Weighting Passages Enhances Accuracy

CRISTINA IOANA MUNTEAN, ISTI-CNR, Italy
FRANCO MARIA NARDINI, ISTI-CNR, Italy
RAFFAELE PEREGO, ISTI-CNR, Italy
NICOLA TONELLOTTO, University of Pisa, Italy
OPHIR FRIEDER, Georgetown University, U.S.A.

We observe that in curated documents the distribution of the occurrences of salient terms, e.g., terms with a high Inverse Document Frequency, is not uniform, and such terms are primarily concentrated towards the beginning and the end of the document. Exploiting this observation, we propose a novel version of the classical BM25 weighting model, called BM25 Passage (BM25P), which scores query results by computing a linear combination of term statistics in the different portions of the document. We study a multiplicity of partitioning schemes of document content into passages and compute the collection-dependent weights associated with them on the basis of the distribution of occurrences of salient terms in documents. Moreover, we tune BM25P hyperparameters and investigate their impact on ad-hoc document retrieval through fully reproducible experiments conducted using four publicly available datasets. Our findings demonstrate that our BM25P weighting model markedly and consistently outperforms BM25 in term of effectiveness by up to 17.44% in NDCG@5 and 85% in NDCG@1, and up to 21% in MRR.

CCS Concepts: • **Information systems → Probabilistic retrieval models**.

Additional Key Words and Phrases: Passage Retrieval, Weighting Models, Salient terms, BM25P, Evaluation.

ACM Reference Format:
Cristina Ioana Muntean, Franco Maria Nardini, Raffaele Perego, Nicola Tonellotto, and Ophir Frieder. 2020. Weighting Passages Enhances Accuracy. *ACM Transactions on Information Systems* 1, 1, Article 1 (January 2020), 25 pages. https://doi.org/10.1145/3428687

1 INTRODUCTION

Passage retrieval, the task of retrieving portions of documents, i.e., passages, relevant to a particular information need, prevails for decades in the information retrieval literature. Since shorter texts tend to exhibit higher homogeneity, retrieving passages instead of complete documents can boost effectiveness for some retrieval tasks [17]. At times, passage retrieval is viewed as an intermediate step in other information retrieval tasks, like question answering and summarization, and often concentrates on devising effective heuristics to split documents in passages [39].

Believing that certain passages pose greater relevance to a given query and that the distribution of salient terms in the content of curated documents is collection dependent, we investigate how passage relevance can be exploited to improve effectiveness on domain-specific retrieval tasks, i.e., for news and web documents collections. Our approach differs substantially from both existing passage retrieval [39] and passage detection [30] efforts, where the aim is to

Work partially supported by the Italian Ministry of Education and Research (MIUR) in the framework of the CrossLab project (Departments of Excellence). Work partially supported by the BIGDATAGRAPES project funded by the EU Horizon 2020 research and innovation programme under grant agreement No. 780751, and by the OK-INSAID project funded by the Italian Ministry of Education and Research (MIUR) under grant agreement No. ARS01_00917.
Authors' addresses: Cristina Ioana Muntean, ISTI-CNR, Italy, cristina.muntean@isti.cnr.it; Franco Maria Nardini, ISTI-CNR, Italy, f.nardini@isti.cnr.it; Raffaele Perego, ISTI-CNR, Italy, r.perego@isti.cnr.it; Nicola Tonellotto, University of Pisa, Italy, nicola.tonellotto@unipi.it; Ophir Frieder, Georgetown University, U.S.A., ophir@ir.cs.georgetown.edu.

either retrieve only highly relevant passages or detect unrelated injected passages from within documents, respectively. In contrast, we focus on improving the effectiveness of a retrieval system in retrieving entire documents. To this end, we exploit passage relevance capitalizing on their keyword density and demonstrate experimentally how this emphasis significantly improves overall retrieval.

Specifically, we introduce a variant of the well-known BM25 weighting model [36], called BM25 Passage (BM25P), improving retrieval effectiveness on different collections of plain, unstructured documents. BM25P takes into account the entire document when assigning a relevance score; however, it distinguishes the importance of different passages by assigning them different weights. BM25P exploits such portions of text by creating a weighted linear combination of term frequencies per passage, improving retrieval effectiveness. To derive the weights, we analyze the density of highly salient terms in the collection of documents, measured in term of Term Frequency (TF), Inverse Document Frequency (IDF) and KL divergence score [5] and observe where they are distributed throughout the content of the documents. This approach is efficient since it is query independent and passage weights are computed at index construction time, i.e., pre-retrieval. We note that BM25P can resemble BM25F where different weights are associated with per-field term statistics [37, 38]. However, BM25F applies to strictly structured documents, while BM25P is likewise applicable to plain, unstructured documents by superimposing onto them a weak structure defined on the basis of the collection-dependent density of salient terms. More so, we show that by exploiting the specific relevance of salient passages, our BM25P weighting model outperforms BM25 on the ad-hoc retrieval task on collections of curated unstructured documents.

We extend a previous contribution aimed at investigating how news documents are structured and where salient terms predominantly appear based on the presence of high IDF terms in passages [10]. There, we introduced BM25P and showed preliminary results of its effectiveness measured on news collections. Furthering those results, we provide here an exhaustive analysis of our weighting model and of the impact of its hyper-parameters on the retrieval performance. We investigate the use of our passage weighting approach in combination with other probabilistic weighting models and different heuristics for the identification of salient terms in documents. To comprehensively assess our approach and all its variations, we present a reproducible evaluation based on four publicly available datasets: `Signal`, `RCV1`, `Aquaint` and `MS-MARCO`. While the first three datasets are collections of news, the fourth is a large scale dataset focused on machine reading comprehension and question answering.

In summary, our contributions are:

- An efficient query-independent method to characterize collections of unstructured documents by the distribution of salient terms in their content. Applied at index construction time, the method is unsupervised and allows to superimpose a weak structure based on keyword densities to plain, unstructured documents;
- BM25P, a novel weighting model that exploits the varying importance of document passages by considering such superimposed structure; we study different ways to partition the document content into portions and to compute the collection-dependent weights associated to them;
- An in-depth analysis of the solution space of our BM25P weighting model showing consistent results across different collections and configurations of its hyper-parameters;
- An exhaustive and reproducible assessment of the effectiveness of BM25P conducted on four publicly available datasets: `Signal`, `RCV1`, `Aquaint` and `MS-MARCO`.
- An experimental investigation of the impact of the proposed passage weighting approach on two additional probabilistic weighting model families, namely, language models (LM) and models based on the divergence from randomness (DFR).

The remainder of the paper is structured as follows: Section 2 discusses related work, while Section 3 analyzes how salient terms predominantly appear in passages and introduces BM25P. Section 4 details our research questions, introduces the experimental settings, and the experimental methodology. The results of our comprehensive evaluation conducted to answers the research questions are discussed in section 5. Finally, we conclude our investigation in Section 6.

2 RELATED WORK

Term Proximity. Many retrieval weighting models were proposed: vector space models, probabilistic models, statistical language models, etc [12]. In all these models, documents are characterized by single term relevance signals w.r.t. a given query. Such relevance signals include in-document term frequency and inverse document frequency. Nevertheless, important relevance signals can be obtained by exploiting the proximity of query terms in a document [43]. Proximity signals reward a document where the matched query terms occur near each other. Several works propose to heuristically incorporate proximity into an existing retrieval model, e.g., through score combinations [6, 31, 35]. Different proximity measures were proposed, e.g., minimum term span, minimum pairwise distance, etc. Tao and Zhai comprehensively examined different measures and concluded that the minimum pairwise distance is most effective [43]. An indirect way to capture proximity is to use high-order n-grams as units to represent text. An early work in this line is the one by Song and Croft, where they introduce bigram and trigram language models and show that they outperform simple unigram language models [42]. More recent work argues that two related retrieval heuristics remains "external" to language modeling: i) proximity heuristics and ii) passage retrieval, which scores a document mainly based on the best matching passage. Lv and Zhai propose a novel positional language model (PLM) which implements both heuristics in a unified language model. The authors define a language model for each position of a document and score a document based on PLM [26]. More recently, He *et al.* improve the Okapi BM25 model by utilizing the term proximity information showing that term proximity enhances the retrieval effectiveness of probabilistic models [16]. Many weighting models also incorporate relevance signals based on the number of co-occurrences of word pairs and their distance in the document [2, 16, 35, 43]. Independently of the proximity weighting model employed, query processing with proximity support requires indexing and storing positional information about term occurrences in the documents. Moreover, the computation of positional features used in the proximity weighting models is expensive in terms of query processing time [29].

Document Structure. While query term proximity in documents can be viewed as fine-grain positional information, document structure provides relevance signals at a coarser grain. Abstracts in scientific papers or titles in Web documents carry summarization information about the contents of the corresponding document [47]. Web and XML documents can be structured via hyper-textual markup language tags; hence, different importance weights can be assigned to different portions of the documents to boost the relevance signals in such portions [14, 51]. Assigning weights to different portions of structured documents was investigated in several works. These weights can be specified by users during query submission [40], even though this process is difficult and can lead to effectiveness reductions w.r.t. corresponding unweighted relevance models [34]. Automatic weight assignment procedures have proved successful in improving effectiveness [34, 50]. Géry *et al.* [15] focus on XML documents and exploit their logical structure (e.g., title, section, paragraph, etc.) as well as other formatting tags (e.g., bold, italic, center, etc.). They take into account the influence of a tag by estimating the probability for this tag to distinguish relevant terms from the others. Then, the relevant terms' weights are integrated in a BM25-like weighting function.

Document Fields. Among all existing relevance models, the *Probabilistic Relevance Framework* [38] led to the development of the BM25 weighting model. BM25 is one of the most commonly deployed weighting models in both academic and industrial information retrieval systems. Despite its simplicity, the BM25 relevance model is still a strong baseline even for advanced neural IR methods [22]. In particular, a finely-tuned BM25 weighting model used in conjunction with pseudo-relevance feedback obtains effectiveness metrics such as average precision and precision at 20 on par with competitive neural relevance models. However, BM25 fails to deal with structured documents. The BM25 relevance signal for individual Web document fields (e.g., title and body) can be computed independently and then combined (typically linearly) to arrive at a final relevance score for the whole document. This method can lead to poor performance due to non-linear saturation of term frequency in the BM25 weighting model. Robertson *et al.* [37] describe a simple way of adapting the BM25 weighting model to deal with structured documents and avoid the term frequency saturation issue. They propose to combine weighted term frequencies before the non-linear term frequency saturation function is applied. This approach, called BM25F, is the de-facto standard for efficient search systems on multi-field Web documents [38].

Document Passages. We mainly focus on news documents that lack a priori subdivision into fields. Hence, we resort to passages composing the documents to obtain proxies for fields. A large body of IR work focuses on passages and their retrieval. It is a well-studied research field originating by pioneering works dating back to the early 1990s [7, 39]. The key observation is that passages can serve as effective proxies for ranking documents. Some works assume the relevance of a document to a query coincides with the highest relevance score among its composing passages w.r.t. the same query, while others combine linearly this score with the overall document relevance score [3, 7, 25, 50, 52]. More recently, learning-to-rank algorithms are deployed in many information retrieval systems [8, 18, 46] to rank documents. These algorithms rely on query-document features to compute the final relevance score of a document w.r.t. the query. Such features can be based on passage-query similarities [49]. Other recent approaches exploit data-driven methods to allow relevance signals at different granularities, e.g., passage, document, etc., to compete with each other for final relevance assessment [13]. A different line of research defines novel retrieval models that actively exploit user reading behaviors showing that user-based reading heuristics have positive impacts on retrieval performance [20, 21]. Our work does not investigate the use of passages for inferring new relevance models. We focus instead on the identification of passages as different sources of relevance signals and on their exploitation in the framework of BM25.

