in V(q)} \left[\frac{(k_+ + 1)f(t,d)}{f(t,d) + k_+\left(1 - b_+ + b_+\frac{|d|}{\Delta}\right)} + \delta_+\right] \cdot \log\left(\frac{N+1}{df(t)}\right) \cdot \frac{(k'_+ + 1)f(t,q)}{k'_+ + f(t,q)}, \tag{3}$$

where k_+, k'_+, b_+ and δ_+ are constants.

3.3 Language modeling (LM) framework

Over the past couple of decades, the LM approach of Ponte and Croft [67] and Hiemstra [35] has been established as a successful retrieval framework. The Kullback-Leiber (KL) divergence formulation with Dirichlet smoothing [46; 94] has been shown to be a highly effective model in the LM framework and is widely used as a baseline in empirical studies in the literature. Therefore we include the KL formulation with Dirichlet smoothing in our current comparative study.

3.3.1 Language model. A language model is a statistical distribution of terms from which texts can be generated. In the query-likelihood retrieval model [67], documents are ranked according to the probability of generating the query terms by the unigram language model estimated for each given document. The approach of [46] views retrieval as a risk minimization problem, with documents being ranked by the minimal KL divergence between the query language model and document language model. One way to estimate the language models is by maximum likelihood (ML), given by relative counts of words. For the ML estimate, the probability of a query term t generated by the language model of document d is of the form $f(t,d)/|d|$. Smoothing is generally employed to adjust the ML estimator, such as assigning a non-zero probability to unseen words, so that the query likelihood given the language model of document d is:

$$p(t \in V(q)|d) = \lambda(d)\frac{f(t,d)}{|d|} + (1 - \lambda(d))\frac{f(t,\mathbb{C})}{|\mathbb{C}|}, \tag{4}$$

with the second component being the smoothing, where $f(t,\mathbb{C})$ is the total term frequency of t in the collection, $|\mathbb{C}|$ is the size (total number of words) of the collection, and $\lambda(d)$ is in general a document-dependent smoothing parameter with value between 0 and 1. A popular method is Dirichlet smoothing [94], which is found to be effective for keyword queries [94], as used in our study. With Dirichlet smoothing, the parameter $\lambda(d)$ of Eq. (4) is $\lambda(d) = |d|/(|d| + \mu)$, where the Dirichlet prior μ is a positive number. In the KL-divergence formulation, the ranking of document d for a query q, $S_{LM}(q,d)$, is given by a sum over query terms $t \in V(q)$:

$$S_{LM}(q,d) = -\sum_{t \in V(q)} p(t|q) \log \frac{p(t|q)}{p(t|d)}, \tag{5a}$$

where

$$p(t|d) = \frac{f(t,d) + \mu\, f(t,\mathbb{C})/|\mathbb{C}|}{|d| + \mu}, \tag{5b}$$

with Eq. (5b) being derived from Eq. (4) by applying Dirichlet smoothing. Together with a ML estimate of the query language model, scoring by Eq. (5a) is rank equivalent to the following:

$$S_{LM}(q,d) \propto \sum_{t \in V(q)} \frac{f(t,q)}{|q|} \log p(t|d), \tag{5c}$$

where $\propto$ is the rank equivalence relation. Comparing with retrieval models that have a TF-IDF ranking function, one notable feature of the LM framework is the apparent absence of an explicit IDF component. However, it has been pointed out that smoothing with a collection-based distribution plays a role similar to the IDF (e.g. [94]).

It has been shown that the query-likelihood can be derived from the log-odds ratio of a document being drawn from the relevant class compared to the non-relevant class [4; 47; 55]. Furthermore as shown by Luk [56], the log-odds ratio is rank equivalent to the probability of relevance. Therefore, the ranking of documents in the query-likelihood language model conforms to the PRP.

3.3.2 Smoothed Pólya Urn Document (SPUD). Because of the success of the LM, various techniques have been developed in the LM framework for further enhancement of performance effectiveness. Recently, a SPUD language model based on the Dirichlet compound multinomial distribution that captures word burstiness, i.e. the tendency of a term to repeat itself in a document, was introduced by [16] and found to be effective, outperforming the traditional LM with Dirichlet smoothing.

The document ranking function of a SPUD model that applies Dirichlet smoothing, $S_{SPUD}(q,d)$, can be written as:

$$S_{SPUD}(q,d) = \sum_{t \in V(q)} \frac{f(t,q)}{|q|} \log p(t|d), \tag{6a}$$

where

$$p(t|d) = \frac{1}{\mu_s|\dot{d}| + 1}\left[\mu_s|\dot{d}|\frac{f(t,d)}{|d|} + \frac{df(t)}{df_{\mathbb{C}}}\right], \tag{6b}$$

with μ_s being a positive constant and $|\dot{d}|$ is the number of distinct terms in d. A difference between the SPUD model and the traditional smoothed LM of Eq. (5a-c) is that the smoothing factor in SPUD is based on $|\dot{d}|$, which is the number of distinct terms in a document, instead of the document length $|d|$. Moreover, the $f(t,\mathbb{C})/|\mathbb{C}|$ component of Eq. (5b) is replaced by $df(t)/df_{\mathbb{C}}$, with $df(t)$ being the document frequency of the term t and $df_{\mathbb{C}}$ is the sum of document frequencies over all distinct terms in the collection.

3.3.3 F3LOG. Fang and Zhai [25] introduced an axiomatic approach to seek effective retrieval functions by varying some established models such as the pivoted normalization, BM25 and LM, such that the variants better satisfy the heuristic constraints of [23]. The derived functions were found to be more stable than the existing retrieval functions with comparable performance [25]. Yang and Fang [90] tested the F3EXP and F3LOG functions that are obtained by applying the axiomatic approach to the LM framework, and found F3LOG to perform well on the Clueweb09 collection. Therefore, we include the F3LOG ranking function in the current study. The ranking function of F3LOG, $S_{F3LOG}(q,d)$, is given by:

$$S_{F3LOG}(q,d) = \sum_{t \in V(q)} (1 + \log(1 + \log(f(t,d)))) \cdot \log\left(\frac{N+1}{df(t)}\right) - \frac{(|d| - |q|) \cdot |q| \cdot s}{\Delta}, \tag{7}$$

where s is a constant.

3.4 Divergence from randomness (DFR)

Amati and van Rijbergen [2] introduced the DFR framework of probabilistic IR models. 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 term frequency, whereby two types of normalization are involved. In general, a DFR term weight of a term t in document d has the form $w(t,d) = Inf_1(t,d) \cdot Inf_2(t,d)$. The first component Inf_1 is the informative content specified by $Inf_1(t,d) = -\log Prob_1(t,d)$, where $Prob_1$ is a probability distribution of term occurrence. The second component is a correction for the risk of accepting the term t as a good descriptor of the document, which is written in terms of a second probability: $Inf_2(t,d) = 1 - Prob_2(t,d)$. The probability $Prob_2(t,d)$ models the aftereffect of sampling: the greater the term frequency of a term in a document, the more it suggests the term to be contributing to discriminate the document. The component Inf_2 represents the first normalization of DFR. A second normalization is then applied, corresponding to a normalization of term frequencies by the document length, such as described by Eq. (8d) below.

3.4.1 PL2. Past works (e.g. [32; 34] and [24]) have found the PL2 instantiation of the DFR framework to be effective. The PL2 model is often used as a retrieval baseline in the literature. Furthermore, it serves as an example of an highly effective probabilistic retrieval model that is not derived based on the PRP.

In the PL2 model, the random distribution used is an approximation to the Poisson distribution (leading to Eq. (8b)), while the ranking score is normalized based on Laplace's law of succession (Eq. (8c)). As for the second term frequency normalization, the 'normalization 2' [2] (Eq. (8d)) is applied, with the assumption that term frequency density is a decreasing function of the document length. The PL2 ranking score, $S_{PL2}(q,d)$, is given by a sum of the weight over the query terms $t \in V(q)$:

$$S_{PL2}(q,d) = \sum_{t \in V(q)} Inf_1(t,d) \cdot Inf_2(t,d) \tag{8a}$$

where

$$Inf_1(t,d) = tfn \cdot \log_2 \frac{tfn}{\lambda} + \left(\lambda + \frac{1}{12 \cdot tfn} - tfn\right) \cdot \log_2 e + 0.5 \cdot \log_2(2\pi \cdot tfn), \tag{8b}$$

$$Inf_2(t,d) = \frac{1}{tfn+1}, \tag{8c}$$

with $\lambda = f(t,\mathbb{C})/N$ = frequency of t in the collection $\mathbb{C}$ / number of documents in the collection, and tfn is a normalization of the term frequency $f(t,d)$ according to:

$$tfn = f(t,d) \cdot \log_2 \left(1 + c\frac{\Delta}{|d|}\right) \tag{8d}$$

where c is a positive constant.

3.4.2 PL3. Because of the success of smoothing techniques applying Dirichlet priors in the language modeling framework, He and Ounis [33] studied the use of Dirichlet priors to the term frequency normalization of other established models, such as BM25 and PL2. The new models with Dirichlet priors normalization were shown to be effective and robust over diverse TREC collections [33]. In particular, the PL3 model is obtained by replacing the TF normalization of PL2 with the Dirichlet prior TF normalization. This variant was tested by Yang and Fang [90] and was found to outperform PL2 on Clueweb09. Therefore it is of interest to compare PL3 with the other retrieval models in this study.

The ranking function of PL3, $S_{PL3}(q,d)$, is the same as the PL2 (Eq. (8)a-c), except the normalized term frequency tfn is replaced by the Dirichlet priors normalization:

$$S_{PL3}(q,d) = \sum_{t \in V(q)} Inf_1(t,d) \cdot Inf_2(t,d) \tag{9a}$$

where

$$Inf_1(t,d) = tfn \cdot \log_2 \frac{tfn}{\lambda} + \left(\lambda + \frac{1}{12 \cdot tfn} - tfn\right) \cdot \log_2 e + 0.5 \cdot \log_2(2\pi \cdot tfn), \tag{9b}$$

$$Inf_2(t,d) = \frac{1}{tfn+1}, \tag{9c}$$

with $\lambda = f(t,\mathbb{C})/N$ = frequency of t in the collection $\mathbb{C}$ / number of documents in the collection, and tfn is a normalization of the term frequency according to:

$$tfn = \frac{f(t,d) + \mu_p \frac{f(t,\mathbb{C})}{|\mathbb{C}|}}{|d| + \mu_p} \cdot \mu_p \tag{9d}$$

where μ_p is a positive constant, $f(t,d)$ and $f(t,\mathbb{C})$ are the occurrence frequency of term t in document d and collection $\mathbb{C}$ respectively, while $|d|$ is the L1 norm (or city-block) document length of d and $|\mathbb{C}|$ is the total number of terms in $\mathbb{C}$.

3.5 Information-based models

Information-based retrieval models are based on the hypothesis that the difference in behavior of a word at the document and collection levels, given by the corresponding probability distribution of the occurrence of the word, indicates the significance of the word in the document [11]. This hypothesis has been shown to work well in the 2-Poisson mixture models, in BM25 and also the DFR models. Clinchant and Gaussier [11] introduced a family of information-based models with retrieval functions that satisfy the heuristic constraints proposed by Fang et al. [23]. In particular, [11] modeled word burstiness (i.e. tendency of a word to repeat itself, once it appears in a document). They developed the mathematical condition for a term frequency distribution to be 'bursty', and showed that such condition is satisfied by certain power-law distributions. Two different bursty power-law distributions were studied in [11]: the Log-logistic distribution (LGD) and Smoothed power-law (SPL) distribution. It is of interest to include information-based models in our comparative study because just as the DFR model, this class of retrieval models perform well [11; 20; 21], while they are not derived from the PRP.

