The Effect of Query Length on Normalisation in Information Retrieval

Ronan Cummins[1] and Colm O'Riordan[2]

[1] Department of Computing Science, University of Glasgow
[2] Department of Information Technology, NUI Galway
ronanc@dcs.gla.ac.uk, colm.oriordan@nuigalway.ie

Abstract. Document length normalisation is known to be a difficult problem in IR, as tuning is often needed to overcome the collection dependence problem known to affect many normalisation schemes. Furthermore, it has been shown in various studies that the most optimal level of normalisation to apply is correlated with query length. In this paper, we confirm this correlation and present experiments which investigates and explains the effect of query length on normalisation.

1 Introduction

Information Retrieval (IR) systems deal with natural language documents and queries and attempt to limit the problem of information overload by returning only those documents that are relevant to a user's need (query). This represents a difficult problem given the presence of synonmy and polysemy in natural language. IR systems typically have to deal with large quantities of potentially semantically ambiguous information.

Many models have been adopted in the information retrieval domain including Boolean models and extensions, vector space models [5] and variations and also probabilistic models [9]. Irrespective of the model adopted, the ability to correctly identify important terms that capture the content of documents and queries is of utmost importance. Much research has been undertaken in the development of good *weighting schemes* that assign suitable weights to terms in the collection. The motivation is to assign low weights to those terms that contain little semantic power (or resolving power) while attaching high weights to those terms that help correctly capture the meaning of documents and queries. There have been many approaches to correctly developing these weighting schemes ranging from heuristic based approaches coupled with empirical analysis to more modern approaches which attempt to learn optimal combinations of sources of evidence [8,3].

Many, if not all, of the developed or learned weighting schemes can be represented as follows:

$$S(Q,D) = \sum_{t \in Q \cap D} (ntf(D) \cdot gw_t(C) \cdot qw_t(Q))$$

L. Coyle and J. Freyne (Eds.): AICS 2009, LNAI 6206, pp. 26–32, 2010.
© Springer-Verlag Berlin Heidelberg 2010

where the similarity ($S()$) between a query Q and document D in a collection C is a function of a normalised term frequency component (a within document score), a global term score (across the collection) and a query-term score (a within query score).

This paper deals with the issue of document normalisation and, in particular, how normalisation is influenced by query length. Document length normalisation is known to be a difficult problem in IR, as tuning is often needed to overcome the collection dependence problem known to affect many normalisation schemes. Furthermore, it has been shown in various studies that the most optimal level of normalisation to apply is correlated with query length [1,2]. In this paper, we confirm this correlation and present experiments which investigates the effect of query length on normalisation. In these experiments, properties of the document collection (such as average document length, standard deviation of document length) are controlled so as to allow experimentation regarding normalisation.

The remainder of the paper is as follows: in the following section, a brief review of some approaches in document normalisation in IR is presented. Details of the properties of the returned set for queries of different length are discussed in Section 3. In Section 4, we describe our experiments which show how the characteristics of the returned sets affect the performance of the weighting scheme adopted. The final section provides some conclusions.

2 Related Work

Document length normalisation is used to help correctly retrieve documents of various lengths. Normalisation is necessary as long documents can otherwise be unfairly promoted. Singhal *et al* discuss two reasons for incorporating normalisation [7]. For long documents, there will be increased probability of repeated occurrences of terms. Hence, the term frequency values will be increased thereby increasing the relevance score of long documents. There is also an increased probability of any given query term being present.

There have been many approaches to normalisation. In the weighting schemes presented by Salton and Buckley [6], a term's frequency in a document is normalised by the frequency of the maximally occurring term in that document. This penalises terms that occur frequently because of the length of the document. A stronger form of normalisation is the cosine normalisation which normalises the score with the normalisation factor of $\sqrt{(w_{t1}^2 + w_{t2}^2 + + \ldots + w_{tn}^2)}$ where n is the number of terms in the document.