Liu *et al.* [24] propose a BM25-based retrieval model rewarding terms based on their location in passages. They assume that the most salient terms in the documents are near the beginning or the end of the sentence, and they weight a term importance w.r.t. its distance from the middle of the sentence. Authors thus extend BM25 with term location information and, to reward terms that are more likely to be nouns, they propose a kernel-based term location retrieval model to capture term placement patterns. This work is different from our approach as we are proposing a novel weighting model, i.e., BM25P, that exploits the varying importance of document passages by considering a superimposed distribution of salient terms in their content. This is not true for the approach of Liu *et al.* as they their focus is on a sentence level, by assuming different term importance when varying the distance from the middle of the sentence. Zhao *et al.* [54] study the effect of rewarding terms according to their locations in documents. They propose BM25-RT, a BM25 improvement where the influence of terms appearing in the initial part of documents is boosted with several shape functions. The weights assigned to the single terms depend only on the re-scaling of the original term weight produced by the chosen function, without taking into account the single term salience w.r.t. the document. They experiment using small web and blog collections and report limited effectiveness improvements. Differently from our work they focus on the initial part of documents only, without investigating the collection-specific distribution of salient terms.

Capitalizing on these observations, we analyze the distribution of the occurrences of highly relevant terms and note that documents belonging to different collections are consistently characterized by areas with different densities of highly relevant terms. We thus exploit this fact to improve the BM25 weighting model.

3 RANKING WITH PASSAGES

Given a user query, a *relevance score* is associated with the indexed documents matching the query. Such relevance score is computed by exploiting a heuristic query-document similarity function, estimating the probability of the document being relevant for the query. Then, the documents retrieved are ranked by their relevance score, and the top documents with the highest scores are returned to the user.

The BM25 scoring function is among the most successful query-document similarity functions, whose roots lie in the *Probabilistic Relevance Framework* [36]. In most information retrieval systems, the relevance score $s_q(d)$ for a document d given a query q follows the general outline given by the best match strategy:

$$s_q(d) = \sum_{t \in q} s_t(q, d), \tag{1}$$

where $s_t(q, d)$ is a term-document similarity function that depends on the number of occurrences of term t in document d and query q, on other document statistics such as document length, and on collection-wide term statistics such as the inverse document frequency (IDF). In particular, in the BM25 weighting model, the relevance score $s_t(q, d)$ is given by:

$$\text{BM25}_t(q, d) = w_q \frac{(k_1 + 1)tf}{k_1 \left((1 - b) + b \frac{dl}{avg\_dl}\right) + tf} w_{IDF}, \tag{2}$$

where dl is the document length, tf is the in-document term frequency, i.e., the ratio between the number $c(t, d)$ of occurrences of term t in document d and dl, avg_dl is the average document length of the collection, w_q is a query-only weight, b and k_1 are parameters (defaults $b = 0.75$, $k_1 = 1.2$). The w_{IDF} component is the IDF factor, which is given by $w_{IDF} = \log \frac{N - N_t + 0.5}{N_t + 0.5}$, where N is the number of documents in the collection, and N_t is the document frequency of term t.

When taking into account the fields that make up a document (e.g., title, headings, abstract and body), each field may be treated as a separate collection of (unstructured) documents over the whole collection, and the relevance score of a document can be computed as a weighted linear combination of the BM25 scores over the individual fields. However, in [37] the authors experimentally showed that such a linear combination of scores has several drawbacks, such as breaking the tf saturation after a few occurrences (a document matching a single query term over several fields could rank higher than a document matching several query terms in one field only), or affecting the document length parameter (when the document length is referred to the actual field weight rather than the whole document). Hence, the authors suggested the BM25F weighting model for structured documents, computing a weighted linear combination of field-based term frequencies and then plugging that combination into the BM25 weighting model (2). The novel tf factor boosts the specific fields without altering collection statistics. The BM25F model is largely adopted in Web search and corporate search technologies, when the document collection is composed by documents with a clear field structure [38].

With unstructured documents such as news or plain texts, relevance signals derived from the term frequency of the query keywords in the different fields are lacking. However, such documents can be divided into arbitrary sections of say a given number of terms or a given percentage of the whole content. The identification of such sections in a document can be carried out automatically, but there is no clear evidence that sections arbitrarily imposed on a generic unstructured document can provide any strong relevance signal.

We hypothesize that in *curated* unstructured documents it is possible to leverage the distribution of keywords in the documents to derive analogous strong relevance signals. To validate our hypothesis on the presence of a weak structure hidden even in unstructured documents and to quantify the impact of some distinguishing document portions (referred to as passages in the following) over other portions, we analyze the density of highly discriminative terms in large document corpora. Several heuristics identify discriminative terms in documents, hereinafter *salient terms*. Intuitively, a salient term appears in only a few documents, i.e., salient terms have a high IDF value. Alternatively, we weight the high IDF terms in a document by their term frequencies, i.e., salient terms have a high TFIDF value. In this Section we illustrate the distribution of the position of the occurrences of salient terms using the IDF heuristics. In Section 5.2 we investigate alternative ways of identifying the top salient terms by using TFIDF or KL divergence score heuristics.

For each document in our four test collections (detailed in Sec. 4), we identify the positions of the occurrences of the top k salient terms (k is a given hyperparameter). Given a document d, we denote with $T^{\uparrow}(d)$ the set of the collection's salient terms occurring in the document. To aggregate such positional information, we evenly split each document d into a set of P passages having about the same length. Then, we compute the distributions of the occurrences of the salient terms (identified with the IDF heuristic) in each of these passages. Finally, we average these values over the entire dataset, giving the distributions shown as heatmaps in Fig. 1 for the top k salient terms, with $k = 5, 10, 15$ and $P = 10$. As demonstrated, for all datasets considered, the first and last parts of collection's documents are more likely to include salient terms than the remaining parts. Moreover, the lower the number k of the top salient terms considered, the more skewed is the probability distribution.

Focusing on curated documents such as news, several news writing guides highlight the need of engaging the reader instantly and summarizing what the story is all about in the opening sentences. The thumbnail rule states that the first sentence(s) should contain all of the *who*, *what*, *when*, *where*, *why* and *how* of the news[1]. On the other hand, no specific rule for closing the news articles is given in writing guides, and the very high likelihood observed even for the last part of the news articles is surprising. Moreover, slight differences in the probabilities are apparent even for the middle passages. Such analysis motivated us to investigate if exploiting this probability distribution, by weighting differently these areas in the documents, can enhance retrieval effectiveness.

Hence, we propose a variant of BM25 called BM25P which uses different weights for the different passages. We divide each document into P passages (P being a hyperparameter), and our proposed BM25P model computes a linear combination tf_P of the term frequencies tf_i in each passage i of the P passages in document d (re-scaled by the hyperparameter α):

$$tf_P = \alpha \sum_i w_i \cdot tf_i. \tag{3}$$

[1] http://handbook.reuters.com.

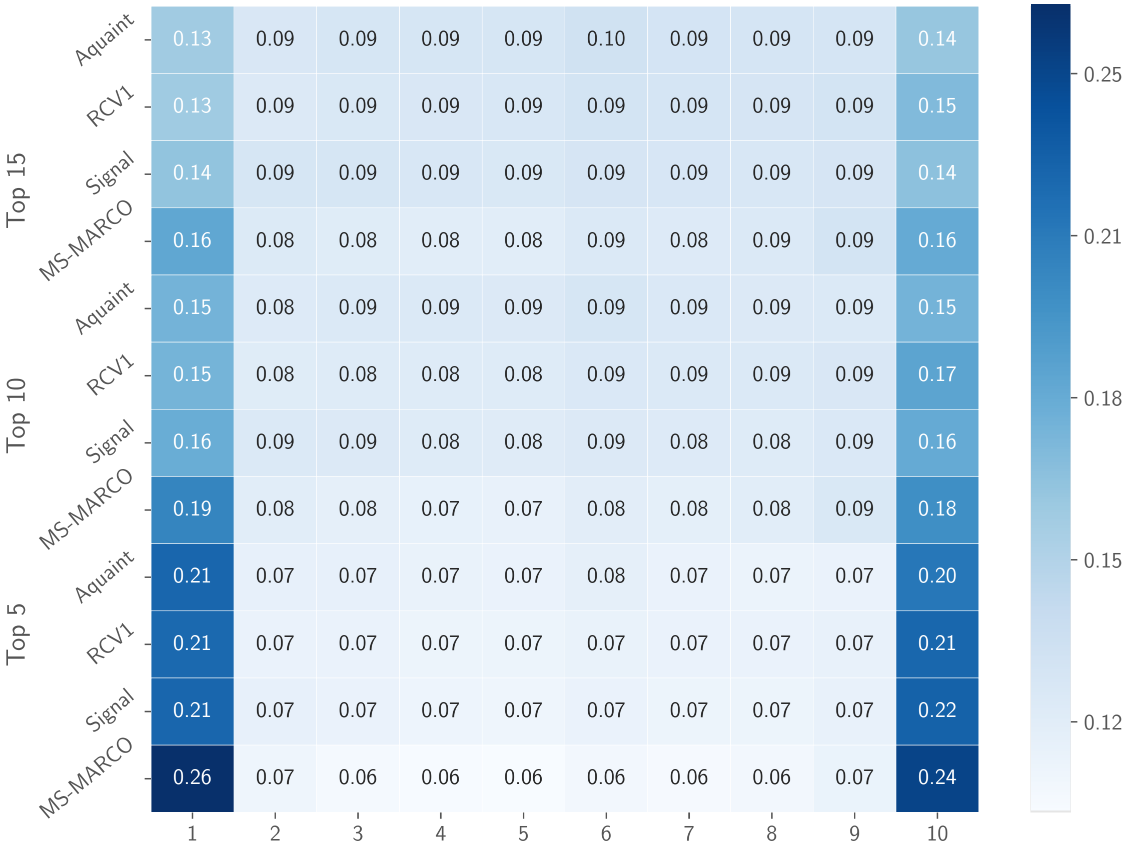


Fig. 1. Probability distribution for the positions of salient terms occurrences in the documents for each of the four collections used.

As suggested in [37] we plugged the term frequency tf_P into the original BM25 formula (Eq. (2)), rather than summing the BM25 scores per passage:

$$\mathrm{BM25P}_t(q,d) = w_q \frac{(k_1+1)tf_P}{k_1\left((1-b)+b\frac{dl}{avg\_dl}\right)+tf_P} w_{IDF}, \tag{4}$$

The empirical probability distribution depicted in Fig. 1 clearly indicates the impact of each passage within the document from the point of view of salient terms. This probability distribution is used to compute the term frequency weights: w_i is directly proportional to the probability distribution of important terms in the i-th passage. Specifically, w_i is computed as the ratio between the number of occurrences of salient terms in the i-th passage P_i and the total number of occurrences of salient terms in the whole document d, i.e.,

$$w_i = \frac{\sum_{\tau\in T^{\uparrow}(d)} c(\tau,P_i)}{\sum_{\tau\in T^{\uparrow}(d)} c(\tau,d)}. \tag{5}$$

We re-scale all weights with the hyperparameter α to amplify the importance of salient terms in impactful passages. Note that BM25P with all passage weights and α set to 1 is equivalent to BM25.

Table 1. Table of symbols.

