Best and Fairest: An Empirical Analysis of Retrieval System Bias

Colin Wilkie, Leif Azzopardi

School of Computing Science, University of Glasgow,
Scotland, UK
{colin.wilkie,leif.azzopardi}@glasgow.ac.uk

Abstract. In this paper, we explore the bias of term weighting schemes used by retrieval models. Here, we consider bias as the extent to which a retrieval model unduly favours certain documents over others because of characteristics within and about the document. We set out to find the least biased retrieval model/weighting. This is largely motivated by the recent proposal of a new suite of retrieval models based on the Divergence From Independence (DFI) framework. The claim is that such models provide the fairest term weighting because they do not make assumptions about the term distribution (unlike most other retrieval models). In this paper, we empirically examine whether fairness is linked to performance and answer the question; is fairer better?

1 Introduction

Retrieval bias in retrieval systems and models has been a long standing problem when developing term weighting schemes [3, 4, 21]. It occurs when a document or group of documents are overly or unduly favoured due to some document feature (such as length, term distribution, etc). Bias is largely seen as problematic, especially, when a retrieval model's weighting function places too much emphasis on certain features. This leads to documents being ranked highly, not because they are relevant, but because of the bias present in the retrieval model's term weighting scheme. For example in [21], pivoted length normalisation was introduced to avoid overly favouring longer documents in the vector space model [19]. In [8], a study of IR heuristics showed how various models violated common sense principles relating to term weighting schemes. By addressing these violations, improvements in performance were obtained. Again linking bias with performance. In this work, instead of directly focusing on performance or looking at whether particular heuristics are violated, we will directly measure the bias associated with different retrieval models and then examine the relationship with performance.

To this end, we conduct a comprehensive empirical analysis across four TREC test collections using seventeen term weighting functions from six families of retrieval models (vector space models, best match models, language models, DFR, Log-Logistic and DFI models) and issue over 80 million queries in the process. We show that there is generally quite a strong correlation between fairness and performance, although this is collection dependent. In the next

section, we shall provide an historical overview of the development of retrieval models. Following this, we detail the experiment methodology we employed. We then report our results and discuss the implications of our findings.

2 Retrieval Models and Term Weighting Schemes

The development of retrieval models has been largely driven by the focus on improving the effectiveness of the retrieval system. Over the years numerous retrieval models have been developed, each leading to subtle and not so subtle differences in the term weighting schemes. For example in [20] Salton tried over 1400 term weighting schemes before arriving at TF.IDF. While reasonably effective, it was later shown that TF.IDF overly favoured the retrieval of long documents and that this bias is detrimental to retrieval performance [21]. Since then, document length normalisation has been a key component in most modern term weighting schemes as a way to mitigate such bias.

Conversely, probabilistic approaches to the term weighting problem have looked towards modelling different aspects of documents, including their term distributions and the distributions of relevance depending on the model in order to obtain a more accurate fit of the data [1, 6, 9, 11, 13, 16, 17]. For example, Harter [11] proposed the 2-Poisson model to represent speciality words and non-speciality words, where speciality words are those which occur densely in elite documents. These speciality words contribute to how informative a document is, while non-speciality words essentially occur at random and do not contribute to how informative a document is. While the 2-Poisson model worked well for particular samples of text, it did not perform very well in practice as it was problematic to estimate the distributions accurately. However, it did lead to the development of more generalised probabilistic models, which became known as Best Match (BM) models [13, 18], from which BM25 has become the most commonly used. Again, document length normalisation was a key feature in ensuring that BM25 did not overly favour shorter or longer documents.

In [17], Ponte and Croft proposed the Language Modelling approach which took an alternative view on estimating the relevance of a document. Language Models are based on generative probabilistic models, and look to estimate the probability of a query, given a document [17]. Under this framework, a mutlinomial distribution is typically used to model the document [23]. This free-er form means that the document data can be fitted more accurately. When the document model is estimated, document statistics are combined/smoothed by the background collection's statistics. Terms appearing in the document are akin to speciality words, while terms in the collection are akin to the non-speciality words that occur at random. Depending on how the document model is smoothed determines what features of a document dominate the term weighting function, and whether any document length normalisation is applied [23]. For instance, Jelinek-Mercer (JM) smoothing does not include any document length normalisation, focusing mainly on the information content of the document to rank documents[1], whereas Bayes smoothing includes document length normalisation.

[1] In [12], it was shown that Language Modelling with Jelinek-Mercer smoothing provides a probabilistic justification for TF.IDF.

