Literature Review of BM25 Extensions

Jie Zheng
University of Illinois at Champaign-Urbana

Abstract: This paper examines various extensions of BM25 in the recent 10 years. Different approach modifies BM25 formulation from different perspective, and achieves better performance for different applications.

Introduction

BM25 is unquestionably among the most effective ranking methods for information retrieval, where BM stands for best matching. Since its inception by Robertson in 1994 [1], researchers have proposed multiple extensions to improve BM25 such as BM25-F, ATIRE BM25, and etc. This paper offers a systematic review of majority BM25 variants in the past 10 years.

BM25 Variations

Table 1a is a summary of BM25 variations. Due to the fact that different researchers might have utilized different notations for the scoring function, all the equations are re-written with the same set of symbols for comparability. Please refer to table 1b for a detailed explanation of the notations used in this review.

BM25

Robertson et al. laid the foundation of BM25 in 1994 [1, 2, 3, 4]. The original BM25 scoring equation (1) mainly consists of 3 parts: inverse document frequency (IDF) factor determined by $df(w)$, term frequency (TF) factor determined by $c(w,d)$, and document length normalization (DLN) factor determined by dl. k_1 and b are parameters setting the scale of TF and DLN, and a good range of k_1 and b are found to be [1.2, 2] and [0.5, 0.8], respectively [4] for various situations. The k_2 component (query length weighting and further document length adjustment) and k_3 component (query term frequency weighting) are frequently left out due to their negligible contribution to the scoring function as proven by multiple experiments.

BM25-F

For structured text data retrieval using BM25, a common approach is to first compute the BM25 score for each field, then combine all the scores linearly to achieve a final score. However, Robertson et al. [5, 6, 4] pointed out that there are several drawbacks with this approach, especially the violation of the non-linearity of TF transformations. This is not desirable because it over-rewards re-occurrences of the same term in different fields. Therefore, they proposed BM25-F, where F denotes fields, such as title, abstract, body, etc... In BM25-F, term frequency at different fields was first linearly combined with corresponding weighting factors for each field, as shown in equation (2). Then the total ranking is computed using the combined term frequency. Robertson et al. theoretically derived that BM25-F holds the nonlinearity of term frequency transformation, and they also empirically demonstrated that BM25-F outperforms linear combinations of score from each field using 2 data sets, Reuters Vol1 and REC dotGov collection [5].

ATIRE BM25

To address the issue of negative IDF in original BM25 formula [7], Trotman et al. [8] introduced a variation of the IDF component in 2012. In the original BM25, when document frequency is more than half of the entire collection, the scoring function would generate a negative IDF. I.e., a document containing a query term that appeared in more than half of the entire collection would rank lower than documents that didn't contain the query term, which is counter-intuitive. While in ATIRE BM25 equation (3), the IDF value never goes negative. That is to say, a document containing a specific query term is always ranked higher than documents without the

query term. Applying supervised learning with the data set of INEX 2008 Wikipedia, k_1 and b are optimized to 0.9 and 0.4 respectively. [8]

BM25L

In 2011, Lv and Zhai [9] observed that BM25 would fail to rank a very long document containing a specific query term more relevant than the documents without the query term. This was caused by the document length normalization in the sub-linear term frequency transformation component of the original BM25 equation. When the document length dl is much larger than the average document length $avdl$, the TF transformation component tends to be zero. Namely, the BM25 scoring function couldn't differentiate a long document including a specific query term from documents where query terms are absent. By adding a shift parameter δ ($\delta > 0$) to the term frequency component as shown in equation (4.2), Lv and Zhai essentially set a non-zero lower bound of the term frequency component for documents containing query terms, thus alleviating over-penalization of very long documents. Text retrieval of 6 data sets proved that BM25L consistently surpassed BM25 and achieved higher MAP, with the shift parameter δ set to 0.5, k_3 set to 1000, k_1 and b in range [0.2,3] and [0.1,0.9] correspondingly [9].

