A Retrievability Analysis: Exploring the Relationship Between Retrieval Bias and Retrieval Performance

Colin Wilkie, Leif Azzopardi
School of Computing Science,
University of Glasgow
Glasgow, United Kingdom
{Colin.Wilkie,Leif.Azzopardi}@glasgow.ac.uk

ABSTRACT

Retrievability provides an alternative way to assess an Information Retrieval (IR) system by measuring how easily documents can be retrieved. Retrievability can also be used to determine the level of retrieval bias a system exerts upon a collection of documents. It has been hypothesised that reducing the retrieval bias will lead to improved performance. To date, it has been shown that this hypothesis does not appear to hold on standard retrieval performance measures (MAP and P@10) when exploring the parameter space of a given retrieval model. However, the evidence is limited and confined to only a few models, collections and measures. In this paper, we perform a comprehensive empirical evaluation analysing the relationship between retrieval bias and retrieval performance using several well known retrieval models, five large TREC test collections and ten performance measures (including the recently proposed PRES, Time Biased Gain (TBG) and U-Measure). For traditional relevance based measures (MAP, P@10, MRR, Recall, etc) the correlation between retrieval bias and performance is moderate. However, for TBG and U-Measure, we find that there is strong and significant negative correlations between retrieval bias and performance (i.e as bias drops, performance increases). These findings suggest that for these more sophisticated, user oriented measures the retrievability bias hypothesis tends to hold. The implication is that for these measures, systems can then be tuned using retrieval bias, without recourse to relevance judgements.

Categories and Subject Descriptors

H.3.4 [**Information Storage and Retrieval**]: Systems and Software: Performance Evaluation

General Terms

Theory, Experimentation

Keywords

Retrievability; Effectiveness; Evaluation; User Measures

1. INTRODUCTION

Traditionally, Information Retrieval systems have been evaluated in terms of performance and efficiency [?, ?]. While performance measures seek to quantify how "good" the system is at retrieving relevant results and efficiency measures seek to quantify how fast results can be retrieved, retrievability evaluates a third and very different aspect of IR systems: how likely a document is to be retrieved [?, ?]. As such retrievability is fundamental to IR because a document cannot be judged relevant if it is never retrieved or ever presented to a user at a sufficiently high rank. Put another way, retrievability precedes relevance [?]. In the most extreme case, if a document is not indexed then the document cannot be retrieved via the retrieval system [?]. However, even if the document is indexed, this does not necessarily mean that a user will be able find it. This is because the retrievability of a document depends upon a number of factors: (i) the ability of the user to pose a good query, (ii) the willingness of the user to examine documents, (iii) the features of the document, (iv) the features of other documents and number of similar documents, (v) the retrieval system/method/model, and (vi) how the documents are indexed. As a result some documents are easy to find, while others are difficult, if not impossible, to find [?]. An open question is how these factors affect retrievability and subsequently performance.

Given the retrievability of each document in a collection it is possible to calculate the retrieval bias of the system. Intuitively, if a system favours the retrieval of a certain type or subset of documents (e.g. retrieving longer documents over shorter documents [?]) then the retrievability of longer documents will be higher and so the retrieval bias of the system will be high. On the other hand, if the retrieval bias of the system is low then it will not unduly favour one document over another due to its characteristics, giving each document a reasonably equal chance of retrieval (i.e. retrieving based on its merits/qualities for being relevant to the queries issued). The implication here, is that such a system is likely to perform better than a biased system because, for any particular document, there is a set of queries which will return that document at a rank high enough for the user to encounter it. Whereas a highly biased system will provide many opportunities to retrieve certain documents (such as the longer documents, given the example above) but few opportunities to retrieve others, making them less likely to be found. Note that this argument assumes that all the documents in the collection have some value, and could be relevant for some information need at some point in time[1]. This argument leads to the *retrievability bias* hypothesis; reducing the retrieval bias of a system will lead to improved retrieval performance. To date the evidence for this hypothesis has been mixed and holds only under certain circumstances. While related work has shown that retrievability can be used to improve performance and efficiency [?, ?, ?], the relationships between retrieval bias and performance has been shown to be complex, non-linear and measure dependent [?, ?, ?, ?].

In this paper we examine the retrieval bias hypotheses by exploring the relationship between retrieval bias and a range of performance measures. We will focus on understanding this relationship in the context of parameter estimation, i.e. how retrieval bias and retrieval performance change as the retrieval model's parameters change. Consequently, if the hypotheses holds in this context, it will be possible to estimate the parameters of a retrieval model without recourse to relevance judgements. To this end, we perform an in-depth study investigating this relationship on four standard retrieval models, five TREC test collections and using ten retrieval performance measures. In doing so, we not only replicate the work and findings previously performed [?, ?], we do so on larger collections and on more retrieval models. The novelty of this work stems from our main contribution, where we examine this relationship against seven other measures, which have not been previously tested, and include three recently proposed measures: PRES [?], Time Biased Gain [?] and U-Measure [?]. The remainder of this paper is as follows: we provide a summary of the related work formally defining retrievability and retrieval bias in Section **??**. Then we describe the method used to explore this relationship in Section **??** before presenting our results in Section **??**. We find that for standard relevancy based measures that a low to moderate correlation exists. This mis-match is attributed to the test collections used which select relevant documents that are significantly longer than other documents in the collection (i.e. some length bias is therefore required) as described in both [?] and [?]. However, with Time Biased Gain and the U-Measure, we find that a strong and significant correlation exists. We attribute this to the fact that these measure account for document length within their estimations. Finally, in Section **??** we conclude with a summary of our findings and contributions along with directions for future work.

