Multinomial Randomness Models for Retrieval with Document Fields

Vassilis Plachouras[1] and Iadh Ounis[2]

[1] Yahoo! Research, Barcelona, Spain
[2] University of Glasgow, Glasgow, UK
vassilis@yahoo-inc.com, ounis@dcs.gla.ac.uk

Abstract. Document fields, such as the title or the headings of a document, offer a way to consider the structure of documents for retrieval. Most of the proposed approaches in the literature employ either a linear combination of scores assigned to different fields, or a linear combination of frequencies in the term frequency normalisation component. In the context of the Divergence From Randomness framework, we have a sound opportunity to integrate document fields in the probabilistic randomness model. This paper introduces novel probabilistic models for incorporating fields in the retrieval process using a multinomial randomness model and its information theoretic approximation. The evaluation results from experiments conducted with a standard TREC Web test collection show that the proposed models perform as well as a state-of-the-art field-based weighting model, while at the same time, they are theoretically founded and more extensible than current field-based models.

1 Introduction

Document fields provide a way to incorporate the structure of a document in Information Retrieval (IR) models. In the context of HTML documents, the document fields may correspond to the contents of particular HTML tags, such as the title, or the heading tags. The anchor text of the incoming hyperlinks can also be seen as a document field. In the case of email documents, the fields may correspond to the contents of the email's subject, date, or to the email address of the sender [9]. It has been shown that using document fields for Web retrieval improves the retrieval effectiveness [17,7].

The text and the distribution of terms in a particular field depend on the function of that field. For example, the title field provides a concise and short description for the whole document, and terms are likely to appear once or twice in a given title [6]. The anchor text field also provides a concise description of the document, but the number of terms depends on the number of incoming hyperlinks of the document. In addition, anchor texts are not always written by the author of a document, and hence, they may enrich the document representation with alternative terms.

The combination of evidence from the different fields in a retrieval model requires special attention. Robertson et al. [14] pointed out that the linear combination of scores, which has been the approach mostly used for the combination of fields, is difficult to interpret due to the non-linear relation between the assigned scores and the term frequencies in each of the fields. Hawking et al. [5] showed that the term frequency

G. Amati, C. Carpineto, and G. Romano (Eds.): ECIR 2007, LNCS 4425, pp. 28–39, 2007.

normalisation applied to each field depends on the nature of the corresponding field. Zaragoza et al. [17] introduced a field-based version of BM25, called BM25F, which applies term frequency normalisation and weighting of the fields independently. Macdonald et al. [7] also introduced *normalisation 2F* in the Divergence From Randomness (DFR) framework [1] for performing independent term frequency normalisation and weighting of fields. In both cases of BM25F and the DFR models that employ normalisation 2F, there is the assumption that the occurrences of terms in the fields follow the same distribution, because the combination of fields takes place in the term frequency normalisation component, and not in the probabilistic weighting model.

In this work, we introduce weighting models, where the combination of evidence from the different fields does not take place in the term frequency normalisation part of the model, but instead, it constitutes an integral part of the probabilistic randomness model. We propose two DFR weighting models that combine the evidence from the different fields using a multinomial distribution, and its information theoretic approximation. We evaluate the performance of the introduced weighting models using the standard .Gov TREC Web test collection. We show that the models perform as well as the state-of-the-art model field-based PL2F, while at the same time, they employ a theoretically founded and more extensible combination of evidence from fields.

The remainder of this paper is structured as follows. Section 2 provides a description of the DFR framework, as well as the related field-based weighting models. Section 3 introduces the proposed multinomial DFR weighting models. Section 4 presents the evaluation of the proposed weighting models with a standard Web test collection. Sections 5 and 6 close the paper with a discussion related to the proposed models and the obtained results, and some concluding remarks drawn from this work, respectively.

2 Divergence from Randomness Framework and Document Fields

The Divergence From Randomness (DFR) framework [1] generates a family of probabilistic weighting models for IR. It provides a great extent of flexibility in the sense that the generated models are modular, allowing for the evaluation of new assumptions in a principled way. The remainder of this section provides a description of the DFR framework (Section 2.1), as well as a brief description of the combination of evidence from different document fields in the context of the DFR framework (Section 2.2).