3.5.1 Log-logistic distribution (LGD). The first power-law distribution used by Clinchant and Gaussier [11] to model burstiness is the log-logistic distribution (LGD). The corresponding scoring function S_{LGD} is given by:

$$S_{LGD}(q,d) = \sum_{t\in V(q)} -f(t,q)\log\left(\frac{\lambda_t^{\beta_l}}{f_l(t,d)^{\beta_l}+\lambda_t^{\beta_l}}\right), \tag{10a}$$

where

$$\lambda_t = df(t)/N, \tag{10b}$$

$$f_l(t,d) = f(t,d)\times\log\left(1+c_l\frac{\Delta}{|d|}\right), \tag{10c}$$

with c_l and β_l being positive constants, and $|d|$ is the length of document d and Δ is the average document length in the collection. The scoring function of Eq. (10a) corresponds to the mean information a document brings to a query. The term frequency normalization in Eq. (10c) employs the 'normalization 2' form of the DFR framework (Eq. (8d)). By using this normalization, the LGD model of Eq. (10a) satisfies all the heuristics constraints [11]. The model based on the LGD performs comparably or better than the strong BM25, language model with Jelinek-Mercer or Dirichlet smoothing, and PL2 baselines in several collections [11].

3.5.2 Smoothed power-law (SPL) distribution. The second power-law distribution studied by [11] to model burstiness is the SPL, which was also shown to be effective. The scoring function of this model, $S_{SPL}(q,d)$, is given by:

$$S_{SPL}(q,d) = \sum_{t\in V(q)} -f(t,q)\log\left(\frac{\lambda_t^{\frac{f_s(t,d)}{f_s(t,d)+1}}-\lambda_t}{1-\lambda_t}\right), \tag{11a}$$

where

$$\lambda_t = df(t)/N, \tag{11b}$$

$$f_s(t,d) = f(t,d)\times\log\left(1+c_s\frac{\Delta}{|d|}\right), \tag{11c}$$

with c_s being a positive constant, and $|d|$ is the length of document d and Δ is the average document length in the collection. Similar to the LGD model, the term frequency normalization (Eq. (11c)) employs the DFR form. With the use of this normalization form, the SPL model satisfies all the heuristic constraints [11].

3.6 Divergence from Independence (DFI)

An approach related to the successful DFR models [2] is the DFI framework introduced by Dinçer and co-workers [20; 21; 43]. In DFR, a speciality word is indicated by its term occurrence distribution diverging from a basic randomness model, which may be given by some probability density function such as Poisson, Hyper-geometric, Bose-Einstein, etc. [2]. In DFI, the occurrence distribution of a speciality word diverges from an independence model instead of a randomness model. The DFI hypothesis is that under independence, the expected frequency of a term t in document d is given by $e(t,d) = f(t,\mathbb{C})\times|d|/|\mathbb{C}|$ (Eq. (12c)), i.e. the occurrence of the term t is distributed in the documents proportionally to the length of the document. Thus DFI is the 'non-parametric counterpart' of DFR, in the sense that the $e(t,d)$ distribution does not contain any parameter, such as the mean or variance that specify the Poisson and other randomness distributions used in DFR.

3.6.1 IRRAc. An instantiation of the DRI approach, named IRRAc in this article, was shown to be effective for tasks that require high recall and high precision, especially with short queries of a few words [20]. The IRRAc method was found to be among the top performers in the 2012 TREC Web track on the Clueweb09 collection [10]. The scoring function of IRRAc, S_{IRRAc}, is given as follows:

$$S_{IRRAc}(q,d) = \sum_{t \in V(q)} f(t,q) \times \Delta(t,d) \times \Lambda(t,d), \tag{12a}$$

where

$$\Delta(t,d) = \left[(f(t,d)+1) \times \log_2 \left(\frac{f(t,d)+1}{\sqrt{e^{+}(t,d)}} \right) \right] - \left[f(t,d) \times \log_2 \left(\frac{f(t,d)}{\sqrt{e(t,d)}} \right) \right], \tag{12b}$$

$$e(t,d) = \frac{f(t,\mathbb{C}) \times |d|}{|\mathbb{C}|}, \tag{12c}$$

$$e^{+}(t,d) = \frac{(f(t,\mathbb{C})+1) \times (|d|+1)}{|\mathbb{C}|+1}, \tag{12d}$$

$$\Lambda(t,d) = \left(\frac{|d| - f(t,d)}{|d|} \right)^{a_i} \times \left(\frac{2}{3} \times \frac{f(t,d)+1}{f(t,d)} \right)^{b_i}, \tag{12e}$$

where a_i and b_i are non-negative constants. The main term weighting component in IRRAc is given by $\Delta(t,d)$ of Eq. (12b), which is the amount of increase in total information by observing term t one more time (i.e. $f(t,d)+1$), given that it occurs $f(t,d)$ times in document d [20]. In Eq. (12c), $e(t,d)$ is the expected frequency of term t in document d under independence. The factor $\Lambda(t,d)$ in Eq. (12e) is a correction applied to deal with spam data in Clueweb09 [20]. This factor consists of a component that promotes documents in which a query term does not fill up the whole document, and a component that favors high term frequency irrespective of the document length. In the original model catered for Clueweb09, the exponents a_i and b_i in Eq. (12e) are fixed (3/4 and 1/4 respectively) [20]. Here, we generalize the model to non-negative values of the exponents to cater for other test collections.

3.7 Pivoted normalization

Within the Vector Space Model, document length normalization has long been known to be an important component in the TF-IDF weight to avoid over-preference of long documents. Singhal et al. [79] introduced a pivoted normalization technique as a correction to common normalization schemes, such as cosine normalization, in order to obtain a weighting function that better matches the probability of relevance of the retrieved documents.

3.7.1 Pivoted unique normalization. While various instantiations of the pivoted normalization technique was tested by Singhal et al. [79], the pivoted unique normalization scheme was found to obtain the best retrieval effectiveness. Thus in this work, we study the scoring function $S_{piv}(q,d)$ for pivoted unique normalization following [79]:

$$S_{piv}(q,d) = \sum_{t \in V(q)} f(t,q) \frac{\frac{1+\log(f(t,d))}{1+\log(f_{avg}(d))}}{1 - b_p + b_p \frac{|\dot{d}|}{\Phi}} \log \left(\frac{N - df(t) + 0.5}{df(t) + 0.5} \right), \tag{13}$$

where b_p is a constant, and $f_{avg}(d)$ is the average term frequency in document d, i.e. $f_{avg}(d) = |d|/|\dot{d}|$, with $|d|$ being the length of the document d and $|\dot{d}|$ being the number of distinct terms in d. Also in Eq. (13), Φ is the average number of distinct terms in a document computed across the collection. This approach employs a relaxed term frequency normalization (Eq. (13)) to avoid over-penalizing long documents as in the case of a simple normalization by the document length $|d|$.

3.7.2 *PIV+.*

Similar to the BM25+ approach (Section 3.2.2), a variant of the pivoted function may be obtained that takes into account proper lower-bounding of the term frequency normalization. In particular, the PIV+ variant is obtained by adding a lower bound to the TF normalization component of the pivoted normalization function [58]. In this comparative study, we include the PIV+ scoring function $S_{p+}(q,d)$, following the version studied by Lv and Zhai [58] and Yang et al. [90]:

$$S_{p+}(q,d) = \sum_{t \in V(q)} f(t,q) \left[\frac{1 + \log(1 + \log(f(t,d)))}{(1 - b_{p+}) + b_{p+} \cdot \frac{|d|}{\Delta}} + \delta_{p+} \right] \log\left(\frac{N+1}{df(t)}\right), \tag{14}$$

where b_p and δ_{p+} are constants.

3.8 Machine generated models

Goswami et al. [28] attempted to find effective retrieval scoring functions by a systematic exploration of the space of all possible scoring functions. They defined a grammar to generate possible scoring functions up to a certain length, with the length being related to the number of elements, variables and operations in a function. Specifically the variables considered are normalized term frequency and normalized document frequency, while the choices of operations include various binary and unary operators [28]. Heuristic constraints [11; 23] are employed to prune the machine generated functions, keeping only those that satisfy the constraints. The effectiveness of the remaining candidate functions are then tested empirically to find the top performing ones over several test collections. The top generated scoring functions are found to perform better than or comparable to established models, including BM25, LM and the Log-logistic (LGD) model [11]. In particular, the scoring functions Number 1 and Number 3 reported in the Conclusion section of [28] are identified as the top performing functions in terms of MAP. In this article, these two functions are denoted as Gos1 and Gos3, respectively. Similar to the Information-based models (Section 3.5), both of these functions employ the 'normalization 2' for the term frequency that is adopted from the DFR framework [2].

3.8.1 *Gos1 scoring function.*

It was shown that Gos1 outperforms the baselines of LM, BM25 and LGD, and is more robust than the baselines over different collections. The scoring function $S_{Gos1}(q,d)$, is given by:

$$S_{Gos1}(q,d) = \sum_{t \in V(q)} f(t,q) e^{\sqrt{\log\left(\frac{f_{G1}(t,d)+\lambda_t}{\lambda_t}\right)}}, \tag{15a}$$

where

$$\lambda_t = df(t)/N, \tag{15b}$$

$$f_{G1}(t,d) = f(t,d) \times \log\left(1 + c_{g1}\frac{\Delta}{|d|}\right), \tag{15c}$$

with c_{g1} being a positive constant, and $|d|$ is the length of document d and Δ is the average document length in the collection.

3.8.2 Gos3 scoring function. The scoring function $S_{Gos3}(q,d)$ is given by:

$$S_{Gos3}(q,d) = \sum_{t \in V(q)} f(t,q)\sqrt{\frac{1}{\sqrt{\lambda_t}} \log\left(f_{G3}(t,d)+1\right)}, \tag{16a}$$

where

$$\lambda_t = df(t)/N, \tag{16b}$$

$$f_{G3}(t,d) = f(t,d) \times \log\left(1 + c_{g3}\frac{\Delta}{|d|}\right), \tag{16c}$$

with c_{g3} being a positive constant, and $|d|$ is the length of document d and Δ is the average document length in the collection.

3.9 MATF-like models

We have also included in our comparative study some other recent models that have been shown to outperform strong baselines such as BM25, LM and PL2. In particular, the Multi-Aspect TF (MATF) model that incorporates two types of term frequency normalization was found to be effective [64]. The MATF approach was subsequently extended to obtain the Maximum Value Distribution (MVD) model [65].

3.9.1 MATF. Paik [64] introduced the MATF model, which is a Vector Space Model formulation with a novel TF-IDF term weighting scheme. The TF-IDF weighting of MATF employs two different within document TF normalizations, whereby one component of the term frequency favors long documents (the Relative Intra-document TF, RITF in Eq. (17c) below), while the other favors short documents (the Length Regularized TF, LRTF in Eq. (17d)). The relative weighting of the two TF components is determined based on the query length (Eq. (17b)). The use of query length in an adaptive term weighting scheme has been studied in some past work [9].

