Compression-Based Document Length Prior for Language Models

Javier Parapar
Information Retrieval Lab
Dept. of Computer Science
University of A Coruña
javierparapar@udc.es

David E. Losada
Dept. of Electronics and Computer Science
Univ. Santiago de Compostela
david.losada@usc.es

Álvaro Barreiro
Information Retrieval Lab
Dept. of Computer Science
University of A Coruña
barreiro@udc.es

ABSTRACT

The inclusion of document length factors has been a major topic in the development of retrieval models. We believe that current models can be further improved by more refined estimations of the document's scope. In this poster we present a new document length prior that uses the size of the compressed document. This new prior is introduced in the context of Language Modeling with Dirichlet smoothing. The evaluation performed on several collections shows significant improvements in effectiveness.

Categories and Subject Descriptors: H.3.3 [Information Search and Retrieval]: Retrieval models

General Terms: Performance, Experimentation.

Keywords: Document Length, Document Priors, Language Models, Compression.

1. INTRODUCTION AND MOTIVATION

Document length is recognized as an important component to achieve state of the art performance in Information Retrieval (IR). For instance, popular IR models, such as BM25 [4], pivoted vector space models [5] or Language Models with Dirichlet smoothing [6], incorporate some form of document length correction.

However, these document length corrections have been often based on very rough estimations of the document's contents, such as the size of the document in bytes or the number of terms in the document. Here, we argue that the scope of a document should be captured using more elaborated measures. In this poster, we propose to apply a document length prior obtained from the size of the compressed document. Our research hypothesis is that, to estimate a document's scope, this compression-based measure is more reliable than either the original size of the document or the number of terms in the document. If two documents have equal size but the compressed size of one of them is much smaller than the other document's compressed size then this seems to indicate that the former document is more verbose than the first one. To the best of our knowledge, this simple idea, which can be easily implemented and evaluated, has not been explored for ad-hoc document retrieval. In contrast, the use of compression-based techniques has attracted a great deal of attention in areas such as classification and clustering [3, 1].

SIGIR'09, July 19–23, 2009, Boston, Massachusetts, USA.
ACM 978-1-60558-483-6/09/07.

Language Modeling (LM) with Dirichlet smoothing is a natural choice to test this compression-based method because it can naturally embed query-independent features through document priors. We have therefore defined different priors based on distinct estimations of documents' scope and tested whether any of these variations outperform the standard Dirichlet approach, which is a very competitive baseline.

In section 2 we define the priors, in section 3 the experiments are reported and the poster concludes with a summary.

2. COMPRESSION-BASED PRIOR

In LM, the probability of a document given a query, $P(d|q)$, is utilized to rank documents and it is often estimated using the Bayes' rule:

$$P(d|q) = \frac{P(q|d) \cdot P(d)}{P(q)} \stackrel{rank}{=} \log P(q|d) + \log P(d) \qquad (1)$$

$P(q)$ can be dropped for document ranking purposes. The prior $P(d)$ encodes a-priori information on documents and the query likelihood, $P(q|d)$, incorporates usually some form of smoothing. In this poster, we are only concerned with unigram Language models and Dirichlet smoothing:

$$P(q|d) = \prod_{i=1}^{n} \frac{tf(q_i, d) + \mu \cdot P(q_i|C)}{|d| + \mu} \qquad (2)$$

where n is the number of query terms, $tf(q_i, d)$ is the raw term frequency of q_i in d, $|d|$ is the total count of terms in the document, and μ is a parameter for adjusting the amount of smoothing applied, where $\mu \geq 0$. $P(q_i|C)$ is the probability of the term q_i occurring in the collection C, and is usually a maximum likelihood estimator computed using the collection of documents.

It can be proved [6] that the query likelihood in a sum-log fashion, $\log P(q|d)$, reduces to :

$$\sum_{i:tf(q_i,d)>0} \log\left(1 + \frac{tf(q_i, d)}{\mu P(q_i|C)}\right) + n \ \log \frac{\mu}{|d| + \mu} \qquad (3)$$

In ad-hoc IR, Dirichlet has been often used with an uniform prior $P(d)$ (i.e. $\log P(d)$ is dropped) because $\log P(q|d)$ is a very competitive model (document length correction is already incorporated by the second addend in eq. 3). Non-uniform priors based on document length have been applied

to improve LM estimations [2] but they were only beneficial to Jelinek-Mercer (JM) smoothing, and not to Dirichlet smoothing. This is because JM smoothing does not provide any length normalization, and so a length-based prior provides some form of correction. In this poster we demonstrate that our novel prior is actually able to produce significant improvements in effectiveness over the standard Dirichlet model. We compare the following non-uniform priors[1]:

$$P(d) = \frac{|d|}{\sum_{d_i \in C} |d_i|}, \quad P(d) = \frac{com(d)}{\sum_{d_i \in C} com(d_i)} \qquad (4)$$

where $com(d)$ is the size (bytes) of the compressed document (zipped) divided by the original size (bytes) of the document. These priors will be referred to as terms prior and zipped prior, respectively.