2. BACKGROUND

In [?], Azzopardi and Vinay introduced the concept of retrievability, a measure that defined how *easily* a document could be retrieved by a particular configuration of an IR system. Formally, retrievability $\mathbf{r}$ of a document $\mathbf{d}$ with respect to an IR system is defined as:

$$\mathbf{r(d)} \propto \sum_{\mathbf{q} \in \mathbf{Q}} \mathbf{O_q . f(k_{dq}, c)}$$

where $\mathbf{q}$ is a query from the very large query set $\mathbf{Q}$, meaning $\mathbf{O_q}$ is the opportunity of the query being chosen from this set. $\mathbf{k_{dq}}$ is the rank at which $\mathbf{d}$ is retrieved given $\mathbf{q}$, and $\mathbf{f(k_{dq}, c)}$ is a utility function which denotes how retrievable the document $\mathbf{d}$ is for the query $\mathbf{q}$ given the rank cutoff $\mathbf{c}$. Retrievability is then calculated by summing over all queries $\mathbf{q}$ in query set $\mathbf{Q}$. Theoretically, $\mathbf{Q}$ represents the universe of all possible queries, but in practice $\mathbf{Q}$ is very large set of queries [?, ?, ?, ?, ?]. The standard measure of retrievability used is a cumulative based approximation, which employs an utility function $\mathbf{f(k_{dq}, c)}$, such that if a document, $\mathbf{d}$, is retrieved in the top $\mathbf{c}$ documents given $\mathbf{q}$, then $\mathbf{f(k_{dq}, c)} = 1$, otherwise $\mathbf{f(k_{dq}, c)} = 0$. This measure provides an intuitive value for each document as it is simply the number of times that the document is retrieved in the top $\mathbf{c}$ documents. Documents falling outside the the top $\mathbf{c}$ are completely ignored, simulating a user who is only willing to pursue the first $\mathbf{c}$ results. Essentially, the more queries that retrieve a document, the more retrievable a document is.

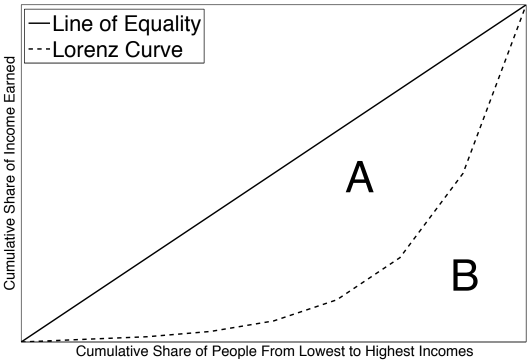

Figure 1: The Lorenz Curve shows inequality within a population. As the area B shrinks, the inequality (i.e. bias) in the population increases.

2.1 Retrieval Bias

To determine the retrieval bias of a system/model given the retrievability scores, a method from Economics and the Social Sciences is used. The Lorenz Curve is used to visualise the inequality within a population given their incomes [?]. This is performed by first sorting the individuals in the population in ascending order of their wealth and then plotting a cumulative wealth distribution. If the wealth in the population is distributed equally then the cumulative distribution would be linear. The extent to which a given distribution deviates from equality is reflected by the skew in the distribution. The more skewed the plot, the greater the amount of inequality, or bias within the population. To summarise the inequality of such distributions the Gini coefficient is commonly used. Given the Lorenz Curve, the Gini coefficient can be calculated by dividing the area marked $\mathbf{A}$ by the area marked $\mathbf{A + B}$ in Figure **??**.

In [?, ?, ?], the Gini coefficient was used to measure the retrieval bias, where the retrievability scores of documents indicates the wealth of the document. If all documents were equally retrievable then the area $\mathbf{A}$ would be zero, and thus the Gini coefficient would be zero (denoting equality within the population, shown as the Line of Equality in Figure **??**). On the other hand if only one document was retrievable and the rest were not, then the area of $\mathbf{B}$ would be zero, and the

[1] An interesting argument arises here concerning the utility of a document. If a document is never going to be retrieved by the system, then why index it? If documents have been indexed by the system that are never going to be retrieved then we introduce inefficiencies and will unnecessarily consume resources. Arguably, these documents should be removed or partitioned.

Gini coefficient would be one (denoting total inequality). Document Retrievability tends to follow the Lorenz Curve when $\mathbf{r}(\mathbf{d})$ is plotted against a bias present in a system.

2.2 Uses and Related Measures

Retrievability - and the theory of - has been used in numerous contexts. For example, retrievability has been used to study search engine bias on the web [?] and within patent collections [?], to improve the efficiency of systems when pruning [?], and it has been related to navigability when tagging information to improve how easily users browsing through the collection could find documents [?]. It is also worth noting that Retrievability is part of a family of measures that approximate how easily information can be found, either from a system point of view, for example, the discoverability [?] and crawlability [?] of content by crawlers and spiders, or from a user point of view, such as navigability [?, ?] and searchability [?].

2.3 Relating Retrievability and Performance

There have been a number of works which suggest that retrievability and retrieval performance are related, and that retrievability can be used to improve both the performance and efficiency of a system [?, ?, ?, ?, ?].

In [?], the authors investigated whether documents with low retrievability could be removed from a collection without harming retrieval performance (in terms of MAP and early precision). They found that with models such as TF.IDF, which exhibited a high retrieval bias on the collection, up to 80% of the least retrievable documents could be removed without significantly reducing retrieval performance. For retrieval models such as BM25 and Language models, where the retrieval bias they exhibit is much less than TF.IDF, only 40-50% of the least retrievable documents could be removed before a significant reduction in performance was observed. These finding suggest that retrievability could be useful in improving the efficiency of the system by removing documents unlikely to be retrieved, and thus unlikely to make a difference to retrieval performance.