The MATF scoring function $S_{MATF}(q,d)$, is given by:

$$S_{MATF}(q,d) = \sum_{t \in V(q)} TFF(t,d) \cdot TDF(t), \tag{17a}$$

where

$$TFF(t,d) = \alpha_{MATF} \cdot g(RITF(t,d)) + (1-\alpha_{MATF}) \cdot g(LRTF(t,d)), \tag{17b}$$

$$RITF(t,d) = \frac{\log(1+f(t,d))}{\log(1+f_{avg}(d))}, \tag{17c}$$

$$LRTF(t,d) = f(t,d) \times \log_2\left(1+\frac{\Delta}{|d|}\right), \tag{17d}$$

$$TDF(t) = g\left(\frac{f(t,\mathbb{C})}{df(t)}\right) \times \log\left(\frac{N+1}{df(t)}\right), \tag{17e}$$

with the function $g(x) = x/(1+x)$, and α_{MATF} in Eq. (17b) is set to be $\alpha_{MATF} = 2/(1+\log_2(1+|q|))$. The RITF factor given by Eq. (17c) signifies a relative term frequency, with $f_{avg}(d)$ being the average term frequency in document d, i.e. $f_{avg}(d) = |d|/|\dot{d}|$, where $|d|$ is the length of the document d and $|\dot{d}|$ is the number of distinct terms in d. The RITF factor is also used in the pivoted unique normalization (Eq. 13). The component $LRTF(t,d)$ of Eq. (17d) employs the 'normalization 2' from the DFR framework (Eq. (8)d). The use of the function $g(x)$ for the two term frequency components in Eq. (17b) ensures that the scoring function conforms to desired heuristic conditions — the properties $g(x)' > 0$ and $g(x)'' < 0$ correspond to the TFC1 and TFC2 conditions, respectively. The component $TDF(t)$ of Eq. (17a) is a term discriminatory factor as given by Eq. (17e). Note that MATF is the retrieval function called NTFIDF tested in [90].

3.9.2 MVD. The Maximum Value Distribution (MVD) model introduced by Paik [65] utilizes the multi-aspect within document term frequency normalization shown to be effective in the MATF model [64]. MVD integrates the multi-aspect tf normalization into a probabilistic framework. Specifically, the MVD term weight is proportional to the probability that the normalized frequency of the term is maximum with respect to its distribution in the set of documents containing the term.

The scoring function $S_{MVD}(q,d)$ of the MVD model is derived to be [65]:

$$S_{MVD}(q,d) = \sum_{t \in V(q)} TFF(t,d) \cdot \log\left(\frac{N+1}{df(t)}\right), \tag{18a}$$

where

$$TFF(t,d) = \alpha_M \cdot G(RITF_k(t,d)) + (1-\alpha_M) \cdot G(LRTF(t,d)), \tag{18b}$$

$$RITF_k(t,d) = \frac{\log(1+f(t,d))}{\log(k_M + f_{avg}(d))}, \tag{18c}$$

$$LRTF(t,d) = f(t,d) \times \log_2\left(1 + \frac{\Delta}{|d|}\right), \tag{18d}$$

and the function G(x) is a mixture of Gumbel $F_g(x)$ and Frechet $F_f(x)$ distributions:

$$G(x) = p_M \cdot F_g(x) + (1-p_M) \cdot F_f(x), \tag{18e}$$

$$F_g(x) = \exp\left(-\exp(-x/\alpha_g)\right), \tag{18f}$$

$$F_f(x) = \exp\left(-\left(\frac{\mu_f}{x}\right)^{\alpha_f}\right). \tag{18g}$$

In Eq. (18b), the component $TFF(t,d)$ is an extension of the MATF weight (Eq. (17b)) to the MVD model, with the mixing parameters α_M of Eq. (18b) being in the range $0 < \alpha_M < 1$. In Eq. (18e), the Gumbel $F_g(x)$ and Frechet $F_f(x)$ functions represent approximations to maximum value models for rare and common terms, respectively. The mixture parameter p_M in Eq. (18e) is specified in terms of the *idf* of the term, with $p_M = \beta_M \cdot idf/(1+\beta_M \cdot idf)$, where β_M is a free parameter such that $\beta_M > 0$. The parameters α_g, α_f and μ_f of Eq. (18f) and (18g) are estimated from the term frequency distributions as described by [65].

3.10 Retrieval models not included in the current comparison

Some notable models that are not tested in this study include the following: (1) Boolean model because it does not provide ranking of the documents as do all the other tested retrieval models; (2) Extended Boolean model (e.g. [66]), as its retrieval effectiveness depends on the definition of the term weights. Hence we do not examine this model until we have found out which term weights are effective in this work; (3) The logistic regression and Pircs models [54]: While these were effective in the study of Luk and Kwok [54] for Chinese retrieval, we find that our implementations of these model perform poorly for the Clueweb09 test collection in our preliminary study; (4) Fuzzy retrieval models such as MMM [85] and Paice [63] models, in which the ranking scores associated with the individual query terms are combined by fuzzy Ordered Weighted Averaging. The score due to each query term first needs to be determined by an appropriate retrieval model, such as those studied in the current work. The use of Fuzzy retrieval models will be examined in future studies; (5) Hybrid models, such as the LDA-BM25, LDA-MATF and LDA-LM models [38; 39], which involve a mixing of a Latent Dirichlet Allocation (LDA) topic modeling component with traditional term-based retrieval models. These hybrid models are not included here because good performance of the traditional base retrieval model implies good performance of the hybrid model, and we focus on comparing the base models in this study; (6) Metasearch techniques, such as Condorcet-fuse [61] and Reciprocal Rank Fusion [14], which combine the retrieval results obtained by different document ranking functions. Like hybrid models, the effectiveness of these metasearch techniques depends on the effectiveness of the base models; (7) Some recent BM25 variants, such as the BM25_{QL} models [37] that introduce a query-length dependence to the term frequency normalization in the traditional BM25. The BM25_{QL} variants are not included in this study because only small differences in performance from the optimized traditional BM25 are observed across the tested datasets (with relative differences less than 1%) [37]. For the largest dataset WT10g tested in [37], the optimized traditional BM25 yielded a MAP value of 0.2050 (similar to our result of 0.2084 (Table 6a)), which was actually higher than the values obtained by the various instances of $\text{BM25}_{QL}(0.1995 - 0.2029)$; (8) neural retrieval models [30], e.g. the deep relevance matching model (DRMM) [29], word-embedding based methods such as the Fisher Vector approach [12], dual embedding space model (DESM) [60], generalized language model [27], and kernel based neural ranking model (k-NRM) [88]. Neural retrieval models generally require external data for training and/or pre-training of word-embedding representations, while the retrieval models described in Sections 3.1 to 3.9 do not require an external corpus. For non-neural retrieval models, an external corpus like Wikipedia may be utilized, for example, for query expansion via pseudo-relevance feedback or otherwise, e.g. [52; 89]. A separate paper is needed for a comparison between such techniques and neural retrieval models, while in this paper we focus on approaches that do not require an external corpus.

4 EXPERIMENTS

This section presents our experimental details. The experimental environment and our training and testing methodology are described in Section 4.1. The procedure of hypothesis testing used in this work is discussed in Section 4.2. Last, the results of our retrieval experiments are presented in Section 4.3. The results allow a comparison of retrieval models within a collection (Section 4.3.1) and across collections (4.3.2). We also compare our results with those in the literature as a reproducibility study (Section 4.3.3).

4.1 Setup and methodology

We compare the effectiveness of retrieval models implemented in our own retrieval system, which was first developed in the 2000's for the purpose of investigating topics like passage-based retrieval [44] or context-based retrieval [86], as such functionalities are not generally supported in open-source systems. Our system will enable our future work such as a study of the effectiveness of the retrieval models in passage-based retrieval with pseudo-relevance feedback.

The TREC test collections used in our experiments are summarized in Table 4. For diversity, both news and webpage collections are used. To test whether the effectiveness of retrieval methods is maintained for different dataset sizes, the collections used in this study span a wide range of sizes, from about 3GB for Disks 4&5 to the terabyte-sized Clueweb09 Category B subset of English-language pages. The Clueweb09 collection contains some spam data. A procedure for removing spam documents was presented by Cormack et al. [15]. However, there appears no standard setting in the literature in this regard — while spam filtering is not applied in some work (e.g. [90; 91]), in the cases where it is applied, there are variations in the threshold used (e.g. [36; 70; 87]). In our comparison of retrieval models reported in this article, spam filtering is not applied on Clueweb09. The performance of the retrieval models is evaluated in terms of MAP, which takes into account both precision and recall, and the precision-oriented NDCG@20. The MAP measure based on

Table 4. Summary of TREC collections

Type	news/federal register; congressional records (CR)	news/federal register			
Dataset	Disks 4&5	Disks 4&5 - CR			
N	556,075	528,153			
Size (GB)	3.27	3			
TREC	6	7	8	Robust 2003	Robust 2004
queries	training	testing			
	301-350	351-400	401-450	601-650	651-700

Type	webpages								
Dataset	WT10g		GOV2			Clueweb09 Category B			
N	1,692,096		25,205,179			50,189,002			
Size (GB)	10		426			≈ 1500			
TREC	9	10	Terabyte			Web track			
			2004	2005	2006	2009	2010	2011	2012
queries	testing				training	testing			training
	451-500	501-550	701-750	751-800	801-850	1-50	51-100	101-150	151-200

the top 1000 retrieved documents is used because it is a widely used metric, thus allowing a comparison of our retrieval results with those in the literature. The NDCG@20 value, which may reflect user preference, is also presented as it is another common metric used in the literature.

Short title queries, each with about 2 to 3 query terms on average, are used in our retrieval because such query lengths are typical in web searches [81]. Stemming is performed on both documents and queries in pre-processing, based on Porter's algorithm [68]. Stop-words are removed, following the Indri stop-word list[5]. The free parameters of each retrieval model (as shown in Table 11 in Appendix A) are calibrated by training on a selected set of 50 title queries (i.e. the development topics [50; 76]) and applied to other sets of queries for testing. Table 4 shows the training and testing sets used in this study. Training is performed separately for the news and webpage collections because of their different document nature. For the news collections, the TREC-6 queries are used for training because the dataset of TREC-6 is a superset used for the other news tracks (Table 4). As earlier 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. 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 should be better trained, so that there is less chance of these systems to miss a relevance document for the last track. Furthermore, for GOV2 the pooling employed by Terabyte-2006 to obtain the list of judged relevant documents was different from previous tracks — the pool for Terabyte-2006 was based on both automatic runs and runs with manually constructed queries, so that there is a reduced bias towards relevant documents containing title query words [7]. As for Clueweb09 Cat-B, training on the 2012 track enables us to compare our test data (for topics of the 2009-2011 tracks) with results found in the literature, such as those of [36] (see Table 9(b)). Separate training is performed for Clueweb09 Cat-B because it has a much larger size and also contains spam while the other two webpage collections do not. Calibration is performed by a grid search of the parameters that maximize the MAP [18].

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

The above methodology of using distinct sets of queries for training and testing has been used by others (e.g. [50; 76; 83]). We apply this calibration methodology instead of cross-validation for the following reasons. First, our approach reflects retrieval systems in practice, as these are generally calibrated beforehand and utilized for retrieval with unseen queries. Second, our methodology uses less training data than cross-validation, especially if there are many testing sets of queries, so that it should be a stronger test of the effectiveness of a retrieval method. Third, our methodology allows the retrieval performance for both the training set and testing sets of queries to be examined, as in Tables 5 to 7. Fourth, our methodology enables checking whether the 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.

The validity of our training and testing methodology is supported by our experiments, as our retrieval results generally match with those reported in the literature and obtained with systems such as Indri, Terrier or Anserini (Section 4.3.3). Otherwise, the performance of our system would be systematically lower than others if our system were significantly off-calibrated for the test data.

4.2 Hypothesis testing

Our study aims to provide some insight regarding the effectiveness of retrieval models and address the research questions stated in the Introduction section. With regard to RQ1, 'Is there a retrieval model that consistently outperforms other models across multiple collections?' we formulate Null Hypothesis Families (NHF) corresponding to each of the evaluation metrics that we employ, i.e. MAP and NDCG@20:

> **NHF 1-MAP**: There is no difference in retrieval effectiveness in terms of MAP averaged over the collection *C* between retrieval model *M* and the best performing model *B*.

and

> **NHF 1-NDCG**: There is no difference in retrieval effectiveness in terms of NDCG@20 averaged over the collection *C* between retrieval model *M* and the best performing model *B*.

