Enhanced News Retrieval: Passages Lead the Way!

Matteo Catena
ISTI-CNR, Pisa, Italy
matteo.catena@isti.cnr.it

Ophir Frieder
Georgetown University, USA
ophir@ir.cs.georgetown.edu

Cristina Ioana Muntean
ISTI-CNR, Pisa, Italy
cristina.muntean@isti.cnr.it

Franco Maria Nardini
ISTI-CNR, Pisa, Italy
francomaria.nardini@isti.cnr.it

Raffaele Perego
ISTI-CNR, Pisa, Italy
raffaele.perego@isti.cnr.it

Nicola Tonellotto
ISTI-CNR, Pisa, Italy
nicola.tonellotto@isti.cnr.it

ABSTRACT

We observe that most relevant terms in unstructured news articles 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 news articles. Our experimentation, conducted using three publicly available news datasets, demonstrates that BM25P markedly outperforms BM25 in term of effectiveness by up to 17.44% in NDCG@5 and 85% in NDCG@1.

CCS CONCEPTS

• **Information systems → Probabilistic retrieval models**;

ACM Reference Format:
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), July 21–25, 2019, Paris, France.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3331184.3331373

1 INTRODUCTION

Passage retrieval, present in literature for decades [10], is the task of retrieving only portions of documents, i.e., passages, relevant to a particular information need. At times, passage retrieval is viewed as an intermediate step in other information retrieval tasks, e.g., question answering and summarization.

Believing that certain passages pose greater relevance to a given query, we investigate how such relevance can be exploited to improve retrieval on a particular domain, specifically news retrieval. We differ from both existing passage retrieval [10] and passage detection [6] efforts, where the aim is to either retrieve only highly relevant passages or detect unrelated injected passages from within documents, respectively. In contrast, our goal is not to answer a query by retrieving single passages or detect injected unrelated passages, but to focus on improving the effectiveness of a retrieval system in retrieving entire news articles. To this end, we exploit passage relevance, capitalizing on their keyword density.

Specifically, we introduce a variant of the well-known BM25 weighting model [7], called BM25 Passage (BM25P), to improve the effectiveness of a news retrieval system. BM25P takes into account the entire news article when assigning a relevance score; however, BM25P distinguishes the importance of different news passages by assigning different weights to different passages. BM25P exploits such portions of text by creating a weighted linear combination of term frequencies per passage, improving the effectiveness of the news retrieval. To derive the weights, we analyze the density of highly discriminative terms in the collection of documents, measured in term of inverse document frequency, and observe where they are distributed throughout the content. This approach is efficient since it is query independent and is applied at index construction time, i.e., pre-retrieval.

The exploitation of term positions in Information Retrieval applications is common. One of the most notable examples related to our work is the BM25F weighting model [9], where term statistics are computed separately for the different fields that make up a document (e.g., title, headings, abstract and body) and then combined together within a BM25-based model. Our proposal resembles closely BM25F, but it considers the positions of the highly relevant terms occurring in the body of unstructured documents.

Term positions are also exploited in the news context for news summarization and classification tasks [2–4]. In news recommendation, the first few sentences and the article title are known to boost the performance of recommender systems. The performance of the system can be further improved by considering the rest of the document, and the best results can be observed when using the whole article text, as in our approach [1, 12]. This result suggests that although news articles tend to concentrate relevant content in the beginning, this does not necessarily imply that the remaining sections of the text can be ignored without hindering accuracy. By making the best of these two observations, we analyze the distribution of the occurrences of highly relevant terms and note that news documents belonging to different collections are consistently characterized by areas with different densities of highly relevant terms. We thus exploit this fact to improve a classical IR weighting model such as BM25. To the best of our knowledge, this is the first contribution in this direction exploiting the in-document distributions of impactful terms within news documents in the BM25 weighting function.

SIGIR '19, July 21–25, 2019, Paris, France

ACM ISBN 978-1-4503-6172-9/19/07…$15.00
https://doi.org/10.1145/3331184.3331373

2 PROBLEM

In modern information retrieval systems, given a user query, a *relevance score* is associated with the query-document pairs. Such relevance score is computed by exploiting a heuristic similarity function, estimating, according to some statistical procedure, the probability of relevance of a document with respect to a query. Then, the documents retrieved are ranked by their score, and the K documents with the highest score 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 [7]. In most IR 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: $s_q(d) = \sum_{t \in q} s_t(q, d)$, 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 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:

$$s_t(q, d) = w_q \frac{(k_1 + 1)tf}{k_1 \left((1 - b) + b\frac{dl}{avg\_dl}\right) + tf} w_{IDF}, \qquad (1)$$

where tf is the in-document term frequency, dl is the document length, 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 [9] the authors proved 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 formula. The novel tf factor boosts the specific fields without altering collection statistics. The BM25F model is considered one of the most successful Web search and corporate search algorithms [8].