2.1 DFR Models

The weighting models of the Divergence From Randomness framework are based on combinations of three components: a randomness model $\mathcal{RM}$; an information gain model $\mathcal{GM}$; and a term frequency normalisation model.

Given a collection D of documents, the randomness model $\mathcal{RM}$ estimates the probability $P_{\mathcal{RM}}(t \in d|D)$ of having tf occurrences of a term t in a document d, and the importance of t in d corresponds to the informative content $-\log_2(P_{\mathcal{RM}}(t \in d|D))$. Assuming that the sampling of terms corresponds to a sequence of independent Bernoulli trials, the randomness model $\mathcal{RM}$ is the binomial distribution:

$$P_{\mathcal{B}}(t \in d|D) = \binom{TF}{tf} p^{tf}(1-p)^{TF-tf} \tag{1}$$

where TF is the frequency of t in the collection D, $p = \frac{1}{N}$ is a uniform prior probability that the term t appears in the document d, and N is the number of documents in the collection D. A limiting form of the binomial distribution is the Poisson distribution $\mathcal{P}$:

$$P_{\mathcal{B}}(t \in d|D) \approx P_{\mathcal{P}}(t \in d|D) = \frac{\lambda^{tf}}{tf!}e^{-\lambda} \quad \text{where} \quad \lambda = TF \cdot p = \frac{TF}{N} \tag{2}$$

The information gain model $\mathcal{GM}$ estimates the informative content $1 - P_{risk}$ of the probability P_{risk} that a term t is a good descriptor for a document. When a term t appears many times in a document, then there is very low risk in assuming that t describes the document. The information gain, however, from any future occurrences of t in d is lower. For example, the term 'evaluation' is likely to have a high frequency in a document about the evaluation of IR systems. After the first few occurrences of the term, however, each additional occurrence of the term 'evaluation' provides a diminishing additional amount of information. One model to compute the probability P_{risk} is the Laplace after-effect model:

$$P_{risk} = \frac{tf}{tf + 1} \tag{3}$$

P_{risk} estimates the probability of having one more occurrence of a term in a document, after having seen tf occurrences already.

The third component of the DFR framework is the term frequency normalisation model, which adjusts the frequency tf of the term t in d, given the length l of d and the average document length $\overline{l}$ in D. *Normalisation 2* assumes a decreasing density function of the normalised term frequency with respect to the document length l. The normalised term frequency tfn is given as follows:

$$tfn = tf \cdot \log_2(1 + c \cdot \frac{\overline{l}}{l}) \tag{4}$$

where c is a hyperparameter, i.e. a tunable parameter. Normalisation 2 is employed in the framework by replacing tf in Equations (2) and (3) with tfn.

The relevance score $w_{d,q}$ of a document d for a query q is given by:

$$w_{d,q} = \sum_{t \in q} qtw \cdot w_{d,t} \quad \text{where} \quad w_{d,t} = (1 - P_{risk}) \cdot (-\log_2 P_{\mathcal{RM}}) \tag{5}$$

where $w_{d,t}$ is the weight of the term t in document d, $qtw = \frac{qtf}{qtf_{max}}$, qtf is the frequency of t in the query q, and qtf_{max} is the maximum qtf in q. If $P_{\mathcal{RM}}$ is estimated using the Poisson randomness model, P_{risk} is estimated using the Laplace after-effect model, and tfn is computed according to normalisation 2, then the resulting weighting model is denoted by PL2. The factorial is approximated using Stirling's formula: $tf! = \sqrt{2\pi} \cdot tf^{tf+0.5}e^{-tf}$.

The DFR framework generates a wide range of weighting models by using different randomness models, information gain models, or term frequency normalisation models. For example, the next section describes how normalisation 2 is extended to handle the normalisation and weighting of term frequencies for different document fields.

2.2 DFR Models for Document Fields

The DFR framework has been extended to handle multiple document fields, and to apply per-field term frequency normalisation and weighting. This is achieved by extending normalisation 2, and introducing *normalisation 2F* [7], which is explained below.