In [?], Bashir and Rauber studied the effect of Pseudo Relevance Feedback (PRF) on performance and Retrievability. They found that standard Query Expansion methods, while increasing performance, also increased the retrieval bias. They discovered that standard query expansion methods were generating specific sets of terms that resulted in the system consistently preferring a particular subset of the collection. To combat the increase in bias, they devised a method of PRF that used clustering; this resulted in a reduction in bias, as well as an increase in performance over other QE techniques. When employing their PRF technique to patent retrieval, they showed that the decrease in bias led to improved recall for prior art search [?] suggesting a direct relationship between retrieval bias and recall-oriented performance metrics. In [?], Pickens *et al* used the theory of retrievability to create reverted indexes that improved both the efficiency and performance of retrieval systems by capitalising on knowing what terms within a document makes that document retrievable [?]. In their experiments on BM25 and PL2, they showed that the reverted index sped up pseudo relevance feedback and also significantly improved retrieval performance over competing algorithms.

Taking a different tack, in [?], Bashir and Rauber compared a number of retrieval models in terms of their retrieval bias as a way to rank different patent retrieval systems submitted to the TREC Chemical Retrieval Track. They found that there was a strong negative correlation between the retrieval bias and recall. Suggesting that retrieval bias could be used as a way to rank system for this task/measure. This idea was further examined by Wilkie and Azzopardi in [?] where they ranked seventeen systems according to their retrieval bias. The authors posed the question, *is fairer better?* They also found a strong negative correlation between retrieval bias and Mean Average Precision (MAP), when ranking systems which was consistent on several test collections. A similar study and analysis was also performed by Bashir and Rauber [?], which produced similar findings. These results suggest that the retrieval models with the least bias are actually the best performing models, providing support for the retrievability bias hypothesis. In our work, rather than looking at ranking systems as is done in [?, ?, ?], we will be focusing on exploring the parameter space of a given retrieval model. It appears that this relationship is much more complicated.

In [?], Azzopardi and Bache discuss the relationship between retrieval bias and performance with respect to the definition of retrievability. The authors point out that a purely random IR system would ensure equal retrievability (resulting in Gini = 0) but this would also result in very poor retrieval performance. Thus it only makes sense to consider retrieval functions which are non-random. They further note that an over fitted retrieval model, which unduly favours the retrieval of a known set of relevant documents for a give set of queries, would be overly biased, but result in very high performance for that set of queries. They conclude that neither extreme is desirable and suggest that there is likely to be a trade-off between retrieval bias and retrieval performance. In a set of preliminary experiments they investigate whether such a trade-off exists. Using the AP and WSJ TREC test collections they found that as retrieval bias decreased it corresponded with an increase in MAP and P@10 for BM25 and the Jelinek Mercer Smoothed Language Model. Contrary to their intuition, their findings suggested that a more useful relationships exists. However, these results were far from conclusive, and very preliminary. More recently, this study was followed up in [?], where Wilkie and Azzopardi examined the relationship between precision based measures and retrieval bias on BM25 and PL2 on a number of TREC news collection. More specifically, they examined the correlation between retrieval bias and retrieval performance (MAP and P@10). Their study revealed a number of interesting observations. Firstly, the relationship between precision based measures and retrieval bias turned out to be non-linear; such that a certain amount of document length normalisation needed to be applied to maximise performance. Once this point was reached, increasing the amount of length normalisation meant performance would begin to degrade rapidly as bias was increased. These results suggest that retrieval bias can be used as an indicator for tuning retrieval systems but that minimising the retrieval bias does not necessarily match with maximum performance on precision based measures (P@10 and MAP). This mis-match was attributed to the fact that in the TREC collections used, relevant documents tend to be longer than the average document. And so less document length normalisation is required to maximise the performance for the measures used, i.e. some bias appears necessary on such collections given these measures.

In [?], it was shown that the TREC pools were not representative of the collection and many longer documents appear in the pools, resulting in more long documents being considered relevant. Therefore, it may be the case that by using test collections that are more representative or measures that incorporate length in their evaluation would lead to a stronger relationships between retrievability bias and retrieval performance. In this work, we go beyond prior work and examine seven measures that have been previously untested, including two new measures which also include mechanisms that account for length, i.e. TBG and U-Measure. Our experiments are performed on a wider variety of test collections and across more retrieval models.

3. EXPERIMENTAL METHOD

3.1 Research Questions

In this study we shall focus on answering the following research question: How do retrieval performance metrics relate to the retrieval bias imposed by systems? We shall investigate this question in the context of estimating the parameters of a retrieval model (i.e. within model analysis as done in [?, ?], rather than a between models analysis as done in [?, ?] which is where different retrieval models are ranked using retrieval bias). We shall also test the Retrievability hypothesis, which we previously stated as follows: that minimising retrieval bias will lead to maximising retrieval performance. Thus, when setting the parameters of a retrieval model, we wish to determine whether it is possible to arrive at an estimate which leads to good retrieval performance by tuning according to the retrieval bias. To this end, we used the following experimental set up.

3.2 Data and Materials

We used five TREC test collections in our experiments: TREC 123 (T123), Aquaint (AQ), WT10G (WT), DotGov (DG) and ClueWeb (CW)[2]. Table **??** details the topics used and the size of each collection. All collections were indexed on Lemur Indri and had stop words removed as well as being Porter stemmed. Using these collections provides good coverage across a range of different sizes, document types and query sets, and includes several sizable web collections.

3.3 Retrieval Models

For the purposes of these experiments we have selected four commonly used retrieval models: Okapi BM25 [?], PL2 [?] from the Divergence From Randomness (DFR) framework, and Language modelling using Bayesian Smoothing (BS) [?] and Jelinek Mercer smoothing (JM)[?]). Using this selection of retrieval models we hope to determine whether the relationship is generalizable across models or differs between them. Our expectation is that for the first three models will behave in a similar fashion as they all have a parameter to control for document length normalisation, while the Jelinek-Mercer Language Model may behave differently.

