Document Normalization Revisited

Abdur Chowdhury
America Online
Reston, Virginia
chowdhury@ir.iit.edu

M. Catherine McCabe
U.S. Government
Washington D.C.
mcatherm@comcast.net

David Grossman, Ophir Frieder
Illinois Institute of Technology
Chicago, IL 60616
{dagr, ophir} @ ir.iit.edu

Abstract
Cosine Pivoted Document Length Normalization has reached a point of stability where many researchers indiscriminately apply a specific value of 0.2 regardless of the collection. Our efforts, however, demonstrate that applying this specific value without tuning for the document collection degrades average precision by as much as 20%.

Categories & Subject Descriptors: H.3.3 Information Search and Retrieval

General Terms: Algorithms, Measurement

Keywords: Information Retrieval, Text Search, Similarity Measure, Relevance Measure, TREC

1. Introduction
The cosine measure normalizes document length so that long documents are not favored simply because they have more terms. Later work empirically showed that, for the TREC collection, longer documents actually have a higher probability of being relevant [1]. One interesting claim from this study was:

> …the slope value is very stable, i.e., the changes in retrieval effectiveness with minor deviations in slope (from the optimal slope value) are very small for all the collections. A constant slope value of 0.20 was effective across collections…

We show that the slope should be recalibrated for fundamentally different document collections. We recomputed the normalization for the web track and obtained a 16% improvement in our baseline run. The resulting effectiveness from this sole modification was an average precision higher than any other group who submitted results to the TREC-10 web track.

2. Calibrating the Slope for Ad Hoc Task
Using pivoted document normalization, we tested slope values from 0.01 to 0.9 and measured the average precision for each value. This was done for the TREC 6-7-8 adhoc tasks (the shaded area in Table 1) that use the TREC disks four and five and the TREC-9 and TREC-10 tasks that use the WT10G collection. For queries over disks four and five, the slope values below 0.4 are indistinguishable. That is, there is only about a 5% difference between the best slope value and the worst (See Table 1). However, for the TREC-9 and TREC-10 tasks, the slope makes a significant difference – up to 36%. For the TREC-9 and TREC-10 tasks, a slope of 0.05 results in the best average precision of 0.2057 and 0.1737, respectively.

SIGIR'02, August 11-15, 2002, Tampere, Finland.
ACM 1-58113-561-0/02/0008.

Table 1: Effect of Slope on Average Precision

Slope	TREC 10	TREC 9	TREC 8	TREC 7	TREC 6
0.01	0.1951	0.1706	0.2290	0.1696	0.2312
0.05	**0.2057**	**0.1737**	0.2370	0.1724	0.24
0.1	0.1944	0.1660	0.2386	0.1728	0.2425
0.2	0.1706	0.1507	0.2303	0.1683	0.2432
0.3	0.1519	0.1353	0.2246	0.1631	0.2326
0.4	0.1297	0.1204	0.2172	0.1562	0.2241
0.5	0.1090	0.1073	0.2059	0.1480	0.2162
0.6	0.0890	0.0969	0.1902	0.1366	0.2017
0.7	0.0657	0.0849	0.1717	0.1206	0.1783
0.8	0.0444	0.0669	0.1500	0.0978	0.1518
0.9	0.0215	0.0335	0.1114	0.0688	0.1204

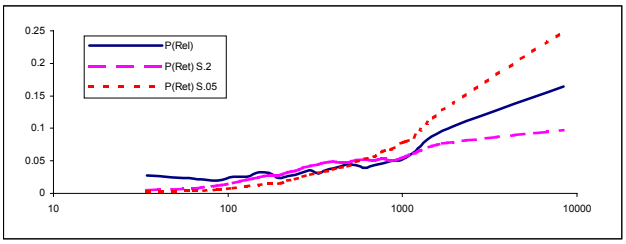

Figure 1: TREC 8

To determine why the variance in slope has a varying degree of impact for different collections, we examine the probability of relevance and the probability of retrieval given different document lengths. As described in [1], we partitioned the document collection into bins of equal size and computed the probability of retrieval and the probability of relevance for each bin. The bin median document size is given on the *x-axis*. These measures are plotted in Figure 1 and Figure 2 using a slope of 0.20 and 0.05, respectively. The slope value is designed so that the retrieval curve will track the relevance curve. For the TREC-8 task (similarly also for the TREC-6 and TREC-7 tasks), both slope values of 0.2 and 0.05 track very closely to the probability of relevance line. This explains the relatively small difference in average precision for the TREC disks four and five.