3. EXPERIMENTS AND RESULTS

The priors described above were evaluated with three ad-hoc collections (TRECs 5, 6 & 8, 50 queries each) and a web collection (WT10g, 100 queries). We applied the Porter stemmer, removed common words using an standard stoplist, and tested the following values for μ: 10, 500, 1000, 2000, 3000, 4000, 5000, 10000, 50000. Although we ran experiments with short queries (title subfield) and long queries (all subfields), we report here only the results for short queries. With long queries, there was no significant difference between priors.

The best MAP and P@10 results for each prior are shown in Table 1 and can be summarized as follows. First, the terms prior is worse than the standard Dirichlet configuration (i.e. uniform prior). This is not surprising as the same outcome was reported in [2]. Dirichlet with an uniform prior already incorporates length correction ($|d|$ in eq. 3) and it does not get further benefits from adding length normalization through a document prior. Second, the zipped prior shows the best performance and, in most of the cases, the improvement over the uniform prior is statistically significant. Furthermore, we studied how performance evolves across all parameter settings and found that the improvement obtained with the zipped prior is very stable. In Figure 1 we show how the performance of the methods evolves against μ. The figure represents only the WT10g collection but the same trends occurred in all the collections with both evaluation measures. These results are very remarkable as they demonstrate that the zipped prior beats the other priors not only in the best case scenario but also in any non-optimal case. For a clearer view, the graph shows only μ values up to 5000 but the trends remain the same for higher μ values.

With long queries, the difference between distinct priors was negligible. No advantage or disadvantage was found with compression. This makes sense because standard Dirichlet tends to retrieve too many long documents when queries are short but this excessive promotion of long documents does not happen with long queries [2] (as n grows the penalization of long documents increases). Therefore, the zipped prior, which penalizes verbose documents, is less useful with long queries.

[1]We also tested a prior based on the number of unique terms and another based on the size in bytes but there were no major differences between these priors and the terms prior.

		P10	
Col.	*unif. prior*	*terms prior*	*zipped prior*
T5	.2960 (1000)	.2820 (10)	**.3160*** (+6.7%) (1000)
T6	.3880 (500)	.3740 (500)	**.4120*** (+6.1%) (500)
T8	.4480 (2000)	.4320 (10)	**.4540** (+1.3%) (500)
WT10g	.3071 (2000)	.2816 (500)	**.3184** (+3.7%) (2000)
		MAP	
Col.	*unif. prior*	*terms prior*	*zipped prior*
T5	.1460 (1000)	.1389 (500)	**.1506*** (+3.1%) (1000)
T6	.2263 (500)	.2210 (10)	**.2311*** (+2.1%) (500)
T8	.2481 (500)	**.2491** (10)	**.2491** (+0.4%) (1000)
WT10g	.2080 (2000)	.1934* (500)	**.2132*** (+2.5%) (3000)

Table 1: Optimal performance (best μ in brackets). Stat. sig. (Wilcoxon $p < 0.05$) differences between each non-uniform prior run and the unif. prior run are starred. The performance of the best model is bolded and its percentage improvement over the unif. prior run is reported in brackets.

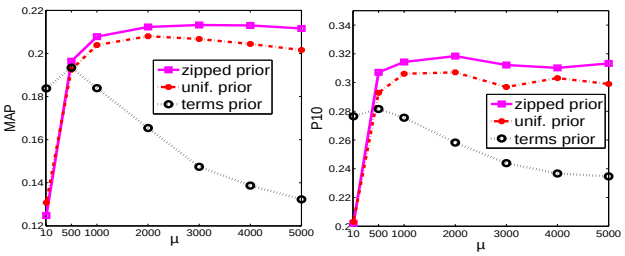

Figure 1: MAP/P10 with varying μ (WT10g).

4. CONCLUSIONS AND FUTURE WORK

In this poster we proposed a novel compression-based prior that significantly outperforms the standard LM Dirichlet model when queries are short. This prior, which can be efficiently implemented, leads to performance improvements that are robust across different parameter settings and test collections. These results are promising and encourage us to further assess the ability of the prior to enhance other LM-based models.

Acknowledgements: This research was co-funded by FEDER, *Ministerio de Ciencia e Innovación* and *Xunta de Galicia* under projects TIN2008-06566-C04-04, 07SIN005206PR, and 2008/068.

5. REFERENCES

[1] R. Cilibrasi and P. Vitanyi. Clustering by compression. *IEEE Trans. on Information Theory*, 51:1523–1545, 2005.

[2] D. Losada and L. Azzopardi. An analysis on document length retrieval trends in language modeling smoothing. *Information Retrieval*, 11:109–138, 2008.

[3] Y. Marton, N. Wu, and L. Hellerstein. On compression-based text classification. In *Proc. ECIR-05*, 300–314, 2005.

[4] S. Robertson, S. Walker, S. Jones, M. HancockBeaulieu, and M. Gatford. Okapi at TREC-3. In *Proc. TREC-3*, 109–127, 1995.

[5] A. Singhal, C. Buckley, and M. Mitra. Pivoted document length normalization. In *Proc. SIGIR-96*, 21–29, 1996.

[6] C. Zhai and J. Lafferty. A study of smoothing methods for language models applied to information retrieval. *ACM Transactions on Information Systems*, 22(2):179–214, 2004.