Parameter Settings

The parameter space explored for each retrieval model was as follows: For BM25, as there are several parameters we could explore, we chose to only alter **b** and keep the **k** and **k1** parameters fixed at default settings. We used 11 parameter settings for **b**, between 0.0 and 1.0 increasing in steps of 0.1 (0,0.1,0.2...1.0). For PL2 we used parameter settings of **c** between 1 and 10 in increments of 1 (1,2,3...10) but also included 0.1 and 100. For BS we used the β parameter settings of 1, 10, 100, 500, 1000, 2000, 3000, 5000 and 10000. And for JM, the λ parameter setting was varied from 0.1 to 0.9 in steps of 0.1 (0.1,0.2,0.3...0.9). These settings cover the range of values typically used for these models.

3.4 Performance Measures

Ten performance measures were used, which we have grouped into four categories: precision based, precision-recall based, recall based and gain based.

Precision Based Measures

In this category, we considered Mean Reciprocal Rank (MRR) and Precision@10 (P@10). We also examined the relationship with other precision based measures (i.e. P@5, P@20, etc) though our findings were similar to those reported for P@10. P@10 was used in [?, ?] and is included to compare with previous work. These precision based measures, interestingly, reflect to some extent the access function used to calculate the scores. For example, the cumulative score is calculated at a cutoff of **c** in the function $\mathbf{f(k_{dq}, c)}$, like P@**c**.

Precision-Recall Based Measures

To consider how retrieval bias relates to measures that consider both precision and recall, we include Mean Average Precision (MAP) and the Binary Preference measure [?] (BPREF). Again, MAP was used in previous studies and is included for completeness as we replicate the results of the previous work. Conversely, we have included BPREF to see if dealing with incompleteness in the judgements, provides a better correlation, when compared to MAP.

Recall Based Performance Measures

The recall measures we shall examine are: Recall@1000, the number of relevant documents retrieved (REL_RET), and PRES [?]. Recall was already explored in [?, ?, ?] to compare different systems. Here we include it to compare within the same model across it's parameter space. We use REL_RET to see whether there is a difference between the retrievability and the absolute recall value. We have included a new and interesting measure proposed by Magdy and Jones called PRES which is a recall based measure that is position sensitive. This means it takes into account the position of the relevant documents retrieved in the ranked list and includes a cut-off which penalises recall beyond that point in the ranked list. This essentially encodes the behaviour that a user will look no further than **c** in the ranked list. The rationale for this cut-off is that it provides a worse case scenario, where if relevant material is presented after the point beyond which the user is willing to look, then they will not find/recall the document similar to the **c** parameter in the retrievability utility function. We set the cut-off in PRES to be 100, like our cut-off for the retrievability utility function. This may mean that we observe a stronger correlation between PRES and retrieval bias than the other measures.

Gain Based Performance Measures In our experiments, we used three gain based measures, the well known and widely used Normalised Cumulative Discounted Gain at 100 (NDCG) [?], along with two recently proposed measures

[2] However, due to space constraints we shall only report on a subset of these collections.

Collections	AQ	T123	DG	WT10G	CW
Docs	1,033,461	1,078,166	1,247,753	1,692,096	50,000,000
TREC Topics	301-400	1-200	551-600	451-550	1-150

Table 1: Collection Statistics

Time Biased gain [?] (TBG) and U-Measure [?]. No previous work has examined the relationship between these measures and retrieval bias, so it will be interesting to determine whether they exhibit a different relationship to the precision/recall based measures. Since TBG and U-Measure are new, it is worth taking some time to explain them, along with the parameters that need to be set in order to use them.

Time Biased Gain In [?], Time Biased Gain was proposed as a way to account for the time it takes to read through and process the result list and to extract relevance from it. The longer a document is the longer it takes to process the document, and so document length is accounted for within the evaluation measure. This means a long document with equal gain to a shorter document will contribute less gain overall as time is wasted reading the document.

The general form of the TBG equation where $\mathbf{G(t)}$ is a gain function over time and $\mathbf{f(t)}$ is the density function is as follows:

$$\mathbf{E[G(t)]} = \int_{\mathbf{0}}^{\infty} \mathbf{G(t)f(t)dt}$$

TBG has a number of parameters that need to be estimated. The $\mathbf{A}$ parameter denotes how long it takes a user to read a word, on average. Essentially, this parameter limits how many documents a user can read in a specified period of time. Increasing $\mathbf{A}$ results in fewer documents being read as it takes longer to read a word. Decreasing $\mathbf{A}$ means the user can read more words and therefore, more documents. User behaviour is simulated in the other parameters, which include: $\mathbf{P(C|R)}$ the probability of clicking a relevant summary, $\mathbf{P(S|R)}$ the probability of saving a relevant document, $\mathbf{P(C|N)}$ the probability of clicking a non-relevant summary, $\mathbf{P(S|N)}$ the probability of saving a non-relevant document, $\mathbf{T_s}$ the time to evaluate a summary, $\mathbf{B}$ a fixed overhead to judge any document (relevant or not) and $\mathbf{H}$, the half life at which gain degrades.

Since it is necessary to calibrate the parameters of TBG, we employed 12 annotators to judge the relevance of 20 snippets and the corresponding documents (if they believed they may be relevant) for 10 topics for each collection. The time spent on each action, and the success of each action was recorded so that we could estimate the required probabilities. Table **??** shows the average values obtained for each collection.

Parameter	T123	AQ	DG	WT	CW
$\mathbf{T_s}$	8.00	6.50	5.10	3.90	4.10
$\mathbf{A}$	0.06	0.03	0.03	0.05	0.05
$\mathbf{B}$	7.30	7.00	7.00	7.10	7.10
$\mathbf{P(C|R)}$	0.63	0.60	0.61	0.57	0.55
$\mathbf{P(C|N)}$	0.42	0.46	0.43	0.56	0.52
$\mathbf{P(S|R)}$	0.72	0.71	0.66	0.62	0.64
$\mathbf{P(S|N)}$	0.38	0.37	0.41	0.43	0.44

Table 2: Time Biased Gain parameter settings.