Turning to Figure 2, with document collection WT10G, we see that both slope values track the relevant curve reasonably well for small document sizes while both diverge for large documents (over 5,400 terms). To identify the cause of a significant difference in average precision found for these two slopes, we then examine the results for only the top 100 documents retrieved.

In Figure 3, we illustrate the results. Here, we can see that the slope value of 0.05 tracks the relevant curve much more closely for large documents. This implies that the cause of a difference in average precision was simply masked in Figure 2 by some low ranking documents that were still retrieved. Further verifying this hypothesis, we examine Figure 4, which shows probabilities for retrieval at only ten documents. Again, we see that the probability of retrieval much more closely tracks the probability of relevance for a slope of 0.05 then a slope of 0.2.

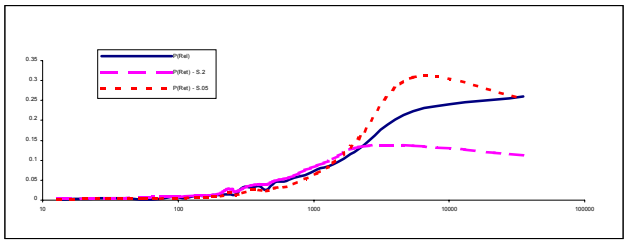

Figure 2: TREC 10 - Top 1000 Retrieved

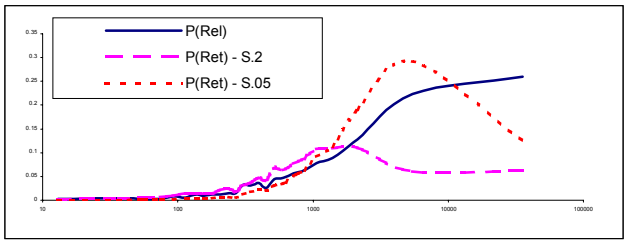

Figure 3: TREC 10 - Top 100 Retrieved

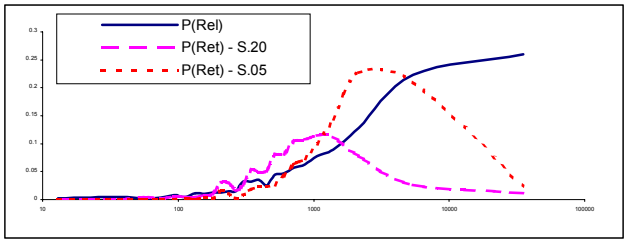

Figure 4: TREC 10 - Top 10 Retrieved

3. Recomputing Probabilistic Normalization

The probabilistic retrieval model also relies on an adjustment for document length [3]. The *b* value of the BM25 strategy is used in a manner quite similar to the slope with sloped cosine. We calibrated the optimal value for *b* across collections and across tasks. We found that, as with the slope, it is critical to adjust the *b* value when significantly changing the document collections as well as when dramatically changing the retrieval task. We find that a slope of 0.25 is 22% better than the values published at 0.75.

Table 2. Average Precision for various BM25 b-values

		Average Precision				
	b	***T10***	***T9***	***T8***	***T7***	***T6***
	0.1	0.208	0.1694	0.2276	0.1711	0.2346
Optimal	**0.25**	**0.213**	**0.1711**	0.2338	0.1734	0.241
	0.5	0.2023	0.1614	0.2327	0.1709	0.2428
Published	**0.75**	0.1786	0.1419	0.2243	0.1662	0.2374
	0.9	0.152	0.1274	0.2166	0.1574	0.2279

4. Conclusion

We have shown that there is a clear need to calibrate, for a specific document collection, the document length component of pivoted document length normalization and the BM25 probabilistic model. We incorporated the new slope value into our TREC-10 system, which uses additional techniques such relevance feedback. Our results improved significantly, yielding an average precision of 0.241 with BM25 and 0.231 with cosine normalization. This performance exceeds the best submission of TREC-10 that had an average precision of .2226.

5. References

[1] Singhal, A., C. Buckley, and M. Mitra, "Pivoted Document Length Normalization," ACM SIGIR 96.

[2] S.E. Robertson, S. Walker, M. Beaulieu, "Okapi at TREC-7: automatic ad hoc, filtering, VLC and interactive," TREC 7, page 253, 1996.