

Information Retrieval

Lab Assignment 2

Paris Mavromoustakos (10407502)

Partick De Kok (5640318)

December 2, 2012

1 Introduction

In the second lab assignment, the goal is to run a small scale IR experiment. That includes the proposal (hypothesis) and the evaluation of TREC's results. In comparison with Assignment 1, we will now emphasize on the interpretation of the output our experiment gives us, rather than the indexing/querying procedure itself. What we expect to obtain from this Assignment is knowledge on how to analyze and evaluate results after running a TREC experiment.

2 Description of the Data

For this experiment we use CSIRO document collections, which are pre-processed and ready to be indexed (trectext format). In this case, the documents consist of website content regarding scientific topics. The collection consists of four parts; Corpus.body is the body of the web pages, corpus.title contains the websites' titles, corpus.metadata and corpus.headers contain the metadata and headers of the web pages.

Moreover, we have 50 TREC Enterprise 2007 topics, which we will apply on our index as Indri queries. Lastly, the relevance assessments file (qrels) is also given and will be used to evaluate the querying output using trec_eval.9.0.

In this experiment we use porter stemming and there is no stopwords removal.

3 Hypothesis

We want to compare 3 different but closely related retrieval models, BM25, BM15 and BM11. The research questions are "Which one of these models performs better?" and "Does one model perform better over all topics or are there per-topic differences, and why?". Before we hypothesize, we find it important to point out the difference between the 3 retrieval models.

BM(Best Match) 25 is a bag-of-words retrieval function that ranks a set of documents based on the query terms appearing in each document, regardless of the inter-relationship between the query terms within a document[1]. BM25 was created as a combination of the BM11 and BM15 formulas[2] and uses the following fraction when computing a document's score:

$$Score(D, Q) = \sum_{i=1}^n IDF(q_i) \times \frac{f(q_i, D) \times (k_1 + 1)}{f(q_i, D) + k_1 \times (1 - b + b \times \frac{|D|}{avgdl})}$$

Where $f(q_i, D)$ is q_i 's term frequency in document D , $|D|$ is document D 's length and $avgdl$ is the average document length in the collection. k_1 and b are free parameters, with $k_1 \in [1.2, 2]$ and $b \in [0, 1][2]$.

As we can see from the formula above, the term b plays an important role as it defines the importance of document length[3]: $b = 1$ means we fully scale the term weight by document length and our the retrieval model transforms into BM11[2], while $b = 0$ means the document length is not taken into consideration (BM15[2]).

BM25's default value for b is 0.75, and additionally for k_1 1.25 is considered a reasonable value[1]. In this experiment, we keep the default value for k_1 and use the 3 different values of b mentioned above.

After studying the 3 retrieval models we hypothesize that BM25 will outperform BM11 and BM15, for the reason that it "balances" between the other two models aiming towards optimization regarding the document length factor. We believe that the extreme values of 0 and 1 for b will give different but still worse results than those of $b = 0.75$, both over all topics, and even per-topic. More, experiments have shown that BM11 outperforms BM15[2], so we expect BM11 to be the second-to-best model regarding performance.

4 Related Work

Numerous books and articles mention the BM25 retrieval model as a combination of BM11 and BM15. "Introduction to Information Retrieval" (Christopher D. Manning, Prabhakar Raghavan & Hinrich Schtze, 2008) discuss BM25 as a probabilistic information retrieval model and mention the different possible values for the b term.

A deeper comparison between the three retrieval models is made in "Modern Information Retrieval" (Ricardo Baeza-Yates, Berthier Ribeiro-Neto, 2006) where the BM25 formula is analyzed in detail. In this book, discussion is presented about the suggested values for k_1 and b and how changing these values affects the model.

5 Results

6 Discussion

7 Conclusion

References

- [1] www.wikipedia.org
- [2] Ricardo Baeza-Yates & Berthier Ribeiro-Neto, *Modern information Retrieval*. Addison Wesley, 2006.
- [3] Christopher D. Manning, Prabhakar Raghavan & Hinrich Schtze, *Introduction to Information Retrieval*. Cambridge University Press, 2008.