The $BM25$ scheme defines similarity between query and document according to:

$$BM25(Q,D) = \sum_{t \in Q \cap D} \left(\frac{tf_t^D \cdot log(\frac{N - df_t + 0.5}{df_t + 0.5}) \cdot tf_t^Q)}{tf_t^D + k_1 \cdot ((1-b) + b \cdot \frac{dl}{dl_{avg}})}\right. \quad (1)$$

where tf_t^D denotes the frequency of a term in a document, tf_t^Q denotes the frequency of a term in a query, N is the number of documents, df_t is the documents containing a term, k_1 and b are constants, dl is the document length and dl_{avg} is

the average document length. In the $BM25$ scheme, the b parameter varies the amount of normalisation used (a higher b value leads to a greater penalisation).

The pivoted normalisation approach [7] defines similarity as:

$$PIV(Q,D) = \sum_{t \in Q \cap D} (\frac{1 + log(1 + log(tf_t^D))}{(1-s) + s.\frac{dl}{dl_{avg}}} \cdot log(\frac{N+1}{df_t}) \cdot tf_t^Q)$$

where s is a normalisation tuning factor. Simlarly to $BM25$, $\frac{dl}{dl_{avg}}$ is the normalisation factor. Chowdhury *et al* show that tuning the normalisation parameters for $BM25$ and the pivoted normalisation scheme can improve performance considerably [1]. In other words, by maintaining the default values, a serious degradation in mean average precision (MAP) can occur. They show that there is a need to tune the normalisation factor for individual collections.

Recent work advocates a normalisation approach which does not require a tuning parameter and achieves high MAP by using $\frac{dl}{\sqrt{dl_{avg}}}$ as the normalisation factor [4].

3 Query Length and Normalisation

In this section, the relationship between the characteristics in the returned set and the length of a query is explored. In order to investigate this relationship a number of queries of different lengths are created and compared against a number of collections. Short, medium and long queries are used and the properties of the returned sets are measured.

For the subsets of the TREC collections used in these experiments (Table 1), we use topics 301 to 450 and create a short query set (title field only), a medium length query set (title and description), and a long query set (title, description and narrative). We use the standard $BM25$ scheme but it should be noted that many term-weighting schemes show similar trends with regard to query length [1,2].

We measured the performance (MAP) of $BM25$ using 9 values of b (0, 0.125 0.25, 0.375, 0.5, 0.625, 0.75, 0.875 and 1) on the collections. Table 1 shows the optimal value of b for each collection for short, medium and long queries for the $BM25$ scheme. It can be seen that in most cases the optimal level of normalisation increases as the query length increases as has been previously reported [1,2].

Table 1. Optimal b per collection for schemes

		$BM25$			
Collections	#Docs	Topics	short	medium	long
LATIMES	131,896	301-450	0.125	0.625	0.875
FBIS	130,471	301-450	0.125	0.25	0.75
FT	210,158	301-450	0.375	0.375	0.625
FR	55,630	301-450	0.75	0.625	0.625