Suppose that a document has k fields. Each occurrence of a term can be assigned to exactly one field. The frequency tf_i of term t in the i-th field is normalised and weighted independently of the other fields. Then, the normalised and weighted term frequencies are combined into one pseudo-frequency tfn_{2F}:

$$tfn_{2F} = \sum_{i=1}^{k} w_i \cdot tf_i \log_2 \left(1 + c_i \cdot \frac{\overline{l_i}}{l_i}\right) \tag{6}$$

where w_i is the relative importance or weight of the i-th field, tf_i is the frequency of t in the i-th field of document d, l_i is the length of the i-th field in d, $\overline{l_i}$ is the average length of the i-th field in the collection D, and c_i is a hyperparameter for the i-th field. The above formula corresponds to normalisation 2F. The weighting model PL2F corresponds to PL2 using tfn_{2F} as given in Equation (6). The well-known BM25 weighting model has also been extended in a similar way to BM25F [17].

3 Multinomial Randomness Models

This section introduces DFR models which, instead of extending the term frequency normalisation component, as described in the previous section, use document fields as part of the randomness model. While the weighting model PL2F has been shown to perform particularly well [7,8], the document fields are not an integral part of the randomness weighting model. Indeed, the combination of evidence from the different fields takes place as a linear combination of normalised frequencies in the term frequency normalisation component. This implies that the term frequencies are drawn from the same distribution, even though the nature of each field may be different.

We propose two weighting models, which, instead of assuming that term frequencies in fields are drawn from the same distribution, use multinomial distributions to incorporate document fields in a theoretically driven way. The first one is based on the multinomial distribution (Section 3.1), and the second one is based on an information theoretic approximation of the multinomial distribution (Section 3.2).

3.1 Multinomial Distribution

We employ the multinomial distribution to compute the probability that a term appears a given number of times in each of the fields of a document. The formula of the weighting model is derived as follows. Suppose that a document d has k fields. The probability that a term occurs tf_i times in the i-th field f_i, is given as follows:

$$P_{\mathcal{M}}(t \in d|D) = \binom{TF}{tf_1 \quad tf_2 \ldots tf_k \quad tf'} p_1^{tf_1} p_2^{tf_2} \ldots p_k^{tf_k} p'^{tf'} \tag{7}$$

In the above equation, TF is the frequency of term t in the collection, $p_i = \frac{1}{k \cdot N}$ is the prior probability that a term occurs in a particular field of document d, and N is the number of documents in the collection D. The frequency $tf' = TF - \sum_{i=1}^{k} tf_i$ corresponds to the number of occurrences of t in other documents than d. The probability $p' = 1 - k\frac{1}{k \cdot N} = \frac{N-1}{N}$ corresponds to the probability that t does not appear in any of the fields of d.

The DFR weighting model is generated using the multinomial distribution from Equation (7) as a randomness model, the Laplace after-effect from Equation (3), and replacing tf_i with the normalised term frequency tfn_i, obtained by applying normalisation 2 from Equation (4). The relevance score of a document d for a query q is computed as follows:

$$\begin{aligned} w_{d,q} &= \sum_{t \in q} qtw \cdot w_{d,t} = \sum_{t \in q} qtw \cdot (1 - P_{risk}) \cdot \Big(-\log_2(P_{\mathcal{M}}(t \in d|D)\Big) \\ &= \sum_{t \in q} \frac{qtw}{\sum_{i=1}^{k} tfn_i + 1} \cdot \Big(-\log_2(TF!) + \sum_{i=1}^{k} \Big(\log_2(tfn_i!) - tfn_i \log_2(p_i)\Big) \\ &\quad + \log_2(tfn'!) - tfn' \log_2(p')\Big) \end{aligned} \tag{8}$$

where qtw is the weight of a term t in query q, $tfn' = TF - \sum_{i=1}^{k} tfn_i$, $tfn_i = tf_i \cdot \log_2(1 + c_i \cdot \frac{\overline{l_i}}{l_i})$ for the i-th field, and c_i is the hyperparameter of normalisation 2 for the i-th field. The weighting model introduced in the above equation is denoted by ML2, where M stands for the multinomial randomness model, L stands for the Laplace after-effect model, and 2 stands for normalisation 2.

Before continuing, it is interesting to note two issues related to the introduced weighting model ML2, namely setting the relative importance, or weight, of fields in the document representation, and the computation of factorials.

