Selective Application of Query-Independent Features in Web Information Retrieval

Jie Peng and Iadh Ounis

Department of Computing Science,
University of Glasgow, G12 8QQ, UK
{pj,ounis}@dcs.gla.ac.uk

Abstract. The application of query-independent features, such as PageRank, can boost the retrieval effectiveness of a Web Information Retrieval (IR) system. In some previous works, a query-independent feature is uniformly applied to all queries. Other works predict the most useful feature based on the query type. However, the accuracy of the current query type prediction methods is not high. In this paper, we investigate a novel approach that applies the most appropriate query-independent feature on a per-query basis, and does not require the knowledge of the query type. The approach is based on an estimate of the divergence between the retrieved document scores' distributions prior to, and after the integration of a query-independent feature. We evaluate our approach on the TREC .GOV Web test collection and the mixed topic sets from TREC 2003 & 2004 Web search tasks. Our experimental results demonstrate that the selective application of a query-independent feature on a per-query basis is very effective and robust. In particular, it outperforms a query type prediction-based method, even when this method is simulated with a 100% query type prediction accuracy.

1 Introduction

Various previous studies have shown that the application of query-independent features, such as PageRank [1] and URL depth [6], can enhance the retrieval effectiveness of a Web IR system [2,5,12]. Most of these studies have mainly focussed on the uniform application of a query-independent feature to all queries. Others use the type of the query such as Homepage finding, Named Page finding, or Topic Distillation, to predict those query-independent features that are most useful for retrieval. For example, the URL type feature is usually effective for the Homepage finding queries [6,8]. In this context, one possible solution is to predict the query type and then apply the most appropriate query-independent feature based on the predicted query type.

However, in a real IR environment, users do not mention the type of their submitted queries. Moreover, the query type prediction is not quite accurate even when it involves binary selections, according to the TREC 2004 Web Track overview [4]. For example, the highest accuracy of query type prediction between Homepage and Named Page finding topics in [4] is 68%.

M. Boughanem et al. (Eds.): ECIR 2009, LNCS 5478, pp. 375–387, 2009.

In this paper, we present a novel method for selecting the most appropriate query-independent feature on a per-query basis, which does not require the knowledge of the query type. For each query and its corresponding top retrieved documents, we propose to estimate the divergence between the retrieved document scores' distributions prior to, and after the integration of the query-independent feature. We then use our proposed decision mechanism, which is based on the distribution of the estimated divergence scores, to selectively apply the most appropriate query-independent feature.

We conduct our experiments on the standard .GOV Web test collection [3]. In order to test our proposed method on a big enough dataset, we mix the topic sets from TREC 2003 & 2004 Web Tracks and separate the dataset into three folds of equal size. We iteratively test our feature selection method on one fold after training on the remaining two folds. Moreover, to build a strong retrieval baseline system, we use a field-based model, namely the PL2F document weighting model. In addition, we experiment with the three query-independent features: PageRank, URL depth and Click Distance.

The objectives of this paper are twofold. Firstly, we examine how important it is to selectively apply query-independent feature on a per-query basis in Web IR. Secondly, we test how effective our proposed method is for the selective application of a query-independent feature on a per-query basis. In particular, we test how effective our proposed method is when the number of candidate features changes. We also show that our approach is more effective than the query type prediction-based method (denoted as QTP), even when this method is simulated with a 100% query type prediction accuracy.

The remainder of this paper is organised as follows. In Section 2, we present the field-based document weighting model, which will be used in this work to rank documents. Section 3 introduces the query-independent features used in this paper, and how they are integrated into a document weighting model. Section 4 describes our proposed method for selectively applying a query-independent feature. We present the experimental setting in Section 5, and analyse the experimental results in Section 6. Finally, we conclude the work in Section 7.

2 Divergence from Randomness Model

Many studies have shown that the overall retrieval performance of a Web IR system can be enhanced when the document structure (or fields), such as the body, the title, and the anchor text of its incoming hyperlinks [14,17]. In particular, Robertson et al. [16] showed improved retrieval effectiveness in TREC Web search tasks when the contribution of each field to the document ranking was controlled by the use of weights.