Symbol	Definition
t	Term
q	Query
d	Document
$T^{\uparrow}(d)$	Salient terms in document d
$s_d(q)$	Relevance score for query q and document d
$s_t(q,d)$	Relevance score for term t in query q and document d
w_q	Query-only weight
k_1, b	BM25 parameters
tf	BM25 in-document term frequency
$c(t,d)$	Number of occurrences of term t in document d
$M_d(t)$	Occurrence probability of term t in the document d's language model
$M^*(t)$	Occurrence probability of term t in the collection's language model
dl	Document length
avg_{dl}	Average document length in the collection
w_{IDF}	IDF weight factor in BM25
N	Number of documents in the collection
N_t	Document frequency of term t
k	Number of salient terms*
P	Number of passages in a document*
P_i	i-th passage in a document
tf_i	Term frequency in the i-th passage
w_i	Weight of the i-th passage
α	BM25P rescaling factor*
tf_P	BM25P in-document term frequency

* denotes a BM25P hyperparameter.

For clarity, Table 1 summarizes all notation used herein.

4 EXPERIMENTAL EVALUATION

We evaluate the BM25P model in operational scenarios. According to our hypothesis, we aim to experimentally assess if assigning different weights to different portions of unstructured documents can enhance the overall retrieval effectiveness. As previously discussed, we cannot compare BM25P with BM25F since our focus is on unstructured documents where text fields are not available. Our reference baseline is thus the BM25 weighting model, still a strong baseline even for most state-of-the-art neural ranking models [22].

In detail, our experiments address the following research questions:

- **RQ1**: Are the first and last passages the most salient?
- **RQ2**: Is the effectiveness of BM25P superior to BM25?
  - A. What is the impact of weighting differently the passages based on the distribution of salient terms?
  - B. How can salient terms be identified?
  - C. What is the impact of the number of passages?
  - D. What is the contribution of each passage?
  - E. Do all passages contribute to some extent?
- **RQ3**: How do different passage segmentation strategies impact BM25P?

- **RQ4**: Does BM25P work for types of documents other than news, e.g., Web documents?
- **RQ5**: How does passage weighting perform when used with other weighting models?

4.1 Datasets

We investigate the research questions by assessing the effectiveness of the BM25P weighting model on four document corpora, three composed of English news articles and one composed of Web documents:

- the AQUAINT Corpus by Linguistic Data Consortium (Aquaint): it includes about one million news articles and 50 manually assessed queries with the correspondent relevance judgments used in 2005 Robust and HARD TREC tracks;
- the Signal Media One-Million News Articles Dataset [11] (Signal): it contains about one million documents collected in September 2015 from different news sources, which include major ones, like Reuters, in addition to local news sources and blogs;
- the Reuters Corpus, Volume 1, version 2 [19] (RCV1): it contains about ~800,000 newswire stories produced by Reuters spanning over a whole year;
- the MS-MARCO Corpus [32] (MS-MARCO): a large-scale dataset focused on machine reading comprehension, question answering, and passage ranking, including about 3.5 millions Web documents and ~5,000 manually assessed queries.

Note that Aquaint provides for each one of the 50 queries both a query topic title and a description. In our tests we used just the topic title field and ignored the description. Signal and RCV1 datasets instead do not provide any evaluation data, i.e., manually assessed queries. Hence, for these two datasets, we adopt the methodology described in [27] and use the news titles as *pseudo-queries*. According to this methodology, there is only one relevant news article for each query, i.e., the article to which the title belongs to. All other articles of the collection are considered to be non-relevant. For each of these two datasets we randomly selected $40,000$ documents to generate the same number of pseudo-queries for each collection. Statistics for the four datasets are summarized in Table 2.

Table 2. Statistics of the four collections used in the experiments.

Dataset	# Queries	avg. Query Length	# Documents	avg. Document Length
Aquaint	50	2.60	1,033,000	249.42
Signal	40,000	6.64	1,000,000	224.22
RCV1	40,000	5.77	804,000	147.38
MS-MARCO	5,193	5.89	3,563,535	671.58

4.2 Experimental Methodology

We removed titles and all collection-specific fields (e.g., source, category, media type, publishing date) from all documents in the datasets and index just the unstructured body of news articles into Terrier positional indices [28]. This type of index provides us with the positions of query term occurrences within the document to differently weight the contribution of matching terms.

We retrieve the top $1,000$ documents for each query from the respective corpus by using BM25 and BM25P. With BM25P, if not otherwise specified, all documents are divided into $P = 10$ passages, thus obtaining 10 passages of about the same length. The frequencies of query terms in each passage are weighted as discussed in Section 3. In the following, we use the distributions of top k salient terms, with $k = 5, 10, 15$.

Once queried, we observe the rank of the relevant documents retrieved and compare the results obtained for BM25P with the ones provided by BM25. To measure the retrieval effectiveness, we consider the NDCG and MRR metrics. NDCG is computed at different cutoffs and is used to evaluate the performance on the Aquaint dataset, where we have multiple relevant documents per query. Conversely, MRR, i.e., the mean of the reciprocal of the rank of the first relevant result, allows us to quantify how good is a given retrieval method in pushing a relevant result towards top positions of the rank, especially for the Signal, RCV1 and MS-MARCO datasets, where only one (a few queries have more than one in the case of MS-MARCO) relevant document per query is provided.

5 RESULTS AND DISCUSSION

We now discuss the results of our experiments conducted to answer the research questions posed in Section 4.

5.1 RQ1: Are the first and last passages the most salient?

Are salient terms uniformly distributed inside documents? If so, then the BM25 model perfectly weights their importance for document retrieval. However, upon further examination of their distribution (using IDF as an indicator of term saliency) we see to the contrary. By looking at the heat maps in Figure 1 we can observe a clear skew in the distribution, with a striking difference between the middle part of the document w.r.t. the first and last passages.

As previously discussed, we tried various methods for assessing the distribution of salient terms, for example we varied the number k of top IDF terms considered when searching their position in the document. We see that when using 5 terms, the distribution is more skewed towards the first and last paragraphs than it is when considering 15 terms for example. This trend is somehow expected since the larger the number of salient terms considered the higher is their collection-wide frequency. We can also see differences between datasets since each collection has its own specificity. Almost independently of the value of k considered, we see that for Aquaint and RCV1 collections, containing older news articles and editorials, the distribution is less skewed than for Signal and MS-MARCO, which are newer collections containing both news and web pages or blogs. MS-MARCO is a different kind of dataset w.r.t. the others, and this difference impacts also the distribution of salient terms. Their density is much higher in the first and last passage, 0.26 and 0.24, respectively, while passages in the middle of the document contain much less in proportion.

Although it may seem that the inner part of the document has a lot less importance, this is relative to the greater importance of the first and last passages. This does not mean that salient terms occur only in the beginning and end of the documents. When considering the effectiveness of using these weights in the BM25P model, in Section 5.2, we identify their necessity to obtain the best retrieval effectiveness.

5.2 RQ2: Is the effectiveness of BM25P superior to BM25?

In Section 5.1, we identified a different distribution of salient terms across passages within documents of different collections. Our method for exploiting such property for improving retrieval is BM25P, introduced in Section 3. To address RQ2, we assess the impact of BM25P and its effectiveness in comparison to BM25. We evaluate all the BM25P hyperparameters, namely k (the number of top IDF terms considered), α (the rescaling factor in the weighting model), and P (the number of passages), to tune for maximum performance.

A. What is the impact of weighting differently the passages? To answer this question and assess whether BM25P achieves a better overall ranking quality with respect to BM25, we conducted experiments on the Aquaint collection.

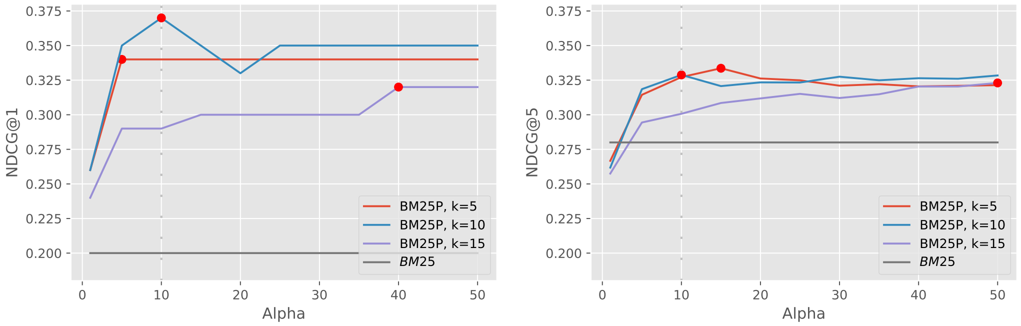


Fig. 2. NDCG@1 and NDCG@5 for BM25 and BM25P, on `Aquaint` as a function of the value of α in the range from 5 to 50.

Table 3 reports the NDCG at different cutoffs measured for BM25 and BM25P with weights computed on top 5, top 10 and top 15 salient terms. The experiment was performed by setting the remaining hyperparameters $\alpha = 10$ and $P = 10$.

Table 3. NDCG at different cutoffs for BM25 and BM25P, on the `Aquaint` collection for $k = \{5, 10, 15\}$, $\alpha = 10$, $P = 10$. We highlight statistically significant differences w.r.t. BM25 with ▲ for p-value < 0.01 and △ for p-value < 0.05 according to the two sample t-test [41].

NDCG	BM25	BM25P		
cutoffs		$k = 5$	$k = 10$	$k = 15$
1	0.200	0.340 $_{+70.00\%}$ △	**0.370** $_{+85.00\%}$ ▲	0.290 $_{+45.00\%}$ △
3	0.291	0.319 $_{+9.58\%}$	**0.335** $_{+15.01\%}$	0.317 $_{+8.78\%}$
5	0.280	0.327 $_{+16.84\%}$	**0.329** $_{+17.44\%}$ ▲	0.301 $_{+7.39\%}$
10	0.270	**0.303** $_{+12.07\%}$	0.298 $_{+10.20\%}$ △	0.291 $_{+7.44\%}$
15	0.269	0.288 $_{+7.30\%}$	**0.296** $_{+9.96\%}$ ▲	0.290 $_{+7.91\%}$ △
20	0.273	0.280 $_{+2.69\%}$	**0.289** $_{+5.81\%}$ △	0.282 $_{+3.35\%}$ △

We highlight that BM25P consistently outperforms BM25. Indeed, BM25P with $k = 10$ is the best setting for the passage weights, with improvements over BM25 that are always statistically significant apart from a single case (NDCG@3). The relative improvement ranges from 5.81% for NDCG@20 to 85% for NDCG@1. Moreover, in three of the six cases, BM25P with $k = 10$ shows statistically significant results with a p-value < 0.01. The other BM25P settings, i.e., with $k = 5$ and $k = 15$, also outperforms BM25, with improvements up to 70% and 45%, smaller than BM25P with $k = 10$, but still statistically significant in many cases.

We further investigate the performance of BM25P by also varying α to assess the impact of this hyper-parameter on the retrieval effectiveness measured in terms of NDCG@1 and NDCG@5. We present the results of this investigation in Figure 2.

Results show that, for $\alpha \geq 10$, BM25P always performs better than BM25. For BM25P with $k = 5$ and BM25P with $k = 10$, the effectiveness does not sensibly increase for α values greater than 10; it tends to remain stable with only small fluctuations, while for BM25P with $k = 15$ the performance tends to increase even if it is not able to outperform the one of BM25P with $k = 10$ for any value of α.

It is worth highlighting that, since the `Aquaint` dataset provides 50 queries only, the achievement of statistically significant improvements is particularly challenging. Therefore, we investigate the robustness of such improvements by testing BM25P also on the `Signal` and `RCV1` datasets. For each one of these datasets we have in fact 40,000 pseudo-queries obtained from the news titles as previously discussed. The results of these additional experiments are reported in Table 4, where we evaluate the retrieval performance in terms of MRR for the `Signal`, `RCV1` and `Aquaint` datasets for varying α and k.

