

Lecture 3: Pivoted Document Length Normalization

Lecturer: Lillian Lee

Scribes: Lakshmi Ganesh, Navin Sivakumar

Abstract

In this lecture, we examine the impact of the length of a document on its relevance to queries. We show that document relevance is positively correlated with document length, and see that relevance scores that use the normalization techniques we've studied so far (L_∞, L_1, L_2) do not capture this correlation correctly. Finally, we present the “pivoted document length normalization” technique introduced by Singhal et al. in [SBM96], which addresses this issue.

1 Introduction

Recall that our approach to representing a document d as a vector \vec{d} is as follows:

- Compute an “unnormalized” vector δ by setting $\delta[j] = TF_d(j) \times IDF_C(j)$;
- Compute a normalization factor $\mathbf{norm}(d)$ (which is typically a function of δ);
- Compute \vec{d} by setting $d[j] = \frac{\delta[j]}{\mathbf{norm}(d)}$.

Observe that increased document length manifests itself in two ways:

1. Increased term counts for a given term, resulting in larger $TF_d(j)$ and therefore larger $\delta[j]$ values;
2. Increased number of non-zero term counts, resulting in more non-zero $TF_d(j)$ and therefore more non-zero $\delta[j]$ values.

It is clear from these observations that without a normalization factor, longer documents will have an advantage when computing relevance scores. Now we will see how normalization corrects for document length.

1.1 Normalization Methods and Document Length

Recall that we studied three normalization methods (based on the L_∞, L_1, L_2 norms). Consider the effect of document length when using a scoring function based on these norms:

- The L_∞ norm *favors* long documents. This is because the L_∞ norm favors documents with many (approximately equal) non-zero $\delta[j]$ values.
- The L_1 norm *penalizes* long documents. This is because the L_1 norm favors documents with one dominant $\delta[j]$, so that longer documents with many non-zero $\delta[j]$ values get penalized.

- The L_2 norm *penalizes* long documents. This is because the L_2 norm increases with increasing number of non-zero $\delta[j]$ values.

What is not clear, however, is whether long documents should be favored or penalized. That is the subject of this lecture. The L_∞ and L_1 normalization methods are not widely used, because of the fact that they favor overly-generic, and overly-specific documents, respectively. The L_2 norm, therefore, is the focus of the rest of this lecture.

2 Effects of Document Length on Document Relevance

Prior to [SBM96], the prevailing view was that document relevance is independent of length and that L_2 normalization exactly negates the effects of document length; hence, the L_2 norm was considered to be a good normalization factor.

The work in [SBM96] challenges these assumptions. Specifically, document relevances and L_2 normalization retrieval scores were analyzed in an annotated dataset to answer the following questions:

1. What is the distribution of relevant documents with respect to length?
2. How does this distribution compare with the distribution of retrieval results using the L_2 norm?

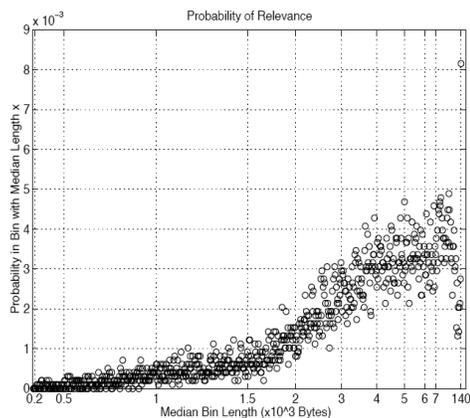
We now provide an outline of the method used in [SBM96]. The data consisted of 50 queries and about 740,000 relevance-annotated documents from TREC. Now consider the following sets:

$$\mathbf{Rel} = \{(q, d) \mid \text{document } d \text{ relevant to query } q\}$$

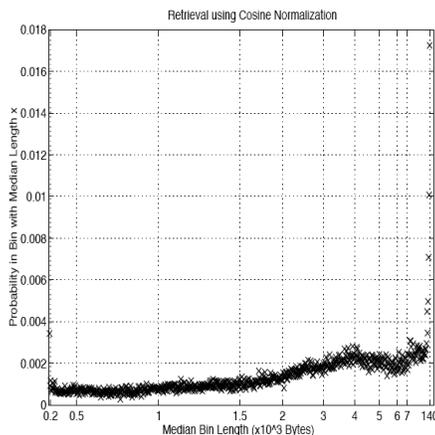
$$\mathbf{Ret} = \{(q, d) \mid \text{document } d \text{ among top 1000 retrieved documents for query } q\}$$