In [1], Amati and van Rijsbergen proposed the Divergence from Randomness (DFR) framework to construct probabilistic term weighting schemes. The framework was also inspired by Harter's 2-Poisson model [11] and further refined the notion of eliteness using semantic information theory and Popper's notion of information content. A DFR model is typically composed of two divergence functions (referred to as $Prob_1$ to characterise the randomness of a term, and $Prob_2$ to model the risk of using the term as a document descriptor [1]) and a normalisation function. Rather than using two Poisson functions, in [1] the authors explored various distributions to characterise the randomness and eliteness of terms. Two of the best performing models they found were the Poisson-Laplace with their second normalisation of term frequency (PL2) and a hyper-geometric model that used Popper's normalisation (DPH). It was argued that the better fit to the underlying term distributions with the document collection, the use of information content and the normalisation of term frequency resulted in superior term weighting schemes (these models have been shown to perform empirically well at subsequent TRECs). It is also worth noting that under the DFR framework it is also possible to instantiate BM25.

In [5], another statistical model was proposed by Clinchant and Gaussier. This model attempted to account for the burstiness of terms within text (and thus obtain a better fit of the data). The term weighting function used a log-logistic distribution to provide a simpler information-based weighting as an alternative way to represent the eliteness of terms. The log-logistic distribution (LGD) model bridges Language Models and the Divergence from Randomness models.

More recently, the Divergence from Independence (DFI) framework has been proposed by Kocabas et al [14]. The DFI framework is most related to the DFR framework where the statistical independence takes place of the randomness. Rather than treating non-speciality words as random, DFI characterises them by their independence i.e. does this term occur independently of this document, or not? This alternative viewpoint means that DFI is essentially the non-parametric counterpart of DFR. No assumptions are made *a priori* about the distribution of underlying data. Therefore, if the data does not fit the prescribed distribution (i.e. Poisson or Laplace) then the non-parametric approach should work better, but even if it does, the non-parametric approach should work just as well. These models are also parameter free - and do not require tuning. In [14], the authors proposed three variants based on different measures of independence (the saturated model of independence, the standardisation model, and normalised Chi-Square distance). In [7], it was shown that the term weighting scheme derived from these different methods performed empirically well (and often the best at TREC). Since these schemes make no assumptions about the distribution of the data, it is contested that they are fairer than other models, and thus better. In this work, we shall test this claim.

To summarise, we have described a range of models that are related but make different assumptions about how to model a term's relevance. Most models are composed of two main parts: one to estimate the value of the information content

of the term, and one to regulate the influence of the document's length. Overly focusing on one part or another, or ignoring one part invariably leads to some form of bias creeping into the retrieval model.

3 Experimental Method

The focus of this study is to assess the level of bias exhibited by each retrieval model/weighting, and then to determine whether there is any relationship between the level of bias and retrieval performance. Specifically, we wish to determine whether the recently proposed DFI models are fairer and better, as they claim to be. To this end, we employed the following methodology. On each test collection, and for each retrieval model/weighting, we measured the bias using retrievability measures, and then measured the corresponding performance using the topics and relevance judgements associated with the collection. Below we shall first outline the collections and topics used, before describing how we estimated the bias and set of retrieval models/weightings we used.

3.1 Data and Materials

For our experiments we report results from four TREC collections: TREC 123 (T123), Aquaint (AQ), WT10G (WT) and DotGov (DG)[2]. These collections are typical of the collections used to test the models outlined in Section 2, where the focus is on creating the best term weighting scheme. These collections, while sufficient, are not so large that it is not possible to estimate the bias associated with each collection given each term weighting scheme (and parameter setting). Table 1 shows the collections used along with their statistics. To report performance we used a number of standard TREC measures: MAP, P@10, NDCG@100 and Mean Reciprocal Rank.

		Collection			
		AQ	**TREC 123**	**DG**	**WT**
TREC Topics		301-400	1-200	551-600	451-550
Number of Documents		1,033,461	1,078,166	1,247,753	1,692,096
Number of Queries		273,245	237,810	337,275	212,201
Avg. Doc. Length	**All**	439	420	1108	617
	Pool	623*	3913*	2056*	6737*
	Relevant	583*	1280*	2175*	2903*
Avg. Doc. Info. Cont.	**All**	2.114	2.194	2.498	2.721
	Pool	2.059*	2.115*	2.183*	2.626*
	Relevant	2.00*	2.120*	2.082*	2.475*

Table 1. Summary of each Collection Statistics. * denotes whether the difference from the whole collection is significant at $p < 0.05$.