The results show that BM25P outperforms significantly BM25 in terms of MRR on all three datasets, confirming once again and on a larger scale the results achieved on the `Aquaint` collection. Indeed, our results also confirm that the best performing setting on this dataset is BM25P with $k = 10$ and $\alpha = 10$. A slightly different result is achieved for the `Signal` and `RCV1` datasets, where the best performing method results to be BM25P with $k = 5$ and $\alpha = 20$, with a small difference from BM25P with $k = 10$. Indeed, on these collections, BM25P, for all k values, always exhibits statistically significant improvements w.r.t. BM25 for $\alpha \geq 10$.

From the numbers in Table 4, we see that MRR is greater for $\alpha \geq 10$ than for $\alpha < 10$. When $\alpha = 10$, the average value of the scaled weights αw_i is equal to 1, i.e., the value of α divided by the number of passages, as explained in Section 3. When $\alpha < 10$, the average value of the scaled weights αw_i becomes lesser than 1, thus penalizing the contribution of tf_P with respect to the document length normalization in the denominator of Eq. (2). Conversely, the mean of the weights is greater than or equal to 1 when $\alpha \geq 10$, and the initial and final passages of the news can get larger weights than the others passages. The impact of α is discussed in detail when assessing the contribution of passages to effectiveness.

Although the importance of the first and last passages is evident, the middle passages of a news article cannot be ignored [4, 45]. Middle passages may get weights greater than 1 for α greater than 10, and this explains why BM25P gets its best results when $\alpha = 20$ for `RCV1` and `Signal`. The best performing setting is BM25P with $k = 10$ and $\alpha = 10$ for `Aquaint` and BM25P with $k = 5$ and $\alpha = 20$ for `Signal` and `RCV1`. A possible explanation of this slight difference is that pseudo-queries of `Signal` and `RCV1` benefit from the skewed probability distribution of BM25P with $k = 5$, which gives a greater importance to the first and last passages and seems to better approximate where the pseudo-queries match. Also the average number of terms in pseudo-queries is twice the number of words in the real-world queries provided with `Aquaint`. We highlight that, for $\alpha \geq 10$ and for all k values tested, the MRR performance of BM25P is better than the MRR performance of BM25, in a statistically significant way. Indeed, results achieved with MRR for `Aquaint` are consistent with the ones discussed for NDCG; namely BM25P with $k = 10$ is the best method and statistically outperforms BM25. BM25P with $k = 5$ and $k = 15$ also behave well on `Aquaint`, but the improvement is

Table 4. MRR for BM25 and BM25P on the three news collections for different values of α. We report statistical significance w.r.t. BM25 with ▲ for p-value < 0.01 and △ for p-value < 0.05.

	Model	k	α						
			1	5	10	20	30	40	50
`Aquaint`	BM25	-	0.485	0.485	0.485	0.485	0.485	0.485	0.485
	BM25P	5	0.459	0.569	0.589△	0.582	0.578	0.578	0.579
	BM25P	10	0.458	0.577△	**0.591▲**	0.578△	0.588△	0.589△	0.586△
	BM25P	15	0.446	0.532	0.540	0.547	0.545	0.558△	0.558△
`Signal`	BM25	-	0.342	0.342	0.342	0.342	0.342	0.342	0.342
	BM25P	5	0.268	0.337	0.351▲	**0.356▲**	0.356▲	0.354▲	0.352▲
	BM25P	10	0.276	0.340	0.350▲	0.353▲	0.352▲	0.351▲	0.349▲
	BM25P	15	0.276	0.339	0.349▲	0.351▲	0.350▲	0.348▲	0.347▲
`RCV1`	BM25	-	0.340	0.340	0.340	0.340	0.340	0.340	0.340
	BM25P	5	0.258	0.344▲	0.363▲	**0.369▲**	0.365▲	0.360▲	0.356▲
	BM25P	10	0.253	0.339	0.356▲	0.360▲	0.356▲	0.351▲	0.347▲
	BM25P	15	0.249	0.334	0.351▲	0.355▲	0.351▲	0.346▲	0.342▲

statistically significant just for few values of α in the case of BM25P with $k = 15$ and $k = 5$. BM25P with $k = 10$ uses top 10 highest IDF terms in each document to create a probability distribution of their positions. Increasing the number of terms for computing the distribution does not necessarily yield better results. We can conclude that 10 salient terms for `Aquaint` and 5 salient terms for `Signal` and `RCV1` achieve the best results, and the distribution flattens as we increase this number (see Figure 1), making it closer to the uniform weighting of BM25.

To gain further insights on the above result, we also perform an analysis of the performance of BM25P against BM25 by fine tuning both models, i.e., by investigating the impact of the parameters k_1 and b. In the analysis reported in Tables 3 and 4, we used the default parameters $b = 0.75$ and $k_1 = 1.2$. We now fine-tune the two parameters k_1 and b of BM25 and BM25P on the four collections by performing a grid search in $[0.3, 0.9]$ with a step of 0.1 for b and $[0.4, 2.0]$ with a step of 0.2 for k_1. The choice to fine tune b and k_1 in these two specific intervals stems from many previous works that perform BM25 optimization [23, 44, 48]. In the case of `Aquaint`, as we have only 50 queries, we were unable to achieve stable results. We found that the same issue was also documented by Robertson *et al.* [44] on a different test collection, where they report instability of results for such a small sample of queries. We successfully fine-tuned b and k_1 for both BM25 and BM25P on `Signal` and `RCV1`. For both datasets, we applied the two weighting functions on a training set of 4,000 queries and evaluated the best combination in terms of MRR on a test set of 1,000 queries. In doing so, we obtained improved performance for both weighting models, proportional with the gains achieved with default parameters. In the following, we assume that all the performance improvements of BM25P over BM25 reported hereinafter are consistent with the fine-tuning outcomes.

The initial results on whether BM25P is more effective in comparison to BM25 and how the number of salient terms impacts the results, prove our hypothesis. A reinforcement of the results from Table 3 and Figure 2 can also be seen in Table 4 commented above. We thus conclude that BM25P outperforms BM25 in terms of NDCG and MRR in retrieving documents belonging to our three news collections.

B. How can salient terms be identified? After analysing how effective BM25P is for retrieving relevant documents based on the occurrence probability of top k IDF terms, a subsequent question surges: can salient terms be identified differently? To address this question we explore two alternative approaches to determine salient terms, namely using TFIDF [33] and KL divergence scores [5].

- **TFIDF**. Besides considering the specificity of the term in the document collection, TFIDF considers also the number of term occurrences per document by multiplying the term in-document frequency with the term inverse document frequency. The resulting distributions are flatter than the ones observed for IDF, namely the first and last passage are assigned relatively smaller weights than in the IDF case.
- **KL divergence scores**. Initially proposed by Carpineto et al. [9], KL divergence scores can be used to identify salient terms in a document as described in [5] and shown in Eq. 6. Specifically, we compute the unigram language model of each document and compare it with the unigram language model of the entire collection. The term contribution to document-collection KL divergence is used to assign a score $Score_{KL}(t)$ to document terms as shown below:

$$Score_{KL}(t) = M_d(t) \log\left(\frac{M_d(t)}{M^*(t)}\right) \tag{6}$$

where $M_d(t)$ is the occurrence probability of term t in the language model of document d, and $M^*(t)$ is the occurrence probability of term t in the language model of the collection. Top scoring terms are considered the most salient for the document and used similarly to the TFIDF and IDF cases. The resulting probability

Table 5. NDCG at different cutoffs for BM25 and BM25P ($\alpha = 10$, $P = 10$) on the `Aquaint` collection, for top k highest IDF, TFIDF and KL divergence scores terms. We highlight statistical significant differences w.r.t. BM25 with $\blacktriangle$ for p-value < 0.01 and $\triangle$ for p-value < 0.05 according to the two sample t-test [41].

NDCG	BM25	BM25P top k IDF			BM25P top k TFIDF			BM25P top k KL score		
Cutoffs		$k = 5$	$k = 10$	$k = 15$	$k = 5$	$k = 10$	$k = 15$	$k = 5$	$k = 10$	$k = 15$
1	0.200	0.340 $^{\triangle}$	**0.370**$^{\blacktriangle}$	0.290$^{\triangle}$	0.350$^{\blacktriangle}$	0.270	0.270	0.340$^{\blacktriangle}$	0.260	0.270
3	0.291	0.319	**0.335**	0.317	0.324	0.314	0.317	0.322	0.314	0.317
5	0.280	0.327	**0.329**$^{\blacktriangle}$	0.301	0.308$^{\triangle}$	0.295$^{\triangle}$	0.295$^{\triangle}$	0.308$^{\triangle}$	0.298$^{\triangle}$	0.295$^{\triangle}$
10	0.270	**0.303**	0.298$^{\triangle}$	0.291	0.291$^{\triangle}$	0.281	0.278	0.293$^{\triangle}$	0.280	0.276
15	0.269	0.288	**0.296**$^{\blacktriangle}$	0.290$^{\triangle}$	0.291$^{\blacktriangle}$	0.283$^{\triangle}$	0.283$^{\blacktriangle}$	0.291$^{\blacktriangle}$	0.282$^{\triangle}$	0.283$^{\blacktriangle}$
20	0.273	0.280	**0.289**$^{\triangle}$	0.282 $^{\triangle}$	0.283$^{\triangle}$	0.277	0.279$^{\triangle}$	0.283$^{\triangle}$	0.278	0.279$^{\triangle}$

distribution is very similar to the TFIDF one, less skewed than that obtained using IDF. As for the other methods, the distribution soon flattens as we increase the number k of salient terms considered.

In Table 5 we compare the NDCG at different cutoffs for BM25 and BM25P ($\alpha = 10$, $P = 10$) obtained on the `Aquaint` dataset when using the three different methods for identifying the top k salient terms where $k = 5, 10, 15$, namely IDF (already presented in Table 3), TFIDF and KL divergence.

We highlight the effectiveness of BM25P regardless of the heuristic used for determining the top-k salient terms, whether IDF, TFIDF or KL divergence scores. BM25P always outperforms BM25, with best results obtained when the top k highest IDF terms are used to determine salient terms. A slightly lower performance is observed for the heuristics based on TFIDF and KL divergence scores. BM25P with $k = 10$ is still the best combination of parameters for $P = 10$ and $\alpha = 10$ in the case of IDF, whereas for both TFIDF and KL divergence score $k = 5$ gives better results than larger values of k.

We also tested various combinations of BM25P based on top k TFIDF and top k KL scores for different values of α. As for IDF, in all the tests BM25P achieves the best results for $\alpha > 10$. For TFIDF, the overall best results are obtained for BM25P with $k = 5$ and $\alpha = 45$. where $NDCG@5 = 0.319$, while for KL divergence score, the best performance is achieved for $k = 5$ and $\alpha = 25$. where $NDCG@5 = 0.320$.

C. What is the impact of the number of passages? We now want to understand what is the impact of the number of passages P on the effectiveness of our BM25P model. Are 10 passages too small or too large a granularity given the average document length? Can we reach a similar or better effectiveness using a different number of passages?