In the above Null Hypothesis Families, *M* is one of the tested models (other than *B*) listed in Table 10, and *C* is one of the collections used for testing (Table 4). The model *B* is the one that gives the highest value of the corresponding metric (MAP or NDCG@20) averaged over all the testing tracks on the collection *C*. In IR, it is common to compare a retrieval model against some specific baseline, generally an established and successful model such as BM25 or LM. Thus, the corresponding null hypothesis would state that there is no difference between the new model and the specific baseline model. Here, we take the novel approach of stating the null hypothesis in terms of the best performing model *B* rather than a specific baseline chosen before the retrieval experiments. This is because we wish to find out whether there exists a retrieval model that consistently out-performs the others (RQ1) and it is not known beforehand which model performs the best. For the answer to RQ1 to be positive, it would require either NHF 1-MAP or NHF 1-NDCG to be rejected for all tested models *M* (other than *B*) and across all the collections *C* used.

The Null Hypothesis Families NHF 1-MAP and NHF 1-NDCG involve many statistical tests corresponding to the retrieval models being tested in this study, leading to the multiple comparisons problem [8]. As a results, it is necessary to apply corrections to the significance levels, such as by the Bonferroni correction or Benjamini-Hochberg procedure [22; 69]. We do not apply the Bonferroni correction [26] because it does not control type 2 errors (acceptance of false null hypothesis, i.e. false negative) and may lead to many discoveries not detected. Instead, we apply the widely-adopted[6] Benjamini-Hochberg procedure [5] which sets the false discovery rate. The Benjamini-Hochberg procedure uses the

[6] The paper by Benjamini and Hochberg is one of the top 25 most cited statistical paper.

p-values of the statistical tests of the whole collection (e.g., Clueweb09) averaged across multiple tracks (e.g., web tracks 2009, 2010 and 2011 used for testing) because each track is considered as a subgroup of the collection. The performance averaged over the tracks used for testing in each collection is reported at the right end of our result tables (Tables 5 to 7). 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 equals 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 which is set to 5% here, corresponding to a statistical significance confidence level of 95%, and $m = 16$ for Null Hypothesis Family NHF 1-MAP and NHF 1-NDCG in this study. The threshold p-value of the Benjamini-Hochberg procedure is reported in the tables. In Tables 5 to 7, the statistical significance indication ('↓') means that the evaluation metric of corresponding model is poorer than the best model with p-value smaller than or equal to the stated FDR p-value threshold. In the statistical significance testing, since the evaluation metric of the tested model must be lower than the best model, a one-tailed test is performed. In this study, the p-values are obtained by the randomization test [80].

4.3 Experimental results

4.3.1 Comparison of retrieval performance within a test collection. Tables 5 to 7 show the MAP and NDCG@20 results of the retrieval models on the various test collections. In the tables, the highest MAP or NDCG@20 values averaged over the testing tracks within each collection are shown in bold. In general the highest MAP and NDCG@20 values within each collection may be obtained by different retrieval models, despite some correlation between the two metrics [31]. However the MATF model [64] yields the highest values of both MAP and NDCG@20, on both the news collection (Disks 4&5 - CR) and the GOV2 collection of webpages, indicating the robustness of the MATF model.

With respect to RQ1, which questions whether any retrieval model consistently outperforms all other methods, we perform the one-tailed randomization test to determine whether the difference in the averaged metric (MAP or NDCG@20) between each retrieval method and the best model within each test collection is statistically significant. While statistical significance at the 95% confidence level is found for many cases (i.e. rejection of NHF 1-MAP or NHF 1-NDCG), in most collections there are some cases for which the difference is not statistically significant (Tables 5 to 7). Hence, we make the following observation:

> **Observation 1**: Within a test collection, the best retrieval model generally does not outperform all other tested methods with statistical significance at the 95% confidence level, either in MAP or NDCG@20.

There are only a few exceptions to the above observation. The exception for MAP is in Clueweb09, for which the MVD model outperforms all others with statistical significance. For NDCG@20, the exceptions are in WT10g with pivoted unique normalization being the best model, and in GOV2 with MATF being the best.

While established models such as BM25, LM and PL2 are commonly chosen as the performance baselines to compare with new retrieval methods, it is observed that these models generally do not attain the top three ranks in either MAP or NDCG@20, for all the tested collections. On the other hand, some recent methods, such as MATF or MVD, consistently yield higher MAP or NDCG@20 values than the standard baselines. Therefore, our results support the recommendation for future studies to include some of the new methods as stronger baselines in the evaluation of new retrieval techniques.

4.3.2 Comparison of retrieval performance across test collections. Tables 5 to 7 indicate that the performance of each retrieval model relative to the others generally varies across the test collections. Such variation may also be seen in

Table 5. Performance of retrieval with title queries on TREC6 (Disks 4&5) collection

	Disks 4&5	Disks 4&5 - CR				
	Performance by track					By collection
	TREC-6	TREC-7	TREC-8	Robust 2003	Robust 2004	7-8,2003-2004
	Topics 301-350	Topics 351-400	Topics 401-450	Topics 601-650	Topics 651-700	Topics 351-450,601-700
	training	testing				testing
(a) MAP						
Ltw1	.1608	.1218	.1782	.1696	.1565	.1565↓
BM25	.2444	.1948	.2681	.2906	.2761	.2573↓
BM25+	.2469	.1938	.2641	.2913	.2747	.2559↓
LM	.2480	.1905	.2651	.2848	.2788	.2547↓
SPUD	.2530	.1966	.2701	.2935	.2856	.2613
F3LOG	.2350	.1857	.2482	.2819	.2636	.2447↓
PL2	.2494	.1921	.2643	.2886	.2798	.2561↓
PL3	.2196	.1729	.2357	.2614	.2419	.2279↓
LGD	.2501	.1908	.2686	.2847	.2767	.2551↓
SPL	.2517	.1869	.2618	.2903	.2811	.2549↓
IRRAc	.2466	.1857	.2667	.2803	.2673	.2499↓
pivoted	.2469	.1944	.2582	.2898	.2780	.2550↓
PIV+	.2422	.1850	.2570	.2684	.2580	.2420↓
Gos1	.2496	.1904	.2679	.2803	.2702	.2521↓
Gos3	.2545	.1996	.2654	.2919	.2806	.2592↓
MATF	.2546	.1984	.2734	.2982	.2863	**.2640**
MVD	.2541	.1941	.2712	.2883	.2782	.2578↓
FDR p-value threshold (NHF 1-MAP)						.0226
(b) NDCG@20						
Ltw1	.2890	.2942	.3437	.2667	.2487	.2885↓
BM25	.4427	.4135	.4658	.3814	.3936	.4137↓
BM25+	.4558	.4156	.4675	.3867	.3933	.4159↓
LM	.4477	.4069	.4602	.3874	.4014	.4140↓
SPUD	.4566	.4153	.4735	.3949	.4065	.4226
F3LOG	.4434	.4034	.4572	.3892	.3867	.4092↓
PL2	.4560	.4176	.4668	.3924	.4046	.4204↓
PL3	.4048	.3853	.4314	.3799	.3688	.3915↓
LGD	.4478	.4130	.4648	.3821	.3886	.4123↓
SPL	.4587	.4143	.4728	.3945	.4058	.4219
IRRAc	.4249	.4084	.4702	.3749	.3869	.4102↓
pivoted	.4603	.4200	.4557	.4012	.4008	.4195↓
PIV+	.4403	.3867	.4587	.3595	.3678	.3933↓
Gos1	.4455	.4131	.4607	.3781	.3843	.4092↓
Gos3	.4692	.4192	.4640	.3964	.4022	.4205↓
MATF	.4727	.4265	.4858	.3962	.4037	**.4282**
MVD	.4711	.4162	.4820	.3918	.3999	.4226↓
FDR p-value threshold (NHF 1-NDCG)						.0377

Note: ↓ indicates poorer performance than the best performing model for the testing queries over the collection (shown in bold) with statistical significance, i.e. with p-value smaller than or equal to the FDR threshold shown in the table.

Table 8, which lists the top six retrieval models in MAP and NDCG@20 for each collection. For example, the table shows that not many retrieval models consistently rank in the top six across all the tested collections. In particular, no single retrieval model consistently performs better than all others across all collections. Thus, we make the following observation:

> **Observation 2**: No single retrieval model consistently outperforms all other tested methods across all the tested collections, either in MAP or NDCG@20.

Table 6. Performance of retrieval with title queries on WT10g and GOV2 collections

	WT10g		GOV2 Terabyte			WT10g	GOV2 Terabyte
	Performance by track					By collection	
	TREC-9	TREC-10	2004	2005	2006	9&10	2004&2005
	Topics 451-500	Topics 501-550	Topics 701-750	Topics 751-800	Topics 801-850	Topics 451-550	Topics 701-800
	testing				training	testing	
(a) MAP							
Ltw1	.1306	.1088	.1256	.1425	.1112	.1197↓	.1341↓
BM25	.2157	.2011	.2656	.3279	.3044	.2084↓	.2971↓
BM25+	.2130	.1997	.2859	.3426	.3036	.2064↓	.3145
LM	.2102	.2050	.2728	.3243	.3083	.2076↓	.2988↓
SPUD	.2115	.2157	.2838	.3403	.3210	.2136	.3123
F3LOG	.2082	.2024	.2377	.2806	.2728	.2053↓	.2593↓
PL2	.2119	.2050	.2676	.3340	.3072	.2084↓	.3011↓
PL3	.1917	.1865	.2408	.2885	.2591	.1891↓	.2649↓
LGD	.2109	.1939	.2701	.3301	.3095	.2024↓	.3004↓
SPL	.2080	.2004	.2690	.3307	.3078	.2042↓	.3002↓
IRRAc	.2053	.1737	.2461	.3148	.2828	.1895↓	.2808↓
pivoted	.2168	.2240	.2672	.3274	.3107	.2204	.2976↓
PIV+	.2147	.1915	.2332	.2863	.2754	.2031↓	.2600↓
Gos1	.2108	.1901	.2661	.3286	.3095	.2004↓	.2976↓
Gos3	.2122	.2034	.2540	.3235	.2902	.2078↓	.2891↓
MATF	.2164	.2265	.2869	.3443	.3221	**.2215**	**.3159**
MVD	.2208	.2148	.2828	.3351	.3199	.2178	.3092
FDR p-value threshold (NHF 1-MAP)						.03115	.0133
(b) NDCG@20							
Ltw1	.2115	.2020	.2765	.3178	.2732	.2068↓	.2974↓
BM25	.3234	.3242	.4110	.4998	.4940	.3238↓	.4559↓
BM25+	.3224	.3325	.4163	.5019	.4997	.3275↓	.4596↓
LM	.3088	.3339	.4154	.4925	.4748	.3214↓	.4543↓
SPUD	.3135	.3426	.4250	.5095	.5025	.3281↓	.4677↓
F3LOG	.3191	.3375	.4127	.4771	.4674	.3283↓	.4453↓
PL2	.3172	.3364	.3995	.5067	.4735	.3268↓	.4536↓
PL3	.2968	.3288	.3879	.4833	.4448	.3128↓	.4361↓
LGD	.3039	.3213	.4144	.4954	.4735	.3126↓	.4553↓
SPL	.3124	.3366	.4093	.5110	.4763	.3245↓	.4606↓
IRRAc	.2956	.2954	.3527	.4770	.4275	.2955↓	.4155↓
pivoted	.3342	.3662	.4318	.5007	.5153	**.3502**	.4666↓
PIV+	.3155	.3206	.4145	.4769	.4553	.3180↓	.4460↓
Gos1	.3023	.3151	.4108	.4978	.4727	.3087↓	.4547↓
Gos3	.3219	.3383	.3966	.4797	.4657	.3301↓	.4386↓
MATF	.3247	.3507	.4519	.5250	.5196	.3377↓	**.4888**
MVD	.3248	.3454	.4380	.5075	.5224	.3351↓	.4731↓
FDR p-value threshold (NHF 1-NDCG)						.0196	.0156

Note: ↓ indicates poorer performance than the best performing model for the testing queries over the collection (shown in bold) with statistical significance, i.e. with p-value smaller than or equal to the FDR threshold shown in the table.

