# Using Parsimonious Language Models on Web Data

Rianne Kaptein[1] Rongmei LI[2] Djoerd Hiemstra[2] Jaap Kamps[1,3]

[1] Archives and Information Studies, Faculty of Humanities, University of Amsterdam
[2] Database Group, University of Twente
[3] ISLA, Informatics Institute, University of Amsterdam

## ABSTRACT

In this paper we explore the use of parsimonious language models for web retrieval. These models are smaller thus more efficient than the standard language models and are therefore well suited for large-scale web retrieval. We have conducted experiments on four TREC topic sets, and found that the parsimonious language model results in improvement of retrieval effectiveness over the standard language model for all data-sets and measures. In all cases the improvement is significant, and more substantial than in earlier experiments on newspaper/newswire data.

**Categories and Subject Descriptors:** H.3 [Information Storage and Retrieval]: H.3.3 Information Search and Retrieval
**General Terms:** Measurement, Experimentation, Performance
**Keywords:** Web Retrieval, Language Models, Parsimonious Language Models

## 1. INTRODUCTION

We examine the use of parsimonious language models for retrieval on a large-scale web data. The parsimonious language model—as introduced by Sparck-Jones et al. [3] and practically implemented in Hiemstra et al. [1]—overcomes some of the weaknesses of the standard language modeling approach. Instead of blindly modeling language use, we should model what language use distinguishes a relevant document from other documents. Words that are common in general English, and words that occur only occasionally in documents, are already well explained by the background corpus, and therefore do not have to be included in a document model. This results in language models with far fewer terms, and when used at indexing time leads to smaller indexes and more efficient retrieval, making them especially attractive for large-scale web retrieval. The decrease in index size should not be at the cost of a loss of retrieval performance, in fact, the parsimonious model may improve performance. We will focus exclusively on the effectiveness here.

## 2. MODELS

In this paper we use a unigram language model. It uses a mixture of the document model with a general collection model as follows, i.e., for a collection $C$, document $D$ and query $q$:

$$P(q|D) = \prod_{t \in q} \left(\lambda P(t|D) + (1-\lambda) P(t|C)\right),$$


*SIGIR'08,* July 20–24, 2008, Singapore.
ACM 978-1-60558-164-4/08/07.

where

$$P_{mle}(t|D) = \frac{tf_{t,D}}{\sum_t tf_{t,D}}$$
$$P_{mle}(t|C) = \frac{\text{doc\_freq}(t,C)}{\sum_{t' \in C} \text{doc\_freq}(t',C)}$$

Instead of using maximum likelihood estimation to estimate the probability $P(t|D)$, it can also be estimated using parsimonious estimation. The parsimonious model concentrates the probability mass on fewer terms than a standard language model. Terms that are better explained by the general language model $P(t|C)$ (i.e. terms that occur about as frequent in the document as in the whole collection) can be assigned zero probability, thereby making the parsimonious language model smaller than a standard language model. The model automatically removes stopwords, and words that are mentioned occasionally in the document [1].

The model is estimated using *Expectation-Maximization*:

$$\text{E-step:} \quad e_t = tf_{t,D} \cdot \frac{\alpha P(t|D)}{\alpha P(t|D) + (1-\alpha) P(t|C)}$$
$$\text{M-step:} \quad P(t|D) = \frac{e_t}{\sum_t e_t}, \text{i.e. normalize the model}$$

In the initial E-step, the maximum likelihood estimates are used to estimate $P(t|D)$. The E-step benefits terms that occur relatively more frequent in the document than in the whole collection. The M-step normalizes the probabilities. After the M-step terms that receive a probability below a certain threshold or pruning factor are removed from the model. In the next iteration the probabilities of the remaining terms are again normalized. The iteration process stops after a fixed number of iterations or when the probability distribution does not change significantly anymore. For $\alpha = 1$, and a threshold of 0, the algorithm produces the maximum likelihood estimate $P_{mle}(t|D)$ as defined before. Lower values of $\alpha$ result in a more parsimonious model. We will denote the resulting estimate by $P_{pars}(t|D)$.

To illustrate the effect of the parsimonious language models, we selected a topic (Terabyte track topic "model railroads") and built three different models of the top 10 results: a standard language model (using maximum likelihood estimation); a standard language model that removes stopwords; and a parsimonious language model. In Table 1 the top ranked terms of all three models are shown. The standard language model that excludes stopwords still contains some words that could be considered as stopwords, like 'm' and 'p'. When a standard stopword list is used there is always a trade-off between being complete and being too aggressive. When the parsimonious model is used, the document is compared to the background corpus to remove all words that do not occur more frequently in the document as in the background corpus, e.g.

**Table 2: Retrieval results on the TREC data sets**

| Dataset | TREC-8 | | | Terabyte '04 | | | Terabyte '05 | | | Terabyte '06 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| # Topics | 50 | | | 49 | | | 50 | | | 50 | | |
| $P(t\|D)$ | MLE | Parsimonious | | MLE | Parsimonious | | MLE | Parsimonious | | MLE | Parsimonious | |
| MAP | 0.2331 | 0.2428 | +4.2%$^{\star}$ | 0.2095 | 0.2206 | +3.3%$^{\star\star\star}$ | 0.2461 | 0.2567 | +4.3%$^{\star}$ | 0.2139 | 0.2374 | +11.0%$^{\star\star\star}$ |
| Bpref | 0.2481 | 0.2571 | +3.6%$^{\star}$ | 0.2926 | 0.3048 | +4.2%$^{\star}$ | 0.3014 | 0.3103 | +3.0%$^{\star\star}$ | 0.3234 | 0.3422 | +5.8%$^{\star\star}$ |
| P@10 | 0.3640 | 0.4040 | +11.0%$^{\star\star}$ | 0.3265 | 0.3714 | +13.8%$^{\star\star\star}$ | 0.4200 | 0.4700 | +11.9%$^{\star\star\star}$ | 0.3300 | 0.3660 | +10.9%$^{\star}$ |

Significance of Pars. over MLE according to t-test, one-tailed, at significance levels 0.05 ($^{\star}$), 0.01 ($^{\star\star}$), and 0.001 ($^{\star\star\star}$).