To answer these questions, we look at the results in Table 6, where we compare BM25P (given $\alpha = 10$ and IDF to identify the $k = 10$ salient terms) with BM25 as a function of the number of passages $P = \{3, 5, 10, 15, 20\}$. For all the values of P tested, our BM25P model achieves better performance than the BM25 model in terms of NDCG at different cutoffs, with improvements varying from 1.3% to 85%. Although we report increases in NDCG for all values of P, when $P = 10$, we have statistically significant improvements for most of the cutoffs. The second best result is obtained with $P = 20$, where the performance in terms of NDCG@1 is equal to the one achieved with $P = 10$ while it significantly drops when considering larger cutoffs. For $P = 15$, again the improvement is statistically significant only for NDCG@1 even if the drop of performance here is limited with respect to the one observed for $P = 20$. Moreover, no statistical significant improvements are observed for BM25P with $P = 5$ and $P = 3$. These results show an interesting fact: dividing the document in 3 equal parts and weighting them differently achieves less advantages than dividing it in a finer grain, such as 10, 15 or 20 passages. Having more passages allows for more sensitivity and a more accurate weighting of the different passages. However it is preferable, according to our results, to split the documents in 10 equal parts rather than in 15 or 20 passage, thus having too finer grain over the document can be counter-productive.

Table 6 presents all results for a fixed $\alpha = 10$. However, it does not reflect the whole picture on the results obtained. Given the results presented above, we now analyse how the number of passages P and the scaling factor α interact with each other. In Figure 3, we vary α from 1 to 50, with steps of 5, for $P = \{3, 5, 10, 15, 20\}$.

For $P = \{10, 15, 20\}$, the best performance is achieved when $\alpha \geq 15$, as in Figures 3 (c), (d), and (e). In these cases, the maximum values of $NDCG@5$ for BM25P, on average, are $\alpha = 28.89$ for $P = 10$, $\alpha = 24.16$ for $P = 15$, and $\alpha = 33.31$ for $P = 20$. On the other hand for small values of P, i.e., $P = 3, 5$, the best results are achieved for smaller values of α. In these cases, the maximum values of $NDCG@5$ for BM25P, on average, are $\alpha = 6.39$ for $P = 3$ and $\alpha = 13.38$ for $P = 5$.

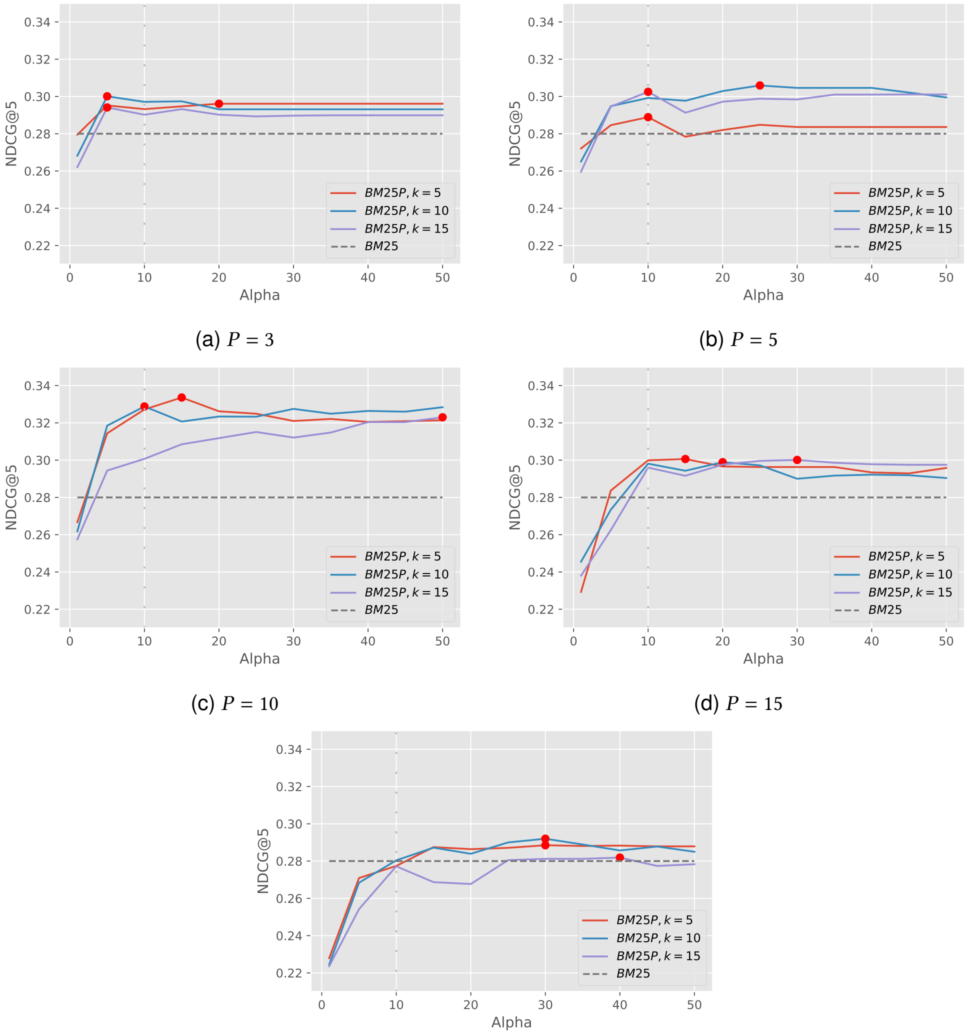
Given the obtained results we can conclude that the re-scaling factor α is important in moderating the impact of the number of passages on the overall distributions, whether boosting or discounting such weights.

D. What is the contribution of each passage? To answer this research question we divide all documents in the `Aquaint` collection into 10 passages, and we build 10 new document collections. The n-th document collection, with n ranging from 1 to 10, is compostед by the first n passages of each document. This experiment allows us to explore the incremental contribution of each passage to the overall effectiveness by comparing BM25 and BM25P applied to increasing portions of the documents. Figure 4 illustrates the results measured with NDCG, with cutoffs 1, 5, 10, and 20. Regarding NDCG@1 and BM25, we see that using the first 30% of the documents results in a higher NDCG@1 value than using the whole documents. With respect to BM25P, the first 30% of the documents produces a similar benefit, resulting in higher NDCG@1 values than BM25. The NDCG gap between BM25 and BM25P is higher for small cutoffs, and decreases as the cutoff increases. For all cutoff values, the effectiveness of both BM25 and BM25P rise quickly when using the initial portions of the documents, then slowly with the central passages, and BM25P has a significant effectiveness improvement over BM25 when the final passage of the documents is included. It is clear that our BM25P model is able to increase the effectiveness over BM25 by boosting the evidence provided by the last passage of documents. Even in this case, the improvement is larger for smaller cutoff values.

Figure 5 illustrates the NDCG contributions of every passage in isolation for both BM25 and BM25P. The results confirm that the effectiveness contributions of the initial and final passages are higher for BM25P than BM25, while BM25 gets higher effectiveness contributions for middle passages. To conclude, the proposed BM25P model is clearly outperforming BM25 for small cutoff values (1, 5, and 10) when the full document is processed, while it is slightly better than BM25 when $k = 20$. In term of effectiveness, the most relevant sections of each document are the initial passages and the last passage.

Table 6. NDCG at different cutoffs for BM25 and BM25P ($\alpha = 10$, $k = 10$) on the `Aquaint` collection. We highlight statistical significant differences w.r.t. BM25 with ▲ for p-value <0.01 and △ for p-value < 0.05 according to the two sample t-test [41].

NDCG	BM25	BM25P				
Cutoffs		$P = 3$	$P = 5$	$P = 10$	$P = 15$	$P = 20$
1	0.200	0.270 +35.00%	0.280 +40.00%	**0.370** +85.00% ▲	0.320 +60.00% △	0.370 +85.00% ▲
3	0.291	0.295 +1.32%	0.300 +2.93%	**0.335** +15.01%	0.306 +4.96%	0.299 +2.51%
5	0.280	0.297 +6.10%	0.299 +6.84%	**0.329** +17.44% ▲	0.298 +6.45%	0.280 +0.13%
10	0.270	0.281 +3.62%	0.280 +3.58%	**0.298** +10.20% △	0.281 +3.73%	0.280 +3.36%
15	0.269	0.285 +6.09%	0.285 +5.81%	**0.296** +9.96% ▲	0.288 +7.02%	0.280 +4.14%
20	0.273	0.278 +1.86%	0.288 +5.44%	**0.289** +5.81% △	0.278 +1.66%	0.271 −0.68%



(a) $P = 3$

(b) $P = 5$

(c) $P = 10$

(d) $P = 15$

(e) $P = 20$

Fig. 3. NDCG@5 for BM25 and BM25P, on `Aquaint`, for different values of P and different values of α.

E. Do all passages contribute to some extent? The previous experiments confirm that the first and last passages contribute the most to the retrieval effectiveness. As highlighted in Figure 1, the first and last document passages have a significantly higher probability of containing occurrences of salient terms. On the other hand, the probability for salient terms to occur in the central passages is not only lower but also almost constant. This characteristic could suggest that some passages in the middle of documents could be ignored since they do not contribute, or contribute only marginally, to the retrieval effectiveness. To gain a clear understanding about this possibility, we conduct an experiment where central passages are left out, and the retrieval is performed based on the passages at the boundaries of the documents

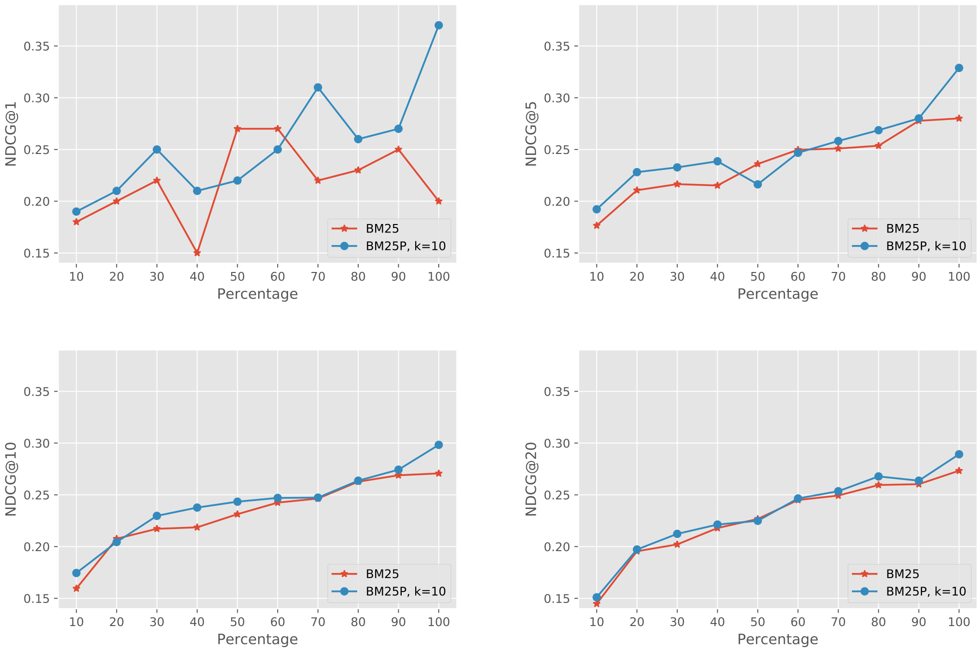


Fig. 4. NDCG at different cutoffs for BM25 and BM25P ($\alpha = 10$, $k = 10$, $P = 10$), on Aquaint, by increasing the number of passages considered.

only. Specifically, on the Aquaint collection, we perform three different runs where only one, two, and three passages at the beginning and the end of each document, respectively, are used for retrieval and compare the results achieved with the ones obtained by using all the passages (values taken from Table 3, for $\alpha = 10$, $k = 10$). Table 7 reports the results of these experiments in terms of NDCG at different cutoffs for both BM25 and BM25P. Despite the relatively high variance of NDCG at small cutoffs, we can observe from the table an increase of NDCG as more passages are considered. As expected at small cutoff values, the effectiveness differences between using the whole content or only first and last passages are small, since the latter are indeed very good indicators of the relevance of a document, but if we aim at large recall or precision values at larger cutoff values, it is better to take into account the whole document content.