Intuitively **Rel** is the set of relevant documents, and **Ret** is the set of retrieved documents. The set **Rel** contained 9805 pairs. The documents were sorted according to increasing length, and then grouped into bins of 1000 documents each, where the bins were labeled by their median document length. For each bin, the fraction of relevant documents and the fraction of retrieved documents contained in the bin was computed.¹ One might wonder why the documents were binned at all - why not simply plot document relevance (or retrieval) against document length? The reason is that such a plot would have very unevenly distributed data points - there would likely be several values of document length for which there would be no corresponding relevance (retrieval) data; additionally, data points would be hard to interpret as it would not be clear whether they represent one document or a thousand. Equal-interval bins would have the same problems. Hence the decision

¹More precisely, the fraction of (q, d) pairs from **Rel** (or **Ret**) such that d is contained in the bin. For example, if 98 relevant documents were contained in a certain bin, then the fraction of relevant documents in that bin would be $\frac{98}{9805}$.



(a) Document Relevance vs Document Length



(b) Document Retrieval vs Document Length

Figure 1: Distribution of Document Relevance and Retrieval w.r.t. Length. Plots from [SBM96].

to bin the documents with an equal number of documents in each bin. The results of this analysis are shown in fig. 1.

Fig. 1a reveals that document relevance increases with document length, refuting the earlier assumption. However, we also see that document retrieval using the L_2 norm increases with document length (fig. 1b). In order to check if L_2 normalization does in fact accurately account for document length, fig. 2 overlays the document relevance and document retrieval trends with respect to document length. To prune out distracting information and reveal the underlying trends, the relevance and retrieval plots are smoothed by grouping each sequence of 24 consecutive bins; each group of 24 bins is plotted as a single point using the average of the median document lengths and the average of the retrieval or relevance probability over the bins.

It is clear from fig. 2 that more long documents are relevant than are retrieved. Hence, L_2 normalization overpenalizes length.

3 Pivoted Document Length Normalization

We see from fig. 2 that L_2 normalization over-retrieves short documents and under-retrieves long documents. Observe that there is a pivot point where the retrieval rate seems to match the relevance rate. In order to correct for this effect, we would like to modify the original normalization factor **norm** to a new factor **norm'** which increases the penalty to documents shorter than the pivot point and decreases the penalty to documents longer than the pivot point. If we denote the pivot point by

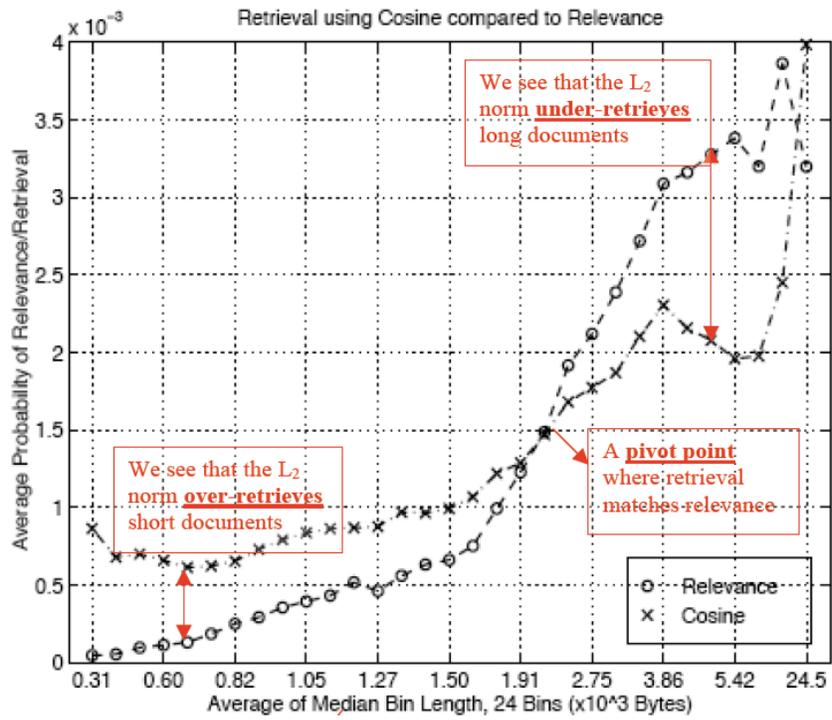


Figure 2: Comparing Document Relevance with Document Retrieval. Plot from [SBM96], with red annotations added.