3.2 Measuring Bias

The retrievability of a document provides an indication of how easily or likely a document is to be retrieved, given a particular configuration of a retrieval system [3]. Intuitively, if a retrieval system has a bias towards longer documents,

[2] We also conducted these experiments on AP and TREC678 where we found similar results and trends.

then we would expect that the retrievability of longer documents to be higher than shorter documents, and vice versa. If a retrieval system tends to retrieve documents with a higher information content, then we would expect to see that the retrievability of such documents would be higher, and vice versa. Formally, the retrievability $r(d)$ of a document d is defined by: $r(d) \propto \sum_{q \in Q} f(k_{dq}, c)$ where q is a query from a large set of queries Q, and k_{dq} is the rank of document d for query q. While there are various measures of retrievability, the simplest is the cumulative based measure which defines $f(k_{dq}, c)$ to be equal to 1 if the document d is retrieved in the top c results for the query q. The retrievability of a document is essentially the count of how many times the document is retrieved in the top c results. The bias that the retrieval model exhibits on the collection can be summarised by using the Gini Coefficient [10], which is commonly used to measure the inequality with a population (usually the wealth of people in a population). In the context of retrievability, if all documents were equally retrievable (i.e the retrieval function fairly retrieve all documents) then the Gini Coefficient would be 0 (denoting equality within the population). Conversely if only one document was retrievable and the rest were not, then the Gini Coefficient would be 1 (denoting total inequality). Usually, documents have some level of retrievability for any given retrieval function, and thus the Gini Coefficient is somewhere between 1 and 0. In [2], it was shown that for a given retrieval model, if a retrieval function was tuned to minimise bias (represented by the lowest Gini Coefficient) then this led to near optimal retrieval performance. Similar relationships were shown in [22] for P@10 and MAP and in [4] for recall based measures. However, here we examine the relationship between retrieval bias and performance across the spectrum of retrieval models and term weighting schemes as opposed to how bias changes when tuning a particular retrieval model (as done in [2, 4, 22]).

Estimating the retrievability of the documents within a collection requires a large number of queries. This is usually done by extracting all the bigrams from the collection and then selecting the most frequent [2–4, 22]. We employed the same method, but only selected those bigrams that appeared at least 20 times. Table 1 shows the total number of queries used on each collection (approximately 250,000 queries per collection). These queries were then issued to the retrieval system with a particular configuration (retrieval model/weighting/parameter setting). Using the cumulative based retrievability with a cut-off $c = 100$, we computed the retrievability of each document and subsequently the Gini Coefficient for each collection given the term weighting scheme.

3.3 Term Weighting Schemes

For our experiments, we used 17 term-weighting schemes stemming from the 6 model/frameworks presented in Section 2. The score assigned to a document given a query is determined by: $s(q, d) = \sum_{t \in q} s(t, d)$ where $s(t, d)$ is the term weighting assigned to term t in document d and q is the query which is composed of sequence of terms. The following schemes were implemented within the Lemur/Indri Framework[3].

[3] See `www.projectlemur.org/`. Code is freely available on GitHub.

The first four term weighting schemes we used were: (i) Term Frequency (TF) where $s(t,d) = n(t,d)$, (ii) Normalised TF (NTF) where $s(t,d) = \frac{n(t,d)}{n(d)}$, (iii) Term Frequency Inverse Document Frequency (TF.IDF) where $s(t,d) = n(t,d).idf(t)$, and (iv) Normalised TF.IDF (NTF.IDF) where $s(t,d) = \frac{n(t,d)}{n(d)}.idf(t)$. Here, $n(t,d)$ is the number of times t occurs in a document d, $n(d)$ is the total number of terms in a document, $idf(t) = \log \frac{N}{df(t)}$ where N is the number of documents in the collection, and $df(t)$ is the number of documents in which t appears. The fifth weighting scheme employed was Pivoted Length Normalisation (PTF.IDF) [21] where $a(d)$ is the average number of terms in d in the collection, and b controls the level of normalisation ($0 < b < 1$):

$$s(t,d) = \frac{n(t,d)}{(1-b) + b.\frac{n(d)}{a(d)}}.idf(t)$$

From the series of Best Match models [13] we employed BM25 ($0 > b > 1$), BM11($b = 0$) and BM15 ($b = 1$):

$$s(t,d) = \frac{(k_1+1).n(t,d)}{\left(k_1.(1-b) + b.\frac{n(d)}{a(d)}\right) + n(t,d)}.idf(t) \tag{1}$$