Table 7. NDCG at different cutoffs for BM25 and BM25P ($\alpha = 10$, $k = 10$) on the Aquaint collection when an increasing number of passages at the boundaries of the document are considered, e.g, xx------xx means retrieval uses only the first and last two passages of each document.

NDCG	x--------x		xx------xx		xxx----xxx		all passages	
Cutoffs	BM25	BM25P	BM25	BM25P	BM25	BM25P	BM25	BM25P
1	0.320	0.340	0.290	0.310	0.250	0.240	0.200	0.370
3	0.278	0.297	0.272	0.293	0.278	0.283	0.291	0.335
5	0.261	0.267	0.272	0.292	0.275	0.285	0.280	0.329
10	0.231	0.237	0.264	0.275	0.268	0.274	0.271	0.298
15	0.218	0.220	0.249	0.255	0.267	0.262	0.269	0.296
20	0.203	0.210	0.239	0.247	0.252	0.256	0.273	0.289

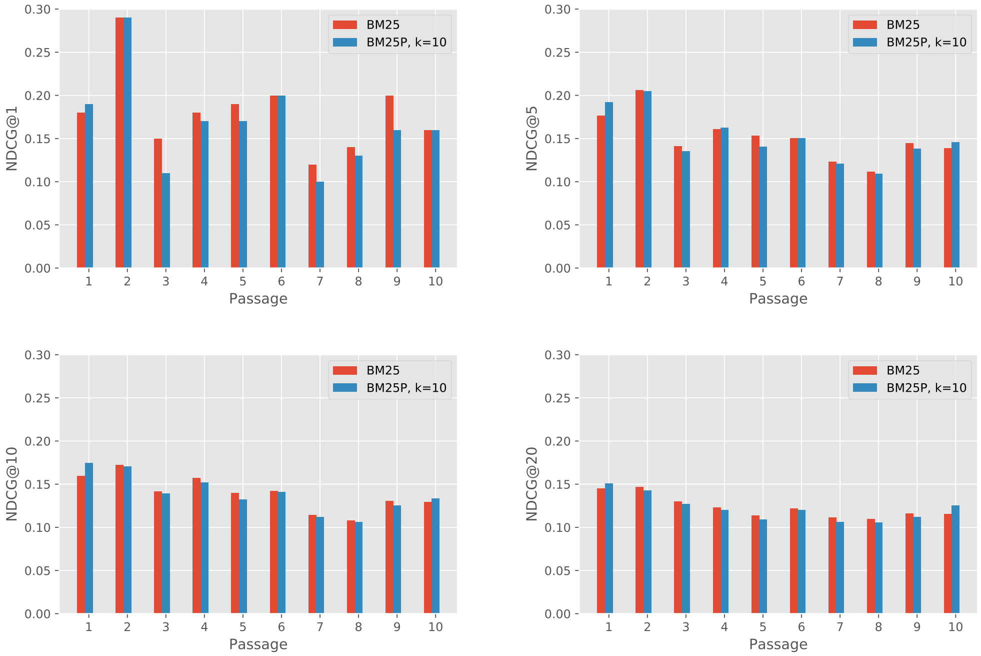


Fig. 5. NDCG@k for BM25 and BM25P ($\alpha = 10$, $k = 10$, $P = 10$) on `Aquaint`, for each passage considered singularly, aka the contribution of each passage when considered alone.

5.3 RQ3: How do different passage segmentation strategies impact BM25P?

We now analyse how the effectiveness of BM25P varies when varying the paragraph splitting heuristic. The heuristic used so far, called "Split P" from now on, divides the document into P equal parts. The length of each passage is thus proportional to the length of the original document. We now investigate two alternative splitting strategies producing passages of fixed length to understand if they can improve the retrieval performance.

To gather the collection-wise statistics used by BM25 and BM25P, we computed the average length of a document in the `Aquaint` collection; the average length is ~249 words. However, the smallest document (after stopword removal) includes just a single word, while the longest one has 8, 861 words. To consider the length distribution in our collection, we look at the 90-th percentile of document lengths and find that 90% of the documents have a length shorter than ~530 words. We consider this threshold of 530 words for experimenting the following fixed-length document splitting strategies:

- "FixedTrunc", where we set the number of passages to 10 and the size of each passage to 53 words, leading to a length of 530 words per document. All documents longer than 530 words are truncated beyond that length. For shorter documents, we keep the fixed passage length of 53 words per passage over ten passages. As a consequence, short documents may have empty passages in their final parts. In this way we end up with a virtual collection in which each document has exactly 530 words.
- "FixedCumul", where we set to 10 the number of passages and to 53 words the length of the first 9 passages if document length is larger than 530 words. In this case the last passage includes all the remaining content of the document. Documents shorter than 530 words are instead managed as in "FixedTrunc".

These settings are consistent with previous findings by Callan [7], who showed that paragraphs have usually a length of 50 words or more.

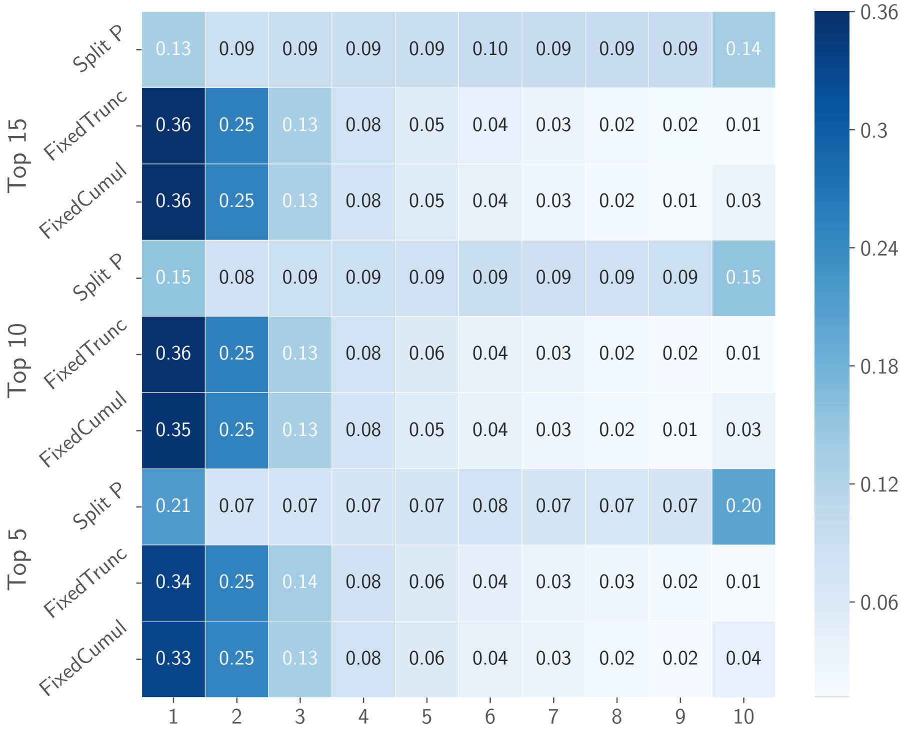


Fig. 6. Probability distribution for the positions of key terms occurrences in the news articles on `Aquaint` for the three passage splitting methods.

Figure 6 shows the probability distribution for the positions of salient term occurrences in the news articles on `Aquaint` for the three passage splitting methods, i.e., Split P ($P = 10$), FixedTrunc and FixedCumul. The Figure provides further evidence of the intuition discussed in Section 3, i.e., starting and closing passages are the most important ones in term of the provided relevance signals. The first strategy tested, i.e., Split P ($P = 10$), dynamically splits the document in ten equal parts, regardless of the varying length of documents. Here, the last part of the document is always significant as the probability increases with a large margin w.r.t. FixedTrunc and FixedCumul. Indeed, the two new strategies that split the document in passages of fixed length appear not able to capture the final yet significant part of the documents. Indeed, FixedTrunc, i.e., the strategy that discards the part of each document above 530 words, is heavily under-performing in the last passages. On the other side, FixedCumul partially addresses the point above as the last passage contains the part of the document above 477 words. Here, weak evidence of the same phenomenon is present as the probability of the last passage is always greater than the one provided by FixedTrunc, yet a fade out effect on the last paragraph still occurs, due to documents which have a smaller length and for which in the final paragraphs there are no salient term occurrences.

Table 8 shows the results of the comparison in terms of NDCG at different cutoffs for these document splitting strategies and BM25. Results show that the variants of BM25P using FixedTrunc and FixedCumul under-perform BM25

Table 8. NDCG at different cutoffs for BM25 and BM25P ($\alpha = 10$) on the `Aquaint` collection. The differences in performance are all statistically significant with p-value < 0.01, except for NDCG@1 where the improvement is not statistically significant.

NDCG	BM25	BM25P FixedTrunc			BM25P FixedCumul		
Cutoffs		$k = 5$	$k = 10$	$k = 15$	$k = 5$	$k = 10$	$k = 15$
1	0.200	$0.230_{+15.00\%}$	$0.230_{+15.00\%}$	$0.230_{+15.00\%}$	$0.230_{+15.00\%}$	$0.230_{+15.00\%}$	$0.230_{+15.00\%}$
3	0.291	$0.218_{-25.08\%}$	$0.209_{-27.91\%}$	$0.202_{-30.58\%}$	$0.215_{-26.30\%}$	$0.212_{-27.11\%}$	$0.208_{-28.72\%}$
5	0.280	$0.218_{-22.00\%}$	$0.215_{-23.20\%}$	$0.208_{-25.46\%}$	$0.217_{-22.66\%}$	$0.212_{-24.10\%}$	$0.212_{-24.16\%}$
10	0.270	$0.211_{-21.81\%}$	$0.207_{-23.42\%}$	$0.203_{-24.86\%}$	$0.213_{-21.22\%}$	$0.207_{-23.54\%}$	$0.207_{-23.41\%}$
15	0.269	$0.201_{-25.02\%}$	$0.198_{-26.38\%}$	$0.196_{-27.03\%}$	$0.203_{-24.40\%}$	$0.200_{-25.62\%}$	$0.199_{-26.19\%}$
20	0.273	$0.203_{-25.51\%}$	$0.201_{-26.32\%}$	$0.198_{-27.21\%}$	$0.206_{-24.53\%}$	$0.202_{-26.00\%}$	$0.202_{-25.98\%}$

for all cutoffs but 1. Indeed, the reduction in performance is statistically significant, with $p < 0.01$. Differently, when considering NDCG@1, the two new strategies outperform BM25 with an improvement of 15%. However, this gain is not statistically significant.

The results above confirm that our variable-length splitting strategy, i.e., Split P ($P = 10$), exploits better the distribution of salient terms throughout an entire document, considering its full length. On the other hand, document splitting strategies based on a fixed length for passages such as FixedTrunc and FixedCumul underperform not only Split P but also the BM25 baseline.

5.4 RQ4: Does BM25P work for types of documents other than news, e.g., Web documents?

Addressing the first three research question demonstrated the performance advantage of BM25P over BM25 for news article retrieval. News articles, generally, are clean, curated, and stylistically uniform. Our fourth research question investigates the effectiveness of BM25P when applied to Web documents, which are significantly more noisy than news. We experimentally assess the performance of BM25P for web document retrieval by using the `MS-MARCO` collection and compare the performance of BM25 and BM25P in terms of MRR by varying k, i.e., the number of the top k salient terms, the criterium used to select them, i.e., top k IDF or top k TFIDF, and α, i.e., the BM25P re-scaling factor. Since for `MS-MARCO` we have on average one relevant document per query, with some exceptions when there are two relevant documents per query, it is a common practice to evaluate retrieval performance in terms of MRR. We conduct the experiment by employing the "Split P" paragraph splitting heuristic using $P = 10$. The results of the investigation are reported in Table 9.

