Introduction to Information Retrieval - Project 2

Nicholas Drummey
University of Southern Maine
Biddeford, ME, USA
nicholas.drummey@maine.edu

Ryan Reed
University of Southern Maine
Parsonsfield, ME, USA
ryan.reed@maine.edu

Figure 1: Blender Stack Exchange Logo, 2022.

ABSTRACT

Stack Exchange is a forum which fosters the discussion of a multitude of field-specific methodologies, tools, techniques, resources, and general Q/A, divided into sub-websites. Three effective and widely used information models are BM25, TF-IDF, and LGD. By taking one of the child sites of Stack Exchange, we receive a collection to analyze and perform the models upon. After which, an evaluation and comparison of those models upon the collection can be performed and discussed.

Also included within the frame of the project are query expansion using the RM3 upon the BM25 model, and the fusion of results for the BM25 and TF-IDF models.

CCS CONCEPTS

• **Information systems** → **Presentation of retrieval results**; **Top-k retrieval in databases**; *Combination, fusion and federated search*; **Retrieval effectiveness**.

KEYWORDS

BM25, TF-IDF, LGD, query expansion, IR system evaluation, IR system comparison

1 INTRODUCTION

The sub site of Stack Exchange chosen for this project was the Blender Stack Exchange. In this sub-forum, the specialization of discussion is on the 3D modeling software called Blender. The choice of this sub-forum as a data collection was based on mutual interest by the team, as both team members are familiar with the software, and the implementation of models generated by Blender into game development engines. An additional combined hope of the team is that through our effort on this project, we will gain further practical knowledge in both the realms of information retrieval and 3D modeling. The Blender Stack Exchange can be accessed at https://blender.stackexchange.com/.

In the previous project, we implemented Boolean and Inverted Index retrieval models. In comparison to the simpler models in the aforementioned project, we construct more complex information retrieval systems in the form of TF-IDF, BM25, and LGD retrieval models within this project.

Size statistics for the Blender Stack Exchange included 185,140 total documents, 83,738 terms, and 10,959,038 total tokens across all documents. The collection documents over the Stack Exchange are all of the posts on the sub-forum, including both question and answer posts.

2 MODEL DESCRIPTIONS

Within this section are descriptions of the models used for the Blender Stack Exchange, including the query expansion model.

2.1 TF-IDF

The TF-IDF, or Term Frequency-Inverse Document Frequency, model uses the statistical measure of which is eponymous with the model, to score the documents within a collection based on a set of query terms. The term frequency of a word in a document is the raw count of the word within the document, and the inverse document frequency is the measure of how frequent a document containing the word appears in the collection. The Formula for the TF-IDF model:

$$tfidf(t, d, D) = tf(t, d) \cdot idf(t, D) \tag{1}$$

Where t is the term, d is the document, and D is the document collection. The formula for term frequency:

$$tf(t, d) = log(1 + freq(t, d)) \tag{2}$$

The formula for inverse document frequency:

$$idf(t, D) = log(\frac{N}{count(d \epsilon D : t \epsilon d}) \quad (3)$$

A higher TF-IDF score implies that the document is more relevant to the query. [4]

2.2 BM25

BM25, or Best Match 25, is a ranking function that ranks documents according to their relevance to a given query. In many ways, it is an improved version of the TF-IDF model, as it uses term frequency and inverse document frequency. Beyond term frequency and inverse document frequency, it incorporates document length & two tuning parameters, one for controlling scaling of document term frequency and one for controlling the scaling by document length. The formula for BM25 is as follows:

$$RSV_d = \sum_{t \epsilon q} log(\frac{N}{df_t}) \cdot \frac{(k_1 + 1) tf_{td}}{k_1((1 - b) + b \cdot (L_d / L_{ave})) + tf_{td}} \quad (4)$$

Where tf_{td} equals the term frequency in document d, L_d and L_{ave} refer to the length of document d and the average document length in the collection respectively. The variables k_1 and b refer to tuning parameters for controlling the scaling of the document term frequency and the scaling by document length respectively.