Weights of fields. In Equation (8), there are two different ways to incorporate weights for the fields of documents. The first one is to multiply each of the normalised term frequencies tfn_i with a constant w_i, in a similar way to normalisation 2F (see Equation (6)): $tfn_i := w_i \cdot tfn_i$. The second way is to adjust the prior probabilities p_i of fields, in order to increase the scores assigned to terms occurring in fields with low prior probabilities: $p_i := \frac{p_i}{w_i}$. Indeed, the assigned score to a query term occurring in a field with low probability is high, due to the factor $-tfn_i \log_2(p_i)$ in Equation (8).

Computing factorials. As mentioned in Section 2.1, the factorial in the weighting model PL2 is approximated using Stirling's formula. A different method to approximate the factorial is to use the approximation of Lanczos to the Γ function [12, p. 213], which has a lower approximation error than Stirling's formula. Indeed, preliminary experimentation with ML2 has shown that using Stirling's formula affects the performance of the weighting model, due to the accumulation of the approximation error from computing the factorial $k + 2$ times (k is the number of fields). This is not the case for the Poisson-based weighting models PL2 and PL2F, where there is only one factorial computation for each query term (see Equation (2)). Hence, the computation of factorials in Equation (8) is performed using the approximation of Lanczos to the Γ function.

3.2 Approximation to the Multinomial Distribution

The DFR framework generates different models by replacing the binomial randomness model with its limiting forms, such as the Poisson randomness model. In this section, we introduce a new weighting model by replacing the multinomial randomness model in ML2 with the following information theoretic approximation [13]:

$$\frac{TF!}{tf_1!tf_2!\cdots tf_k!tf'!}{p_1}^{tf_1}{p_2}^{tf_2}\cdots {p_k}^{tf_k}{p'}^{tf'} \approx \frac{1}{\sqrt{2\pi TF}^{k}}\frac{2^{-TF\cdot D\left(\frac{tf_i}{TF},p_i\right)}}{\sqrt{p_{t1}p_{t2}\cdots p_{tk}p'_t}} \tag{9}$$

$D\left(\frac{tf_i}{TF},p_i\right)$ corresponds to the information theoretic divergence of the probability $p_{ti} = \frac{tf_i}{TF}$ that a term occurs in a field, from the prior probability p_i of the field:

$$D\left(\frac{tf_i}{TF},p_i\right) = \sum_{i=1}^{k}\left(\frac{tf_i}{TF}\log_2\frac{tf_i}{TF\cdot p_i}\right) + \frac{tf'}{TF}\log_2\frac{tf'}{TF\cdot p'} \tag{10}$$

where $tf' = TF - \sum_{i=1}^{k} tf_i$. Hence, the multinomial randomness model $\mathcal{M}$ in the weighting model ML2 can be replaced by its approximation from Equation (9):

$$w_{d,q} = \sum_{t\in q} qtw \cdot \frac{\frac{k}{2}\log_2(2\pi TF)}{\sum_{i=1}^{k} tfn_i + 1} \cdot \Bigg(\sum_{i=1}^{k}\left(tfn_i\log_2\frac{tfn_i/TF}{p_i} + \frac{1}{2}\log_2\frac{tfn_i}{TF}\right) + tfn'\log_2\frac{tfn'/TF}{p'} + \frac{1}{2}\log_2\frac{tfn'}{TF}\Bigg) \tag{11}$$

The above model is denoted by M_DL2. The definitions of the variables involved in the above equation have been introduced in Section 3.1.

It should be noted that the information theoretic divergence $D\left(\frac{tf_i}{TF},p_i\right)$ is defined only when $tf_i > 0$ for $1 \leq i \leq k$. In other words, $D\left(\frac{tf_i}{TF},p_i\right)$ is defined only when there is at least one occurrence of a query term in all the fields. This is not always the case, because a Web document may contain all the query terms in its body, but it may contain only some of the query terms in its title. To overcome this issue, the weight of a query term t in a document is computed by considering only the fields in which the term t appears.

The weights of different fields can be defined in the same way as in the case of the weighting model ML2, as described in Section 3.1. In more detail, the weighting of fields can be achieved by either multiplying the frequency of a term in a field by a constant, or by adjusting the prior probability of the corresponding field.

An advantage of the weighting model M$_\text{D}$L2 is that, because it approximates the multinomial distribution, there is no need to compute factorials. Hence, it is likely to provide a sufficiently accurate approximation to the multinomial distribution, and it may lead to improved retrieval effectiveness compared to ML2, due to the lower accumulated numerical errors. The experimental results in Section 4.2 will indeed confirm this advantage of M$_\text{D}$L2.