The results show that, even on significantly less curated documents like the Web documents of `MS-MARCO`, BM25P improves the performance of BM25 (Table 9). The best performance improvement is achieved when $\alpha = 5, 10$ for both top k IDF or top k TFIDF. In the case of BM25P using top k IDF, the gain achieved by BM25P ranges from 2.05% to 3.5%. In particular, when $\alpha = 5$, BM25P achieves the best improvement of up to 3.5% ($k = 10$). The above improvements are all statistically significant and, in most cases, the statistical significance is with pvalue < 0.01. However the improvement, although significant, is smaller than in the case of `Aquaint` for example; however, for `MS-MARCO`, we compute the metric on more than 5,000 queries whereas for `Aquaint` we only have 50 queries.

Web documents differ from news articles; we thus also wanted to see whether using the top k TFIDF terms instead of IDF can boost performance. As expected, using top k TFIDF terms instead of IDF, BM25P shows the same trend in performance (Table 9), but with a higher relative improvement and more statistically significant results, In this case when considering top 5 terms, BM25P with TFIDF shows more improvement than with IDF, and both are better than BM25. However, in both cases, for $\alpha \geq 20$, BM25P loses effectiveness and performs worse than BM25. For top k TFIDF,

Table 9. Effectiveness of BM25 and BM25P in terms of MRR on the MS-MARCO collection by varying k, α and the salient term selection criterium, i.e., top k IDF or top k TFIDF. We report statistical significance w.r.t. BM25 with ▲ for p-value < 0.01 and △ for p-value < 0.05.

Model		α			
		1	5	10	20
BM25	-	0.251	0.251	0.251	0.251
BM25P	top 5 IDF	0.220	$0.257^{\triangle}_{+2.19\%}$	0.254	0.245
	top 10 IDF	0.219	$0.260^{\blacktriangle}_{+3.50\%}$	$0.257^{\blacktriangle}_{+2.25\%}$	0.250
	top 15 IDF	0.217	$0.260^{\blacktriangle}_{+3.35\%}$	$0.257^{\blacktriangle}_{+2.05\%}$	0.249
BM25P	top 5 TFIDF	0.217	$0.259^{\blacktriangle}_{+3.20\%}$	$0.258^{\blacktriangle}_{+2.48\%}$	0.250
	top 10 TFIDF	0.215	$0.259^{\blacktriangle}_{+2.68\%}$	$0.257^{\blacktriangle}_{+1.97\%}$	0.248
	top 15 TFIDF	0.213	$0.258^{\blacktriangle}_{+2.40\%}$	$0.256^{\blacktriangle}_{+1.48\%}$	0.247

best performance is achieved by BM25P for all values of k and for $\alpha = 5, 10$, where the improvement on BM25 ranges from 1.48% to 3.2%, and it is always statistically significant for a p-value < 0.01. We thus conclude that BM25P is an effective retrieval method for Web documents. Also in such scenario, it represents a valid alternative to the standard BM25.

5.5 RQ5: How does passage weighting perform when used with other weighting models?

Finally, we investigate the impact of passage weighting when employed with other scoring functions. We perform this analysis to evaluate if our approach for passage weighting is general and effective also on different models. In the following, we experiment with two different probabilistic weighting model families, namely, language models (LM) and models based on the divergence from randomness (DFR).

In Dirichlet-smoothed LM weighting model modeling [53], the maximum likelihood of a term t occurring in a document d is smoothed to the collection-wise language model. Applying a log transformation to convert the product of probabilities into a relevance score as in (1), the LM's relevance score $LM_t(q, d)$ is given by:

$$LM_t(q, d) = w_q \log\left((1 - \lambda)\frac{tf}{dl} + \lambda\frac{F_t}{T_c}\right), \tag{7}$$

where T_C is the number of tokens in the collection, F_t is the frequency of the term t in the collection, and $\lambda = \frac{\mu}{\mu + dl}$ is the Dirichlet smoothing, with $\mu = 2500$.

The DLH13 weighting model [1] is a generalization of the parameter-free hyper-geometric DFR model in a binomial case, whose relevance score $DFR_t(q, d)$ is given by:

$$DFR_t(q, d) = \frac{w_q}{tf + 0.5}\left(tf \log_2\left(\frac{tf \cdot N \cdot avg\_dl}{dl \cdot F_t}\right) + \frac{1}{2}\log_2\left(2\pi tf\left(1 - \frac{tf}{dl}\right)\right)\right). \tag{8}$$

By plugging the re-scaled linear combination tf_P of the term frequencies tf_i in each passage i of the P passages in documents d as in (4), and summing up the single term contributions among the query terms, we obtain the corresponding LMP and DFRP weighting models.

In Table 10 we report the NDCG values at different cutoffs for the BM25, LM and DFR weighting models, without and with passage weighting as in (4). For each passage-enhanced weighting model, we report the highest effectiveness measure obtained by varying the α hyper-parameter, i.e., 10 for BM25P, 15 for LMP and 5 for DFRP. For all cutoff values

Table 10. NDCG at different cutoffs for different weighting models without and with passage weighting on the `Aquaint` collection for the best k value per model, best α value per model, $P = 10$. We highlight statistically significant differences w.r.t. corresponding model without passage weighting with ▲ for p-value < 0.01 and △ for p-value < 0.05 according to the two sample t-test [41].

NDCG@	BM25	BM25P	LM	LMP	DFR	DFRP
1	0.200	**0.370**▲	0.360	**0.420**	0.260	**0.360**△
3	0.291	**0.335**	0.388	**0.433**	0.298	**0.347**△
5	0.280	**0.329**▲	0.367	**0.397**	0.300	**0.342**△
10	0.270	**0.298**△	0.344	**0.358**	0.288	**0.318**△
15	0.269	**0.296**△	0.324	**0.337**	0.283	**0.312**▲
20	0.273	**0.289**△	0.312	**0.324**	0.281	**0.306**▲

tested, all passage-enhanced weighting models clearly outperform the corresponding original weighting models. For BM25P, LMP and DFRP the relative improvement ranges from +5.8% to +85.0%, from +3.8% to +16.7%, and from +8.6% to +38.5%, respectively. We observed the same behavior on similar tests conducted on the `Signal` and `RCV1` collections and not reported here for brevity. We can thus conclude that we experimentally measured an effectiveness boost when deploying passage weighting with different probabilistic weighting models such as BM25, LM and DFR.

6 CONCLUSION

For news articles, we observed that a common stylistic feature is the preponderance of occurrences of salient terms at the beginning and at the end of the article. By capitalizing on this stylistic feature to improve the effectiveness of news retrieval, we proposed BM25P, a variant of the well-known BM25 weighting model that considers salient term distribution variations among the different passages of the document. In BM25P such distribution information is used to assign different weights to the occurrences of query terms, depending on which passage they appear in, boosting or reducing the importance of certain passages in the document, typically giving greater importance to the first and last passages. This distinguishes BM25P from the traditional BM25, which does not consider the position of the occurrences in the document but weights uniformly all of them. To test the effectiveness of our approach, we used four different datasets: the `Aquaint` collection with the 50 assessed queries from the 2005 Robust and HARD TREC tracks, the `Signal` and `RCV1` corpora for which we synthetically generated a large number of pseudo-queries and relevance judgments, and the `MS-MARCO` collection that allowed us to assess if weighting passages differently works also for different kinds of documents and queries, i.e., web documents that are generally more noisy and less curated than news articles and queries in the domain of question answering.

Our experiments showed that, by differently weighting passages, BM25P markedly improves NDCG and MRR with respect to using BM25 on all evaluated collections. Our exhaustive experiments covered the entire solution space of the proposed weighting model and showed consistent results across different configurations of its hyperparameters and the different strategies tested to split the document content in passages and to compute the collection-dependent weights associated to each of them. Moreover, we showed that our passage weighting approach is consistent and provides performance advantages also if applied to probabilistic weighting models different from BM25, namely language models (LM) and models based on the divergence from randomness (DFR).

We observed that BM25P significantly improves BM25 for NDCG on `Aquaint` with percentages up to 85% for small cutoffs, while the MRR computed on `Signal` and `RCV1` increases of 4.1% and 8.5% respectively and on `Aquaint` with up to 21%. BM25P resulted to outperform BM25 also on `MS-MARCO`, with statistically significant improvements ranging from 1.48% to 3.2% thus demonstrating that it represents a valid alternative to BM25 for Web document retrieval as well.

More importantly, these consistent improvements come for free since BM25P is totally unsupervised and very efficient, just requiring the computation of a few collection-dependent weights at index construction time.

As future work we plan to study the impact of adaptively varying the number of passages weighted – here statically set independently of the length of the specific document considered – and the use of our BM25P model in conjunction with BM25F for retrieving semi-structured documents. Reasonably, we expect to have orthogonal advantages from weighting differently the different fields of structured documents and the passages within each one of these fields and one can think to a kind of BM25PF weighting model where the two approaches coexist and possibly combine their strengths. In the same line of investigation we will focus on approaches that can learn query-specific or document-specific patterns for which specific passages of the document can be more important and informative.

ACKNOWLEDGMENTS

We acknowledge the efforts of the anonymous reviewers whose detailed comments helped to improve the article. We also acknowledge our respective colleagues who provided useful insights.

REFERENCES

[1] Giambattista Amati. 2006. Frequentist and Bayesian Approach to Information Retrieval. In *Proceedings of the 28th European Conference on Advances in Information Retrieval (ECIR'06)*. Springer-Verlag, Berlin, Heidelberg, 13–24. https://doi.org/10.1007/11735106_3

[2] Jing Bai, Yi Chang, Hang Cui, Zhaohui Zheng, Gordon Sun, and Xin Li. 2008. Investigation of Partial Query Proximity in Web Search. In *Proceedings of the 17th International Conference on World Wide Web (WWW '08)*. ACM, New York, NY, USA, 1183–1184. https://doi.org/10.1145/1367497.1367717

[3] Michael Bendersky and Oren Kurland. 2010. Utilizing Passage-based Language Models for Ad Hoc Document Retrieval. *Inf. Retr.* 13, 2 (April 2010), 157–187.

[4] Toine Bogers and Antal van den Bosch. 2007. Comparing and Evaluating Information Retrieval Algorithms for News Recommendation. In *Proceedings of the 2007 ACM Conference on Recommender Systems (RecSys '07)*. Association for Computing Machinery, New York, NY, USA, 141–144. https://doi.org/10.1145/1297231.1297256

[5] Stefan Büttcher and Charles L. A. Clarke. 2006. A Document-Centric Approach to Static Index Pruning in Text Retrieval Systems. In *Proceedings of the 15th ACM International Conference on Information and Knowledge Management (CIKM '06)*. Association for Computing Machinery, New York, NY, USA, 182–189. https://doi.org/10.1145/1183614.1183644

[6] Stefan Büttcher, Charles L. A. Clarke, and Brad Lushman. 2006. Term Proximity Scoring for Ad-Hoc Retrieval on Very Large Text Collections. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '06)*. Association for Computing Machinery, New York, NY, USA, 621–622. https://doi.org/10.1145/1148170.1148285

[7] James P. Callan. 1994. Passage-Level Evidence in Document Retrieval. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '94)*. Springer-Verlag, Berlin, Heidelberg, 302–310.