Therefore, in order to obtain a strong baseline system, we apply a field-based Divergence From Randomness (DFR) weighting model. In particular, we use the PL2F field-based model [10], which was shown to be effective on the TREC Web test collections [15]. Using the DFR PL2F model, the relevance score of a document D for a query Q is given by:

$$score(D,Q) = \sum_{t \in Q} qtw \cdot \frac{1}{tfn+1}\big(tfn \cdot \log_2 \frac{tfn}{\lambda} + (\lambda - tfn) \cdot \log_2 e + 0.5 \cdot \log_2(2\pi \cdot tfn)\big) \quad (1)$$

where λ is the mean and variance of a Poisson distribution, given by $\lambda = F/N$; F is the frequency of the query term t in the whole collection, and N is the number of documents in the whole collection. The query term weight qtw is given by qtf/qtf_{max}; qtf is the query term frequency; qtf_{max} is the maximum query term frequency among the query terms.

In PL2F, tfn corresponds to the weighted sum of the normalised term frequencies tf_f for each used field f, known as *Normalisation 2F* [10]:

$$tfn = \sum_f \left(w_f \cdot tf_f \cdot \log_2(1 + c_f \cdot \frac{avg\_l_f}{l_f}) \right), (c_f > 0) \quad (2)$$

where tf_f is the frequency of term t in field f of document D; l_f is the length in tokens of document D in field f, and avg_l_f is the average length of the field across all documents; c_f is a hyper-parameter for each field, which controls the term frequency normalisation; the importance of the term occurring in field f is controlled by the weight w_f. The values of these parameters are obtained by training as will be explained in Section 5.

3 Query-Independent Features

We use three widely used query-independent features, namely PageRank, URL depth and Click Distance, which have been shown to particularly enhance the retrieval performance of a Web IR system on some of the TREC Web search tasks [5,8]. For the integration of a query-independent feature into a document weighting scheme, we use the $FLOE$ method [5], which has been shown to be an effective approach for transforming a query-independent feature score into a document relevance score.

3.1 PageRank (PR)

Documents in the Web are connected through hyper-links. A hyper-link is a connection between a source and a target document. There is a simple assumption that a hyper-link from document A to document B stipulates that the document A's author considers document B to be valuable. A high number of incoming links often indicates that many documents' authors consider the given document to be of a high quality. PageRank [1] extends this idea by not only counting the number of incoming links to a document, but also by taking the quality of incoming links into account. The PageRank feature score of a given document is computed as follows:

$$Score(D)_{PR} = (1 - \lambda_{PR}) + \lambda_{PR} \cdot \sum_{i=1}^{n} \frac{Score(D_i)_{PR}}{c(D_i)} \quad (3)$$

where D_i is a Web document linking to page D, $c(D_i)$ is the number of outgoing links from document D_i, and λ_{PR} is a damping factor. In this paper, we use the default setting $\lambda_{PR} = 0.85$ [1].

3.2 URL Depth (UD)

A Uniform Resource Locator (URL), which contains a string of symbols, defines the unique location of a document on the Web. The string of symbols can be divided into many components by the symbol '/'. For example: the URL `www.firstgov.gov/topics/science.html` can be divided into 3 components, which are *www.firstgov.gov*, *topics* and *science.html*. The URL depth feature score for a given document is defined as follows:

$$Score(D)_{UD} = Num_{component} \tag{4}$$

where $Num_{component}$ is the number of components after the division.

3.3 Click Distance (CD)

Click Distance is a link metric which measures the number of minimum clicks it takes to reach a web document from a given root [5]:

$$Score(D)_{CD} = Num_{click} \tag{5}$$

where Num_{click} is the number of clicks from the root to document D. For example, if it takes 6 clicks from the root to go to page A and 2 clicks from the root to go to page B, then page B has a smaller Click Distance than page A.

