Review of BM25 and its variants

Introduction

BM25 [1] is one of the most widely-used ranking functions in the "bag of words" text retrieval. A ranking function is used to estimate how relevant the documents are given a search query. The ''BM" in BM25 stands for "best matching". People also call it Okapi BM25 since it was first used in the Okapi information retrieval system, implemented at London's City University in the 1980s and 1990s [2]. BM25 is developed based on the probabilistic model, but can also fit in the vector space model framework. It includes the within-document term frequency information and document length normalization. Although decades old, it is still one of the most effective ranking formulas.

Over the years, researchers have developed some variants based on the original BM25, such as BM25F [3] and BM25+ [4]. These variants have improvements upon the original BM25 to serve specific goals. The next section will show some typical variants.

Formulas of BM25 and it variants

BM25

Robertson et al. [1] developed the original formulation of BM25, with the formula as follows:

$$\sum_{t \in q} \log\left(\frac{N - df_t + 0.5}{df_t + 0.5}\right) \cdot \frac{tf_{td}}{k_1 \cdot \left(1 - b + b \cdot \left(\frac{L_d}{L_{avg}}\right)\right) + tf_{td}} \quad , \qquad (1)$$

where N is the total number of documents in the collection, df_t is the number of documents containing term t, tf_{td} is the term frequency of term t in document d. L_d is the length of document d, L_{avg} is the average document length in the collection. k_1 is a parameter, b is the slope parameter in pivoted normalization. In this formula, the log component is the IDF (inverse document frequency) weight of the query term.

Since in equation (1) the IDF component is negative when $df_t > N/2$, it is often written as the following equation (2) in practice to avoid this problem, such as in the Apache Lucene search engine software [5].

$$\sum_{t \in q} \log\left(1 + \frac{N - df_t + 0.5}{df_t + 0.5}\right) \cdot \frac{tf_{td}}{k_1 \cdot \left(1 - b + b \cdot \left(\frac{L_d}{L_{avg}}\right)\right) + tf_{td}} \quad . \qquad (2)$$

In the ATIRE search engine developed by Trotman et al. [6], they use the BM25 in the following form:

$$\sum_{t \in q} \log\left(\frac{N}{df_t}\right) \cdot \frac{(k_1 + 1) \cdot tf_{td}}{k_1 \cdot \left(1 - b + b \cdot \left(\frac{L_d}{L_{avg}}\right)\right) + tf_{td}} \quad . \qquad (3)$$

This form also avoids the negative value problem, and it adds the multiplier ($k_1 + 1$) to the term frequency tf_{td} .

BM25F

Robertson and Zaragoza in 2009 developed the BM25F [3] ranking function, where F stands for fields. This variant of BM25 is used for documents with structures.
The BM25F formula is as follows:

$$\sum_{i \in q} \log \frac{N}{df_i} \cdot \frac{(k_1+1)\tilde{tf}_i}{k_1+\tilde{tf}_i} \quad (4)$$

where

$$\tilde{tf}_i = \sum_{z=1}^{Z} v_z \frac{tf_{zi}}{B_z}, \quad B_z = \left((1-b_z)+b_z \frac{len_z}{avlen_z}\right), \quad 0 \le b_z \le 1$$

Here i is the query term, df_i is the number of documents containing term i, tfz_i is the term frequency of term i in zone z. L_d is the length of document d, $lenz$ is length of zone z, $avlenz$ is the average zone length. v_z is zone weight. bz is the slope parameter in pivoted normalization in zone z.

A document can be considered as of many zones, such as the title, the abstract, the author, the body and the anchors. If we simply use BM25 for each zone and combine the zone scores, it implies that that the same term from different zones is different and is independent of each other, which is not reasonable. Instead, BM25F first combine the frequency counts of terms in all the fields, and then apply BM25. This has the advantage of avoiding over-counting the first occurrence of the term in each zone, which contributes a large weight.

BM25+

In 2011, Lv and Zhai published some BM25 variants, including BM25L [9], BM25+ [8] and BM25-adpt [7], among which BM25+ is the most widely used.

BM25+ addresses the problem of over penalization of long documents by BM25. It simply adds a small constant to the term frequency normalization formula. The equation for BM25+ is as below:

$$\sum_{t \in q} \log\left(\frac{N+1}{df_t}\right) \cdot \left(\frac{(k_1+1)\cdot tf_{td}}{k_1 \cdot \left(1-b+b\cdot\left(\frac{L_d}{L_{avg}}\right)\right)+tf_{td}} + \delta\right) \quad (5)$$

where δ is a small constant.

BM25L [9] also addresses the issue that BM25 penalizes longer documents too much compared to shorter ones. Instead, the small constant δ is added to the term c_{td}, with
$c_{td} = tf_{td}/(1-b+b\cdot(L_d/L_{avg}))$.
The formula for BM25L is as below:

$$\sum_{t \in q} \log\left(\frac{N+1}{df_t+0.5}\right) \cdot \frac{(k_1+1)\cdot(c_{td}+\delta)}{k_1+c_{td}+\delta} \quad (6)$$

BM25-adpt [7] is the adaptive term frequency normalization for BM25. Unlike BM25 where the parameter k_1 is term-independent, BM25-adpt sets it in a term-specific way. It uses an information gain measure to estimate the contributions of repeated term occurrences.

Conclusion

BM25 has various extensions. Most of its extensions is as simple as the original BM25, but can handle some specific requirements such as avoiding negative IDF, not over-penalizing the long documents, treating documents with structures as of different zones. These variants of BM25 improve the original algorithm and make it more adaptable to the fast-developing search engines of text retrieval purposes.

References

1. Robertson, S.E., Walker, S., Jones, S., Hancock-Beaulieu, M., Gatford, M.: Okapi at TREC-3. In: Proceedings of the 3rd Text Retrieval Conference (TREC-3), pp. 109–126, Gaithersburg (1994)

2. "Okapi BM25" Wikipedia, Wikimedia Foundation, 24 February 2021, https://en.wikipedia.org/wiki/Okapi_BM25.

3. Robertson, S., Zaragoza, H.: The probabilistic relevance framework: BM25 and beyond. Found. Trends Inf. Retrieval 3(4), 333–389 (2009)

4. Lv, Y., Zhai, C.: Lower-bounding term frequency normalization. In: Proceedings of the 20th ACM International Conference on Information and Knowledge Management (CIKM 2011), pp. 7–16, Glasgow (2011)

5. Kamphuis C., de Vries A.P., Boytsov L., Lin J. (2020) Which BM25 Do You Mean? A Large-Scale Reproducibility Study of Scoring Variants. In: Jose J. et al. (eds) Advances in Information Retrieval. ECIR 2020. Lecture Notes in Computer Science, vol 12036. Springer, Cham.

6. Trotman, A., Jia, X.F., Crane, M.: Towards an efficient and effective search engine. In: SIGIR 2012 Workshop on Open Source Information Retrieval, pp. 40–47, Portland (2012)

7. Lv, Y., Zhai, C.: Adaptive term frequency normalization for BM25. In: Proceedings of the 20th ACM International Conference on Information and Knowledge Management (CIKM 2011), pp. 1985–1988, Glasgow (2011)

8. Lv, Y., Zhai, C.: Lower-bounding term frequency normalization. In: Proceedings of the 20th ACM International Conference on Information and Knowledge Management (CIKM 2011), pp. 7–16, Glasgow (2011)

9. Lv, Y., Zhai, C.: When documents are very long, BM25 fails! In: Proceeding of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2011), pp. 1103–1104, Beijing (2011)