BM25+

Built upon their work on BM25L [9], Lv and Zhai [10] further generalized their approach to remediate failure in scoring lengthy documents, a common problem existed in many retrieval models including BM25, PL2, Dirichlet Prior Method (Dir), pivoted length normalization (Piv) … By simply adding a positive parameter δ to the term frequency factor of the original BM25 scoring function as in equation (5), the retrieval precision was substantially improved especially for lengthy queries across diverse data sets. The parameter δ explicitly sets the lower bounds of the term frequency component, or intuitively, an enforced score boost for documents with $c(w,d) > 0$. Lv and Zhai also theoretical proved that BM25+ effectively addressed the common problem of over-penalization of long documents. Based on experiments, setting δ=1 in the BM25+ equation achieves satisfactory mean average precision (MAP) and efficiency for all the cases. Furthermore, Lv and Zhai applied the same concept – enforced lower bound of term frequency factor – to PL2, Dir, and Piv and derived PL2+, Dir+, and Piv+ ranking functions correspondingly, all with considerable advances in text retrieval, particularly for long documents.

BM25-adp

BM25 has remained one of the most popular methods for text retrieval for 20 years and different lines of work have been performed to improve BM25 with a fixed and term-independent k_1. Lv and Zhai in 2011 [11] suggested an adaptive approach to derive the term-specific k_1 based on information gain measure IG_w^t as defined in equation (6.1). IG_w^t is the information contribution of matching a specific term w from t to $t+1$ times, and $df_t(w)$ was the rounded term frequency $c'(w,D)$ after document length normalization as in equation (6.2). Then $\widehat{k_1}(w)$ was directly calculated by associating term frequency normalization component in BM25 with the information gain measure IG_w^t as in equation (6.4), which can be efficiently done offline. In addition to the term-specific $\widehat{k_1}(w)$, BM25-adp equation (6.4) replaces the inverse document frequency component of BM25 with IG_w^1. Lv and Zhai conducted experiments with various data sets and proved that BM25-adp was more effective and efficient compared to standard BM25.

BM25T

Historically, the probabilistic interpretation of k_1 remains under-defined and thus leading to the challenge of optimizing k_1, especially in absence of training data. Lv and Zhai filled in this gap in 2012 [12] by introducing BM25T, which automatically computes the k_1 parameter based on a log-logistic model without the need of training data. The "T" in BM25T denotes term, and k_1 is term-specific which is different from the traditional BM25 [1, 2, 3]. This scheme provides a probabilistic view of k_1 "as the scale parameter of the log-logistic model", and the sub-linear term frequency component in traditional BM25 was analytically proven to be a special case of

the log-logistic model. This can be intuitively interpreted as following: for a specific query term w, the information contributed by matching the query term $c(w,d)$ times in document d is related to the term frequency of other documents in the collection containing the query term w, i.e., the elite set defined in [12]. However, BM25T might be problematic and lead to imprecise estimation of $k_1(w)$ in absence of term frequency values, for example, text retrieval of a query containing a rare term. To resolve this data sparseness issue, Lv and Chai further proposed two more approaches, BM25C and BM25Q, where C stands for collection and Q stands for query correspondingly. In BM25C, $k_1(w)$ across all the query terms for collection C was averaged to generate a term-independent but collection-specific k_1; while in BM25Q, $k_1(w)$ across the same query was averaged to generate a query-specific k_1. With 4 data sets (Robust04, WT2G, WT10G, and news collection AP) of different length, genre, and homogeneity, Lv and Zhai experimentally measured that BM25T, BM25C, BM25Q consistently deliver better or comparable results to an optimized BM25 tuned with training data.

BM25-CTF

Jimenez et al. [13] observed that IDF in BM25 often suffer from data sparseness in spite of its effectiveness. In addition, collection term frequency (CTF) could contribute to document ranking from a different perspective for terms with same or similar IDF. Intuitively, comparing 2 terms with same IDF, the one with higher CTF appears more frequently in a document on average, thus less useful in information retrieval. By integrating CTF into the TF and IDF component of BM25, Jimenez et al. derived a BM25-CTF function with Boosted-TF and Boosted-IDF factors as shown in equation 8.1. The Boosted-TF factor utilizes the "divergence measure from randomness" [13], i.e., the ratio of observed $c(w,d)$ and its estimate with an adjusting factor $C(d)$. The Boosted-IDF factor is the product of inverse collection term frequency (CTF), IDF as in classic BM25, and Poisson-IDF. Testing with TREC-1 to TREC-8 proves that BM25-CTF provides substantial improvement in MAP with negligible computational cost overhead comparing to classic BM25.

