

# Automatic Document Prior Feature Selection for Web Retrieval

Jie Peng, Craig Macdonald, Iadh Ounis  
Department of Computing Science  
University of Glasgow, Scotland, UK  
{pj,craigm,ounis}@dcs.gla.ac.uk

## ABSTRACT

Document prior features, such as Pagerank and URL depth, can improve the retrieval effectiveness of Web Information Retrieval (IR) systems. However, not all queries equally benefit from the application of a document prior feature. This paper aims to investigate whether the retrieval performance can be further enhanced by selecting the best document prior feature on a per-query basis. We present a novel method for selecting the best document prior feature on a per-query basis. We evaluate our technique on the TREC .GOV Web test collection and its associated TREC 2003 Web search tasks. Our experiments demonstrate the effectiveness and robustness of our proposed selection method.

**Categories and Subject Descriptors:** H.3.3 [Information Storage & Retrieval]: Information Search & Retrieval

**General Terms:** Performance, Experimentation

**Keywords:** Prior Feature Selection

## 1. INTRODUCTION

Several previous studies have shown that integrating document prior features, such as Pagerank or URL depth, into a document weighting scheme can improve the retrieval performance of a Web IR system [1, 2]. However, not all queries benefit equally from the application of a given feature.

In this paper, we present a novel method for selecting the best document prior feature on a per-query basis. For a given query and its corresponding top retrieved documents, we propose to estimate the divergence between the retrieved document scores distribution prior to, and after the integration of the document prior feature. We then observe that the divergence distribution can be fitted by a Gaussian distribution. Based on this observation, we use a Bayesian decision mechanism to decide which document prior feature is the best for a given query. In this paper, we firstly assess whether it is indeed important to apply the best document prior feature on a per-query basis; secondly, we examine the effectiveness of our proposed document prior feature selection method.

## 2. PRIOR FEATURE SELECTION

The distribution of retrieval scores has been studied to predict the effectiveness of search engines [6]. Instead, we use the divergence of retrieval scores to predict when a doc-

ument prior feature should be applied. There are several different ways to estimate the divergence between the document scores distribution prior to, and after the integration of the document prior feature. For example, Kullback-Leibler divergence [3] and Jensen-Shannon divergence [4]. Jensen-Shannon divergence has an upper bound ( $\leq 1$ ) while Kullback-Leibler does not. To limit the problem of sparseness, we use the Jensen-Shannon divergence:

$$JS(X, Y) = \sum_i x_i \cdot \log_2 \frac{x_i}{\frac{1}{2} \cdot x_i + \frac{1}{2} \cdot y_i} \quad (1)$$

where for the top retrieved documents of a given query,  $X = \{x_i\}$ ,  $Y = \{y_i\}$  and  $x_i$  and  $y_i$  are the relevance score of document  $i$  prior to, and after the integration of a given document prior feature, respectively. It is easy to verify that  $JS(X, Y) \neq JS(Y, X)$ . Following [7], we use the symmetric Jensen-Shannon divergence, resulting in divergence scores that are in the range (0, 2]:

$$SJS(X, Y) = JS(X, Y) + JS(Y, X) \quad (2)$$

We examine the distribution of the divergence of retrieval scores prior to, and after the integration of the document prior feature, measured using Equation (2). As an example, for the TREC 2003 Web Track mixed task dataset, Figure 1 shows a histogram of divergence scores distribution for Pagerank for those queries for which Pagerank led to a better retrieval performance than without the integration of Pagerank. The plot also shows a maximum-likelihood fit using a Gaussian distribution, suggesting that the divergence scores can be fitted using a Gaussian distribution. The maximum-likelihood fit involves the setting of the mean and variance, which needs to be set using training data. Note that we would obtain the same observation, i.e. a Gaussian fit, if all the queries of the TREC 2003 Web Track mixed task dataset were plotted.