From the space of Language Models we implemented three different smoothing methods: Laplace smoothing (LP), which doesn't take into consideration document length normalisation and is similar to TF/NTF models (see Equation 2, where α is the level of smoothing parameter, and V is the number of unique terms in the collection); Jelinek Mercer Smoothing (JM) which is similar to the TF.IDF/NTF.IDF models, again without any explicit document length normalisation (see Equation 3), where $p(t,d)$ is the maximum likelihood estimate of the probability of a term appearing in a document, i.e. $p(t,d) = \frac{n(t,d)}{n(d)}$, and $p(t)$ is the the maximum likelihood estimate of the probability of a term appearing in the collection; and Bayes Smoothing with Dirichlet Prior (BS) which is often the best performing Language Model [23] and has a β parameter which controls the amount of length normalisation implicitly (see Equation 4).

$$s(t,d) = \frac{n(t,d) + \alpha}{n(d) + V.\alpha} \tag{2}$$

$$s(t,d) = \lambda.p(t,d) + (1-\lambda).p(t) \tag{3}$$

$$s(t,d) = \frac{n(t,d) + \beta.p(t)}{n(d) + \beta} \tag{4}$$

From the Divergence from Randomness framework [1], we choose two of the best performing models, DPH and PL2. DPH is parameter-free model:

$$s(t,d) = \frac{\big(1-p(t,d)\big)^2}{n(t,d)+1}.\left(\frac{n(t,d).N}{n(t)}.\log\Big(\frac{n(t,d).a(d)}{n(d)}\Big) + \frac{1}{2}\log\Big(2\pi.n(t,d).\big(1-p(t,d)\big)\Big)\right) \tag{5}$$

where $n(t)$ is the number of times term t appears in the collection. While PL2 has a parameter c to control document length normalisation:

$$s(t,d) = \frac{1}{n_2(t,d,c)+1}.\Bigg(n_2(t,d,c).log_2\Big(\frac{n_2(t,d,c)}{\Lambda(t)} + \Big(\Lambda(t) + \frac{1}{12.n_2(t,d,c)} - n_2(t,d,c)\Big).\log_2 e + \frac{1}{2}log_2(2\pi.n_2(t,d,c)\Bigg) \quad (6)$$

where the second normalisation component $n_2(t,d,c)$ is equal to $n(t,d).\log_2\big(1+\frac{c.a(d)}{n(d)}\big)$ and the mean of the Poisson distribution is defined by $\Lambda(t) = \frac{\sum_d n(t,d)}{N}$. As previously mentioned the Log Logistic Distribution (LGD) model [5] bridges DFR and Language models and is defined as:

$$s(t,d) = \log_2\left(\frac{\frac{d(t)}{N} + \log_2\big(n(t,d).(1+\frac{c.a(d)}{n(d)})\big)}{\log_2\big(n(t,d).(1+\frac{c.a(d)}{n(d)})\big)}\right) \quad (7)$$

where the c parameter controls the amount of smoothing ($c > 0$). Finally, we explored three of the best Divergence from Independence models [7]): DFIa referred to as irra12a in [7] based on the saturated model of independence (see Equation 8), and DFIb referred to as irra12b in [7] which is based on the standardisation model (see Equation 9):

$$s(t,d) = \log_2(1 + \frac{(n(t,d) - e(t,d))^2}{e(t,d)}) \quad (8)$$

$$s(t,d) = \log_2(1 + \frac{(n(t,d) - e(t,d))}{\sqrt{e(t,d)}}) \quad (9)$$

where $e(t,d) = \frac{n(t).n(d)}{N.a(d)}$. The third model DFIc based on the normalised Chi-Square measure of independence (referred to as irra12c in [7]):

$$s(t,d) = \left(\big(n(t,d)+1\big).\log_2\left(\frac{n(t,d)+1}{\sqrt{e_p(t,d)}}\right) - n(t,d).\log_2\left(\frac{n(t,d)}{\sqrt{e(t,d)}}\right)\right).\Delta(t,d) \quad (10)$$

where:

$$\Delta(t,d) = \left(\frac{n(d)-n(t,d)}{n(d)}\right)^{\frac{3}{4}} \times \left(\frac{n(t,d)+1}{n(t,d)}\right)^{\frac{1}{4}}$$

and:

$$e_p(t,d) = \frac{\big(n(t)+1\big).\big(n(d)+1\big)}{N.a(d)} + 1$$

Using this selection of retrieval models and term weighting-schemes we hypothesised that we would observe a reduction in bias as the models evolved over time from TF and TF.IDF to the more sophisticated DFR/DFI models. While some of the models are parameter-free (i.e. TF, NTF, TF.IDF, NTF.IDF, DPH, DFIa,DFIb and DFIc), we were required to estimate the free parameter for the other models. To estimate the parameters, we explored a parameter sweep, and selected the setting that resulted in the fairest or least biased model according to the Gini Coefficient. Setting the model this way requires no recourse to relevance judgements, and means that we can determine whether a model is capable of