3.4 The *FLOE* Method

Craswell et al. [5] proposed the *FLOE* method for transforming a query-independent feature score into a per-document relevance score. The method allocates a query-independent feature score for each document D as follows:

$$score(D, Q) = score_{QD}(D, Q) + score_{QI}(D) \tag{6}$$

where $score_{QD}(D, Q)$ is the query-dependent relevance score of D given a query Q and can be estimated by a document weighting scheme, such as PL2F in this paper; $score_{QI}(D)$ is the query-independent relevance score for a given document D, estimated by the $FLOE$ using a query-independent feature. $score(D, Q)$ is the final relevance score of document D given the query Q.

Craswell et al. proposed two different versions of the *FLOE* method. In this paper, we denote them as *FLOE1* and *FLOE2*, respectively. The two versions of *FLOE* are defined as follows:

$$FLOE1(S, w, k, a) = w \cdot \frac{S^a}{k^a + S^a} \tag{7}$$

$$FLOE2(S, w, k, a) = w \cdot \frac{k^a}{k^a + S^a} \tag{8}$$

where S is the query-independent feature score, w, k and a are parameters. With the same w, k, and a settings, in Equation (7), a document with a higher query-independent feature score attains a higher relevance score after the transformation, while in Equation (8), a document with a higher query-independent feature score attains a lower relevance score after the transformation. For example, PageRank scores are mapped using $FLOE1$, as a document that has a high PageRank score is usually considered to be a high-quality document; On the contrary, URL depth scores are transformed using $FLOE2$ as documents with shorter URL depth are usually seen as more authoritative than pages with longer URL depth.

4 Feature Selection

In this section, we propose a novel method for selectively applying the most appropriate query-independent feature on a per-query basis. The distribution of retrieval scores has been applied to predict the effectiveness of a search engine [11]. In this paper, we use the divergence between the retrieved document scores' distributions, prior to and after the integration of the query-independent feature, to predict which query-independent feature should be applied, independently of the query type.

URL type feature has shown its effectiveness in Homepage finding task [8] and it is computed based on two distributions: one is the distribution of the number of documents in the relevance assessment set with different URL type; another one is the distribution of the number of documents in the test collection with different URL type. Inspired by this idea, we propose a decision mechanism, which is also based on two different distributions of the estimated divergence scores, to selectively apply the most appropriate query-independent feature. The details of the method are provided in the following sections.

4.1 Divergence between Probability Distributions

There are several different ways to estimate the divergence between the document scores distribution prior to, and after the integration of the query-independent feature. In this paper, we use Jensen-Shannon divergence [9], given as follows:

$$JS(X, Y) = \sum_{i=1}^{n} x_i \cdot \log_2 \frac{x_i}{\frac{1}{2} \cdot x_i + \frac{1}{2} \cdot y_i} \tag{9}$$

where for the top n retrieved documents of a given query, $X = \{x_i\}$, $Y = \{y_i\}$ and x_i and y_i are the relevance scores of document i prior to, and after the integration of a given query-independent feature, respectively. It is easy to verify that $JS(X, Y) \neq JS(Y, X)$. In order to avoid the issue of the ordering of X and Y, we use the symmetric Jensen-Shannon (SJS) divergence:

$$SJS(X, Y) = JS(X, Y) + JS(Y, X) \tag{10}$$

4.2 Decision Mechanism

For a given query Q, assume that we have k query-independent candidate features: f_1, f_2, ..., f_k, and we need to apply the most appropriate one. For this purpose, we describe the decision mechanism of our selective application method as follows:

- First, on the training dataset, we use the SJS divergence estimation method to estimate the f_ϕ's divergence score for each query. Note that one divergence score will be estimated for each query on each given feature f_ϕ.
- Second, we put all of the estimated f_ϕ's divergence scores into a bin (note that the estimated divergence scores for different query-independent features will be put into different bins) and divide the bin into several equal size sub-bins, according to the logscale of the divergence scores of f_ϕ. Each sub-bin corresponds to an interval of divergence scores. We denote the interval of each sub-bin of feature f_ϕ as $S_\chi(f_\phi)$. Note that the number of sub-bins is an important parameter which needs an appropriate setting.
- Third, for each sub-bin, it contains two important numbers: one is the number of queries, whose divergence scores are in the interval of this sub-bin, we denote it as $c(S_\chi(f_\phi))$; another is the number of queries for which f_ϕ led to a better retrieval performance than all the other query-independent features in the interval of this sub-bin, we denote it as $c(S_\chi(f_\phi, BEST))$. Note that the above three steps are completed on a training dataset.
- Finally, on the test dataset, with the given query Q, we use the SJS divergence estimation method to estimate the divergence score between the top retrieved document scores distribution prior to, and after the integration of a feature f_ϕ. The resulting divergence score is then allocated into the corresponding interval of feature f_ϕ's sub-bin. The probability of f_ϕ being the most appropriate query-independent feature for this given query Q is defined as follows:

$$P(f_\phi|Q) = \frac{c(S_\chi(f_\phi, BEST))}{c(S_\chi(f_\phi))} \tag{11}$$

  We apply feature f_ϕ if and only if it has the highest $P(f_\phi|Q)$ score compared with all other features. Note that the computational cost of our proposed feature selection method is very cheap as we only compute the divergence of the top n retrieved documents. n is a parameter that needs an appropriate setting.

As an example, we selectively apply the most appropriate query-independent feature between PageRank and URL depth on the .GOV test collection, using the title-only mixed topics from the TREC 2003 Web Track. In this dataset, there are 350 queries in total. Based on our retrieval system setting, there are 74 queries where PageRank is the most appropriate query-independent feature, 91 queries where the URL depth is the most appropriate query-independent feature and 185 queries where both PageRank and URL depth produce the same retrieval performance. In this example, we set the number of the top retrieved documents, namely n in Equation (9), to 1000 and the number of sub-bins to 5. From Table 1, we can see that, in some intervals, such as $S_\chi = S_3$ and $S_\chi = S_5$, the probability

Table 1. Example of the probability of PageRank ($f_\phi = PR$) and URL depth ($f_\phi = UD$) being the most appropriate query-independent feature in each interval S_χ on the TREC 2003 Web Track, respectively

	$c(S_\chi(PR, BEST))$	$c(S_\chi(PR))$	$P(PR\|Q)$	$c(S_\chi(UD, BEST))$	$c(S_\chi(UD))$	$P(UD\|Q)$
$S_\chi = S_1$	1	7	0.1428	1	2	0.5
$S_\chi = S_2$	12	64	0.1875	3	9	0.3333
$S_\chi = S_3$	52	239	0.2175	46	199	0.2311
$S_\chi = S_4$	4	31	0.1290	38	132	0.2878
$S_\chi = S_5$	5	9	0.5555	3	8	0.3750
total	74	350		91	350	

Table 2. Details of the number and percentage of topics associated to each topic type for TREC 2003 and TREC 2004 Web Tracks

	TREC 2003			TREC 2004		
	HP	NP	TD	HP	NP	TD
Number of topics	150	150	50	75	75	75
Percentage	42.9%	42.9%	14.2%	33.3%	33.3%	33.3%

of PageRank ($f_\phi = PR$) being the most appropriate query-independent feature when allocated in $S_\chi = S_5$ is higher than in $S_\chi = S_3$ even though $S_\chi = S_3$ has a higher number of $c(S_\chi(f_\phi, BEST))$. This shows that our decision mechanism is based on the distributions from both $c(S_\chi(f_\phi, BEST))$ and $c(S_\chi(f_\phi))$. A similar phenomenon is also observed for the URL depth feature ($f_\phi = \text{UD}$). Assume that the divergence scores of PageRank and URL depth for a given query Q are allocated into interval S_5 and S_4, respectively, which means that the probabilities of PageRank and URL depth being the most appropriate query-independent feature for this given query are equal to 0.5555 and 0.2878, respectively. In this case, we apply PageRank as it has higher $P(f_\phi|Q)$ score ($0.5555 > 0.2878$).