4 Experimental Evaluation

In this section, we evaluate the proposed multinomial DFR models ML2 and $\mathrm{M_D}$L2, and compare their performance to that of PL2F, which has been shown to be particularly effective [7,8]. A comparison of the retrieval effectiveness of PL2F and BM25F has shown that the two models perform equally well on various search tasks and test collections [11], including those employed in this work. Hence, we experiment only with the multinomial models and PL2F. Section 4.1 describes the experimental setting, and Section 4.2 presents the evaluation results.

4.1 Experimental Setting

The evaluation of the proposed models is conducted with the .Gov TREC Web test collection, a crawl of approximately 1.25 million documents from the .gov domain. The .Gov collection has been used in the TREC Web tracks between 2002 and 2004 [2,3,4]. In this work, we employ the tasks from the Web tracks of TREC 2003 and 2004, because they include both informational tasks, such as the topic distillation (td2003 and td2004, respectively), as well as navigational tasks, such as named page finding (np2003 and np2004, respectively) and home page finding (hp2003 and hp2004, respectively). More specifically, we train and test for each type of task independently, in order to get insight on the performance of the proposed models [15]. We employ each of the tasks from the TREC 2003 Web track for training the hyperparameters of the proposed models. Then, we evaluate the models on the corresponding tasks from the TREC 2004 Web track.

In the reported set of experiments, we employ $k = 3$ document fields: the contents of the <BODY> tag of Web documents (b), the anchor text associated with incoming hyperlinks (a), and the contents of the <TITLE> tag (t). More fields can be defined for other types of fields, such as the contents of the heading tags <H1> for example. It has been shown, however, that the body, title and anchor text fields are particularly effective for the considered search tasks [11]. The collection of documents is indexed after removing stopwords and applying Porter's stemming algorithm. We perform the experiments in this work using the Terrier IR platform [10].

The proposed models ML2 and $\mathrm{M_D}$L2, as well as PL2F, have a range of hyperparameters, the setting of which can affect the retrieval effectiveness. More specifically, all three weighting models have two hyperparameters for each employed document field: one related to the term frequency normalisation, and a second one related to the weight of that field. As described in Sections 3.1 and 3.2, there are two ways to define the weights of fields for the weighting models ML2 and $\mathrm{M_D}$L2: (i) multiplying the normalised frequency of a term in a field; (ii) adjusting the prior probability p_i of the i-th field. The field weights in the case of PL2F are only defined in terms of multiplying the normalised term frequency by a constant w_i, as shown in Equation (6).

In this work, we consider only the term frequency normalisation hyperparameters, and we set all the weights of fields to 1, in order to avoid having one extra parameter in the discussion of the performance of the weighting models. We set the involved hyperparameters c_b, c_a, and c_t, for the body, anchor text, and title fields, respectively, by directly optimising mean average precision (MAP) on the training tasks from the Web track of TREC 2003. We perform a 3-dimensional optimisation to set the values of the hyperparameters. The optimisation process is the following. Initially, we apply a simulated annealing algorithm, and then, we use the resulting hyperparameter values as a starting point for a second optimisation algorithm [16], to increase the likelihood of detecting a global maximum. For each of the three training tasks, we apply the above optimisation process three times, and we select the hyperparameter values that result in the highest MAP. We employ the above optimisation process to increase the likelihood that the hyperparameters values result in a global maximum for MAP. Figure 1 shows the MAP obtained by ML2 on the TREC 2003 home page finding topics, for each iteration of the optimisation process. Table 1 reports the hyperparameter values that resulted in the highest MAP for each of the training tasks, and that are used for the experiments in this work.