Based on Observations 1 and 2, the answer to the research question RQ1 is negative. The first implication of this result is that in the evaluation of new retrieval techniques, it is desirable to use multiple baselines for comparison by selecting several retrieval models known to perform well in multiple collections. This would avoid the issue of weak baselines [3; 42]. Second, a new retrieval technique should be tested in multiple collections to confirm its robustness across collections. Table 8 shows that the PL2 and BM25/BM25+ models can consistently attain the top six rank among retrieval models across the tested collections, confirming the effectiveness of these models. However, these models

Table 7. Performance of retrieval with title queries on Clueweb09 Cat-B collection

	Clueweb09 Cat-B Web track				
	Performance by track				By collection
	2009	2010	2011	2012	2009-2011
	Topics 1-50	Topics 51-100	Topics 101-150	Topics 151-200	Topics 1-150
	testing			training	testing
(a) MAP					
Ltw1 (-k 11)	.0661	.0798	.0509	.0917	.0654↓
BM25	.0965	.1111	.0832	.1170	.0967↓
BM25+	.0988	.1129	.0874	.1160	.0995↓
LM	.0904	.1037	.0879	.1082	.0939↓
SPUD	.0893	.1091	.0837	.1172	.0938↓
F3LOG	.0848	.0982	.0763	.1109	.0851↓
PL2	.0960	.1085	.1012	.1108	.1018↓
PL3	.0878	.1030	.0801	.1093	.0901↓
LGD	.0945	.1073	.0952	.1088	.0989↓
SPL	.0973	.1092	.1041	.1060	.1034↓
IRRAc	.0907	.1024	.1071	.1338	.1000↓
pivoted	.0762	.1013	.0832	.1135	.0867↓
PIV+	.0799	.0948	.0755	.1104	.0832↓
Gos1	.0906	.1047	.0906	.1112	.0952↓
Gos3	.0920	.1088	.0922	.1187	.0975↓
MATF	.0899	.1140	.1067	.1218	.1034↓
MVD	.1005	.1200	.1056	.1341	**.1085**
FDR p-value threshold (NHF 1-MAP)					.0096
(b) NDCG@20					
Ltw1 (-k 11)	.1676	.1357	.1186	.1074	.1407↓
BM25	.2534	.1732	.1792	.1437	.2023↓
BM25+	.2536	.1759	.1887	.1409	.2065↓
LM	.2195	.1385	.1655	.1277	.1750↓
SPUD	.2169	.1566	.1558	.1308	.1767↓
F3LOG	.2098	.1463	.1584	.1318	.1718↓
PL2	.2225	.1459	.1836	.1265	.1845↓
PL3	.2182	.1492	.1652	.1215	.1779↓
LGD	.2188	.1429	.1768	.1248	.1800↓
SPL	.2316	.1539	.1904	.1249	.1925↓
IRRAc	.2636	.2139	.2142	.1960	**.2308**
pivoted	.1832	.1717	.1554	.1247	.1701↓
PIV+	.2065	.1469	.1603	.1273	.1716↓
Gos1	.2093	.1345	.1685	.1223	.1712↓
Gos3	.2209	.1556	.1689	.1332	.1822↓
MATF	.2294	.2070	.2064	.1423	.2144
MVD	.2474	.2279	.2095	.1707	.2282
FDR p-value threshold (NHF 1-NDCG)					.0204

Note: ↓ indicates poorer performance than the best performing model for the testing queries over the collection (shown in bold) with statistical significance, i.e. with p-value smaller than or equal to the FDR threshold shown in the table.

generally do not give the best performance. On the other hand, the MATF and MVD models are found to perform within the top four in both MAP and NDCG@20 across the tested collections, with MATF being the top performer in either MAP or NDCG@20 for several collections. This suggests that MATF and MVD are good candidates as baselines in retrieval experiments, in support of the recommendation in Section 4.3.1.

A somewhat surprising finding in our experiments is that the retrieval models conforming to the PRP, such as BM25 and LM, do not necessarily yield the best retrieval results, for all datasets (Table 8). An exception is the SPUD

Table 8. Summary of the top six ranking of retrieval models in MAP and NDCG@20 on various test collections

MAP	
Disks 4&5 - CR (news / federal register)	**MATF***, SPUD*, Gos3, MVD, BM25, PL2
WT10g	**MATF***, pivoted*, MVD*, SPUD*, BM25, PL2
GOV2	**MATF***, BM25+*, SPUD*, MVD*, PL2, LGD
Clueweb09 Cat-B	**MVD***, MATF, SPL, PL2, IRRAc, BM25+

NDCG@20	
Disks 4&5 - CR (news / federal register)	**MATF***, SPUD*, MVD, SPL*, PL2, Gos3
WT10g	**pivoted***, MATF, MVD, Gos3, F3LOG, SPUD
GOV2	**MATF***, MVD, SPUD, pivoted, LGD
Clueweb09 Cat-B	**IRRAc***, MVD*, MATF*, BM25+, BM25, SPL

Note: * indicates no statistically significant difference from the best performing model over the collection (shown in bold), i.e. with p-value larger than the corresponding FDR threshold shown in Table 5-7.

model, which appears among the top performing models in both MAP and NDCG@20 for several datasets. The SPUD model differs from LM by modelling word burstiness [16]. The results thus suggest even for models conforming to PRP, retrieval performance may be enhanced by better estimates of the probability of relevance.

4.3.3 Reproducibility study. As a reproducibility study, we compare our retrieval results with those in the literature, focusing on the webpage collections, which are larger than the news collections. For a fair comparison, we select some representative works which report retrieval experiments that use the same testing data (collections and tracks) as in this work (Table 4). Specifically, we compare with Goswami et al. [28] for WT10g and GOV2 datasets (Table 9(a)), and with Huston and Croft [36] for Clueweb09 Cat-B (Table 9(b)). Leading retrieval models representing different frameworks are employed in [28] and [36], including the BM25 and LM [28; 36], PL2 [36] and the log-logistic (LGD) model [28]. These leading retrieval models are widely used as baselines in IR, so comparing with the results of [28; 36] both serves to reaffirm how well these important baseline values can be reproduced in retrieval experiments, and to support the validity of our current comparative study. This is achieved as Table 9(a) and (b) show that for the retrieval models across the various webpage test collections, our retrieval results generally corroborate with those in the literature, with relative differences in MAP within about 4.5%. For Clueweb09 Cat-B (Table 9(b)), our MAP and NDCG@20 values are generally lower than those obtained by [36]. One possible reason for our lower values may be that spam filtering was used by [36] but not in our retrieval. The differences are smaller in MAP (less than 4.2%) than in NDCG@20 (9.3-21%). This may be due to the sensitivity of the high ranked (top 20) documents retrieved depending on whether spam filtering is used. Our lower MAP and NDCG@20 values compared with [36] in this case are less likely to be due to the use of off-calibrated parameters on our test data, as our MAP and NDCG@20 values on Clueweb09 Cat-B are better than other works in the literature that also did not specify the use of spam filtering, such as those of Yang and Fang [90] (see our Table 9(c)) and Yang et al. [91] (see our Table 9(d)).

Recently, Yang and Fang [90] and Yang et al. [91] have conducted comprehensive reproducibility studies of retrieval models. As a reference, it is of interest to compare our retrieval results with these other studies. In both [90] and [91], there is no separation of training and testing data. In [90], topics of several tracks on each collection are combined and the various retrieval models are optimized by a grid search method [90]. Thus we believe it is fair to compare the combined retrieval results of our training and testing sets with the results of [90] and [91], as presented in Table

Table 9. Comparison of our retrieval results (MAP/NDCG@20) with those in the literature

(a) Comparison with Goswami et al. [28]

	WT10g, TREC 9&10 (Topics 451-550)		GOV2, Terabyte 2004-05 (Topics 701-800)	
	MAP		MAP	
	Goswami et al. [28]	Our results [Table 6(a)]	Goswami et al. [28]	Our results [Table 6(a)]
BM25	.2057	.2084 (+1.3%)	.3017	.2971 (-1.5%)
LM	.2084	.2076 (-0.4%)	.2976	.2988 (+0.4%)
LGD	.2029	.2024 (-0.2%)	.2909	.3004 (+3.3%)
Gos1	.1920	.2004 (+4.4%)	.2910	.2976 (+2.3%)
Gos3	.2087	.2078 (-0.4%)	.2978	.2891 (-2.9%)

(b) Comparison with Huston & Croft [36]

	Clueweb09 Cat-B, Web track 2009-2011 (Topics 1-150)			
	MAP		NDCG@20	
	Huston & Croft [36]	Our results [Table 7(a)]	Huston & Croft [36]	Our results [Table 7(b)]
BM25	.099	.0967 (-2.3%)	.223	.2023 (-9.3%)
LM	.098	.0939 (-4.2%)	.221	.1750 (-20%)
PL2	.105	.1018 (-3.0%)	.233	.1845 (-21%)

(c) Comparison with Yang & Fang [90]

	GOV2, Terabyte 2004-06 (Topics 701-850)		Clueweb09 Cat-B, Web Track 2009-12 (Topics 1-200)	
	MAP		MAP	
	Yang & Fang [90]	Our results [Table 6(a)]	Yang & Fang [90]	Our results [Table 7(a)]
BM25	.297	.2995 (+0.8%)	.089	.1019 (+15%)
LM	.299	.3108 (+3.9%)	.090	.0975 (+8.3%)
PL2	.303	.3032 (+0.1%)	.089	.1041 (+17%)
LGD	.300	.3034 (+1.1%)	.086	.1014 (+18%)
SPL	.299	.3027 (+1.2%)	.093	.1041 (+12%)

(d) Comparison with Yang et al. [91]

	WT10g, TREC 9&10 (Topics 451-550)			
	MAP			
	Indri [91]	Terrier [91]	Anserini [91]	Our results [Table 6(a)]
BM25	.1955	.2136	.2012	.2084 (+6.6%, -2.4%, +3.6%)
LM	.1915	.2111	.2034	.2076 (+8.4%, -1.7%, +2.1%)
PL2	.2012	.2129	.1889	.2084 (+3.6%, -2.1%, +10.3%)
SPL	.1947	-	.1726	.2042 (+4.9%, -, +18.3%)

	GOV2, Terabyte 2004-06 (Topics 701-850)			
	MAP			
	Indri [91]	Terrier [91]	Anserini [91]	Our results [Table 6(a)]
BM25	.2970	.3050	.3030	.2993 (+0.8%, -1.9%, -1.2%)
LM	.2995	.2976	.2954	.3018 (+0.8%, +1.4%, +2.2%)
PL2	.3029	.3076	.3067	.3029 (0%, -1.5%, -1.2%)
SPL	.3017	-	.3070	.3025 (+0.3%, -, -1.5%)

	Clueweb09 Cat-B, Web Track 2010-2012 (Topics 51-200))			
	NDCG@20			
	Indri [91]	Terrier [91]	Anserini [91]	Our results [Table 7(b)]
BM25	.1390	.1456	.1422	.1654 (+19%, +14%, +16%)
LM	.1164	.1287	.1227	.1439 (+24%, +12%, +17%)
PL2	.1223	.1312	.1274	.1520 (+24%, +16%, +19%)
SPL	.1209	-	.1273	.1564 (+29%, -, +23%)

Note: The percentages are the relative difference between our results and the corresponding values in the literature. In (d), the relative differences shown are with respect to Indri, Terrier and Anserini, respectively.