	AQ					DG				
Model	**Gini**	**MAP**	**P@10**	**NDCG**	**MRR**	**Gini**	**MAP**	**P@10**	**NDCG**	**MRR**
TF	0.979	0.054	0.165	0.132	0.264	0.987	0.011	0.018	0.021	0.074
TF.IDF	0.977	0.071	0.180	0.148	0.296	0.987	0.014	0.026	0.029	0.094
NTF	0.971	0.034	0.057	0.039	0.122	0.975	0.017	0.032	0.031	0.067
NTF.IDF	0.967	0.048	0.086	0.061	0.175	0.971	0.024	0.038	0.040	0.099
PTF.IDF	0.956	0.063	0.122	0.094	0.227	0.970	0.049	0.088	0.087	0.160
BM25	**0.544**	0.162	0.316	0.263	0.478	**0.614**	0.167	**0.222**	**0.277**	**0.479**
BS	0.581	0.140	0.290	0.240	0.457	0.637	0.151	0.210	0.257	0.455
JM	0.669	0.125	0.253	0.206	0.436	0.666	0.108	0.172	0.213	0.410
LP	0.572	0.127	0.253	0.204	0.436	0.632	0.109	0.166	0.208	0.406
LGD	0.576	0.145	0.310	0.249	0.448	0.715	0.130	0.188	0.228	0.427
PL2	0.605	0.169	0.331	0.281	0.503	0.803	**0.182**	0.220	0.271	0.440
DPH	0.548	**0.181**	**0.369**	**0.315**	**0.564**	0.869	0.149	0.196	0.244	0.425
DFIa	0.612	0.173	0.349	0.297	0.519	0.870	0.153	0.194	0.243	0.426
DFIb	0.607	0.173	0.349	0.297	0.530	0.868	0.154	0.196	0.245	0.426
DFIc	0.610	0.135	0.274	0.226	0.439	0.807	0.139	0.196	0.235	0.410
Correlation		-0.941*	-0.909*	-0.897*	-0.925*	-	-0.686*	-0.759*	-0.765*	-0.808*

	T123					WT10G				
Model	**Gini**	**MAP**	**P@10**	**NDCG**	**MRR**	**Gini**	**MAP**	**P@10**	**NDCG**	**MRR**
TF	0.997	0.020	0.040	0.042	0.107	0.996	0.023	0.059	0.064	0.140
TF.IDF	0.997	0.025	0.048	0.050	0.122	0.996	0.030	0.085	0.084	0.169
NTF	0.977	0.017	0.031	0.030	0.060	0.979	0.010	0.006	0.008	0.020
NTF.IDF	0.974	0.024	0.048	0.045	0.085	0.975	0.015	0.022	0.017	0.051
PTF.IDF	0.957	0.047	0.165	0.164	0.277	0.972	0.031	0.083	0.069	0.115
BM25	**0.575**	0.122	0.301	0.307	0.478	0.670	0.094	0.183	0.176	0.313
BS	0.593	0.111	0.285	0.292	0.461	0.657	0.098	0.167	0.160	0.265
JM	0.664	0.049	0.253	0.253	0.401	0.680	0.058	0.128	0.127	0.260
LP	0.586	0.048	0.256	0.253	0.387	**0.653**	0.056	0.126	0.124	0.250
LGD	0.787	0.105	0.228	0.229	0.370	0.816	0.096	0.120	0.107	0.218
PL2	0.656	0.122	0.311	0.312	0.474	0.776	0.098	0.187	0.189	0.318
DPH	0.837	**0.128**	0.319	0.327	**0.490**	0.872	**0.121**	**0.226**	**0.228**	**0.409**
DFIa	0.781	0.126	0.319	0.326	0.484	0.859	0.116	0.194	0.198	0.379
DFIb	0.770	**0.128**	**0.323**	**0.330**	0.489	0.854	0.116	0.194	0.198	0.379
DFIc	0.690	0.111	0.278	0.285	0.449	0.754	0.118	0.194	0.201	0.385
Correlation		-0.624*	-0.807*	-0.801*	-0.811*	-	-0.554*	-0.597*	-0.570*	-0.563*

Table 2. Performance values along with Gini scores for each model. The final rows report the correlation between each performance measure and Gini. * denote whether the correlation is significant at $p < 0.05$.

being the fairest. Furthermore, prior work has shown that setting the model this way is close to optimal [22], we therefore believe that this is an appropriate way to determine which model is the fairest, and to see if it also is the best.