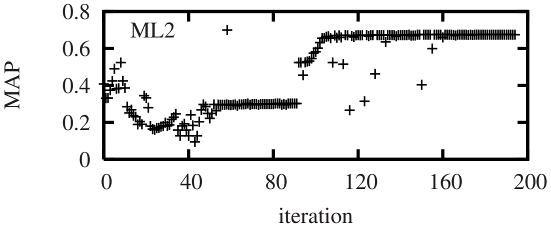

Fig. 1. The MAP obtained by ML2 on the TREC 2003 home page finding topics, during the optimisation of the term frequency normalisation hyperparameters

The evaluation results from the Web tracks of TREC 2003 [3] and 2004 [4] have shown that employing evidence from the URLs of Web documents results in important improvements in retrieval effectiveness for the topic distillation and home page finding tasks, where relevant documents are home pages of relevant Web sites. In order to provide a more complete evaluation of the proposed models for these two types of Web search tasks, we also employ the length in characters of the URL path, denoted by $URLpathlen$, using the following formula to transform it to a relevance score [17]:

$$w_{d,q} := w_{d,q} + \omega \cdot \frac{\kappa}{\kappa + URLpathlen} \tag{12}$$

where $w_{d,q}$ is the relevance score of a document. The parameters ω and κ are set by performing a 2-dimensional optimisation as described for the case of the hyperparameters c_i. The resulting values for ω and κ are shown in Table 2.

4.2 Evaluation Results

After setting the hyperparameter values of the proposed models, we evaluate the models with the search tasks from TREC 2004 Web track [4]. We report the official TREC evaluation measures for each search task: mean average precision (MAP) for the topic distillation task (td2004), and mean reciprocal rank (MRR) of the first correct answer for both named page finding (np2004) and home page finding (hp2004) tasks.

Table 1. The values of the hyperparameters c_b, c_a, and c_t, for the body, anchor text and title fields, respectively, which resulted in the highest MAP on the training tasks of TREC 2003 Web track

ML2			
Task	c_b	c_a	c_t
td2003	0.0738	4.3268	10.8220
np2003	0.1802	4.7057	8.4074
hp2003	0.1926	310.3289	624.3673
$\mathrm{M_DL2}$			
Task	c_b	c_a	c_t
td2003	0.2562	10.0383	24.6762
np2003	1.0216	9.2321	21.3330
hp2003	0.4093	355.2554	966.3637
PL2F			
Task	c_b	c_a	c_t
td2003	0.1400	5.0527	4.3749
np2003	1.0153	11.9652	9.1145
hp2003	0.2785	406.1059	414.7778

Table 2. The values of the hyperparameters ω and κ, which resulted in the highest MAP on the training topic distillation (td2003) and home page finding (hp2003) tasks of TREC 2003 Web track

ML2		
Task	ω	κ
td2003	8.8095	14.8852
hp2003	10.6684	9.8822
$\mathrm{M_DL2}$		
Task	ω	κ
td2003	7.6974	12.4616
hp2003	27.0678	67.3153
PL2F		
Task	ω	κ
td2003	7.3638	8.2178
hp2003	13.3476	28.3669

Table 3 presents the evaluation results for the proposed models ML2, $\mathrm{M_DL2}$, and the weighting model PL2F, as well as their combination with evidence from the URLs of documents (denoted by appending U to the weighting model's name). When only the document fields are employed, the multinomial weighting models have similar performance compared to the weighting model PL2F. The weighting models PL2F and $\mathrm{M_DL2}$ outperform ML2 for both topic distillation and home page finding tasks. For the named page finding task, ML2 results in higher MRR than $\mathrm{M_DL2}$ and PL2F.

Using the Wilcoxon signed rank test, we tested the significance of the differences in MAP and MRR between the proposed new multinomial models and PL2F. In the case of the topic distillation task td2004, PL2F and $\mathrm{M_DL2}$ were found to perform statistically significantly better than ML2, with $p < 0.001$ in both cases. There was no statistically significant difference between PL2F and $\mathrm{M_DL2}$. Regarding the named page finding task np2004, there is no statistically significant difference between any of the three proposed models. For the home page finding task hp2004, only the difference between ML2 and PL2F was found to be statistically significant ($p = 0.020$).

Regarding the combination of the weighting models with the evidence from the URLs of Web documents, Table 3 shows that PL2FU and $\mathrm{M_DL2U}$ outperform ML2U for td2004. The differences in performance are statistically significant, with $p = 0.002$ and $p = 0.012$, respectively, but there is no significant difference in the retrieval effectiveness between PL2FU and $\mathrm{M_DL2U}$. When considering hp2004, we can see that PL2F outperforms the multinomial weighting models. The only statistically significant difference in MRR was found between PL2FU and $\mathrm{M_DL2FU}$ ($p = 0.012$).