U-Measure

In [?], Sakai and Dou proposed a new, user based evaluation metric called U-Measure. U-Measure is designed to estimate the amount of gain a user obtains when reading through documents in the ranked list. In U-Measure, it is assumed users will read the snippet of every document in the list and will always read relevant documents and never read non-relevant documents (i.e. it assumes that $\mathbf{P(C|R) = 1}$ and $\mathbf{P(C|N) = 0}$). Once a user begins to read a relevant document, they will only read a certain percentage of the document before returning to the ranked list to read through the remaining results. As a user reads further down the ranked list, the amount of gain they receive from a document decays and there is a cut-off that indicates when a user will stop reading results.

U-Measure can be configured to reflect different users by altering two parameters. The first of these parameters defines what portion of a relevant document the user will read. Adjusting this to a higher value means more time is spent on relevant documents and as such, less documents will be read overall. Setting this parameter to lower values means users will receive less gain per document but will be able to read more documents. The second parameter defines how far the user will read. This parameter, set on a character limit, provides the point at which the user will stop reading through results and close that session. Higher values mean users will read more documents and will therefore be likely to receive more gain from the session. For these experiments we used the parameter settings suggested by Sakai and Dou [?] where the character limit is 132000 and the percentage of a document read is 20%. These settings allow for roughly 100 documents to be read, per topic, by the user in the collections we have used.

The general form of the U-Measure equation is as follows:

$$\mathbf{U} = \frac{\mathbf{1}}{\mathbf{N}} \sum_{\mathbf{pos \cdot 1}}^{|\mathbf{tt}|} \mathbf{g(pos)D(pos)}$$

In this equation, $\mathbf{N}$ is a normalisation factor while **pos** is the offset position in **tt** and $\mathbf{D(pos)}$ is a decay factor based on position.

While TBG and U-Measure appear similar, some key differences affect the outcome of these measures. The main difference being TBG includes probabilities for a user to read a relevant or non-relevant document, U-Measure assumes the perfect user who will always read relevant documents and never read non relevant documents. Another key difference is that TBG assumes a user will always read the entirety of any document they click on, conversely, U-Measure dictates that a user will always read a fixed percentage of the documents they click. These important differences make each of these measures subtly different. However, in contrast to all the other measures described, they differ in that they account for the length of the document in the measure, and the amount of gain is proportional to how much effort/time is required to extract that gain given the ranked list. We pre-

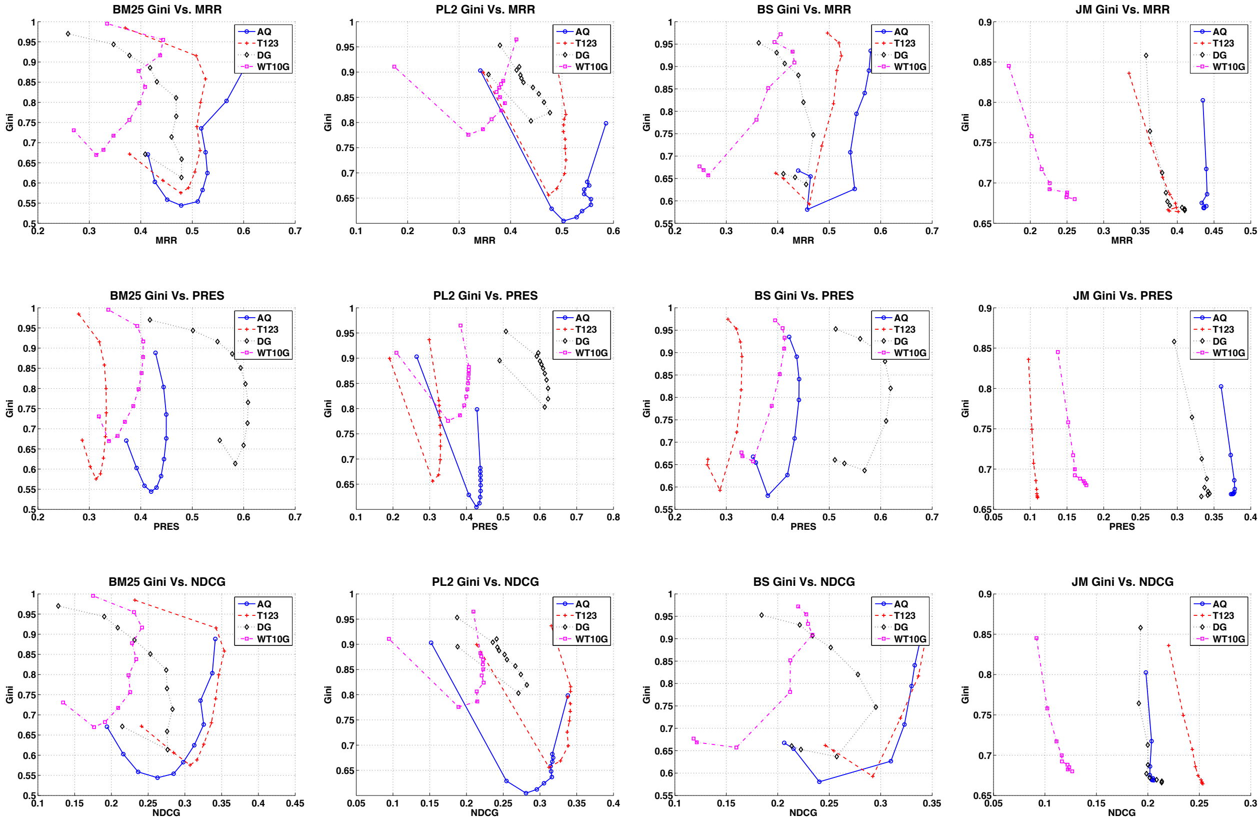

Figure 2: Plots of Gini vs. performance measure for each model on AQ, T123, DG and WT10G. The relationship between retrieval bias and performance is clearly not linear given these plots.

viously mentioned in the Section ?? that it was suggested the test collections tend to house longer relevant documents, but if the user has to go spend more time in order to extract that gain, then on these measures we may observe a stronger relationship with retrieval bias.