Similar to the TF-IDF model, higher scores indicate higher relevancy with the query. [5]

2.3 LGD

The LGD, or log-logistic distribution, model is a more advanced model that is a combination of two instances of informative content, which are derived from probability distributions. The more the distribution of a word within a document deviates from the average distribution of that word in a collection, the more likely the word is significant for the associated document. This pretext forms the basis of information based information retrieval models, such as LGD.[3] To begin explaining this mathematically, consider the following formula:

$$RSV_{q,d} = \sum_{w \epsilon q \cap d} x_w^q [log((\frac{N_w}{N}) + t_w^d) - log(\frac{N_w}{N})] \quad (5)$$

where the parameters of a word frequency probability distribution are selected as the second term frequency normalization ($t_w{}^d$) and the document frequency. Higher LGD ranking indicates higher relevancy to the associated query.[2]

2.4 RM3

RM3, or relevance model 3, is a technique of query expansion based upon blind feed-back. First, a collection of documents must be ranked by a ranking function, in which a top set of documents become evidence, or feedback, for the relevance model. The relevance model then takes this evidence, and captures the behavior, or pattern, of the returned documents and adapts the associated query based upon the correspondingly computed weights of terms in the evidence. Often, for competitions, the term weights are normalized to sum to one, for this project, this was not chosen.[1] The features for relevance model 3 is as follows:

$$\sum_{q_i \epsilon Q \cap D} log(idf(q_i)) \quad (6)$$

Table 1: IR Model Search Results

Model	P@5	nDCG@100
TF-IDF	0.450	0.858
BM25	0.430	0.851
LGD	0.490	0.871
TF-IDF/BM25 Fusion	0.45	0.875
RM3-BM25	0.390	0.883

Table 2: Shared Top-3 Results Between All Models (Query = "weathering effects")

Rank	url	Relevance (0-2)
0	URL	2
1	URL	2
2	URL	0

3 EVALUATION

Within this section is the discussion of search results, and the comparison of results between different models.

3.1 Search Results

Table 1 features the nDCG@100 and Precision@5 results of each models run. The log-logistic distribution model performed the overall best, with the highest precision of all model and performed close to the best nDCG@100 result. The worst performing model based upon these metrics was the best-match 25 model following relevance model 3 query expansion, although it had the highest nDCG@100, query expansion caused a loss of around 0.06. TF-IDF, BM25, and the BM25/TF-IDF fusion, all performed similarly within a close grouping.

Table 2 features the same top-3 results returned from each of the systems as examples. Interestingly, the results greatly differ after the top-3, although there is patterns within all of the models where the ranking is the same for a multitude of documents, albeit with obviously different scoring for each system. The fusion of the term-frequency and best-match 25 models can be seen as an improvement upon both, with the fusion having a higher nDCG@100 than its 'parent models', but inheriting the higher precision of the TF-IDF model. In the next section, we present visualizations and comparisons of the systems and the recorded measures.

3.2 Model Comparisons

Although the models performed similarly, visualization can aid in finding smaller differences and put the data into a better view. In figure 2, we can see the precision at cut 5 for all models, and a clear difference from the 'mean' with models RM3 and LGD. RM3, as briefly previously discussed, had much more negative deviation

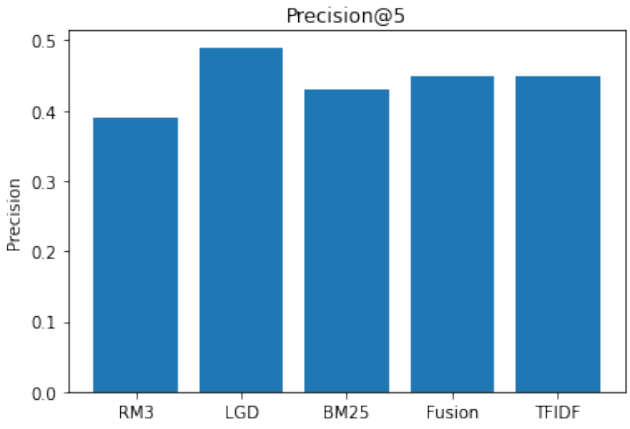

Figure 2: A bar graph of the precision at cut 5 for each model.

from the resulting precisions than the other models. LGD, however, offered the highest precision rate at top-5. There was minor deviance between BM25 and the other two 'mean' models, the TF-IDF/BM25 fusion and the TF-IDF model, but the decrease in precision was minor in comparison to the overall precision rates. Table