9(c) and (d). In [91], retrieval is performed with several open-source systems, including Indri, Terrier and Anserini. It is observed that for retrieval on WT10g and GOV2, our MAP values match well with those of [90] and the various open-source systems in [91]. For Clueweb09 Cat-B our results are generally somewhat higher, in terms of MAP (Table 9(c)) or NDCG@20 (Table 9(d)). This may be because retrieval is hard for the topics on Clueweb09 Cat-B, such as with generally low MAP values (around 0.09) obtained with the various retrieval models in [90], or low NDCG@20 values (below 0.15) in [91]. Table 9(d) indicates that on Clueweb09 Cat-B, the NDCG@20 values obtained by our system are quite significantly higher than the open-source systems of [91], though MAP values of the open-source systems are not available for comparison. As spam filtering was not applied in all of these cases, the results suggest NDCG@20 to be sensitive for Clueweb09. Also, as the systems were trained by optimizing MAP, the differences between the systems may be larger in NDCG@20 values, as Table 9(b) shows smaller relative differences in MAP than NDCG@20 between our system and that of [36] on Clueweb09.

An observation of Table 9(d) is that variations in performance among different implementations of retrieval models such as provided by different open-source toolkits can be as large as the differences between different retrieval models. Some insights that may be drawn from this observation include: (1) For a fair comparison of the effectiveness of different retrieval models, it is important to conduct the comparison on the same platform, under the same system characteristics such as stop-word removal and the stemming algorithm used. This enables any observed difference in performance to be attributed to the different models rather than system variations; (2) Even though the top retrieval models are derived based on very different approaches, they are very competitive, attaining similar values of evaluation metrics like MAP or NDCG. This also shows that it is very difficult to raise the level of retrieval effectiveness (e.g. [3]).

5 ANALYSIS OF EXPERIMENTAL RESULTS AND DISCUSSION

For the purpose of an analysis and discussion of our experimental results, we first summarize the characteristics of the tested retrieval models in Section 5.1. In Section 5.2, we discuss our results with respect to heuristic constraints. Last, in Section 5.3, we seek to identify some important features in an effective retrieval ranking function so as to answer our research question RQ2.

5.1 Analytical comparison of retrieval ranking functions

Table 10 summarizes some characteristics of the ranking functions of the retrieval models tested in the current study. Some models conform to the PRP [74], such as the classical probabilistic BM25, as documents are ranked according to their estimated probability of relevance. As mentioned in Section 3.3.1, the LM and SPUD models also conform to the PRP. Many retrieval models are designed based on the heuristic constraints of Fang et al. [23]. These include models obtained in the axiomatic approach [25], in which a base model is modified to better satisfy heuristic constraints, such as the BM25+, F3LOG, PL3 and PIV+ models. The models Gos1 and Gos3 are functions generated within a certain function space and pruned using heuristic constraints. Retrieval models that invoke heuristic constraints are indicated in Table 10.

Some past works [23; 24] have evaluated many of the leading retrieval models, including BM25, LM, PL2 and pivoted normalization, in terms of whether they satisfy the heuristic constraints. So far, analytical evaluation based on retrieval heuristics is lacking for some of the new retrieval models. We provide such an evaluation for the Ltw1, MATF and MVD models in Appendix B. Overall, the majority of the models included in this study satisfy all the heuristic constraints either unconditionally, or conditionally under appropriate settings of the model parameters (e.g. [11; 16; 24; 28; 58] and

Table 10. Characteristics of the ranking functions of retrieval models

Model	PRP	invokes constraints	Factors appearing in ranking function						tf ($f(t,d)$) normalization					
			$f(t,d)$	$df(t)$	$\lvert d\rvert$	$\lvert\dot{d}\rvert$	$f(t,\mathbb{C})$	$\lvert q\rvert$	$\log(tf)$	$\frac{tf}{tf+K}$	$\frac{tf}{avg.tf}$	$\lvert d\rvert$	$\lvert\dot{d}\rvert$	DFR '2'
Ltw1			✓	✓					✓					
BM25	✓		✓	✓	✓					✓		✓		
BM25+*	✓	✓	✓	✓	✓					✓		✓		
LM	✓		✓		✓		✓		✓			✓		
SPUD*	✓	✓	✓	✓		✓			✓		✓		✓	
F3LOG		✓	✓	✓	✓			✓	✓					
PL2			✓		✓		✓		✓					✓
PL3		✓	✓		✓		✓		✓			✓		
LGD		✓	✓	✓	✓				✓					✓
SPL*		✓	✓	✓	✓				✓					✓
IRRAc*			✓		✓		✓		✓			✓		
pivoted*			✓	✓	✓	✓			✓		✓		✓	
PIV+		✓	✓	✓	✓				✓			✓		
Gos1		✓	✓	✓	✓				✓					✓
Gos3		✓	✓	✓	✓				✓					✓
MATF*		✓	✓	✓	✓	✓	✓	✓	✓		✓			✓
MVD*		✓	✓	✓	✓	✓			✓		✓			✓

Note: Retrieval models that attain either the top MAP or NDCG@20 value in any of the tested collections, and models that do not differ from these top models with statistical significance, are marked with *.

Appendix B.2 and B.3). The only exception is the Ltw1 model, which only satisfies the TFC1, TFC2 and TDC constraints, but not the ones involving document length normalization.

Despite the diverse theoretical bases or motivation behind the various retrieval models, their ranking functions invariably contain factors made up of term or document statistics. Table 10 summarizes the factors that appear in the ranking functions. All ranking functions contain the term frequency $f(t,d)$ of query terms, reflecting the intuition that higher occurrence of a query term is a strong indication of relevance. In all tested retrieval models, the term frequency appears either in a logarithmic function or in the form $tf/(tf+K)$, both of which satisfy the TFC1 and TFC2 constraints. With respect to the TDC constraint, the majority of ranking functions contain document frequency $df(t)$, or more precisely $1/|df(t)|$, which serves as a discriminatory factor that indicates the importance of a query term. Ranking functions not containing $df(t)$ generally have instead a $f(t,\mathbb{C})/|\mathbb{C}|$ factor, which has the effect of term discrimination. Regarding the length normalization constraints, the majority of ranking scores decreases with increasing document length $|d|$, which serves as a normalization factor that reduces the bias towards long documents. Document length normalization is applied in the retrieval models in various ways, including: (1) dividing by a function of $|d|$; (2) dividing by a function of $|\dot{d}|$, where $|\dot{d}|$ is the number of distinct terms in the document; or (3) the 'normalization 2' of the DFR framework (Eq. 8d).

5.2 Justification of heuristic constraints

Tables 5-7 show that Ltw1 performs particularly poorly in MAP and NDCG@20 compared with other tested models. One notable difference between Ltw1 and the other models is the lack of document-length normalization in Ltw1. Thus, the results confirm the importance of the length normalization constraints.

Our results confirm the good performance of the machine-generated models of [28]. In particular, the Gos3 model is among the top performing models on some collections (Table 8). This indicate that the best machine-generated models of [28], which are designed to conform to heuristic constraints, can rival theoretically motivated models, such as those based on PRP. The results thus justify the heuristic constraints to be good guidelines to obtain effective ranking functions.

On the other hand, we find that models obtained by the axiomatic approach, such as BM25+, PL3, F3LOG and PIV+, do not necessarily perform better than the corresponding base models. The BM25+ model can yield better MAP than the parent BM25, such as in the GOV2 and Clueweb09 Cat B collections, but this better effectiveness is not consistently observed in all collections. Previously, Kamphuis et al. [40] also did not find any improvement of BM25+ over the traditional BM25, for retrieval on several newswire collections including the Disks4&5 - CR. In our experiments, the PL3, F3LOG and PIV+models generally yield poorer MAP and NDCG@20 than their base models in all collections. It appears that a sufficiently optimized base model can perform comparably with its variants obtained by the axiomatic approach. Therefore, our results indicate that better satisfaction of heuristic constraints is not sufficient to guarantee good retrieval effectiveness.

5.3 Important features of an effective document ranking function

In Table 10, the retrieval models that attain either the top MAP or NDCG@20 value in any of the tested collections, and models that do not differ from these top models with statistical significance, are marked with an asterisk (*). In order to answer the research question RQ2, we seek to identify some common features shared by these top performing retrieval models. We highlight several important features in Section 5.3.1. Section 5.3.2 then discusses the usage of these features for effective retrieval.

5.3.1 Important features. Some features identified as contributing to effective retrieval are as follows.

(1) *A logarithmic functional form of the term frequency component.* As shown in Table 10, the term frequency component of all the tested retrieval models employ either a $tf/(tf+K)$ functional form (as in BM25 and BM25+) or some logarithmic function of the term frequency (as in all the other models). Both of such functional forms provides a sub-linear damping of the term frequency, conforming to the TF1 and TF2 heuristic constraints (Section 2.2). Comparing the two functional forms, Table 8 indicates that models using the logarithmic form (e.g. MATF, MVD or SPUD) are more robust in attaining top performance across two or more collections, in both MAP and NDCG@20.

(2) *Normalization of the term frequency by the average term frequency in a document.* The average term frequency is given by $f_{avg}(d) = |d|/|\dot{d}|$. The normalization of the term frequency by $f_{avg}(d)$ was first introduced by Singhal et al. [79] in pivoted unique normalization. Various forms of the f_{avg}-based normalization appear in the top performing models of our study. For example, the SPUD model contains a logarithm of the direct ratio $|\dot{d}| \cdot f(t,d)/|d| = f(t,d)/f_{avg}(d)$ (Eq. (6a-b)). In pivoted unique normalization (Eq. (13)), the term frequency component is $(1+\log(f(t,d))/(1+\log(f_{avg}(d))$. In the MATF model, f_{avg}-based normalization is employed in the Relative Intra-document TF (RITF) component, given by $\log(1+f(t,d))/\log(1+f_{avg}(d))$ (Eq. (17c)), and similarly in the MVD model (Eq. (18c)). Thus in all these top performing models, f_{avg}-based normalization is used in a sub-linear logarithmic functional form of the term frequency (i.e. item (1) above).

As pointed out by Singhal et al. [79], the average term frequency may be taken as a representation of the verbosity of a document. Normalization by $f_{avg}(d)$ is thus a type of verbosity normalization [62] that deals with high term frequencies being due to an author tending to use more words or the effect of word burstiness as in the SPUD model [16].

(3) *Normalization either based on the number of discrete terms in a document or by the DFR 'normalization 2'.* Based on the LNC1 heuristic, document length normalization is crucial for an effective retrieval function, as shown by the poor performance of the Ltw1 model. However, a simple normalization based on document length may over-penalize long documents [79]. As mentioned before, Table 8 indicates that the SPUD, MATF and MVD models are robust in attaining top performance across two or more collections, in MAP and NDCG@20. This suggests that with regard to document length normalization, the favorable methods are either based on the discrete number of terms $|\dot{d}|$ (as in SPUD) or by the DFR 'normalization 2' (as in MATF and MVD). In fact, both of these methods avoid over-penalizing long documents, in line with the LNC2 and TF-LNC constraints [23].

Apart from the SPUD model, normalization based on the discrete number of terms in a document, $|\dot{d}|$, is also employed in pivoted unique normalization (Eq. (13)). The quantity $|\dot{d}|$ may be taken as a crude measure of scope [16]. The scope hypothesis is that some documents may be longer because they cover many different topics [75]. With the normalization of $1/(\mu_s|\dot{d}| + 1)$ in SPUD (Eq. (6b)), instead of the normalization by $|d|$ as in LM (Eq. (4b)), long documents are only penalized for the occurrence of distinct terms, which may correspond to a wider scope.