Conclusion

In this paper, we evaluated the major variants of BM25, incl. BM25F, ATIRE BM25, BM25L, BM25+, BM25-adp, BM25-T, and BM25-CTF. Even though some approach has achieved better results for certain applications, there is no clear winner according to empirical comparison with TREC collections [14]. A continuation of improvement in BM25, in combination with recent advances in machine learning and artificial intelligence could benefit information retrieval.

BM25	Robertson el al.	$f(q,\ d) = \sum_{w\in q} log\ \frac{N-df(w)+0.5}{df(w)+0.5} \cdot \frac{(1+k_1)\cdot c(w,d)}{k_1\cdot\left(1-b+b\frac{dl}{avdl}\right)+c(w,d)} \cdot \frac{(1+k_3)c(w,q)}{k_3+c(w,q))} + k_2 \cdot \lvert q\rvert \frac{avdl-dl}{avdl+dl}$	(1)
BM25-F	Robertson el al.	$c'(w,\ d) = \sum_{f=1}^{K} v_f\ c(w,f)$	(2)
ATIRE BM25	Trotman et al.	$f(q,\ d) = \sum_{w\in q} log\ \frac{N}{df(w)} \cdot \frac{(1+k_1)\cdot c(w,d)}{k_1\cdot\left(1-b+b\frac{dl}{avdl}\right)+c(w,d)}$	(3)
BM25L	Lv and Zhai	$f(q,\ d) = \sum_{w\in q} log\ \frac{N+1}{df(w)+0.5} \cdot f'(q,D) \cdot \frac{(1+k_3)c(w,q)}{k_3+c(w,q))}$	(4.1)
		$f'(q,D)\ = \begin{cases} \frac{(1+k_1)[c'(w,d)+\delta]}{k_1+[c'(w,d)+\delta]} & if\ \ c'(w,d) > 0 \\ 0 \quad otherwise \end{cases}$	(4.2)
		$c'(w,\ d) = \frac{c(w,d)}{1-b+b\frac{dl}{avdl}}$	(4.3)
BM25+	Lv and Zhai	$f(q,\ d) = \sum_{w\in q} log\ \frac{N+1}{df} \cdot [\frac{(1+k_1)\cdot c(w,d)}{k_1\cdot\left(1-b+b\frac{dl}{avdl}\right)+c(w,d)} + \delta] \cdot \frac{(1+k_3)c(w,q)}{k_3+c(w,q))}$	(5)
BM25-adp	Lv and Zhai	$IG_w^t = -log_2\left(\frac{df(w)+0.5}{N+1}\right) + log_2(\frac{df_{t+1}(w)+0.5}{df_t(w)+1})$	(6.1)

		$df_t(w) = \begin{cases} \lvert\{D \lvert c'(w,D) \geq t - 0.5\}\rvert \quad if\ t \in \{2,3,\dots\} \\ df(w) \quad if\ t = 1 \\ N \quad if\ t = 0 \end{cases}$	(6.2)
		$f(q,\ d) = \sum_{w \in q} IG_w^1 \ \cdot [\frac{(1+\widehat{k_1}(w))\cdot c\prime(w,d)}{\widehat{k_1}(w)+c\prime(w,d)}] \cdot \frac{(1+k_3)c(w,q)}{k_3+c(w,q))}$	(6.3)
		$\widehat{k_1}(w) = arg\min_{k_1} \left(\frac{IG_w^t}{IG_w^1} - \frac{(1+k_1)\cdot t}{k_1+t}\right)^2$	(6.4)
		$c'(w,\ d) = \frac{c(w,d)}{1-b+b\frac{dl}{avdl}}$	(6.5)
BM25T	Lv and Zhai	$f(q,\ d) = \sum_{w \in q} log\ \frac{N+1}{df(w)} \cdot [\frac{(1+\widehat{k_1}(w))\cdot c\prime(w,d)}{\widehat{k_1}(w)+c\prime(w,d)}] \cdot \frac{(1+k_3)c(w,q)}{k_3+c(w,q))}$	(7.1)
		$c'(w,\ d) = \frac{c(w,d)}{1-b+b\frac{dl}{avdl}}$	(7.2)
		$\widehat{k_1}(w) = arg\min_{k_1} \left(g(k_1) - \frac{1}{\lvert C_w\rvert}\sum_{D \in C_w} \log(c'(w,D)+1)\right)^2$	(7.3)
		$g(k_1) = \begin{cases} \frac{k_1}{k_1-1}\log(k_1) \\ 1 \quad otherwise \end{cases} if \quad k_1 \neq 1$	(7.4)
BM25-CTF	Jimenez et al.	$f(q,\ d) = \sum_{w \in q} bidf(w) \cdot \frac{(1+k_1)\cdot btf(w,d)}{k_1\cdot\left(1-b+b\frac{dl}{avdl}\right)+btf(w,d)} \cdot \frac{(1+k_3)btf(w,q)}{k_3+btf(w,q))}$	(8.1)
		$btf(w,d) = \frac{dl}{\sum_{w' \epsilon d'} \frac{c(w',d)}{c(\widehat{w'},d)}} \cdot \frac{c(w,d)}{c(\widehat{w},d)}$	(8.2)
		$bidf(w) = ictf(w) \cdot idf(w) \cdot pidf(w)$	(8.3)
		$ictf(w) = \log(\frac{M}{ctf(w)})$	(8.4)
		$idf(w) = \log(\frac{N-df(w)+0.5}{df(w)+0.5})$	(8.5)
		$pidf(w) = \log(\frac{\widehat{df(w)}}{\widehat{ctf(w)}} + 1)$	(8.6)