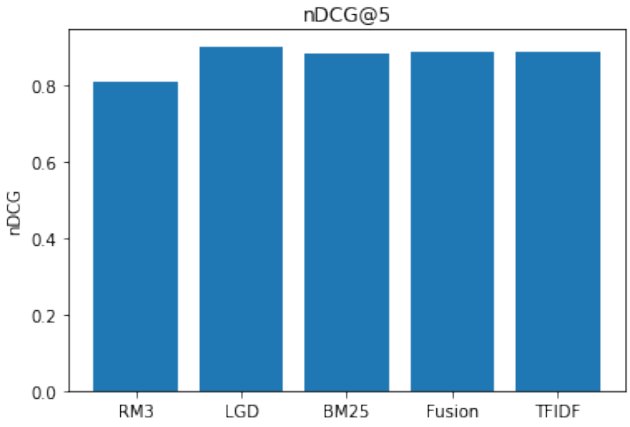

Figure 3: A bar graph of the nDCG at cut 5 for each model.

Z gives us the nDCG@5 for all models, and theres a much stronger commonality between the values. For the most part, the nDCG@5 follows the same pattern as precision@5, with a strong account of mean results. Much like precision, LGD performed better than all others, while RM3 gave the worst nDCG@5. While precision has already been a topic, an interesting insight into the performance of these models is the precision drop off as k increases. This can be reworded into the precision drop off rate over the size of a collection. Table Z gives us a visualization of how these models suffered between 5 documents, and 25 documents, and the precision drop off rate is drastic for each model. Interestingly, there appears to be a nearly identical drop off rate for all models, impacted mainly by initial precision@5.

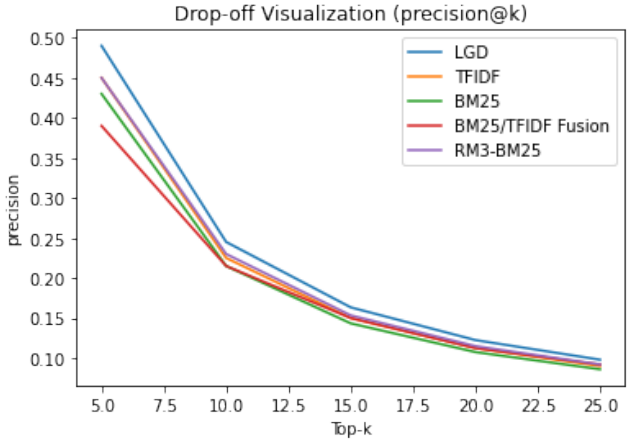

Figure 4: A graph of precision drop off rate for each model.

4 CONCLUSION

In conclusion, after utilizing all five models and analyzing the runs upon the collection, there is one model that performed admirably and one that failed to provide effective results. LGD, or the log-logarithmic distribution, had the highest overall precision and nDCG, while the relevance model 3 query expansion only negatively impacted the best-match 25 model. Despite this, all five models performed similarly with moderate levels of precision and high nDCG values. It is worth noting that the patterns of similar results, both in ranking and evaluation measures, reinforce the importance of data analysis in information retrieval to discover abnormalities between models, and determine the best fit model for a collection.

As for experience with the project, all members of the team found it to be highly beneficial to our understanding of IR model comparison, as well as the journey into more advanced models than project one. Throughout this project we've put in practice with important tools, such as PyTerrier and the trec_eval software. An important discipline gained by teammates during this project, was the exploration into scientific papers, primarily on LGD and RM3, as only the source papers provided information about these subjects. Other notable skills that we feel we've grown in over the course of this project include report writing with overleaf, presentation, and Blender as we worked over the collection further.

REFERENCES

[1] Nasreen Abdul-Jaleel and James Allan. 2004. UMass at TREC 2004: Novelty and HARD. (March 2004). https://trec.nist.gov/pubs/trec13/papers/umass.novelty.hard.pdf
[2] Stéphane Clinchant and Eric Gaussier. 2010. Information-Based Models for Ad Hoc IR. (Jan. 2010). https://doi.org/10.1145/1835449.1835490
[3] Stéphane Clinchant and Eric Gaussier. 2011. A Log-logistic Model for IR. *Springer Verlag* 14, 1 (2011), 5–25. https://doi.org/10.1007/s10791-010-9143-7
[4] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2009. *An Introduction to Information Retrieval.* Cambridge University Press, Cambridge, England.
[5] Wikipedia. 2022. *Okapi BM25.* Retrieved November 10, 2022 from https://en.wikipedia.org/wiki/Okapi_BM25#The_ranking_function

Received 10 November 2022