With BM25 we used 11 parameter settings for b between 0.0 and 1.0 increasing in steps of 0.1 (where BM11 is when $b = 0$ and BM15 is when $b = 1$). For PL2 and LGD we set parameter c to values between 1 and 10 but also included 0.1 and 100 to test the extremes. For Bayes (BS) and Laplace (LP), we set their respective smoothing parameters (β and α) to: 1, 10, 100, 500, 1000, 2000, 3000, 5000 and 10000 while on Jelinek Mercer and PTF.IDF we set their respective parameters (λ and c) between 0.1 and 0.9 increasing in steps of 0.1.

Overall, this resulted in 81 different configurations given the 17 term weighting schemes. With four collections and approximately 250,000 queries per collection, this amounted to well over 80 million queries being issued to generated the data for the results reported here.

4 Results and Analysis

Table 2 shows the bias (expressed by the Gini Coefficient, where lower is fairer) along with the performance associated with each model for each of the collections. For each measure we calculated the Pearson's correlation with Gini (where

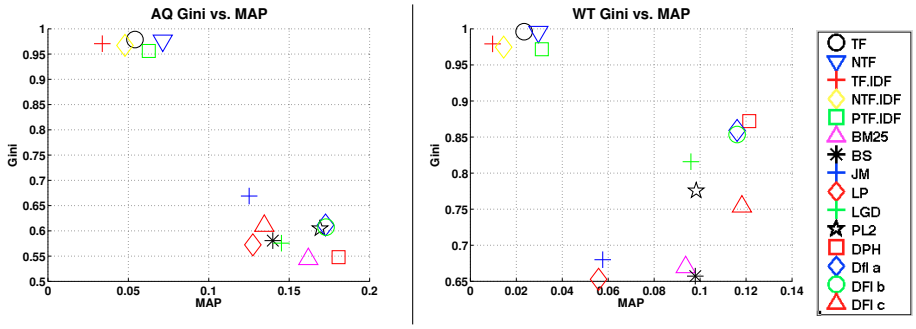

Fig. 1. AQ (Left) & WT (Right): Gini vs. MAP across the range of retrieval models.

an $*$ denotes whether the result was statistically significant at $p < 0.005$). For all the collections and measures we see that a moderate to high negative correlation exists between bias and performance (and is significant in most cases). To provide an impartial presentation of the results, Figure 1 shows a plot of Gini versus MAP for AQ (where the strongest correlations were observed) and WT (where the weakest correlations were observed). On AQ, we see that the fairest model also obtains the best MAP, while on WT the fairest model is mediocre at best. From the plots and tables, we can also see that the DFI models were not the fairest models, but they are all reasonably effective (and the best performing model on T123, and second best on AQ and WT) with the DFIb term weighting scheme performing the best out of all DFI models. It is also apparent from the plots that there are two main groups: the TF/TF.IDF based models which don't explicitly perform document length normalisation and the other models, which do. Within this second group, however, the relationship is bit more complicated, as can be seen from the WT plot in Figure 1. On WT, without the TF/TF.IDF models the correlation would appear to be positive.

To explore the relationship between models further and to see what document features they make more or less retrievable, we plotted the retrievability of documents versus document length (and versus average information content of a document[4]). This was done by sorting the documents according to the length or information content, then grouping documents into buckets we calculated the average length (information content) and the average retrievability. The results are plotted in Figure 2 for both AQ and WT.

The first observation we can make here is to see that on TF.IDF (blue triangles) longer documents are much more retrievable than shorter documents. However, when TF is normalised (NTF.IDF shown as yellow triangles), the trend is reversed and shorter documents become highly retrievable. It is clear that most of the bias associated with these models stems from the lack of length normalisation, similar and corresponding patterns were observed for TF and NTF. Of note, is the erratic shape of the plot of information content versus retrievability for TF.IDF (blue triangles), suggesting that the term weighting is not particular robust or consistent when compared to other models.

[4] This was calculated by summing the TF.IDF scores of the all the terms in the document and then dividing by the document's length.