Assume that we have two document prior features  $f_1$  and  $f_2$ <sup>1</sup>. For a given query, we propose to automatically select the most effective document prior feature. For this purpose, we describe the Bayesian decision mechanism, which will be used as our document prior feature selection decision mechanism. For a given query, the probability of feature  $f_j$  being the most effective document prior feature with a given divergence *Score* is defined as follows:

$$P(f_j | Score) = \frac{P(f_j) \cdot P(Score | f_j)}{P(Score)} \quad (3)$$

<sup>1</sup>Note that the proposed selection method can naturally expand to select the most effective document prior feature amongst  $n$  document prior features, where  $n > 2$ .

where  $P(f_j)$  is the prior probability of feature  $f_j$  being the most effective document prior feature for a given query. Using a training set, it can be computed as the number of queries for which the application of  $f_j$  led to the most effective retrieval performance divided by the total number of queries;  $P(\text{Score}|f_j)$  is the probability of obtaining divergence *Score* when the most effective document prior feature is  $f_j$ . As we mentioned above, this distribution can be fitted by a Gaussian distribution, given as follows:

$$P(\text{Score}|f_j) = \frac{1}{\sigma\sqrt{2\pi}} \cdot \exp\left(-\frac{(\text{Score} - \mu)^2}{2\sigma^2}\right) \quad (4)$$

where  $\mu$  and  $\sigma$  are the mean and variance of the Gaussian distribution, set using training. We do a simple normalisation on  $P(\text{Score}|f_j)$  as different document prior features have different divergence distributions, which might result in different  $\sigma$  and different range of  $P(\text{Score}|f_j)$ . We normalise  $P(\text{Score}|f_j)$  as follows:

$$P(\text{Score}|f_j) = \lambda \cdot \frac{1}{\sigma\sqrt{2\pi}} \cdot \exp\left(-\frac{(\text{Score} - \mu)^2}{2\sigma^2}\right) \quad (5)$$

where  $\lambda$  is the parameter that controls the range of  $P(\text{Score}|f_j)$ ;  $P(\text{Score}) = \sum_{j=1}^n P(f_j) \cdot P(\text{Score}|f_j)$ ;  $n$  is the number of document prior features involved in the feature selection;  $P(\text{Score})$  can be ignored as it does not affect the final decision making.

	Sample 1	Sample 2	Sample 3
	MAP		
Baseline	0.5441*	0.5631*	0.5694*
+ PR	0.5896*(8.4%)	0.6164*(9.5%)	0.6090*(6.9%)
+ URL	0.5954*(9.4%)	0.6152*(9.3%)	0.6125*(7.6%)
+S(PR, URL)	0.6234*(14.6%)	0.6380*(13.3%)	0.6412*(12.6%)
+M(PR, URL)	<b>0.6460</b> (18.7%)	<b>0.6688</b> (18.8%)	<b>0.6701</b> (17.7%)

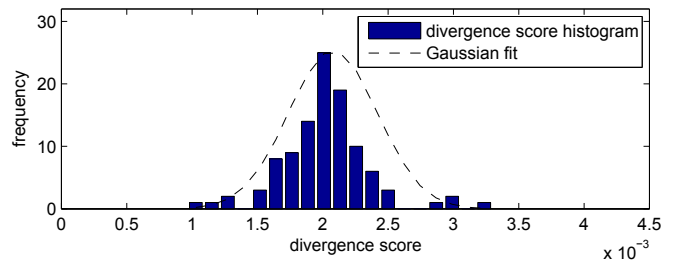
**Table 1: MAP on test dataset. Values in parenthesis denote percentage improvement over baseline. The best runs in each column are highlighted in bold. Values statistically different from the best in column are denoted \* (Wilcoxon Matched-Pairs Signed-Ranks Test,  $p < 0.05$ ).**