3.5 Retrieval Bias

To compute the retrieval bias, we first generated the retrievability scores $\mathbf{r(d)}$ for each document using the methodology used in previous work [?, ?, ?, ?, ?]. The procedure is as follows: (i) Extract all the bigrams that occurred within the collection at least 10 times. (ii) For each retrieval model and parameter setting, issue this set of bigrams as queries. For each query $\mathbf{O_q}$ (the chance the query will be issued) is $\mathbf{\frac{1}{Q}}$ meaning all queries are equally likely to be issued. (iii) Given the results, computed the $\mathbf{r(d)}$ with $\mathbf{c = 100}$ for all documents. (iv) Repeat for each model and parameter setting.

On each collection we extracted between 200,000 & 300,000 bigrams to use as queries. However, on ClueWeb, we had to use a reduced set because of the time it takes to run queries against a collection of that size. For ClueWeb, we used 50,000 queries. In [?], Wilkie and Azzopardi found that it was possible to estimate the retrieval bias using a smaller portion of queries, as the parameter setting where the bias was minimised was the same regardless of the number of queries used.

During the course of this analysis, we issued these sets of queries on 4 retrieval models, each with around 10 parameter settings, on 5 test collections, totalling approximately 50 million queries.

Retrievability Measures We computed the retrievability scores for cumulative scoring with a cut-off of $\mathbf{c = 100}$. For this work, we only used one cutoff as in [?] it was found that while varying c would result in different retrievability scores, the same trend across the parameter space was found i.e. the point at which retrieval bias was minimised was the same regardless of c.

4. RESULTS AND ANALYSIS

To analyse the data for each parameter setting, we plotted the points corresponding to the retrieval performance and the retrieval bias (as surmised by the Gini coefficient, referred to as Gini). For clarity, on each of the plots, we have excluded ClueWeb as the Gini values were in a much more confined space than on the other collections[3]. However, the same trend was observed as on the collections shown.

We also calculated the Pearson's correlation between each performance measure and the retrieval bias to determine if there was a linear relationship or not. Table ?? shows the correlations for each collection and each retrieval model. Stars (*) indicate whether the relationship was statistically significant when $p < 0.05$. Note that for the retrieval bias hypothesis to hold, we would expect to see a strong negative

[3] The reasons for confined Gini values is due to the size of the ClueWeb and the smaller set of queries used. However this does not affect the relationships we found, as shown by the Table of correlations.