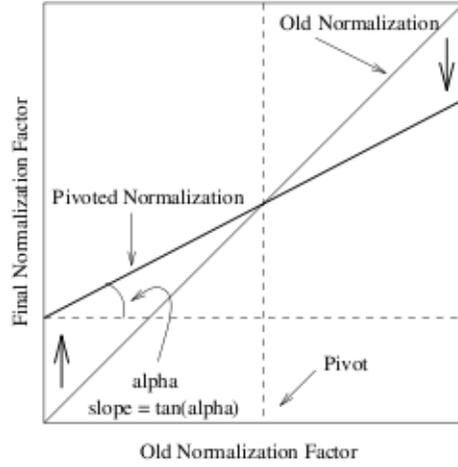


Figure 3: A visual comparison between the new pivoted normalization factor and original normalization factor as functions of the original normalization. From [SBM96].

p , we can summarize the requirements as follows:

$$\begin{aligned} \mathbf{norm}'(d) &> \mathbf{norm}_2(d), \text{ if } \mathbf{norm}_2(d) < p \\ \mathbf{norm}'(d) &< \mathbf{norm}_2(d), \text{ if } \mathbf{norm}_2(d) > p \\ \mathbf{norm}'(d) &= \mathbf{norm}_2(d), \text{ if } \mathbf{norm}_2(d) = p \end{aligned}$$

Fig. 3 depicts these requirements graphically. We see that we can visualize the new normalization factor as a “clockwise rotation” of the original normalization factor about the pivot point. The proposed new normalization factor \mathbf{norm}' is a linear function of the L_2 normalization, so we have:

$$\mathbf{norm}'(d) = m\mathbf{norm}_2(d) + b,$$

for some parameters m, b . To see why this is reasonable, note that the slope parameter m determines the angle by which we rotate the normalization line in fig. 3; a slope $m = 1$ leaves normalization factor unrotated, while a slope $m = 0$ rotates the line to horizontal. Note that this argument also demonstrates that the parameter m should be in the interval $(0, 1)$. The parameter b can then be chosen to shift the new normalization line vertically in order to have it intersect the original line at the pivot point. In order to make this more explicit, we can use the fact that \mathbf{norm}' and \mathbf{norm}_2 should agree at the pivot point p to rewrite expression defining the new normalization factor:

$$p = mp + b \Rightarrow b = (1 - m)p.$$

Substituting this expression for b gives

$$\mathbf{norm}'(d) = m\mathbf{norm}_2(d) + (1 - m)p.$$

Note that this gives another interpretation of \mathbf{norm}' as an interpolation between \mathbf{norm}_2 and the pivot point p .

Unfortunately, both the slope m and the pivot point p are unknown parameters in the above formulation. The following steps show how to reduce the expression so that it contains only a single parameter. Observe that the quantity $(1 - m)p$ is independent of the document d and positive. Thus, we can divide the expression by $(1 - m)p$ to obtain the following expression, which is equivalent under rank to $\mathbf{norm}'(d)$:

$$\frac{m}{(1 - m)p} \mathbf{norm}_2(d) + 1$$

This expression is convenient, since the parameters m and p have been grouped, allowing us to eliminate a parameter. To see why this is the case, suppose that m', p' give the optimal values of the parameters. Then for any other choice of pivot p'' , there is a choice m'' such that

$$\frac{m'}{(1 - m')p'} = \frac{m''}{(1 - m'')p''}.$$

Thus p can be set to an arbitrary value; choosing the pivot p to be the average L_2 normalization factor $\overline{\mathbf{norm}_2}$ gives the final form for the pivoted normalization factor:

$$\mathbf{norm}''(d) = \frac{m}{1 - m} \times \frac{\mathbf{norm}_2(d)}{\overline{\mathbf{norm}_2}} + 1.$$

To understand why this choice is reasonable, recall that the pivot parameter specifies the point where $\mathbf{norm}'' = \mathbf{norm}_2$; according to [SBM96], it is reasonable to believe that the average $\overline{\mathbf{norm}_2}$ gives the length of an “appropriately-sized” document which does not need renormalization.

4 Problems

1. Prove that we are free to choose an arbitrary pivot parameter in the pivoted normalization. More precisely, suppose m and p are the optimal parameters, where $m \in (0, 1)$ and $p > 0$. For an arbitrary positive p' , find $m' \in (0, 1)$ such that

$$\frac{m}{(1 - m)p} = \frac{m'}{(1 - m')p'}.$$

2. In lecture, we claimed that the fact that pivoted normalization retrieval results match the distribution of relevant documents more closely than L_2 normalization retrieval results was evidence that pivoted normalization is an improvement over L_2 normalization. This exercise explores whether that line of reasoning is plausible.²

²In [SBM96], the claim above is used to motivate the technique of pivoting, but the results are then evaluated under average precision in order to confirm that the quality of retrieval results is improved by pivoting.