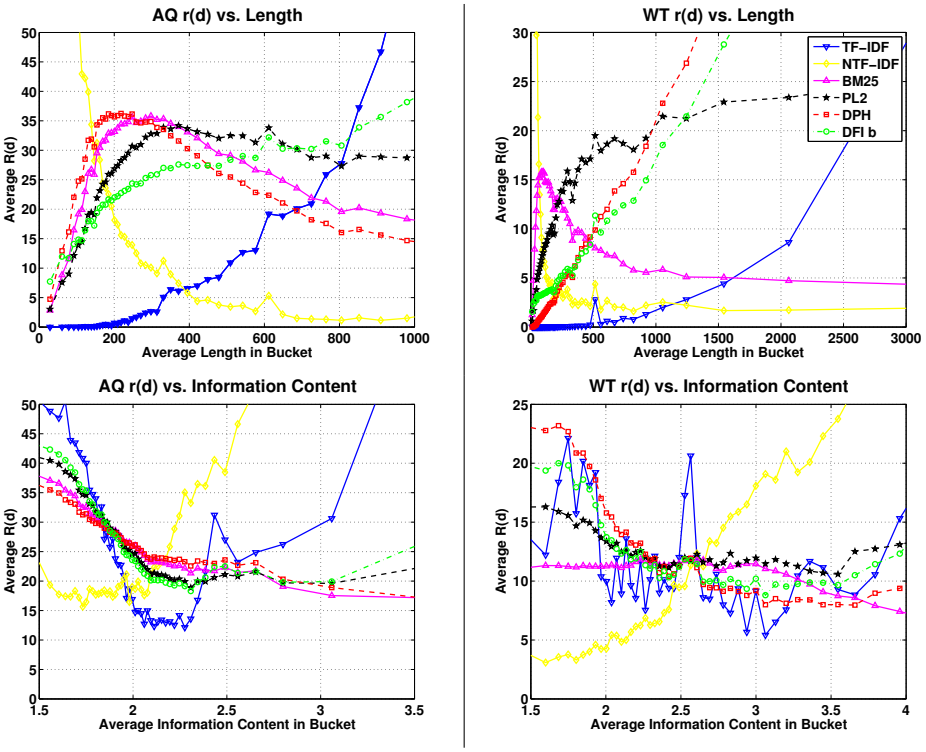

Fig. 2. AQ (Left) & WT (Right) $R(d)$ vs length (top), $R(d)$ vs average information content (bottom).

As previously mentioned, BM25 is consistently the fairest model. If we examine the length plots then we can see that across the different lengths BM25 (purple triangles) tends not to overly favour longer documents when compared to the other retrieval models (though has a tendency to favour average length documents). In terms of information content, BM25 provides about the same level of retrievability. Given these plots, it is clear why BM25 is the fairest.

On AQ, we see that DPH (red squares) tends to be very similar to BM25 and very fair in terms of length. However, on WT, DPH clearly favours documents proportional to their length such that longer documents are much more retrievable. If we examine Table 1, we can see that relevant documents in WT are also much longer - so this bias seems to improve the performance. PL2 (black stars) shows a similar bias toward longer documents, however it is somewhat mitigated when compared to DPH because it has a parameter that can be tuned.

With respect to the Divergence From Independence Models, we have plotted the best performing model, DFIb (green circles). Much like the DFR models, on AQ the model is quite fair across document lengths. However, when applied to the web collections which have much more varied lengths, the DFIb (and the other DFI models) tend to make longer documents more retrievable. On the information content plots, DFIb also favours documents which have lower information content. This perhaps, is to be expected as longer documents tend to have lots of non-informative words which reduces their overall average information content.

On WT, in particular, models that favour documents that are longer and that contain less information content on average, tend to perform much better than the fairer models (i.e. DPH, PL2 and DFI all outperform BM25). To examine why this is the case we examined the length and information content of documents in the collection, the pool of judged documents the relevant documents. Table 1 reports the mean values for each collection. To determine whether relevant documents and pooled documents were different, either in terms of length or information content, compared to other documents in the collection, we performed un-paired t-tests. We found that relevant/pooled documents in these test collections were longer and had lower information content (this was significantly so). These results suggest that the pools are not representative of the collection, a finding which was also shown in [15]. Interestingly however, as the difference between the average document length (and average information content) of the collection and the relevant/pooled documents becomes larger, the lower the correlation between fairness and performance. It is an open question whether relevant documents are actually longer and/or lower in information content, or whether this is an artefact of the test collection creation process. Nonetheless, we observe that when the relevant documents are more like the collection, fairer is better, as witnessed on the AQ collection.

5 Summary and Conclusions

In this paper, we have measured the retrieval bias of a spectrum of retrieval model/weightings to determine which model is the fairest. While we have observed that there is strong correlation between fairness and performance, tailoring the model to the nuances of the test collection invariably leads to better performance at the expense of making certain documents less retrievable. Without test collections which are representative of the underlying documents, it is hard to definitely say whether doing so is a good thing or bad thing. However, without knowing what documents are likely to be relevant (or what their characteristics are) in advance, the most sensible way to select a model/weighting is to choose the one that is the fairest, then as usage data is obtained to tune the system accordingly. In this sense, BM25 generally exhibits the least bias on the collection, while delivering competitive retrieval performance. This is quite remarkable given all the subsequent models developed. This work prompts further research questions: (i) how do we optimise performance given such biases, (ii) how do we make more representative test collections, (iii) what is the impact of such biases on future sets of queries (i.e. what if shorter and more informative documents were more likely to be relevant), and (iv) if the performance measures took into account document length and utility (such as Time Biased Gain and the U-Measure), would fairer lead to better?