The DFR 'normalization 2' follows the form $tfn = f(t,d) \cdot \log_2(1 + c\Delta/|d|)$ (Eq. (8d)). Here, the logarithmic form of the $|d|$ normalization takes into account the increase in term frequency does not follow a linear relationship with the document length. Thus, with the DRF 'normalization 2' form, long documents are penalized but with a diminishing effect, as employed in the LRTF components of MATF (Eq. (17d)) and MVD (Eq. (18d)). A retrieval function based on the 'normalization 2' form satisfies the LNC2 constraint, as well as the TF-LNC constraint for small $f(t,d)$ [28].

5.3.2 Usage of the important features. We briefly discuss the usage of the identified important features. In particular, we draw the following insights by comparing different retrieval models that use the features differently.

(1) *The individual features alone are not sufficient for effective retrieval.* While our results support the use of a $\log(f(t,d))$ form (item (1) of Section 5.3.1) that satisfies the TF1 and TF2 heuristics, it does not guarantee good performance because the overall performance depends on other factors in the retrieval model. For example, while the Ltw1 model uses a $\log(f(t,d){+}1)$ form (Eq. (1)), same as the top performing MATF, the Ltw1 model generally performs poorly due to the lack of document length normalization as required by the LNC1 heuristic. Thus, it is apparent that the LNC1 constraint is particularly important for an effective retrieval function. Likewise, we expect it is also crucial to include a term discrimination factor following the TDC constraint, as all our tested models contain such a factor.

(2) *The features may be combined to enhance retrieval effectiveness.* For example, both the traditional Dirichlet-smoothed LM (Eq. (5a-c)) and the SPUD model (Eq. (6a-b)) employ a logarithmic function of the term frequency. The SPUD model differs from LM by having the additional normalization features: normalization of the term frequency by the avarage term frequency, and normalization based on $|\dot{d}|$ (i.e. items (2) and (3) of Section 5.3.1, respectively). Thus, the addition of the important features can enhance retrieval effectiveness.

The top performing MATF (Eq. (17a-e)) and MVD (Eq. (18a-g)) models also employ all of the three identified features of Section 5.3.1. In particular, these two models contain a mixture of features items (2) and (3) via the RITF and LRTF components, respectively. Thus, the good performance of MATF and MVD further supports a combination of the important features to provide an effective ranking function.

6 CONCLUSION

We have performed retrieval experiments with an extensive list of term-independence retrieval models, on several TREC news and webpage test collections. Our study includes established retrieval models as well as some promising recent retrieval methods, all of which are implemented in our own retrieval system. Thus, these retrieval methods can be tested and compared under the same environment.

With respect to our research question RQ1, no single retrieval model is found to consistently outperform all other tested methods with statistical significance across all collections used. Several retrieval models which are commonly used as IR baselines, namely BM25, LM and PL2, generally do not rank in the top three in either MAP or NDCG@20 across all collections. On the other hand, some recent retrieval models, such as MATF and MVD, consistently yield higher MAP and NDCG@20 than the common baselines. Therefore the MATF and MVD models are good candidates to be included as baselines in future evaluation of new retrieval methods. Moreover, new methods should be tested in multiple collections to confirm their robustness across collections. A notable feature of the MATF model is that it is non-parametric, i.e. the model has no free parameters. Thus, an advantage of MATF is that model calibration is not required when it is applied on different collections.

Overall, our results justify heuristic constraints [23] to be good guidelines to effective ranking functions. However, we also find that better satisfaction of heuristic constraints is not a guarantee to superior retrieval performance. To answer RQ2, we seek the common features that appear in the ranking functions of the highest performing models. In this regard, our results support the following as important features: (1) a logarithmic function of the term frequency; (2) term frequency normalization by the average term frequency; and (3) normalization either based on the number of discrete terms in a document or by the DFR 'normalization 2'.

Finally, our work also serves as a reproducibility study for leading established and recent retrieval models. The retrieval results of the models implemented in our system match those reported in the literature, obtained on systems such as Indri [90], Terrier [91] or Anserini [91]. Our results thus support the use of these open-source systems to provide reproducible baselines.

APPENDIX

A CALIBRATED RETRIEVAL MODEL PARAMETERS

Table 11 summarizes the range over which the various retrieval model parameters are searched and the optimized values obtained for three test collections.

B ANALYTIC EVALUATION

Analytic evaluation based on retrieval heuristics is up to now unavailable in the literature for some of the retrieval models tested in the current study. Here we provide such evaluation for the models Ltw1 (B.1), MATF (B.2) and MVD (B.3).

B.1 Ltw1

From Eq. (1), as the ranking function $S_{Ltw1}(q, d)$ varies as the logarithm of $f(t, d)$, the the term frequency heuristics TFC1 and TFC2 are always satisfied. The term discrimination TDC heuristic is also easily seen to be always satisfied. The lack of document length consideration in the LTW1 model means that the length related constraints are not satisfied.

Table 11. Range of model parameters in grid search and their optimized values on various test collections

Retrieval model	free parameters and grid search range	Optimized value		
		Disks 4&5 - CR	WT10g/GOV2	Clueweb09 Cat-B
Ltw1	-	-	-	-
BM25	$k : 0.1-4$, $b : 0.1-1.0$	$k : 0.6$, $b : 0.4$	$k : 0.8$, $b : 0.3$	$k : 3$, $b : 0.15$
BM25+	$k_+ : 0.5-2000$, $b_+ : 0.05-4$, $\delta_+ : 0.1-10$	$k_+ : 0.8$, $b_+ : 0.45$, $\delta_+ : 1$	$k_+ : 0.8$, $b_+ : 0.45$, $\delta_+ : 1$	$k_+ : 1500$, $b_+ : 1$, $\delta_+ : 0.2$
LM	$\mu : 100-4000$	$\mu : 400$	$\mu : 700$	$\mu : 2000$
SPUD	$\mu_s : 0.0001-0.01$	$\mu_s : 0.003$	$\mu_s : 0.0015$	$\mu_s : 0.0004$
F3LOG	$s : 0.001-0.05$	$s : 0.01$	$s : 0.008$	$s : 0.008$
PL2	$c : 1-15$	$c : 10$	$c : 8$	$c : 12$
PL3	$\mu_p : 100-30000$	$\mu_p : 3000$	$\mu_p : 1500$	$\mu_p : 3000$
LGD	$c_l : 1-15$, $\beta_l : 0.2-3.0$	$c_l : 4$, $\beta_l : 1.0$	$c_l : 4$, $\beta_l : 0.9$	$c_l : 10$, $\beta_l : 2.0$
SPL	$c_s : 1-12$	$c_s : 6$	$c_s : 7$	$c_s : 5$
IRRAc	$a_i : 0.1-4$,$b_i : 0.0001-0.4$	$a_i : 1.2$, $b_i : 0.0005$	$a_i : 0.8$, $b_i : 0.05$	$a_i : 1.8$, $b_i : 0.1$
pivoted	$b_p : 0.0001-0.05$	$b_p : 0.025$	$b_p : 0.02$	$b_p : 0.0005$
PIV+	$b_{p+} : .001-200$, $\delta_{p+} : .0001-20$	$b_{p+} : 100$, $\delta_{p+} : 0.05$	$b_{p+} : 0.015$, $\delta_{p+} : 1$	$b_{p+} : 1$, $\delta_{p+} : 0.0005$
Gos1	$c_{g1} : 2-40$	$c_{g1} : 5$	$c_{g1} : 4$	$c_{g1} : 20$
Gos3	$c_{g2} : 2-60$	$c_{g2} : 8$	$c_{g2} : 8$	$c_{g2} : 40$
MATF	-	-	-	-
MVD	$k_M : 0.5-5$, $\alpha_M : 0.5-20$, $\beta_M : 0.5-1.0$	$k_M : 0.6$, $\alpha_M : 4$, $\beta_M : 0.8$	$k_M : 0.7$, $\alpha_M : 4$, $\beta_M : 0.8$	$k_M : 3$, $\alpha_M : 12$, $\beta_M : 0.7$

B.2 MATF

From Eq. (17a-d), it is clear that $S_{MATF}(q,d)$ is monotonic increasing function of $f(t,d)$, so that TFC1 is always satisfied. In Eq. (17b), the form of the function $g(x) = x/(1+x)$ ensures TFC2 is satisfied. The TDC heuristic is satisfied with the term discrimination factor $TDF(t)$ of Eq. (17e), which is a combination of the usual IDF factor and a factor based on the average elite set term frequency, $f(t,\mathbb{C})/df(t)$. The LCN1 heuristic is satisfied due to the document length normalization of Eq. (17d). Thus, the ranking function satisfies $\partial S_{MATF}(q,d)/\partial f(t,d) > 0$ (TFC1), $\partial^2 S_{MATF}(q,d)/\partial f(t,d)^2 < 0$ (TFC2) and $\partial S_{MATF}/\partial df(t) < 0$. Furthermore, $S_{MATF}(q,d)$ is based on the DRF normalization (Eq. (17d)). Goswami et al. [28] showed that under these conditions, the ranking function also satisfies the LNC2 heuristic. Moreover, for small $f(t,d)$, the TF-LNC heuristic is also satisfied [28].

B.3 MVD

The MVD model is an extension of the MATF model. One difference is in the term frequency factor of the ranking function, with the form $G(x)$ of Eq. (18e)) instead of the form $g(x) = x/(1+x)$ in MVD. Nonetheless, it can be seen from Eq. (18e-g)) that $G'(x) > 0$ and $G''(x) < 0$. Thus, the TC1 and TC2 heuristics are satisfied by MVD. The TDC heuristic is satisfied with the IDF factor in Eq. (18a)), while LNC1 is satisfied with the length normalization in Eq. (18d)). Similar to the discussion in B.2, by the work of Goswami et al. [28], the LNC2 and TF-LNC heuristics are also satisfied by MVD.

ACKNOWLEDGMENTS

We would like to thank Jiaul H. Paik for his help and sharing his code for parameter estimation for the Frechet distribution in the MVD model. This work was supported in part by the HK PolyU project P0030932.

REFERENCES

[1] Amati, G., Carpineto, C., and Romano, G. Comparing weighting models for monolingual information retrieval. In *Workshop of the Cross-Language Evaluation Forum for European Languages* (2003), pp. 310–318.

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

[3] Armstrong, T. G., Moffat, A., Webber, W., and Zobel, J. Improvements that don't add up: ad-hoc retrieval results since 1998. In *Proceedings of the 18th ACM Conf. on Information and Knowledge Management (CIKM'09)* (2009), pp. 601–610.

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

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

[6] Bennett, G., Scholer, F., and Uitdenbogerd, A. A comparative study of probabilistic and language models for information retrieval. In *Proceedings of the 19th Australasian Database Conference (ADC2008)* (2008), pp. 65–74.

[7] Büttcher, S., Clarke, C. L. A., and Soboroff, I. The trec 2006 terabyte track. In *Proceedings of the 2006 Text Retrieval Conference* (2006), Gaithersburg, MD: NIST Special Publication 500-272.

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

[9] Chung, T., Luk, R., Wong, K., Kwok, K., and Lee, D. Adapting pivoted document-length normalization for query size: experiments in chinese and english. *ACM Transactions on Asian Language Information Processing 5*, 3 (2006), 245–263.

[10] Clarke, C. L., Craswell, N., and Voorhees, E. M. Overview of the trec 2012 web track. In *Proceedings of the Twenty-first Text REtreival Conference (TREC-2012)* (2012), Gaithersburg, MD: NIST Special Publication 500-298.