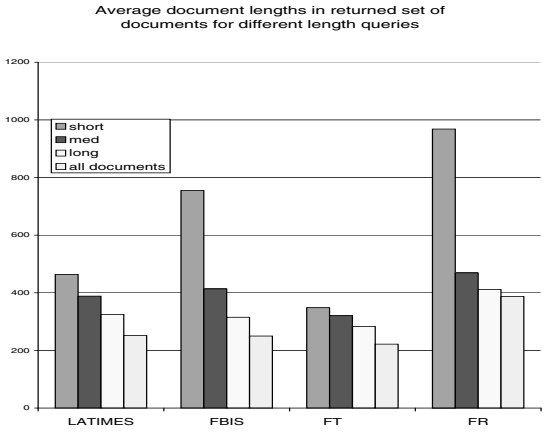

Fig. 1. Average length of returned documents

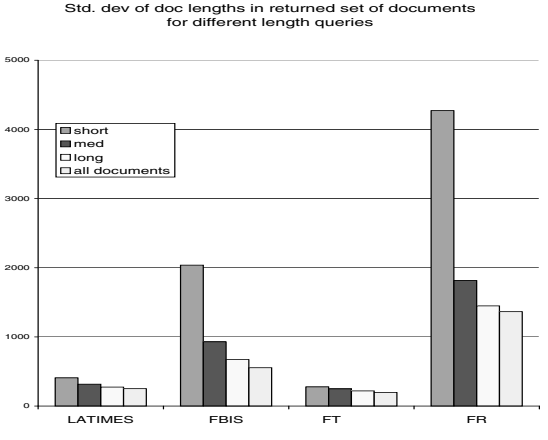

Fig. 2. Standard deviation of returned documents

An analysis of the characteristics of the *entire* returned set of documents for each set of queries (short, medium and long) on the collections shows that short queries return longer documents on average, while medium and long queries return documents that are closer to the average length document in the entire collection (Figure 1). Furthermore, Figure 2 shows that the standard deviation of document length for the returned set of documents is also greater for the shorter queries, indicating that there is a greater variation in the lengths of the documents returned. When the queries are longer, the sets of returned documents become such large samples of the document collections, that the average document length and standard deviation are very similar to those of the documents in the entire collection. The fact that the deviation in document length

in the returned set is higher for shorter queries is an important factor in the normalisation to be applied and is investigated next.

4 Experiment: Deviation in Document Length

The normalisation function in many term-weighting schemes (and in the $BM25$ scheme) is comprised of a ratio of a specific document length to the average document length ($\frac{dl}{dl_{avg}}$). However, this ratio gives us relatively little information about the distribution of document lengths in the collection or the expected deviation of a document length from the average document length. For collections with a high deviation of document length, this ratio ($\frac{dl}{dl_{avg}}$) will vary considerably from very low values (for the shorter documents) to very high values (for the longer documents). Furthermore, from the previous experiment, it was determined that short queries (which return documents sets with a high standard deviation) require a smaller value of b for optimal performance. Therefore, when this ratio varies considerably (short queries), we need a low level of normalisation (otherwise this ratio would be overly influential with regard to the retrieval of documents). Ultimately, this suggests that the optimal value of b may be inversely related to the deviation in the length of the documents in the collection.

To test this hypothesis, we devise an experiment which changes the properties of a collection while keeping the queries constant. A small sample of documents from the LATIMES collection is used with the medium length topics (301-350). A small collection of 8,598 LATIMES documents with an average document length of 28 and a standard deviation of 10 (i.e. low deviation given the average document length) is created. The performance (MAP) for various values of b on this collection is measured. The characteristics of the collection is then modified by adding a small sample of extremely long documents (526 documents). This dramatically changes the properties of the collection by adding only a few documents. The collection now consists of 9,124 documents with an average document length of 95 and a standard deviation of 272 (i.e. high deviation given the average document length). The performance (MAP) for the same values of b is recalculated. From Figure 3, it can be seen that when the standard deviation is low (little variation in document length) normalisation is less important (an expected result) but still has a benefit at higher levels of b. When the standard deviation increases, the optimal level of b drops sharply because using a high value of b severely penalises the longer documents (some of which are relevant to at least one of the topics). For the ad hoc retrieval task if any part of a longer document is relevant to the topic, the entire document is deemed relevant. Usually, a high standard deviation indicates a number of very long documents, due to a lower bound (of zero) on the distribution of document lengths.

Figure 4 shows the results from a similar experiment using a different 3,710 LATIMES documents with an average length of 151 and standard deviation of 12. A further 48 longer documents are added changing the average documents length to 184 and the standard deviation to 287. The keys in both figures show