Table 3. Evaluation results for the weighting models ML2, M_DL2, and PL2F on the TREC 2004 Web track topic distillation (td2004), named page finding (np2004), and home page finding (hp2004) tasks. ML2U, M_DL2U, and PL2FU correspond to the combination of each weighting model with evidence from the URL of documents. The table reports mean average precision (MAP) for the topic distillation task, and mean reciprocal rank (MRR) of the first correct answer for the named page finding and home page finding tasks. ML2U, M_DL2U and PL2FU are evaluated only for td2004 and hp2004, where the relevant documents are home pages (see Section 4.1).

Task	ML2	M_DL2	PL2F
		MAP	
td2004	0.1241	0.1391	0.1390
		MRR	
np2004	0.6986	0.6856	0.6878
hp2004	0.6075	0.6213	0.6270
Task	ML2U	M_DL2U	PL2FU
		MAP	
td2004	0.1916	0.2012	0.2045
		MRR	
hp2004	0.6364	0.6220	0.6464

A comparison of the evaluation results with the best performing runs submitted to the Web track of TREC 2004 [4] shows that the combination of the proposed models with the evidence from the URLs performs better than the best performing run of the topic distillation task in TREC 2004, which achieved MAP 0.179. The performance of the proposed models is comparable to that of the most effective method for the named page finding task (MRR 0.731). Regarding the home page finding task, the difference is greater between the performance of the proposed models with evidence from the URLs, and the best performing methods in the same track (MRR 0.749). This can be explained in two ways. First, the over-fitting of the parameters ω and κ on the training task may result in lower performance for the test task. Second, using field weights may be more effective for the home page finding task, which is a high precision task, where the correct answers to the queries are documents of a very specific type.

From the results in Table 3, it can be seen that the model M_DL2, which employs the information theoretic approximation to the multinomial distribution, significantly outperforms the model ML2, which employs the multinomial distribution, for the topic distillation task. As discussed in Section 3.2, this may suggest that approximating the multinomial distribution is more effective than directly computing it, because of the number of computations involved, and the accumulated small approximation errors from the computation of the factorial. The difference in performance may be greater if more document fields are considered.

Overall, the evaluation results show that the proposed multinomial models ML2 and M_DL2 have a very similar performance to that of PL2F for the tested search tasks. None of the models outperforms the others consistently for all three tested tasks, and the weighting models M_DL2 and PL2F achieve similar levels of retrieval effectiveness. The next section discusses some points related to the new multinomial models.

5 Discussion

This section discusses (i) the advantages of the proposed multinomial models compared to the existing field-based weighting models, and (ii) the use of normalisation 2 (or normalisation 2F) for weighting fields in any of the field-based DFR weighting models.

The proposed models result in similar retrieval effectiveness to that of PL2F (Equation (6)), and also provide a new approach to the combination of evidence from the fields, compared to PL2F or BM25F [17], where a weighted sum aggregates term frequencies. Indeed, by employing multinomial distributions, the combination of fields takes place in the probabilistic weighting model. Hence, the weight of a term in a document depends explicitly on the distribution of term frequencies in the different fields.

A second advantage of the multinomial models over PL2F or BM25F is that they allow for a more principled approach to the weighting of fields, rather than just multiplying term frequencies by a constant. As suggested earlier, in the case of ML2 and $\mathrm{M_D}$L2, the prior probability of each field can be used as a weight for that field. The same approach cannot be applied to PL2F, because the randomness model does not consider document fields.

Normalisation 2 is primarily used for normalising the frequency of terms in a document, or in the document fields. In addition, it can also be used to weight the document fields, possibly avoiding the introduction of additional hyperparameters. Indeed, from the equation of normalisation 2: $tfn_i = tf_i \cdot \log_2 \left(1 + c_i \cdot (\overline{l_i}/l_i)\right)$, where $c_i \in (0, +\infty)$, it can be seen that applying a very high value for a particular document field, such as the title field, results in weak term frequency normalisation, and also multiplies the original term frequency. In this way, it may not be necessary to employ separate hyperparameters for field weights, thus reducing the imposed training overhead.

Overall, the proposed multinomial models offer a novel and effective way to combine document fields in a theoretically driven approach. Their introduction in the DFR framework can also generate a family of new weighting models, by combining different information gain or term frequency normalisation models.