[8] Gabriele Capannini, Claudio Lucchese, Franco Maria Nardini, Salvatore Orlando, Raffaele Perego, and Nicola Tonellotto. 2016. Quality versus efficiency in document scoring with learning-to-rank models. *Information Processing & Management* 52, 6 (2016), 1161 – 1177. https://doi.org/10.1016/j.ipm.2016.05.004

[9] Claudio Carpineto, Renato de Mori, Giovanni Romano, and Brigitte Bigi. 2001. An Information-Theoretic Approach to Automatic Query Expansion. *ACM Trans. Inf. Syst.* 19, 1 (Jan. 2001), 1–27. https://doi.org/10.1145/366836.366860

[10] Matteo Catena, Ophir Frieder, Cristina Ioana Muntean, Franco Maria Nardini, Raffaele Perego, and Nicola Tonellotto. 2019. Enhanced News Retrieval: Passages Lead the Way!. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'19)*. ACM, New York, NY, USA, 1269–1272. https://doi.org/10.1145/3331184.3331373

[11] David Corney, Dyaa Albakour, Miguel Martinez, and Samir Moussa. 2016. What do a Million News Articles Look like?. In *Proc. NewsIR'16 Workshop*. CEUR-WS, 42–47.

[12] Bruce Croft, Donald Metzler, and Trevor Strohman. 2009. *Search Engines: Information Retrieval in Practice* (1st ed.). Addison-Wesley Publishing Company, USA.

[13] Yixing Fan, Jiafeng Guo, Yanyan Lan, Jun Xu, Chengxiang Zhai, and Xueqi Cheng. 2018. Modeling diverse relevance patterns in ad-hoc retrieval. In *The 41st international ACM SIGIR conference on research & development in information retrieval*. 375–384.

[14] Michael Fuller, Eric Mackie, Ron Sacks-Davis, and Ross Wilkinson. 1993. Structured Answers for a Large Structured Document Collection. In *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '93)*. ACM, New York, NY, USA, 204–213. https://doi.org/10.1145/160688.160720

[15] Mathias Géry and Christine Largeron. 2012. BM25T: a BM25 extension for focused information retrieval. *Knowledge and Information Systems* 32, 1 (01 Jul 2012), 217–241. https://doi.org/10.1007/s10115-011-0426-0

[16] Ben He, Jimmy Xiangji Huang, and Xiaofeng Zhou. 2011. Modeling term proximity for probabilistic information retrieval models. *Information Sciences* 181, 14 (2011), 3017–3031.

[17] Marcin Kaszkiel and Justin Zobel. 2001. Effective ranking with arbitrary passages. *JASIST* 52, 4 (2001), 344–364.

[18] F. Lettich, C. Lucchese, F. M. Nardini, S. Orlando, R. Perego, N. Tonellotto, and R. Venturini. 2019. Parallel Traversal of Large Ensembles of Decision Trees. *IEEE Transactions on Parallel and Distributed Systems* 30, 9 (Sep. 2019), 2075–2089. https://doi.org/10.1109/TPDS.2018.2860982

[19] David D. Lewis, Yiming Yang, Tony G. Rose, and Fan Li. 2004. RCV1: A New Benchmark Collection for Text Categorization Research. *J. Mach. Learn. Res.* 5 (2004), 361–397.

[20] Xiangsheng Li, Yiqun Liu, Jiaxin Mao, Zexue He, Min Zhang, and Shaoping Ma. 2018. Understanding reading attention distribution during relevance judgement. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*. 733–742.

[21] Xiangsheng Li, Jiaxin Mao, Chao Wang, Yiqun Liu, Min Zhang, and Shaoping Ma. 2019. Teach machine how to read: reading behavior inspired relevance estimation. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 795–804.

[22] Jimmy Lin. 2019. The Neural Hype and Comparisons Against Weak Baselines. *SIGIR Forum* 52, 2 (Jan. 2019), 40–51. https://doi.org/10.1145/3308774.3308781

[23] Aldo Lipani, Mihai Lupu, Allan Hanbury, and Akiko Aizawa. 2015. Verboseness fission for BM25 document length normalization. In *Proceedings of the 2015 International Conference on the Theory of Information Retrieval*. 385–388.

[24] Baiyan Liu, Xiangdong An, and Jimmy Xiangji Huang. 2015. Using Term Location Information to Enhance Probabilistic Information Retrieval. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '15)*. ACM, New York, NY, USA, 883–886. https://doi.org/10.1145/2766462.2767827

[25] Xiaoyong Liu and W. Bruce Croft. 2002. Passage Retrieval Based on Language Models. In *Proceedings of the Eleventh International Conference on Information and Knowledge Management (CIKM '02)*. Association for Computing Machinery, New York, NY, USA, 375–382. https://doi.org/10.1145/584792.584854

[26] Yuanhua Lv and ChengXiang Zhai. 2009. Positional language models for information retrieval. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*. 299–306.

[27] Sean MacAvaney, Andrew Yates, Kai Hui, and Ophir Frieder. 2019. Content-Based Weak Supervision for Ad-Hoc Re-Ranking. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'19)*. Association for Computing Machinery, New York, NY, USA, 993–996. https://doi.org/10.1145/3331184.3331316

[28] Craig Macdonald, Richard McCreadie, Rodrygo LT Santos, and Iadh Ounis. 2012. From puppy to maturity: Experiences in developing Terrier. *Proceedings of the OSIR Workshop at the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '12)*. (2012), 60–63.

[29] Craig Macdonald, Nicola Tonellotto, and Iadh Ounis. 2017. Efficient & Effective Selective Query Rewriting with Efficiency Predictions. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '17)*. ACM, New York, NY, USA, 495–504. https://doi.org/10.1145/3077136.3080827

[30] Saket Mengle and Nazli Goharian. 2009. Passage detection using text classification. *JASIST* 60, 4 (2009), 814–825.

[31] Christof Monz. 2004. Minimal span weighting retrieval for question answering. In *Proceedings of the SIGIR Workshop on Information Retrieval for Question Answering*, Vol. 2.

[32] Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. MS MARCO: A Human Generated MAchine Reading COmprehension Dataset. (November 2016). https://www.microsoft.com/en-us/research/publication/ms-marco-human-generated-machine-reading-comprehension-dataset/

[33] Juan Ramos et al. 2003. Using TF-IDF to determine word relevance in document queries. In *Proceedings of the first instructional conference on machine learning*, Vol. 242. Piscataway, NJ, 133–142.

[34] Joaquin Rapela. 2001. Automatically Combining Ranking Heuristics for HTML Documents. In *Proceedings of the 3rd International Workshop on Web Information and Data Management (WIDM '01)*. ACM, New York, NY, USA, 61–67. https://doi.org/10.1145/502932.502945

[35] Yves Rasolofo and Jacques Savoy. 2003. Term proximity scoring for keyword-based retrieval systems. In *European Conference on Information Retrieval*. Springer, 207–218.

[36] Stephen Robertson, S. Walker, S. Jones, M. M. Hancock-Beaulieu, and M. Gatford. 1995. Okapi at TREC-3. In *Overview of the Third Text REtrieval Conference (TREC-3)*. Gaithersburg, MD: NIST, 109–126. https://www.microsoft.com/en-us/research/publication/okapi-at-trec-3/

[37] Stephen Robertson, Hugo Zaragoza, and Michael Taylor. 2004. Simple BM25 Extension to Multiple Weighted Fields. In *Proceedings of the Thirteenth ACM International Conference on Information and Knowledge Management (CIKM '04)*. Association for Computing Machinery, New York, NY, USA, 42–49. https://doi.org/10.1145/1031171.1031181

[38] Stephen E. Robertson and Hugo Zaragoza. 2009. The Probabilistic Relevance Framework: BM25 and Beyond. *Found. Trends Inf. Retr.* 3, 4 (April 2009), 333–389.

[39] Gerard Salton, J. Allan, and Chris Buckley. 1993. Approaches to Passage Retrieval in Full Text Information Systems. In *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '93)*. Association for Computing Machinery, New York, NY, USA, 49–58. https://doi.org/10.1145/160688.160693

[40] T. Schlieder and H. Meuss. 2002. Querying and ranking XML documents. *Journal of the American Society for Information Science and Technology* 53, 6 (2002), 489–503. https://doi.org/10.1002/asi.10060

[41] Mark D. Smucker, James Allan, and Ben Carterette. 2007. A Comparison of Statistical Significance Tests for Information Retrieval Evaluation. In *Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management (CIKM '07)*. Association for Computing Machinery, New York, NY, USA, 623–632. https://doi.org/10.1145/1321440.1321528

[42] Fei Song and W. Bruce Croft. 1999. A General Language Model for Information Retrieval. In *Proceedings of the Eighth International Conference on Information and Knowledge Management (CIKM '99)*. Association for Computing Machinery, New York, NY, USA, 316–321. https://doi.org/10.1145/319950.320022

[43] Tao Tao and ChengXiang Zhai. 2007. An Exploration of Proximity Measures in Information Retrieval. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '07)*. ACM, New York, NY, USA, 295–302. https://doi.org/10.1145/1277741.1277794

[44] Michael Taylor, Hugo Zaragoza, Nick Craswell, Stephen Robertson, and Chris Burges. 2006. Optimisation methods for ranking functions with multiple parameters. In *Proceedings of the 15th ACM international conference on Information and knowledge management*. 585–593.

[45] Anastasios Tombros and Mark Sanderson. 1998. Advantages of Query Biased Summaries in Information Retrieval. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '98)*. Association for Computing Machinery, New York, NY, USA, 2–10. https://doi.org/10.1145/290941.290947

[46] Nicola Tonellotto, Craig Macdonald, and Iadh Ounis. 2018. Efficient Query Processing for Scalable Web Search. *Foundations and Trends in Information Retrieval* 12, 4–5 (2018), 319–492. http://dx.doi.org/10.1561/1500000057

[47] Andrew Trotman. 2005. Choosing Document Structure Weights. *Inf. Process. Manage.* 41, 2 (March 2005), 243–264. https://doi.org/10.1016/j.ipm.2003.10.003

[48] Andrew Trotman, Antti Puurula, and Blake Burgess. 2014. Improvements to BM25 and language models examined. In *Proceedings of the 2014 Australasian Document Computing Symposium*. 58–65.

[49] Mengqiu Wang and Luo Si. 2008. Discriminative Probabilistic Models for Passage Based Retrieval. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '08)*. ACM, New York, NY, USA, 419–426. https://doi.org/10.1145/1390334.1390407

[50] Ross Wilkinson. 1994. Effective Retrieval of Structured Documents. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '94)*. Springer-Verlag New York, Inc., New York, NY, USA, 311–317. http://dl.acm.org/citation.cfm?id=188490.188591

[51] J. E. Wolff, H. Florke, and A. B. Cremers. 2000. Searching and browsing collections of structural information. In *Proceedings IEEE Advances in Digital Libraries 2000*. 141–150. https://doi.org/10.1109/ADL.2000.848377

[52] Zhijing Wu, Jiaxin Mao, Yiqun Liu, Min Zhang, and Shaoping Ma. 2019. Investigating Passage-level Relevance and Its Role in Document-level Relevance Judgment. In *Proceedings of the 42Nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'19)*. ACM, New York, NY, USA, 605–614. https://doi.org/10.1145/3331184.3331233

[53] Chengxiang Zhai and John Lafferty. 2004. A Study of Smoothing Methods for Language Models Applied to Information Retrieval. *ACM Trans. Inf. Syst.* 22, 2 (April 2004), 179–214. https://doi.org/10.1145/984321.984322

[54] Jiashu Zhao, Jimmy Xiangji Huang, and Shicheng Wu. 2012. Rewarding Term Location Information to Enhance Probabilistic Information Retrieval. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '12)*. ACM, New York, NY, USA, 1137–1138. https://doi.org/10.1145/2348283.2348507