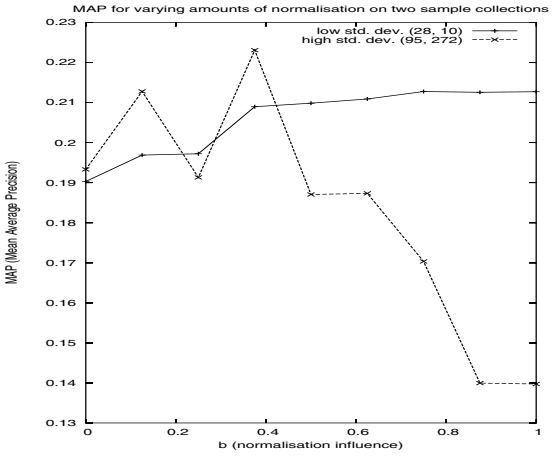

Fig. 3. Δ of deviation with varying b

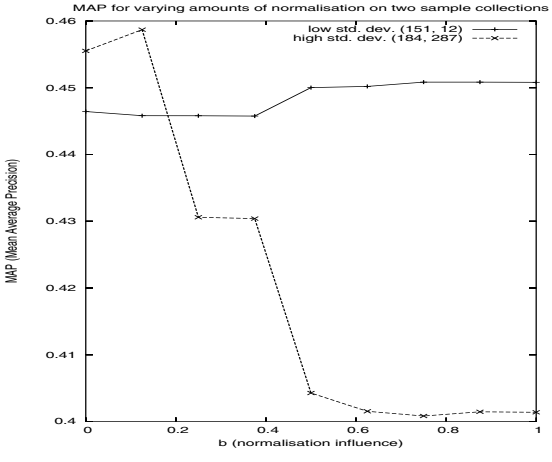

Fig. 4. Δ of deviation with varying b

the average document length and standard deviation respectively in parentheses for each collection. It can seen that this phenomenon is the same (albeit contrived and exaggerated in these cases) as that which was observed when using different length queries. It is important to note that in these experiments the same medium-length queries are being used. Therefore, the effect of query length is not responsible for the change in the optimal level of normalisation to apply. For short queries (which return a set with a higher deviation in document length), a lower level of b is beneficial. Hence, we argue that these two phenomena are actually related and ultimately, it is the deviation of documents lengths that influences the normalisation parameter in term-weighting schemes.

5 Conclusion

Some approaches have attempted to incorporate query length into the term-weighting function [2] and, given the preliminary studies, this would seem a logical approach. However, as these experiments have shown, query length only brings about the change in the returned document set distribution, which is the underlying reason certain collections need different normalisation parameter settings.

Acknowledgements

The first author of this work is funded by an IRCSET-Marie Curie International Mobility Fellowship in Science, Engineering and Technology.

References

1. Chowdhury, A., Catherine McCabe, M., Grossman, D., Frieder, O.: Document normalization revisited. In: SIGIR 2002: Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Tampere, Finland, pp. 381–382. ACM Press, New York (2002)
2. Chung, T.L., Luk, R.W.P., Wong, K.F., Kwok, K.L., Lee, D.L.: Adapting pivoted document-length normalization for query size: Experiments in Chinese and English. ACM Transactions on Asian Language Information Processing (TALIP) 5(3), 245–263 (2006)
3. Cummins, R., O'Riordan, C.: Evolving local and global weighting schemes in information retrieval. Information Retrieval 9(3), 311–330 (2006)
4. Cummins, R., O'Riordan, C.: Evolved term-weighting schemes in information retrieval: an analysis of the solution space. Artificial Intelligence Review, 51–68 (November 2007)
5. Salton, G., Wong, A., Yang, C.S.: A vector space model for automatic indexing. Commun. ACM 18(11), 613–620 (1975)
6. Salton, G., Buckley, C.: Term-weighting approaches in automatic text retrieval. Information Processing & Management 24(5), 513–523 (1988)
7. Singhal, A., Buckley, C., Mitra, M.: Pivoted document length normalization. In: SIGIR 1996: Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Zurich, Switzerland, pp. 21–29. ACM Press, New York (1996)
8. Trotman, A.: Learning to rank. Information Retrieval 8, 359–381 (2005)
9. Turtle, H., Croft, W.B.: Inference networks for document retrieval. In: SIGIR 1990: Proceedings of the 13th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 1–24. ACM, New York (1990)