6 Conclusions

In this work, we have introduced two new weighting models that combine document fields for Information Retrieval. While field-based weighting models, such as PL2F [7], or BM25F [17], combine evidence from fields in the term frequency normalisation component, we take a different approach. In the context of the DFR framework [1], we employ multinomial randomness models, and model the document fields in the probabilistic retrieval model. The first model, ML2, employs directly the multinomial distribution to assign a relevance score to documents, and the second model, $\mathrm{M_D}$L2, uses an information theoretic approximation of the multinomial distribution.

We have performed experiments in the context of the .Gov TREC Web test collection. The evaluation results show that the new models perform as well as PL2F for a range of Web search tasks, such as topic distillation, named page finding and home page finding. In particular, for the topic distillation task, the model $\mathrm{M_D}$L2 performs as well as PL2F, and significantly outperforms ML2, suggesting that it is more effective to approximate the multinomial distribution, than to compute it directly.

The proposed multinomial models represent a novel and effective approach to the combination of document fields, which is achieved in a principled way within a probabilistic framework. As a result, one of their advantages is that, for example, they allow for the investigation of the weighting of fields in terms of the prior probabilities of each field.

References

1. Amati, G., van Rijsbergen, C.J.: Probabilistic models of information retrieval based on measuring divergence from randomness. ACM TOIS **20** (2002) 357–389
2. Craswell, N., Hawking, D.: Overview of TREC-2002 web track. In: Proceedings of TREC-2002, Gaithersburg, MD, USA (2002)
3. Craswell, N., Hawking, D., Wilkinson, R., Wu, M.: Overview of the TREC-2003 web track. In: Proceedings of TREC-2003, Gaithersburg, MD, USA (2003)
4. Craswell, N., Hawking, D.: Overview of TREC-2004 web track. In: Proceedings of TREC-2004, Gaithersburg, MD, USA (2004)
5. Hawking, D., Upstill, T., Craswell, N.: Toward better weighting of anchors. In: Proceedings of the 27th annual international ACM SIGIR conference on Research and Development in Information Retrieval, ACM Press (2004) 512–513
6. Jin, R., Hauptmann, A.G., Zhai, C.X.: Title language model for information retrieval. In: Proceedings of the 25th annual international ACM SIGIR conference on Research and Development in Information Retrieval, ACM Press (2002) 42–48
7. 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: Proceedings of the Cross Language Evaluation Forum (CLEF) 2005. (2005)
8. Macdonald, C., He, B., Plachouras, V., Ounis, I.: University of Glasgow at TREC 2005: Experiments in Terabyte and Enterprise Tracks with Terrier. In: Proceedings of TREC-2005, Gaithersburg, Maryland USA (2005)
9. Macdonald, C., Ounis, I.: Combining fields in known-item email search. In: SIGIR ’06: Proceedings of the 29th annual international ACM SIGIR conference on Research and Development in Information Retrieval, ACM Press (2006) 675–676
10. Ounis, I., Amati, G., Plachouras, V., He, B., Macdonald, C., Lioma, C.: Terrier: A High Performance and Scalable Information Retrieval Platform. In: Proceedings of ACM SIGIR’06 Workshop on Open Source Information Retrieval (OSIR) (2006)
11. Plachouras, V.: Selective Web Information Retrieval. PhD thesis, Department of Computing Science, University of Glasgow (2006)
12. Press, W., Teukolsky, S., Vetterling, W., Flannery, B.: Numerical Recipes in C: The Art of Scientific Computing. Cambridge University Press (1992)
13. Rényi, A.: Foundations of probability. Holden-Day (1970)
14. Robertson, S., Zaragoza, H., Taylor, M.: Simple BM25 extension to multiple weighted fields. In: Proceedings of the 13th ACM Conference on Information and Knowledge Management (CIKM’04), ACM Press (2004) 42–49
15. Soboroff, I.: On evaluating web search with very few relevant documents. In: Proceedings of the 27th annual international ACM SIGIR conference on Research and Development in Information Retrieval, ACM Press (2004) 530–531
16. Yuret, D.: From Genetic Algorithms To Efficient Optimization. Master Thesis, MIT, A.I. Technical Report No. 1569. (1994)
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, Gaithersburg, MD, USA (2004)