### 3. EXPERIMENTS AND ANALYSIS

We use the standard .GOV Web test collection, and its corresponding TREC 2003 Web track mixed topics and relevance assessments. As a training dataset, we sample 80% of the TREC 2003 Web Track mixed topics task: 120 homepage finding topics, 120 named page finding topics, and 40 topic distillation topics, randomly chosen from the 350 topics available in this task. Our test dataset is the remaining 20% of the TREC 2003 Web track mixed topics task (30 homepage finding topics, 30 named page finding topics and 10 topic distillation topics). We repeat this sampling 3 times. The evaluation measure used in all our experiments is mean average precision (MAP). The range of  $i$  in Equation (1) is (0,1000] as TREC usually requires 1000 retrieved documents for each query.

For indexing and retrieval, we use the Terrier IR platform<sup>2</sup>, and apply standard stopwords removal. In addition, to boost early precision, we apply a light version of Porter’s stemming algorithm for English. We index the body, anchor text and titles of documents as separate fields and use the PL2F field-based Divergence From Randomness (DFR) weighting model [5] as a baseline retrieval system. The parameters of the PL2F document weighting model are set by

<sup>2</sup><http://ir.dcs.gla.ac.uk/terrier>



**Figure 1: Histogram of divergence scores distribution split in equal divergence score ranges and the Gaussian fit of pagerank for TREC 2003 Web Track.**

optimising MAP on the training dataset, using a simulated annealing procedure.

We experiment with two document prior features, namely Pagerank (PR) and URL depth (URL) [1]. Firstly, we assess the maximum performance that could be achieved by manually choosing the optimal document prior feature for each query. From Table 1, we can see that the manual document prior feature selection  $M(PR, URL)$  can lead to a significant improvement over the PL2F baseline as well as systems where a given document prior feature (e.g. Pagerank) has been applied uniformly to all queries. This suggests that a document prior feature selection on a per-query basis can significantly enhance the retrieval performance of a Web IR system.

Our automatic feature selection method  $S(PR, URL)$  also leads to similar marked improvements over the PL2F baseline as well as the uniform application of a given document prior feature, such as Pagerank or URL depth. Its overall performance, while naturally lower than the manual  $M(PR, URL)$  method, is still fairly comparable. Note that the above two results are consistent across all our three random samplings, suggesting that our proposed feature selection method is robust.

### 4. CONCLUSIONS

We investigated the retrieval performance achieved with document prior feature selection on a per-query basis on the TREC 2003 Web Track, using a standard TREC Web test collection. We showed that the appropriate selection of a document prior feature selection on a per-query basis can significantly enhance the retrieval performance. Moreover, we observed that our proposed automatic document prior feature selection method consistently and markedly increases the retrieval performance over baselines that only use a single type of document prior feature uniformly on all queries. In the future, we plan to apply a more thorough cross-validation, to assess the robustness of our method.

### 5. REFERENCES

- [1] J. Kamps, G. Mishne, M. de Rijke. Language Models for Searching in Web Corpora. In *Proceedings of TREC 2004*.
- [2] W. Kraaij, T. Westerveld, D. Hiemstra. The Importance of Prior Probabilities for Entry Page Search. In *Proceedings of SIGIR 2002*, pages 27-34.
- [3] S. Kullback. *Information Theory and Statistics*. Jown Wiley & Sons, New York, USA, 1959.
- [4] J. Lin. Divergence Measures Based on the Shannon Entropy. *IEEE Transactions on Information Theory*, 37:145-151, 1991.
- [5] C. Macdonald, V. Plachouras, B. He, C. Lioma, I. Ounis. University of Glasgow at WebCLEF 2005: Experiments in per-field normalisation and language specific stemming. In *Proceedings of CLEF 2005*.
- [6] R. Manmatha, T. Rath, F. Feng. Modeling Score Distributions for Combining the Outputs of Search Engines. In *Proceedings of SIGIR 2001*, pages 267-275.
- [7] V. Plachouras. Selective Web Information Retrieval. *Phd Thesis*. Univ of Glasgow, 2006.