With unstructured documents we lack the strong relevance signals derived from the term frequency of the query keywords in the different fields available in Web document. However, we formulate the hypothesis that in *curated* unstructured documents such as news articles it is possible to leverage the distribution of keywords in the documents to derive analogous strong relevance signals. To validate our hypothesis on the structure of news articles and to quantify the impact of some distinguishing document portions (referred to as passage in the following) over other portions, we analyze the density of highly discriminative terms in large news corpora. We consider a term as discriminative, hereinafter *key term*, if it appears in only a few documents, i.e., it has a high IDF value.

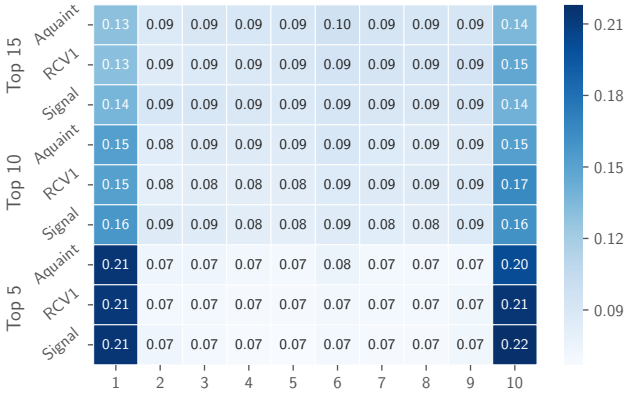

Figure 1: Probability distribution for the positions of key terms occurrences in the news articles of the three collections used.

For each news article in our three test collections (detailed in Sec. 3), we identify the positions of the occurrences of the k terms with the highest IDF. To aggregate such positional information, we evenly split each news article d into a set $P(d)$ of 10 passages having about the same length[1]. Then, we compute the distributions of the occurrences of the key terms 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 5, 10 and 15 key terms. As demonstrated by the plots, independently from the datasets considered, the first and last parts of news articles are more likely to include key terms than the remaining parts. Moreover, the lower the number k of the top key terms considered, the more skewed the probability distribution. The higher likelihood of key terms occurring in the opening passages was expected. 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[2]. 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 news article, can enhance retrieval effectiveness.

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

$$tf_P = \sum_{i \in P(d)} \alpha w_i \cdot tf_i. \qquad (2)$$

As suggested in [9] we plugged the term frequency tf_P into the original BM25 formula (Eq. (1)), rather than summing the BM25 scores per passage. The empirical probability distribution depicted

[1] Tests conducted with different values of $|P(d)|$ are not discussed for lack of space.
[2] http://handbook.reuters.com.

in Fig. 1 gives us a clear indication of the impact of each passage within the document from the point of view of important 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. We re-scale all weights with the hyperparameter α to amplify the importance of highly relevant terms in impactful passages. In the following we will use the distributions of top-5, top-10 and top-15 key terms as different passage weighting methods to be plugged into BM25P, which we henceforth refer to as $BM25P_5$, $BM25P_{10}$, and $BM25P_{15}$. Note that BM25P with all passage weights and α set to 1 is equivalent to BM25.

https://www.overleaf.com/project/5c506da7b0bc603b37fb19de

3 EXPERIMENTAL SETUP

The experimental assessment of the proposed weighting model relies on the following corpora of English news articles:

- the AQUAINT Corpus by Linguistic Data Consortium (`Aquaint`),
- the Signal Media One-Million News Articles Dataset (`Signal`),
- the Reuters Corpus, Volume 1, version 2 (`RCV1`).

The 2005 Robust and HARD TREC tracks provide 50 queries and their associated relevance judgements for the `Aquaint` dataset. The `Signal` and `RCV1` datasets do not provide any evaluation data. Hence, for these two datasets, we adopt the methodology described in [5] 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. Specifically, 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 three datasets are summarized in Table 1.

Table 1: Statistics of the three collections used.

Dataset	# Queries	avg. QLen	# Docs	avg. DocLen
`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

For each dataset, we index the unstructured body of news articles (by ignoring titles and all collection-specific fields such as source, category, media type, and publishing date) into positional indices, with Terrier. This type of index provides us with the positions of query term occurrences within the document, to differently weight the contribution of matching terms.

With the query relevance data built as detailed above, *we investigate if, by weighting news passages differently, our proposed BM25P model is able to improve retrieval effectiveness w.r.t. BM25.* We answer this research question by retrieving the top 1,000 documents for each query from the respective news corpus by using BM25 and BM25P. With BM25P, documents are virtually divided into 10 passages weighted as discussed above.