**Table 1: Top ranked terms of topic "model railroads"**

| Standard LM | | Standard LM No stopwords | | Parsimonious LM | |
|---|---|---|---|---|---|
| Term | $P_{mle}(t)$ | Term | $P_{mle}(t)$ | Term | $P_{pars}(t)$ |
| the | 0.0440 | m | 0.0203 | museum | 0.0527 |
| and | 0.0289 | museum | 0.0143 | railroad | 0.0402 |
| of | 0.0261 | p | 0.0119 | train | 0.0344 |
| a | 0.0205 | railroad | 0.0112 | tel | 0.0280 |
| m | 0.0140 | train | 0.0093 | trains | 0.0233 |
| in | 0.0130 | www | 0.0089 | adults | 0.0185 |
| to | 0.0125 | hours | 0.0085 | museums | 0.0169 |
| for | 0.0118 | ca | 0.0084 | depot | 0.0142 |

the word "www" is very common in the .GOV2 corpus and does therefore not occur in the parsimonious model. The parsimonious model does not only remove all standard stopwords, but also the corpus specific stopwords.

# 3. EXPERIMENTS

## 3.1 Experimental Set-up

We test our models on four TREC datasets, Web track TREC-8 (WT2g collection of 250K documents) and Terabyte tracks 2004, 2005 and 2006 (.GOV2 collection of 25M documents) [4]. Using parsimonious language models at indexing time can significantly reduce the index size, but in order to experiment with all parameters we choose to use parsimonious models at retrieval time. For efficiency reasons, we only rerank top 1,000 results of the standard language model. We use the standard language model as described in Section 2, where $P(t|D)$ is calculated using either maximum likelihood estimation, $P_{mle}(t|D)$, or according to the parsimonious model, $P_{pars}(t|D)$. Stopwords are not removed.

In ad hoc retrieval, the standard value of the smoothing parameter $\lambda$ in the language model is 0.15. In the TREC Terabyte tracks, it is known that the .GOV2 collection requires little smoothing [2], i.e. a value of 0.9 for $\lambda$ gives the best results. Experiments on the TREC-8 web data confirm that also the small Web data collection requires substantially less smoothing, so for both datasets we use a value of 0.9 for $\lambda$.

For the parsimonious model we have to set the parameters $\alpha$ and the threshold parameter. We set the threshold parameter at 0.0001, i.e. words that occur with a probability less than 0.0001 are removed from the index. We set $\alpha = 0.1$ for the parsimonious model, based on initial experiments with a part of the topic set.

## 3.2 Results

The results of the models on the different topic sets are summarized in Table 2. A number of observations present themselves: We see that the use of the parsimonious language model leads to the improvement of retrieval effectiveness on all four data-sets. In fact, we see a substantial improvement on all three measures: mean average precision (MAP) increases with 3% to 11%; binary preference (Bpref) increases with 3% to 6%; and precision at rank 10 (P@10) increases with 11% to 14%.[1] The fact that both early precision (P@10) and overall precision (MAP) improve signals that the parsimonious models have a beneficial effect on both precision and recall. Moreover, all the improvements on all four data-sets and three measures are statistically significant, signalling that these beneficial effects apply to a large fraction of the topics. Furthermore, additional experiments show that a larger improvement can be attained when stemming is used, i.e. increases in MAP up to 14%.

# 4. CONCLUSIONS

From our experiments, we can conclude that the parsimonious language model is to be preferred over the standard language model. The parsimonious model produces smaller document models (and hence reduces the index) and obliviates the need for stopword lists. Retrieval results of the parsimonious model are superior to the standard language model, over a range of measures and four TREC data sets.

Earlier experiments [1] found only moderate improvements in MAP of around 3% for the TREC 7 and TREC 8 adhoc track using newspaper/newswire data. We find improvements in MAP in the range 3% to 11% for Web data. Arguably, the larger Web collections are more susceptible to the parsimonious language model.

**Acknowledgments** This research was supported by the Netherlands Organization for Scientific Research (NWO, under project # 612.066.513).

# REFERENCES

[1] D. Hiemstra, S. Robertson, and H. Zaragoza. Parsimonious language models for information retrieval. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 178–185. ACM Press, New York NY, 2004.

[2] J. Kamps. Effective smoothing for a terabyte of text. In *The Fourteenth Text REtrieval Conference (TREC 2005)*. National Institute of Standards and Technology. NIST Special Publication, 2006.

[3] K. Sparck-Jones, S. Robertson, D. Hiemstra, and H. Zaragoza. Language modelling and relevance. In W. Croft and J. Lafferty, editors, *Language Modeling for Information Retrieval*, pages 57–71. Kluwer Academic Publishers, 2003.

[4] TREC. Text REtrieval Conference, 2008. `http://trec.nist.gov/`.

[1] We use binary preference mainly as a safe-guard. The parsimonious models are potentially retrieving documents not part of the original assessment pool. Since Bpref and MAP are in agreement, we have no reason to distrust the MAP scores.