Table 1a. Formulation of BM25 variants

N	Number of documents in the collection
$c(w,d)$	Term frequency, total occurrence of the term w within the document d
$c^{'}(w,d)$	Combined term frequency of the term w within the document d composed of multiple fields
$c(w,\mathrm{f})$	Term frequency of the term w within the field f (f could be title, abstract, …) within a document
v_f	Weight factor of term frequency of field f
$c(w,\mathrm{q})$	Query term frequency, total occurrence of the term w within the query q
$df(\mathrm{w})$	Document frequency, number of documents containing the term w

dl	Document Length
$avdl$	Average document length
$\|q\|$	Query length, number of query terms
k_1, k_2, k_3, b	Parameters used in BM25 functions
δ	Shift parameter in BM25L
$ctf(w)$	Collection term frequency, total occurrence of the term w within the entire collection
M	Number of terms in the entire collection

Table 1b. Notation Table

References

1. S. Robertson, S. Walker, S. Jones, M. Beaulieu, M. Gatford. Okapi at TREC-3. *TREC-3*, pages 109-126, 1994.
2. S. Robertson, S. Walker, M. Beaulieu, M. Gatford, A. Payne. Okapi at TREC-4. *TREC-4*, pages 73-96, 1995.
3. S. Robertson and S. Walker. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *SIGIR '94*, pages 232–241, 1994.
4. S. Robertson and H. Zaragoza. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends In Information Retrieval*, 3(4): pages 333–389, 2009.
5. S. Robertson, H. Zaragoza, and M. Taylor. Simple BM25 Extension to Multiple Weighted Fields. In *CIKM '04*, pages 42–49, 2004.
6. H. Zaragoza, N. Craswell, M. Taylor, S. Saria, and S. Robertson, Microsoft Cambridge at TREC-13: Web and HARD tracks. In *Proceedings of TREC-2004*.
7. H. Fang, T. Tao, and C. Zhai. A formal study of information retrieval heuristics. In *SIGIR '04*, pages 49–56, 2004.
8. A. Trotman, X. Jia, M. Crane, Towards an Efficient and Effective Search Engine. SIGIR 2012 Workshop on Open, *Source Information Retrieval*, pages 40-47, Portland 2012.
9. Y. Lv and C. Zhai. When documents are very long, BM25 fails! *SIGIR*, pages1103-1104, 2011.
10. Y. Lv and C. Zhai. Lower-bounding term frequency normalization. *CIKM*, pages 7-16, 2011.
11. Y. Lv and C. Zhai. Adaptive term frequency normalization for BM25. *CIKM*, pages 1985-1988, 2011
12. Y. Lv and C. Zhai, A log-logistic model-based interpretation of TF normalization of BM25. *ECIR*, pages 244-255, 2012.
13. S. Jimenez, S. P. Cucerzan, F. A. Gonzalez, A. Gelbukh and G. Duenas. BM25-CTF: Improving TF and IDF factors in BM25 by using collection term frequencies. Journal of Intelligent & Fuzzy Systems, vol. 34, no. 5, pages 2887-2899, 2018.
14. A. Trotman, A. Puurula, and B. Burgess. Improvements to BM25 and language models examined. In *Proceedings of ACM ADCS '14*, pages 58–65, 2014.