- (a) Consider a set of 10 relevant documents with the following length distribution:

Document length	1	2	3	4
Number of documents	1	2	3	4

Construct a set A of 50 retrieved documents such that:

- The set A is intuitively a good retrieval result;
- The distribution of document lengths in A does not match the distribution of document lengths of relevant documents.

Now describe a way to modify A so that the distribution of document lengths matches the relevant documents more closely and yet seems to harm retrieval performance. (You may assume that for any given document length, the corpus contains many more documents which are not relevant than those which are relevant.)

- (b) Do you think it is likely that the improved correlation with the distribution of length of relevant documents when using pivoted normalization instead of L_2 normalization is a case of the situation described in part (a) (in other words, do you think that pivoting might decrease the quality of retrieval results, despite improving the match between the lengths of retrieved documents and relevant documents)? Why or why not?
3. Recent results [LAB08] suggest that the positive correlation between document relevance and document length demonstrated in [SBM96] was misleading as the data from TREC was not completely labeled for relevance, and unlabeled documents were treated as non-relevant.³ It seems that the question of how document length and relevance are related is open again.
- (a) How might this incomplete labeling induce a bias in the observed relationship between relevance and document length?
- (b) **A thought exercise (no solution provided):** Keeping in mind the issue discussed in part (a), how might you go about determining the actual relationship between document length and relevance?

5 Solutions

1. Straightforward algebraic manipulations allow us derive the following expression for m' in terms of m, p, p' :

$$m' = \frac{p'}{p(1-m) + p'm} m.$$

Since $m \in (0, 1), p > 0, p' > 0$, we have $p(1-m) + p'm > 0$ and so $m' > 0$. Now observe that

$$m' = \frac{p'}{p(1-m) + p'm} m < \frac{p'}{p'm} m = 1.$$

³In the next lecture, we will explore in some detail the issue of incomplete labeling.

2. (a) We present one of many possible solutions. Construct A with the following length distribution

Document length	1	2	3	4
Total number of documents	15	15	15	5
Number of relevant documents	1	2	3	4

by including all the relevant documents, and then adding other documents which are not relevant to satisfy the sizes above (i.e. 14 extra documents of length 1, 13 of length 2, etc.).

Now modify A as follows:

- Remove 10 documents of length 1, including the 1 relevant document;
- Remove 5 documents of length 2, including the 2 relevant documents;
- Add 15 documents of length 4 which are not relevant.

This modification has the net result of exchanging 3 relevant documents for 3 documents which are not relevant, most likely an undesirable effect. However, the distribution of document lengths in the set is now:

Document length	1	2	3	4
Total number of documents	5	10	15	20
Number of relevant documents	0	0	3	4

which exactly matches the distribution of sizes of relevant documents.

- (b) If we believe that documents of the same length have approximately the same norm, then for any fixed document length, pivoting rescales all scores of documents of that length by approximately the same factor. Hence, pivoting does not cause much change in the ranking of a document relative to those of the same length. Consequently, if pivoting causes additional documents of a certain length to be retrieved, the first documents added will tend to be the highest ranked documents (relative to same length documents); similarly, if pivoting removes documents of a certain length from the retrieval set, the first documents removed will be among the lowest ranked of that length. Therefore, if we believe that the underlying scoring function (e.g. dot product scoring) accurately ranks same length documents relative to each other, then at any given length, pivoting should tend to either add the most relevant non-retrieved documents or remove the least relevant retrieved documents, avoiding the situation described above, where relevant documents were preferentially removed over non-relevant documents.
3. (a) Note that all unlabeled documents are treated as non-relevant. Therefore, if documents of a certain length are underrepresented in the set of documents that are labeled, then we can expect that a relatively large number of *relevant* documents of that length are unlabeled (hence deemed non-relevant). Thus, when we observe that (according to the relevance labels) long documents are more likely to be relevant than short documents, it

could be that we are in fact observing the effect of long documents being more likely to have labels than short documents.

References

- [LAB08] David E. Losada, Leif Azzopardi, and Mark Baillie. Revisiting the relationship between document length and relevance. In *CIKM '08: Proceeding of the 17th ACM Conference on Information and Knowledge Management*, pages 419–428, New York, NY, USA, 2008. ACM.
- [SBM96] Amit Singhal, Chris Buckley, and Mandar Mitra. Pivoted document length normalization. In *SIGIR '96: Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 21–29, New York, NY, USA, 1996. ACM.