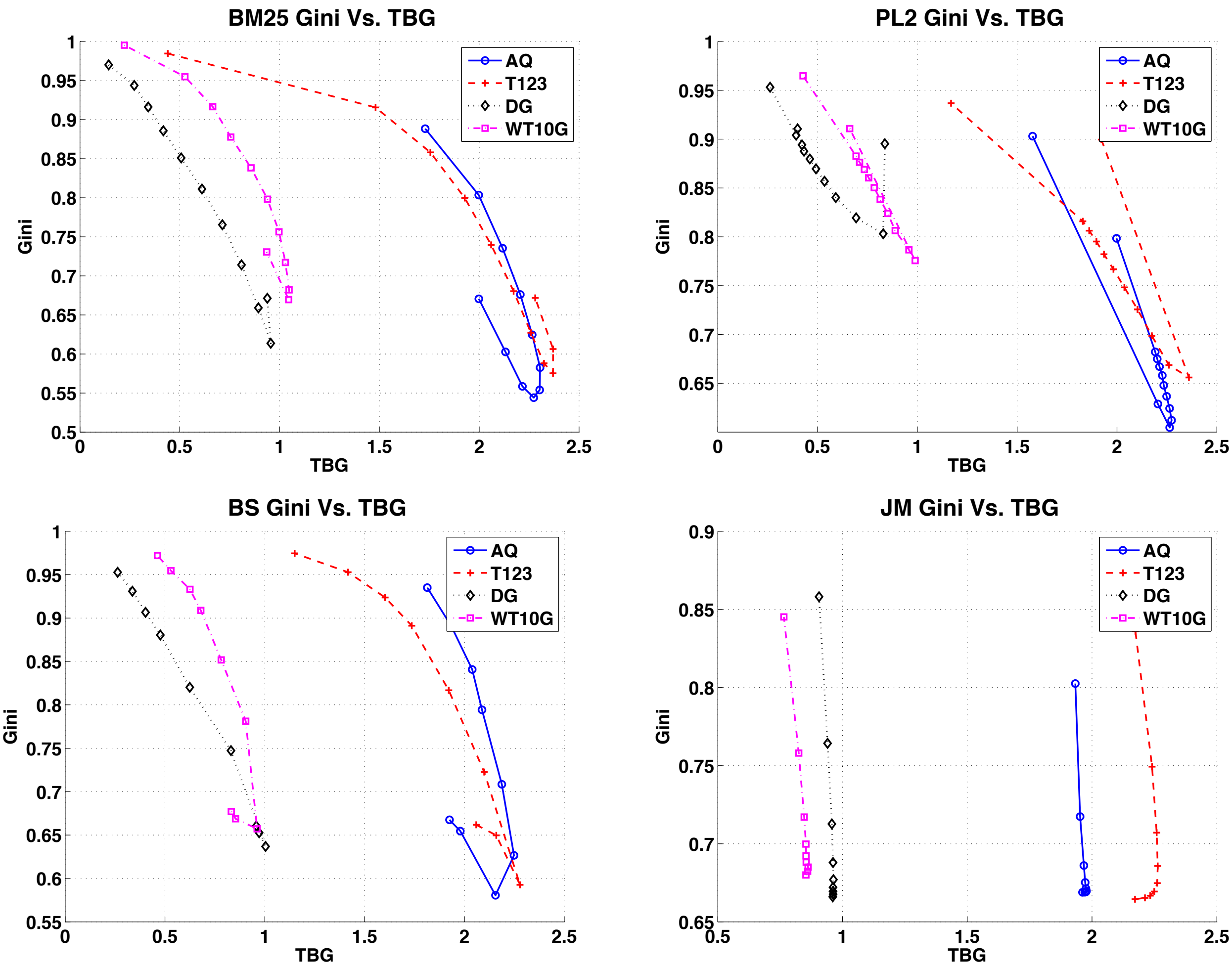

Figure 3: Plots of Gini vs. TBG for each model on AQ, T123, DG and WT10G. We see that as retrieval bias is reduced there is a corresponding increase in TBG in most cases.

correlation between performance and bias, such that as bias decreases, performance increases.

4.1 Precision Based Measures

In Figure **??**, the first row shows the plots of MRR against Gini on each collection for each retrieval model. We see similar trends between BM25 and BS when evaluating performance with MRR. The predominate trend on these two plots is that there often positive correlations between bias and MRR. However, these findings do tell us that to maximise MRR in these cases, bias is actually necessary. It is clear in these plots that the point of minimum Gini does not correspond to the point of best performance. The correlations for PL2, BM25 and BS are moderate to high while JM achieves very high negative correlations (except on AQ). When examining the JM plots we see that there is a clear negative correlation and that lowering bias increases performance but on AQ the line is almost vertical suggesting that lowering bias has no impact on performance. When bias has no impact on performance we argue that using the parameter setting with the least bias would be advisable to allow for changes in the collection.

For P@10, we observe very similar findings to MRR. These results have been replicated from [**?**]. Again, we see that the JM Language Model provides strong negative correlations even for the AQ collection. However, the results are rather mixed on the other models - where we again see both positive and negative correlations. These findings, to some extent, back up the preliminary results reported in [**?**]. However, here we also show that positive correlations exist (something not shown/known previously). These findings suggests that the relationships with precision measures (such as MRR and P@10) is highly conditional - and may or may not hold.

4.2 Precision-Recall Based Measures

With respect to MAP, we find that the correlation with retrieval bias is much the same as for P@10 and MRR. While there are a number of cases where there are strong, significant, negative correlations, there are also a number of cases where positive correlations are observed (though more often than not, not significant). The best match up was again on the JM language model for most collections. Our results on BPREF, are similar to, but slightly better than on MAP. This suggests that accounting for incompleteness in the measures does lead to improved correlations with retrieval bias. These findings re-confirm past findings that there is not a strong relationship between MAP/BPREF and retrieval bias.