5 Experimental Environment

We use the standard .GOV Web test collection, and its corresponding TREC 2003 & 2004 Web Tracks title-only topics and relevance assessment sets. For the TREC 2003 and TREC 2004 Web Tracks, there are three different topic types, namely Homepage (HP) finding topics, Named Page (NP) finding topics and Topic Distillation (TD) topics. From Table 2, we can see that the percentages of each topic type are different across the TREC 2003 and TREC 2004 datasets. This means that there is a possible bias problem, especially on the TD topics if we train on the TREC 2003 dataset and test on the TREC 2004 queries. In order to avoid this bias problem and assess our proposed method on a big enough training and test datasets, we mix the TREC 2003 and TREC 2004 Web Track topics and relevance assessment sets, respectively. We use a 3-fold cross-validation process by separating the mixed datasets into three folds of equal size, each fold contains 41 Topic Distillation topics, 75 Homepage finding topics and 75 Named Page finding topics. We iteratively test our feature selection method on one fold after training on the remaining two folds.

For indexing and retrieval, we use the Terrier IR platform[1] [13], and apply standard stopwords removal. In addition, to boost early precision, we apply the first two steps of the 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 DFR document weighting model [10], as described in Section 2. We experiment with the three query-independent features introduced in Section 3, namely PageRank, URL depth and Click Distance. While for obvious length constraints, this paper concentrates on the aforementioned three features, it is straightforward to expand the work using another set of features.

The evaluation measure used in all our experiments is the mean average precision (MAP). The parameters that are related with the PL2F document weighting model and the FLOE methods are set by optimising MAP on the training dataset, using a simulated annealing procedure [7]. We use $FLOE1$ for PageRank and $FLOE2$ for the URL depth and Click Distance. The number of the top retrieved documents, namely n in Equation (9), and the number of bins in Section 4.2 are also set by optimising MAP over the training dataset, using a large range of different value settings. For the Click Distance feature, we use `firstgov.gov` as the root. The maximum Click Distance is 46 in the .GOV collection. For those documents that cannot be reached from the root, we assume a Click Distance of 47.

In our experiments, we mainly conduct four different kinds of evaluations:

- Firstly, we assess how important it is to selectively apply a query-independent feature on a per-query basis in Web IR.
- Secondly, we test how effective our proposed method is for selectively applying one query-independent feature out of two candidate features.
- Thirdly, as the number of candidate features increases, the selective application becomes more challenging. We further investigate how effective our proposed method is for selectively applying a query-independent feature when there are more than two candidate features.
- Finally, as described in the introduction (see Section 1), we use the QTP method as an alternative baseline approach to apply the most appropriate query-independent feature. In order to compare our proposed method to a strong QTP method, we simulate an optimal 100% accuracy for this method, meaning that the simulated QTP method knows with certainty the query type before applying a query-independent feature.

We report the obtained results, and their analysis in the next section.

6 Discussion

Table 3 provides the MAP upper bounds that can be achieved by manually and selectively applying a query-independent feature on a per-query basis, first when there are two possible candidate features (columns 6-8), and second when we use

[1] `http://ir.dcs.gla.ac.uk/terrier`

Table 3. The MAP upper bounds, highlighted in bold, which are achieved by the manual selective application of query-independent features on each test fold

	MAP							
	PL2F	+PR	+UD	+CD	+(PR|UD)	+(PR|CD)	+(UD|CD)	+(PR|UD|CD)
Fold 1	0.6113	0.6430	0.6399	0.6284	**0.6887**$*\diamond\star$	**0.6721**$*\diamond\bullet$	**0.6745**$*\star\bullet$	**0.6992**$*\diamond\star\bullet$
Fold 2	0.5488	0.5802	0.5740	0.5668	**0.6250**$*\diamond\star$	**0.6146**$*\diamond\bullet$	**0.6150**$*\star\bullet$	**0.6436**$*\diamond\star\bullet$
Fold 3	0.5587	0.5792	0.5858	0.5806	**0.6221**$*\diamond\star$	**0.6049**$*\diamond\bullet$	**0.6154**$*\star\bullet$	**0.6322**$*\diamond\star\bullet$