References

1. Amati, G., Van Rijsbergen, C.J.: Probabilistic models of ir based on measuring the divergence from randomness. ACM Trans. on Info. Sys. pp. 357–389 (2002)
2. Azzopardi, L., Bache, R.: On the relationship between effectiveness and accessibility. In: Proc. of the 33rd international ACM SIGIR. pp. 889–890 (2010)

3. Azzopardi, L., Vinay, V.: Retrievability: An evaluation measure for higher order information access tasks. In: Proc. of the 17th ACM CIKM. pp. 561–570 (2008)
4. Bashir, S., Rauber, A.: On the relationship bw query characteristics and ir functions retrieval bias. J. Am. Soc. Inf. Sci. Technol. 62(8), 1515–1532 (2011)
5. Clinchant, S., Gaussier, E.: Bridging language modeling and divergence from randomness models: A log-logistic model for ir. In: Proc. of the 2nd International Conference on Theory of Information Retrieval. pp. 54–65. ICTIR '09 (2009)
6. Crestani, F., Lalmas, M., Van Rijsbergen, C.J., Campbell, I.: Is this document relevant? probably: a survey of probabilistic models in information retrieval. ACM Computing Survey 30(4), 528–552 (1998)
7. Dinçer, B.T., Kocabas, I., Karaoglan, B.: Irra at trec 2010: Index term weighting by divergence from independence model. In: TREC (2010)
8. Fang, H., Tao, T., Zhai, C.: A formal study of information retrieval heuristics. In: Proc. of the 27th ACM SIGIR conference. pp. 49–56. SIGIR '04 (2004)
9. Fuhr, N.: Probabilistic models in ir. Computer Journal 35(3), 243–255 (1992)
10. Gastwirth, J.: The estimation of the lorenz curve and gini index. The Review of Economics and Statistics 54, 306–316 (1972)
11. Harter, S.P.: A probabilistic approach to automatic keyword indexing. part i. on the distribution of specialty words in a technical literature. Journal of the American Society for Information Science 26(4), 197–206 (1975)
12. Hiemstra, D.: A probabilistic justification for using tf.idf term weighting in information retrieval. International Journal on Digital Libraries 3(2), 131–139 (2000)
13. Jones, K.S., Walker, S., Robertson, S.E.: A probabilistic model of information retrieval: development and comparative experiments (parts 1 and 2). Information Processing and Management 36(6), 779–808 (2000)
14. Kocabas, I., Dinçer, B.T., Karaoglan, B.: A nonparametric term weighting method for information retrieval based on measuring the divergence from independence. Information Retrieval pp. 1–24 (2013)
15. Losada, D.E., Azzopardi, L., Baillie, M.: Revisiting the relationship between doc. length and relevance. In: Proc. of the 17th ACM CIKM'08. pp. 419–428 (2008)
16. Maron, M.E., Kuhns, J.L.: On relevance, probabilistic indexing and information retrieval. Journal of the ACM 7(3), 216–244 (1960)
17. Ponte, J.M., Croft, W.B.: A language modeling approach to information retrieval. In: Proc. of the 21st ACM SIGIR conference. pp. 275–281. SIGIR '98 (1998)
18. Robertson, S.E., Walker, S.: Some simple effective approx. to the 2-poisson model for probabilistic weighted retrieval. In: Proc. of ACM SIGIR94. pp. 232–241 (1994)
19. Salton, G., Wong, A., Yang, C.S.: A vector space model for automatic indexing. Communications of the ACM 18(11), 613–620 (1975)
20. Salton, G.: Automatic Information Organization and Retrieval. (1968)
21. Singhal, A., Buckley, C., Mitra, M.: Pivoted document length normalization. In: Proce. of the 19th ACM SIGIR conference. pp. 21–29. SIGIR '96 (1996)
22. Wilkie, C., Azzopardi, L.: Relating retrievability, performance and length. In: Proc. of the 36th ACM SIGIR conference. pp. 937–940. SIGIR '13 (2013)
23. Zhai, C., Lafferty, J.: A study of smoothing methods for language models applied to ad hoc ir. In: Proc. of the 24th ACM SIGIR. pp. 334–342 (2001)

Acknowledgements: This work is supported by the EPSRC Project, *Models and Measures of Findability* (EP/K000330/1).