[11] Clinchant, S., and Gaussier, E. Information-based models for ad hoc ir. In *Proceedings of the 33rd Ann. Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval* (2010), pp. 234–241.

[12] Clinchant, S., and Perronnin, F. Aggregating continuous word embeddings for information retrieval. In *Proceedings of the workshop on continuous vector space models and their compositionality* (2013), pp. 100–109.

[13] Cooper, W., Chen, A., and Gey, F. Full text retrieval based on probabilistic equations with coefficients fitted by logistic regression. In *Proceedings of the TREC-2 Conference* (1994), Gaithersburg, MD: NIST Special Publication SP, pp. 57–66.

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

[15] Cormack, G. V., Smucker, M. D., and Clarke, C. L. A. Efficient and effective spam filtering and re-ranking for large web datasets. *Information Retrieval 14*, 5 (2011), 441–465.

[16] Cummins, R., Paik, J. H., and Lv, Y. A pólya urn document language model for improved information retrieval. *ACM Transactions on Information Systems 33*, 4, Article 21 (2015), 1–34.

[17] Dai, S., Diao, Q., and Zhou, C. Performance comparison of language models for information retrieval. In *Proceedings of IFIP International Conference on Artificial Intelligence Applications* (2005), pp. 721–730.

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

[19] Dang, E. K. F., Wu, H. C., Luk, R. W. P., and Wong, K. F. Building a framework for the probability ranking principle by a family of expected weighted rank. *ACM Transactions on Information Systems 27*, 4, Article 20 (2009), 1–37.

[20] Dinçer, B. T. Irra at trec 2012: Divergence from independence (dfi). In *Proceedings of the Twenty-first Text REtreival Conference (TREC-2012)* (2012), Gaithersburg, MD: NIST Special Publication 500-298.

[21] Dinçer, B. T., Kocabaş, I., and Karaoğlan, B. Irra at trec 2010: Index term weighting by divergence from independence model. In *Proceedings of the Nineteenth Text REtreival Conference (TREC-2010)* (2010), Gaithersburg, MD: NIST Special Publication 500-294.

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

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

[24] Fang, H., Tao, T., and Zhai, C. Diagnostic evaluation of information retrieval models. *ACM Transactions on Information Systems 29*, 2, Article 7 (2011), 1–42.

[25] Fang, H., and Zhai, C. An exploration of axiomatic approaches to information retrieval. In *Proceedings of the 28th Ann. Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval* (2005), pp. 480–487.

[26] Fuhr, N. Some common mistakes in ir evaluation, and how they can be avoided. *ACM SIGIR Forum 51*, 3 (2017), 32–41.

[27] Ganguly, D., Roy, D., Mitra, M., and Jones, G. J. Word embedding based generalized language model for information retrieval. In *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval* (2015), pp. 795–798.

[28] Goswami, P., Gaussier, E., and Amini, M.-R. Exploring the space of information retrieval term scoring functions. *Information Processing and*

Management 53 (2017), 454–472.

[29] Guo, J., Fan, Y., Ai, Q., and Croft, W. B. A deep relevance matching model for ad-hoc retrieval. In *Proceedings of the 25th ACM international on conference on information and knowledge management* (2016), pp. 55–64.

[30] Guo, J., Fan, Y., Pang, L., Yang, L., Ai, Q., Zamani, H., Wu, C., Croft, W. B., and Cheng, X. A deep look into neural ranking models for information retrieval. *Information Processing and Management 57*, 6 (2020), 102067.

[31] Gupta, S., Kutlu, M., Khetan, V., and Lease, M. Correlation, prediction and ranking of evaluation metrics in information retrieval. In *Proceedings of the European Conference on Information Retrieval (ECIR 2019)* (2019), pp. 636–651.

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

[33] He, B., and Ounis, I. A study of the dirichlet priors for term frequency normalization. In *Proceedings of the 28th Ann. Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval* (2005), pp. 465–471.

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

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

[36] Huston, S., and Croft, W. B. A comparison of retrieval models using term dependencies. In *Proceedings of the 23rd ACM Conf. on Information and Knowledge Management (CIKM'14)* (2014), pp. 111–120.

[37] Jian, F., Huang, J. X., Zhao, J., and He, T. A new term frequency normalization model for probabilistic information retrieval. In *Proceedings of the 41st Ann. Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval* (2018), pp. 1237–1240.

[38] Jian, F., Huang, J. X., Zhao, J., He, T., and Hu, P. A simple enhancement for ad-hoc information retrieval via topic modelling. In *Proceedings of the 39th Ann. Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval* (2016), pp. 733–736.

[39] Jian, F., Huang, J. X., Zhao, J., Ying, Z., and Wang, Y. A topic-based term frequency normalization framework to enhance probabilistic information retrieval. *Computational intelligence 36*, 2 (2020), 486–521.

[40] Kamphuis, C., de Vries, A. P., Boytsov, L., and Lin, J. Which bm25 do you mean? a large-scale reproducibility study of scoring variants. In *Proceedings of the European Conference on Information Retrieval (ECIR 2020)* (2020), pp. 28–34.

[41] Khankasikam, K. A comparison of information retrieval models applied to thai digital library. In *The 2nd International Conference on Computer and Automaton Engineering (ICCAE)* (2010), pp. 335–338.

[42] Kharazmi, S., Scholer, F., Vallet, D., and Sanderson, M. Examining additivity and weak baselines. *ACM Transactions on Information Systems 34*, 4, Article 23 (2016), 1–18.

[43] Kocabaş, I., Dinçer, B. T., and Karaoğlan, B. A nonparametric term weighting method for information retrieval based on measuring the divergence from independence. *Information Retrieval 17* (2014), 153–176.

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

[45] Kwok, K. A network approach to probabilistic information retrieval. *ACM Transactions on Information Systems*, 13 (1996), 325–353.

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

[47] Lafferty, J., and Zhai, C. X. Probabilistic relevance models based on document and query generation. In *Language modeling for information retrieval* (2001), Dordrecht: Springer, pp. 1–10.

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

[49] Lavrenko, V., and Croft, W. B. Relevance-based language model. In *Proceedings of the 24th Ann. Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval* (2001), pp. 120–127.

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

[51] Lee, D., Chuang, H., and Seamons, K. Document ranking and the vector-space model. *IEEE Software 14*, 2 (1997), 67–75.

[52] Li, Y., Luk, R. W. P., Ho, E. K. S., and Chung, F. L. Improving weak ad-hoc queries using wikipedia asexternal corpus. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval* (2007), pp. 797–798.

[53] Losee, R. M. Comparing boolean and probabilistic information retrieval systems across queries and disciplines. *Journal of the American Society for Information Science 48*, 2 (1997), 143–156.

[54] Luk, R., and Kwok, K. A comparison of chinese document indexing strategies and retrieval models. *ACM Transactions on Asian Language Information Processing 1*, 3 (2002), 225–268.

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

[56] Luk, R. W. P. Why is information retrieval a scientific discipline. *Foundations of Science* (2021), To appear.

[57] Lv, Y., and Zhai, C. X. A comparative study of methods for estimating query language models with pseudo feedback. In *Proceedings of the 18th ACM Conf. on Information and Knowledge Management (CIKM'09)* (2009), pp. 299–306.

[58] Lv, Y., and Zhai, C. X. Lower-bounding term frequency normalization. In *Proceedings of the 20th ACM Conf. on Information and Knowledge*

Management (CIKM'11) (2011), pp. 7–16.

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

[60] Mitra, B., Nalisnick, E., Craswell, N., and Caruana, R. A dual embedding space model for document ranking. *arXiv preprint arXiv:1602.01137* (2016).

[61] Montague, M., and Aslam, J. A. Condorcet fusion for improved retrieval. In *Proceedings of the 11th ACM Conf. on Information and Knowledge Management (CIKM'02)* (2002), pp. 538–548.

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

[63] Paice, C. Soft evaluation of boolean search queries in information retrieval systems. *Information Technology: Research and Development 3*, 1 (1984), 33–42.

[64] Paik, J. H. A novel tf-idf weighting scheme for effective ranking. In *Proceedings of the 36th Ann. Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval* (2013), pp. 343–352.

[65] Paik, J. H. A probabilistic model for information retrieval based on maximum value distribution. In *Proceedings of the 38th Ann. Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval* (2015), pp. 585–594.

[66] Pohl, S., Moffat, A., and Zobel, J. Efficient extended boolean retrieval. *IEEE Transactions on Knowledge and Data Engineering 24*, 6 (2012), 1014–1024.

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

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

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

[70] Raviv, H., Kurland, O., and Carmel, D. Document retrieval using entity-based lanugage models. In *Proceedings of the 39th Ann. Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval* (2016), pp. 65–74.

[71] Robertson, S., Walker, S., and Beaulieu, M. Okapi at trec-7: automatic ad hoc, filtering, vlc and interactive track. In *Proceedings of the Seventh Text REtreival Conference (TREC-7)* (1999), Gaithersburg, MD: NIST Special Publication 500-242, pp. 253–264.

[72] Robertson, S., Walker, S., Jones, S., Hancock-Beauleiu, M., and Gatford, M. Okapi at trec-3. In *Proceedings of the Third Text REtreival Conference (TREC-3)* (1995), Gaithersburg, MD: NIST Special Publication 500-226, p. 109.

[73] Robertson, S., and Zaragoza, H. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends in Information Retrieval 3*, 4 (2009), 333–389.

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

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

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

[77] Salton, G., and Buckley, C. Term-weighting approaches in automatic text retrieval. *Information Processing and Management 24*, 5 (1988), 513–523.

[78] Savoy, J. Comparative study of monolingual and multilingual search models for use with asian languages. *ACM Transactions on Asian Language Information Processing 4*, 2 (2005), 163–189.

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

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

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

[82] Tao, T., and Zhai, C. X. Regularized estimation of mixture models for robust pseudo-relevance feedback. In *Proceedings of the 29th Ann. Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval* (2006), pp. 162–169.

[83] Trotman, A., Puurula, A., and Burgess, B. Improvements to bm25 and language models examined. In *Proceedings of the 2014 Australasian Document Computing Symposium* (2014), p. 58.

[84] Turtle, H. R., and Croft, W. B. A comparison of text retrieval models. *The Computer Journal 35*, 3 (1992), 279–290.

[85] Waller, W. G., and Kraft, Donald, H. A mathematical model of a weighted boolean retrieval system. *Information Processing and Management 15* (1979), 235–245.

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

[87] Xiong, C., and Callan, J. Query expansion with freebase. In *Proceedings of the 2015 ACM International Conference on the Theory of Information Retrieval* (2015), pp. 111–120.

[88] Xiong, C., Dai, Z., Callan, J., Liu, Z., and Power, R. End-to-end neural ad-hoc ranking with kernel pooling. In *Proceedings of the 40th International ACM SIGIR conference on research and development in information retrieval* (2017), pp. 55–64.

[89] Xu, Y., Jones, G. J. F., and Wang, B. Query dependent pseudo-relevance feedback based on wikipedia. In *Proceedings of the 32nd international ACM*

SIGIR conference on Research and development in information retrieval (2009), pp. 59–66.

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

[91] Yang, P., Fang, H., and Lin, J. Anserini: Reproducible ranking baselines using lucene. *ACM Journal of Data and Information Quality 10*, 4, Article 16 (2018), 1–20.

[92] Zhai, C. X., and Lafferty, J. Model-based feedback in the language modeling approach to information retrieval. In *Proceedings of the 10th ACM Conf. on Information and Knowledge Management (CIKM'01)* (2001), pp. 403–410.

[93] Zhai, C. X., and Lafferty, J. Two-stage language models for information retrival. In *Proceedings of the 25th Ann. Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval* (2002), pp. 49–56.

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