Once queries have been processed, we observe the rank of the relevant documents retrieved and compare the results obtained for BM25P with the BM25 ones. To measure retrieval effectiveness, we consider NDCG@k and MRR metrics. NDCG@k is used to evaluate the performance on the `Aquaint` dataset, where we have multiple relevant documents per query. Conversely, MRR, as 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 rank positions, especially for the `Signal` and `RCV1` datasets, where only one relevant document per query is known. We also evaluate the baseline BM25 and the weighting methods proposed for BM25P, i.e., $BM25P_5$, $BM25P_{10}$, and $BM25P_{15}$, for different values of the α hyper-parameter.

Table 2: NDCG at different cutoffs for BM25 and BM25P ($\alpha = 10$) on the `Aquaint` collection. We highlight statistical significant differences w.r.t. BM25 with ▲ for $p < 0.01$ and △ for $p < 0.05$ according to the two sample t-test [11].

NDCG@k	BM25	$BM25P_5$	$BM25P_{10}$	$BM25P_{15}$
NDCG@1	0.200	0.310 $_{+55.0\%}^{\triangle}$	**0.370** $_{+85.0\%}^{\blacktriangle}$	0.290 $_{+45.0\%}^{\triangle}$
NDCG@3	0.291	0.303 $_{+4.12\%}$	**0.335** $_{+15.01\%}$	0.317 $_{+8.78\%}$
NDCG@5	0.280	0.288 $_{+3.03\%}$	**0.329** $_{+17.44\%}^{\blacktriangle}$	0.301 $_{+7.39\%}$
NDCG@10	0.270	0.271 $_{+0.11\%}$	**0.298** $_{+10.20\%}^{\triangle}$	0.291 $_{+7.44\%}$
NDCG@15	0.269	0.271 $_{+0.65\%}$	**0.296** $_{+9.96\%}^{\blacktriangle}$	0.290 $_{+7.91\%}^{\triangle}$
NDCG@20	0.273	0.268 $_{-2.11\%}$	**0.289** $_{+5.81\%}^{\triangle}$	0.282 $_{+3.35\%}^{\triangle}$

4 EXPERIMENTAL RESULTS

The experiments conducted aim to assess whether BM25P achieves a better overall ranking quality with respect to BM25. Table 2 reports the NDCG at different cutoffs measured on the `Aquaint` dataset for BM25, $BM25P_5$, $BM25P_{10}$, and $BM25P_{15}$. All these tests were performed with $\alpha = 10$. We highlight that BM25P consistently outperforms BM25. Indeed, $BM25P_{10}$ results 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 five of the six cases, $BM25P_{10}$ shows statistically significant results with p-values of $p < 0.01$. The other proposed methods, i.e., $BM25P_5$ and $BM25P_{15}$, also improve NDCG over BM25, although with smaller relative benefits. In only one case, NDCG@20 with $BM25P_5$, our weighting model has a lower NDCG than BM25, but the difference is not statistically significant.

We further investigate the performance of BM25P against BM25 on the `Aquaint` dataset, by varying α to assess the impact of this hyper-parameter on the retrieval effectiveness measured in terms of 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 $BM25P_5$ and $BM25P_{10}$, the effectiveness does not sensibly increase for α values greater than 10, while for $BM25P_{15}$ the performance tends to increase even if it is not able to outperform the one of $BM25P_{10}$ for any value of α. In conclusion $BM25_{10}$ with $\alpha = 10$ is the best weighting model in terms of NDCG@5.

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 3, where we evaluate the retrieval performance in terms of MRR for the `Signal`, `RCV1` and `Aquaint` datasets.

The results show that BM25P performs significantly better than BM25 on all three datatsets, thus confirming the results achieved by

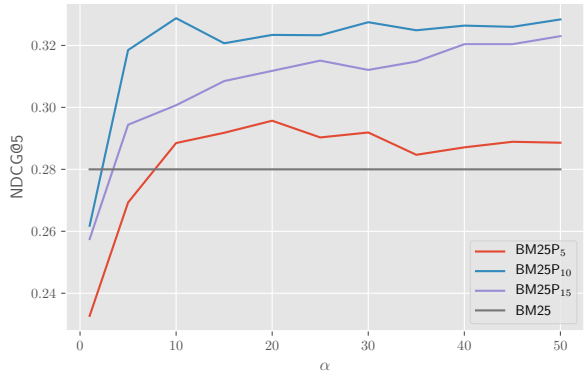

Figure 2: NDCG@5 for BM25 and BM25P, on Aquaint, for BM25P and for different values of α from 5 to 50.

