Comparison Between BM25 and Variants

Kefan Chen

Background

BM25 is one of the most classic score functions that we learnt and used in our text analysis tasks. It conveys the meaningful ideas of several important factors and features that we always need to deal with or consider when applying score function. There have been a great many BM25 variants, either changing the way regarding dealing with certain relevant ideas, or incorporating new features to satisfy specific scenarios.

In this review, we will compare BM25 and its variants to see and investigate their differences, thinking about the reasons why there would be these changes, and how these differences could be useful or applied to certain tasks.

Comparisons

BM25

$$\sum_{t\in q} log(\frac{N-df_t+0.5}{df_t+0.5}) * \frac{tf_{td}}{k_1*(1-b+b*(\frac{L_d}{L_{avg}}))+tf_{td}}$$

The above is the most classic and original formula of the BM25 score function. As also discussed in the lecture, we can see that the formula specifically considers the following aspects in its formula.

- Document Frequency $log(\frac{N-df_t+0.5}{df_t+0.5})$. Here the formula will penalize the term which appears frequently in more documents. This is because if a term tends to appear in more documents, then it is more likely to be a common word instead of a meaningful word that we can use to match a specific document, so we do not need to put much weight on this term.
- Term Frequency tf_{td}. If a term appears more times in a document, then this term is more likely to be more related to this term. So we should consider assigning more weights to this relation. However, the rewarding should not be linear, since the first occurrence or the first occurrences usually have much more meaning compared to later repetitive appearance. So we can see the formula here actually rewards more score for initial appearances.
- Document Length $\frac{L_d}{L_{abg}}$. The formula also considers the document length, since the longer the document is, it is more likely to include more terms as well. Based on this understanding, we should reward shorter documents and penalize longer documents.

Lucene Variant

$$\sum_{t\in q} log(1 + \frac{N-df_t+0.5}{df_t+0.5}) * \frac{tf_{td}}{k_1*(1-b+b*(\frac{L_{dlossy}}{L_{avg}}))+tf_{td}}$$

The Lucene Variant is based on the original BM25 formula and slightly changes how a few relevant factors are treated in the score function mainly with the following difference.

- In the original BM25 formula, suppose for a specific word, it appears in more than half of the documents in the collection, its IDF component $log(\frac{N-df_t+0.5}{df_t+0.5})$ would be negative. However this might not make sense to be a penalty for the overall matching score. So the Lucene Variant changes this component to $log(1 + \frac{N-df_t+0.5}{df_t+0.5})$, which means the value is always positive at least, no matter how frequently a term appears in the whole collection.
- In the original BM25 formula, it uses the actual document length to do a long document penalty. But here in Lucene Variant, it uses L_{dlossy}instead so that the document length component computation would be much cheaper and could be pre-computed more easily, though it also means it might not be as accurate and detailed as the original BM 25 formula.

BM25+

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

BM25+ is a meaningful variant which mainly improves the deficiency of the unfair document length penalty for long documents in the original BM25 formula. Its detailed changes include the following.

- The term frequency is modified to add δ value, which serves as a bonus for terms which appear at least once in the documentation. This could effectively reduce the bias applied to the long document.
- The document frequency part is modified to avoid negative values. This is similar to the first point we discussed in Lucene Variant.

BM25L

$$\sum_{t\in q} log(\frac{N+1}{df_t+0.5}) * \frac{(k_1+1)*(\frac{tf_{td}}{1-b+b*(\frac{L_d}{L_{avg}})}+\delta)}{k_1+\frac{tf_{td}}{1-b+b*(\frac{L_d}{L_{avg}})}+\delta}$$

BM25L actually expresses a very similar idea regarding how to improve based on the original BM25 formula compared to the BM25+ variant discussed above, though via a different way.

- The term frequency part is differently formulated then modified in order to increase the score for long documents.
- The document frequency part is modified to avoid negative values. This is similar to the first point we discussed in Lucene Variant.

BM25F

BM25F is actually a different variant compared to the above variants. The above variants mainly introduce different methods to deal with the factors in the original BM25 formula. BM25F instead incorporates other information to specifically account for structured documents.

For a structured document, there are always multiple fields embedded in the document, for example, title, abstract, heading, body, etc. If we are evaluating conference papers, then the abstract section would definitely include the most important information and keywords. So it would be natural thinking that we should be able to assign more weights to fields that we believe would be more important.

BM25F essentially discussed a possible way to achieve it. BM25F will weight the term frequency according to the weights of the fields, then combine all the fields resulting in pseudo frequencies. For example, if we believe the abstract field is twice as important as the body field, we would want to assign double weights to abstract field vs body field, which means we can double the abstract field to achieve this purpose.

Summary

As we can see from the comparisons above, all variants have their idea and formula based on the classic scoring function BM25, but vary on the approach to deal with some details and incorporate other points. This comparison does not cover all the possible variants, as we can still easily find other variants online trying to improve on other details or using a different way to improve the details.

There is no single best form of BM25 variant that we can always use in our text analysis work. When doing a specific job or achieving a specific use case, we always need to consider what would be relatively important factors we have to specifically consider for our use case, thus choosing the most suitable BM25 variant, or trying to see which one could potentially outperform others for our task. We are also hoping to see more novel ideas in the new BM25 variant and their applications as well.

Reference

1. Wikipedia. Okapi BM25. https://en.wikipedia.org/wiki/Okapi_BM25.

2. Wikipedia. Lossy Compression. https://en.wikipedia.org/wiki/Lossy_compression.
3. Stephen Robertson, Hugo Zaragoza. The Probabilistic Relevance Framework: BM25 and Beyond.
4. Yuanhua Lv, ChengXiang Zhai. Lower-Bounding Term Frequency Normalization.
5. Edel Garcia. A Tutorial on the BM25F Model.
6. Chris Kamphuis, Arjen P. de Vries, Leonid Boytsov, Jimmy Lin. Which BM25 Do You Mean? A Large-Scale Reproducibility Study of Scoring Variants.
7. Yaël Champclaux, Taoufiq Dkaki, Josiane Mothe. Enhancing High Precision by Combining Okapi BM25 with Structural Similarity in an Information Retrieval System.
8. Hugo Zaragoza, Nick Craswell, Michael Taylor, Suchi Saria, and Stephen Robertson. Microsoft Cambridge at TREC-13: Web and HARD tracks.