all three features (column 9). In each row, values that are statistically different from PL2F, PL2F+PR, PL2F+UD and PL2F+CD are marked with $*$, $\diamond$, $\star$ and $\bullet$, respectively (Wilcoxon Matched-Pairs Signed-Ranks Test, $p < 0.05$). Tables 4 & 5 show the MAP obtained by applying our proposed selective application method when there are two and more than two candidate features, respectively. The best retrieval performance in each row is highlighted in bold. The symbol † denotes that our approach applies the most appropriate query-independent feature for a statistically significant number of queries, according to the Sign Test ($p < 0.05$). The symbol $*$ denotes that the MAP obtained by using our method is statistically better than the one achieved by the PL2F baseline, as well as all the systems where a query-independent feature has been uniformly applied to all queries, according to the Wilcoxon Matched-Pairs Signed-Ranks Test ($p < 0.05$). Table 6 shows the comparison between our proposed method and the QTP method. The best retrieval performance and the highest prediction accuracy in each row is highlighted in bold and in italic, respectively. In Tables 4 - 6, *Number* reports the number of queries for which the selected query-independent feature has been correctly applied (denoted *Pos.*), using the manual upper bound approach as a ground truth. Conversely, the column *Neg.* reports the number of queries for which the system has failed to apply the most appropriate feature. The column *Neu.* reports the number of queries where all query-independent features produced the same MAP.

Firstly, we assess how important it is to selectively apply a query-independent feature on a per-query basis in Web IR, by estimating the upper bounds performances of the selective application method. This allows to estimate the extent to which it is indeed possible to enhance the retrieval performance of a Web IR system when the most appropriate query-independent feature is applied on a per-query basis. From Table 3, it is clear that using a manual selective method leads to significant increases in performances compared to the PL2F baseline as well as systems where a query-independent feature was applied uniformly to all queries. We also observe that the upper bounds of the selective application among three query-independent features are markedly higher than the selective application between any two of them, although not significantly so. This suggests that the selective application of a query-independent feature on a per-query basis is very important for a Web IR system, and that the retrieval performance could be further improved when the number of query-independent features increases.

Secondly, we test how effective our proposed automatic method is for selectively applying a query-independent feature when there are two candidate

Table 4. Evaluation of our automatic selective application between two query-independent features

Selective Application between PR and UD							
	MAP				Number		
	PL2F	PL2F+PR	PL2F+UD	Selective	Pos.	Neg.	Neu.
Fold 1	0.6113	0.6430	0.6399	**0.6641**†∗	65	28	98
Fold 2	0.5488	0.5802	0.5740	**0.5979**†∗	63	38	90
Fold 3	0.5587	0.5792	0.5858	**0.6049**†∗	69	29	93
Selective Application between PR and CD							
	MAP				Number		
	PL2F	PL2F+PR	PL2F+CD	Selective	Pos.	Neg.	Neu.
Fold 1	0.6113	0.6430	0.6284	**0.6515**†	54	34	103
Fold 2	0.5488	0.5802	0.5668	**0.5914**†	67	38	86
Fold 3	0.5587	0.5792	0.5806	**0.5911**†	58	28	105
Selective Application between UD and CD							
	MAP				Number		
	PL2F	PL2F+UD	PL2F+CD	Selective	Pos.	Neg.	Neu.
Fold 1	0.6113	0.6399	0.6284	**0.6477**†	56	33	102
Fold 2	0.5488	0.5740	0.5668	**0.5875**†	69	36	86
Fold 3	0.5587	0.5858	0.5806	**0.5994**†	64	28	99

features. We compare our proposed method to the PL2F baseline, as well as the method that applies a query-independent feature uniformly to all queries. From Table 4, we can see that, for the three different combinations, namely PR|UD, PR|CD and UD|CD, our proposed approach can always markedly improve the PL2F baseline and that of the systems where a query-independent feature is uniformly applied. In particular, for the selective application between PageRank and URL depth, the improvement is constantly statistically significant on each fold. Moreover, we also observe that a statistically significant number of queries have been applied with the most appropriate query-independent feature on all possible combinations and on all folds. This suggests that our proposed approach is an effective method for selecting the most appropriate feature from any two candidate features.