BM25P in terms of NDCG@k on the Aquaint collection. Indeed, results also confirm that the best performing method on this dataset is $BM25P_{10}$ when $\alpha = 10$. A slightly different result is achieved for the Signal and RCV1 datasets, where the best performing method results to be $BM25P_5$. Indeed, on these collections, $BM25P_5$, $BM25P_{10}$ and $BM25P_{15}$ always show statistically significant improvements w.r.t. BM25 with $p < 0.01$ for $\alpha \geq 10$.

Table 3 reports the MRR while varying the value of α. MRR is higher 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. 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.(1). 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 best performing setting is $BM25P_{10}$ ($\alpha = 10$) for Aquaint and $BM25P_5$ ($\alpha = 20$) 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 $BM25P_5$, which gives a larger importance to the first and last passages and seems to better approximate where the pseudo-queries match. Indeed, results achieved with MRR for Aquaint are consistent with the ones discussed for NDCG@k; namely $BM25P_{10}$ is the best method and statistically outperforms BM25. $BM25P_5$ and $BM25P_{15}$ also behave well on Aquaint, but the improvement is statistically significant just for few values of α in the case of $BM25P_{15}$. $BM25P_{10}$ uses top 10 highest IDF terms in each document to create a probability distribution of their positions. We also look at top 15 in the case of $BM25P_{15}$, but increasing the number of terms for computing the distribution does not yield better results. We can conclude that 10 terms for Aquaint and 5 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.

5 CONCLUSIONS

For news articles, we observed that a common stylistic feature is the preponderance of occurrences of key terms (i.e., terms with an high IDF value) at the beginning and at the end of the article. We proposed BM25P, a variant of BM25, which considers key term distribution variations among the different passages of the news. 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. Our experiments showed that, by differently weighting news passages, BM25P markedly improves NDCG and MRR with respect to using BM25. In particular, we observed that BM25P significantly improves 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.

As future work we plan to study the impact of (adaptively) varying the number of passages weighted – here set equal to 10 – and the use of our BM25P model in conjunction with BM25F for retrieving semi-structured news articles.

Acknowledgements. This paper is supported by the EU H2020 BIGDATAGRAPES (grant agreement Nº780751).

Table 3: MRR for BM25 and BM25P on the three collections for different values of α. We report statistical significance w.r.t. BM25 with ▲ for $p < 0.01$ and △ for $p < 0.05$.

		α						
	Model	1	5	10	20	30	40	50
Aquaint	BM25	0.485	0.485	0.485	0.485	0.485	0.485	0.485
	$BM25P_5$	0.438	0.518	0.547	0.548	0.544	0.554	0.554
	$BM25P_{10}$	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▲

REFERENCES

[1] Toine Bogers and Antal van den Bosch. 2007. Comparing and Evaluating Information Retrieval Algorithms for News Recommendation. In *Proc. RecSys*. ACM, 141–144.
[2] Jose M. Chenlo and David E. Losada. 2014. An empirical study of sentence features for subjectivity and polarity classification. *Inf. Sc.* 280 (2014), 275 – 288.
[3] Dipanjan Das and André F.T. Martins. 2007. A survey on automatic text summarization. *Lit. Survey for the Lang. and Stat. II course at CMU* 4 (2007), 192–195.
[4] Chin-Yew Lin. 1999. Training a Selection Function for Extraction. In *Proc. CIKM*. ACM, 55–62.
[5] Sean MacAvaney, Andrew Yates, Kai Hui, and Ophir Frieder. 2019. Content-Based Weak Supervision for Ad-Hoc Re-Ranking. In *SIGIR 2019*.
[6] Saket Mengle and Nazli Goharian. 2009. Passage detection using text classification. *JASIST* 60, 4 (2009), 814–825.
[7] Stephen E. Robertson, Steve Walker, Susan Jones, Micheline Hancock-Beaulieu, and Mike Gatford. 1996. Okapi at TREC-3. 109–126.
[8] Stephen E. Robertson and Hugo Zaragoza. 2009. The Probabilistic Relevance Framework: BM25 and Beyond. *Found. Trends Inf. Retr.* 3, 4 (April 2009), 333–389.
[9] Stephen E. Robertson, Hugo Zaragoza, and Michael Taylor. 2004. Simple BM25 Extension to Multiple Weighted Fields. In *Proc. CIKM*. ACM, 42–49.
[10] Gerard Salton, James Allan, and Chris Buckley. 1993. Approaches to Passage Retrieval in Full Text Information Systems. In *Proc. SIGIR*. ACM, 49–58.
[11] Mark D. Smucker, James Allan, and Ben Carterette. 2007. A Comparison of Statistical Significance Tests for Information Retrieval Evaluation. In *Proc. CIKM*. ACM, 623–632.
[12] Anastasios Tombros and Mark Sanderson. 1998. Advantages of Query Biased Summaries in Information Retrieval. In *Proc. SIGIR*. ACM, 2–10.