Thirdly, as the number of candidate features increases, the selective application method raises more challenges. We further investigate how effective our proposed method is for selectively applying the most appropriate query-independent feature when there are more than two candidate features. In particular, we select the most appropriate query-independent feature out of the three used PR, UD, and CD features. The evaluation results from Table 5 show that our approach can constantly make a significant improvement over PL2F and that of the systems where a query-independent feature was uniformly applied. The observation is upheld on each fold. Moreover, we also observe that a statistically significant number of queries have been applied with the most appropriate query-independent feature on all folds. In addition, comparing the best MAP results that can be obtained in each fold in Tables 4 & 5, we can see that the retrieval performance obtained by using our

Table 5. Evaluation of our automatic selective application among more than two query-independent features

Selective Application among PR, UD and CD								
	MAP					Number		
	PL2F	PL2F+PR	PL2F+UD	PL2F+CD	Selective	Pos.	Neg.	Neu.
Fold 1	0.6113	0.6430	0.6399	0.6284	**0.6653**†∗	61	38	92
Fold 2	0.5488	0.5802	0.5740	0.5668	**0.5994**†∗	68	45	78
Fold 3	0.5587	0.5792	0.5858	0.5806	**0.6128**†∗	65	37	89

Table 6. Comparison between our proposed method and the QTP method

Selective Application among PR, UD and CD										
	Our Proposed Method					The QTP Method				
	Pos.	Neg.	Neu.	Accuracy	MAP	Pos.	Neg.	Neu.	Accuracy	MAP
Fold 1	61	38	92	*61.6%*	**0.6653**	55	44	92	55.6%	0.6588
Fold 2	68	45	78	*60.2%*	**0.5994**	67	46	78	59.3%	0.5967
Fold 3	65	37	89	*63.7%*	**0.6128**	63	39	89	61.8%	0.6077

proposed approach can be further improved when there are more than two candidate query-independent features. This is encouraging, as this suggests that our proposed automatic approach remains effective and robust even when the number of candidate features increases. Overall, while the results obtained in Tables 4 & 5 are naturally lower than the upper bounds performances in Table 3, they are nevertheless roughly reasonably comparable.

Finally, as mentioned in Section 1, we use the alternative QTP method, to apply the most appropriate query-independent feature on a per-query basis. We train the QTP method using the same training procedure described in Section 5, by identifying the most effective feature for a given query type. We compare our proposed method to the optimal QTP method, by simulating an ideal 100% accuracy in detecting the query type. From Table 6, we can see that our proposed method constantly outperforms the QTP method in both accuracy and MAP on all folds. This particularly stresses the effectiveness and robustness of our approach compared to the QTP method, given that the query type prediction in a practical system is usually much lower than 100% (See Section 1). It also suggests that queries which have the same type do not necessarily equally benefit from the application of a given query-independent feature since the MAP value obtained from the QTP method is not equal to the value of the upper bound on each fold, even though the accuracy of the query type prediction is simulated equal to 100%.

7 Conclusions

In this paper, we have proposed a novel method for the selective application of a query-independent feature on a per-query basis. We have tested our proposed

approach on the TREC .GOV Web test collection and the mixed topic sets from the TREC 2003 & 2004 Web Tracks.

We have obtained very encouraging experimental results. First, we showed that the retrieval performance can be significantly improved by an optimal selective application of a query-independent feature. This indicates that the selective application of the query-independent feature on a per-query basis can indeed significantly enhance the retrieval performance of a Web IR system.

Second, using our proposed automatic selective application method, and any two query-independent features, we observed that the most appropriate feature has been applied for a statistically significant number of queries. The improvement in MAP was statistically significant when the selective application occurred using PageRank and URL depth.

Third, as the number of candidate features increases, the selective application raises more challenges. Therefore, we further investigated how effective our proposed method is for selectively applying the most appropriate query-independent feature when there are more than two candidate features. The experimental results showed that our proposed approach can constantly make a significant improvement in MAP over a strong field-based document ranking model, as well as that of the systems where a query-independent feature was uniformly applied. We also observed that the most appropriate query-independent feature has been applied in a statistically significant number of queries.

Finally, we compared our proposed method to a simulated QTP method, which has an ideal 100% accuracy on the query type prediction. We observed that our proposed method constantly outperforms the QTP method in all folds. This suggests that our proposed selective application approach is effective and robust.

Acknowledgements. We thank Craig Macdonald & Ben He for their helpful comments and feedback on the paper.

References

1. Brin, S., Page, L.: The anatomy of a large-scale hypertextual Web search engine. In: Proceedings of WWW 1998, Amsterdam, The Netherlands (1998)
2. Cai, D., He, X.F., Wen, J.R., Ma, W.Y.: Block-level Link Analysis. In: Proceedings of SIGIR 2004, Sheffield, United Kingdom (2004)
3. Craswell, N., Hawking, D.: Overview of the TREC 2002 Web Track. In: Proceedings of TREC 2002, USA (2002)
4. Craswell, N., Hawking, D.: Overview of the TREC 2004 Web Track. In: Proceedings of TREC 2004, USA (2002)
5. Craswell, N., Robertson, S., Zaragoza, H., Taylor, M.: Relevance Weighting for Query Independent Evidence. In: Proceedings of SIGIR 2005, Salvador, Brazil (2005)
6. Kamps, J., Mishne, G., de Rijke, M.: Language Models for Searching in Web Corpora. In: Proceedings of TREC 2004, USA (2004)
7. Kirkpatrick, S., Gelatt, C., Vecchi, M.: Optimization by simulated annealing. Science 220(4598) (1983)

8. Kraaij, W., Westerveld, T., Hiemstra, D.: The Importance of Prior Probabilities for Entry Page Search. In: Proceedings of SIGIR 2002, Tampere, Finland (2002)
9. Lin, J.: Divergence Measures Based on the Shannon Entropy. IEEE Transactions on Information Theory, 37 (1991)
10. Macdonald, C., Plachouras, V., He, B., Lioma, C., Ounis, I.: University of glasgow at webCLEF 2005: Experiments in per-field normalisation and language specific stemming. In: Peters, C., Gey, F.C., Gonzalo, J., Müller, H., Jones, G.J.F., Kluck, M., Magnini, B., de Rijke, M., Giampiccolo, D. (eds.) CLEF 2005. LNCS, vol. 4022, pp. 898–907. Springer, Heidelberg (2006)
11. Manmatha, R., Rath, T., Feng, F.: Modeling Score Distributions for Combining the Outputs of Search Engines. In: Proceedings of SIGIR 2001, USA (2001)
12. Metzler, D., Strohman, T., Zhou, Y., Croft, W.B.: Indri at TREC 2005: Terabyte Track. In: Proceedings of TREC 2005, USA (2005)
13. Ounis, I., Amati, G., Plachouras, V., He, B., Macdonald, C., Lioma, C.: Terrier: A High Performance and Scalable Information Retrieval Platform. In: Proceedings of OSIR 2006, Seattle, USA (2006)
14. Plachouras, V.: Selective Web Information Retrieval. PhD thesis, Univesity of Glasgow (2006)
15. Plachouras, V., Ounis, I.: Multinomial randomness models for retrieval with document fields. In: Amati, G., Carpineto, C., Romano, G. (eds.) ECIR 2007. LNCS, vol. 4425, pp. 28–39. Springer, Heidelberg (2007)
16. Robertson, S., Zaragoza, H., Taylor, M.: Simple BM25 extension to multiple weighted fields. In: Proceedings of CIKM 2004, Washington DC, USA (2004)
17. Zaragoza, H., Craswell, N., Taylor, M., Saria, S., Robertson, S.: Microsoft Cambridge at TREC-13: Web and HARD tracks. In: Proceedings of TREC 2004, USA (2004)