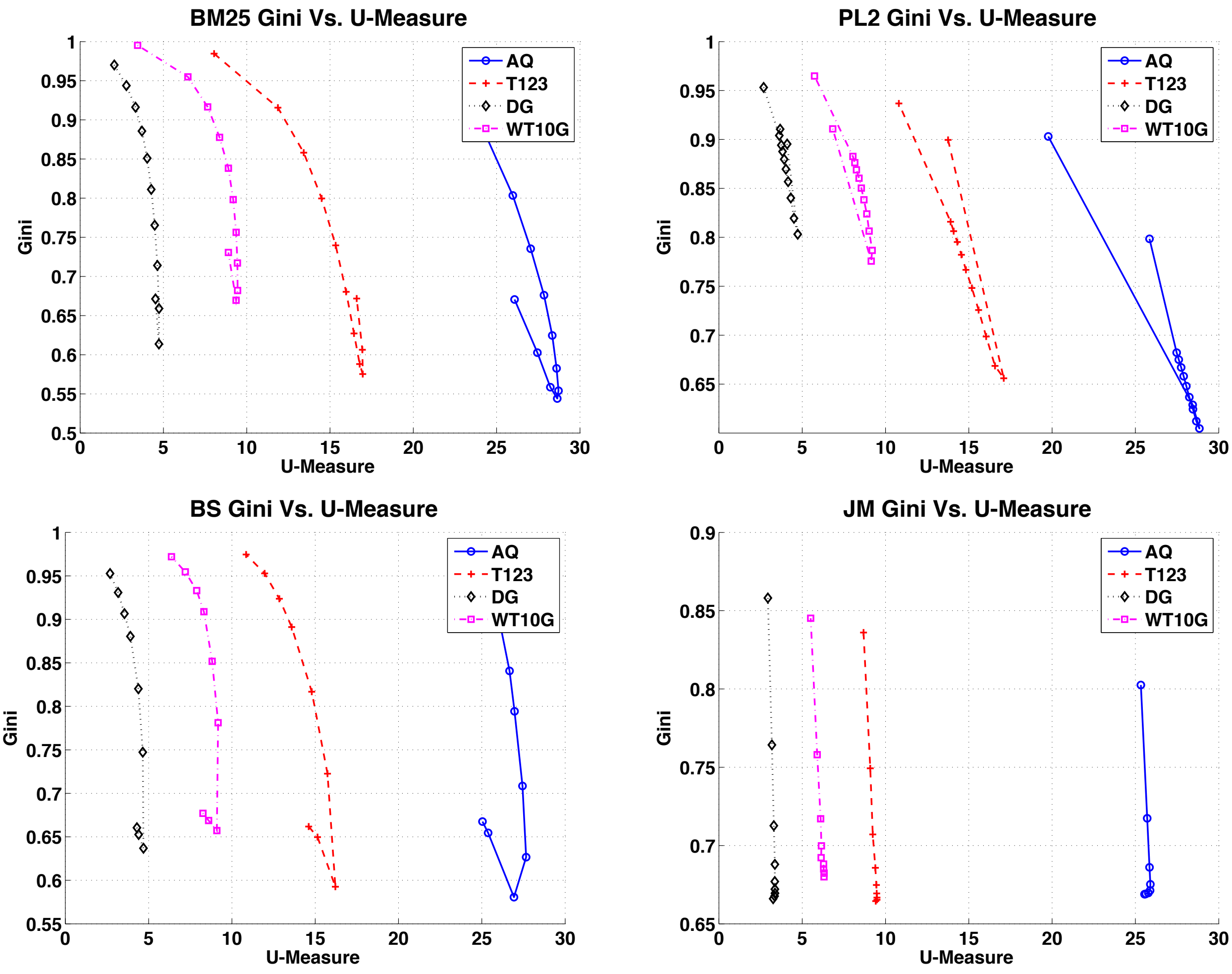

Figure 4: Plots of Gini vs. U-Measure for each model on AQ, T123, DG and WT10G. We see that reducing retrieval bias leads to increases in U-Measure.

4.3 Recall Based Measures

In [?], they found there was a strong correlation between recall and retrieval bias when ranking different retrieval models. In the context of estimating parameters for a particular retrieval model, we wonder if we will find a similar relationship. For the recall (and also REL_RET, denoted as R_R in the Table) measures we observe quite a mixed relationship. Once again there are both positive and negative correlations across BM25, PL2 and BS. However, on JM there are reasonably strong negative correlations, which are also significant on most collections. So, while the relationship tends to hold across models as shown in [?] for recall, here we show that this relationship only holds for JM.

Figure **??** presents a series of plots for PRES against Gini for each retrieval model on the second row. Again, we can see quite similar patterns between BM25, PL2 and BS for PRES where the correlations are often positive, indicating that higher levels of bias actually lead to improved performance. However, on JM there is a very strong negative correlation between PRES and Gini. It would seem that the lack of length normalisation in JM, is responsible for the differences observed between models and suggests that the type of parameter within a model may exhibit different relationships with retrieval bias. This however is left for future work.

4.4 Gain Based Measures

Now we turn our attention to the relationship between retrieval bias and the gain based measures. Row 3 of Figure **??** shows the relationship between retrieval bias and NDCG. On inspection the plots here are fairly similar to those of MRR, and so too are the correlations (where we see a mixture of positive and negative correlations, and that for the JM Language Model there are strong significant negative correlations across all collections and four match up perfectly). Again, it appears that a very mixed relationship exists between retrieval bias and NDCG.

Figure **??** shows the plots for TBG while Figure **??** shows the plots for U-Measure. What is quite striking about these plots is that for most of the models (BM25, PL2 and BS) there appears to be very strong negative correlations. On inspection of the correlation values in Table **??**, we can see that for TBG all correlations are moderate to high, negative, and all but three are statistically significant. Interestingly, the poorest correlations are observed on the JM model, which for other measures showed the strongest correlations. This strengthens the suggestion that the difference between JM

		Precision		**Precision-Recall**		**Recall**			**Gain**		
Model	**Coll.**	**MRR**	**P@10**	**BPREF**	**MAP**	**R_R**	**Recall**	**PRES**	**NDCG**	**TBG**	**U-Meas.**
BM25	**AQ**	0.66*	0.51	0.40	0.27	0.14	0.57	0.25	0.57	-0.89*	-0.92*†
	T123	-0.12	-0.03	-0.32	0.02	0.06	-0.03	-0.17	-0.03	-0.91*†	-0.94*†
	DG	-0.77*	-0.77*	-0.59*	-0.63*	-0.51*	-0.71*	-0.63*	-0.71*	-0.99*†	-0.90*
	WT	0.60*	0.25	0.14	0.44	0.60*	0.38	0.39	0.38	-0.94*	-0.84*
	CW	-0.61*	-0.08	0.18	0.16	-0.26	-0.40	0.04	-0.40	-0.94*†	-0.97*
PL2	**AQ**	-0.56	-0.55	-0.73*	-0.78*	-0.86*	-0.68	-0.79*	-0.56	-0.96*	-0.95*
	T123	-0.42	-0.51	-0.55	-0.48	-0.50	-0.29	-0.53	-0.44	-0.89*†	-0.95*†
	DG	-0.76*†	-0.87*†	-0.84*	-0.84*	-0.53	-0.10	-0.65*	-0.83*	-0.73*	-0.95*†
	WT	-0.01	-0.26	-0.20	-0.14	-0.17	-0.18	-0.21	-0.25	-0.99*†	-0.95*
	CW	-0.62*	0.49	0.76*	0.74*	0.56	0.60	0.70	0.20	-1.00*†	-0.83*
BS	**AQ**	0.78*	0.68*	0.56	0.55	0.63	0.63	0.64	0.70*	-0.64	-0.30
	T123	0.79*	0.78*	0.52	0.72*	0.79*	0.75*	0.69*	0.80*	-0.94*†	-0.90*†
	DG	-0.57	-0.31	-0.24	-0.20	0.36	0.14	0.20	-0.27	-1.00*†	-0.85*†
	WT	0.94*	0.86*	0.86*	0.88*	0.93*	0.88*	0.90*	0.89*	-0.90*†	-0.71*
	CW	-0.72*†	-0.51	0.16	0.07	0.07	-0.37	0.10	-0.65	-0.55*†	-0.64*†
JM	**AQ**	-0.13	-0.91*†	-0.97*	-0.45	-0.37	-0.85*†	-0.93*	-0.85*†	-0.94*	-0.64
	T123	-0.97*†	-0.98*	-0.98*	-0.99*	-1.00*†	-0.99*†	-0.98*	-0.99*†	-0.38	-0.99*
	DG	-0.86*†	-0.50†	-0.98*	-0.99*†	-0.98*	-0.73*	-0.97*	-0.73*	-0.98*	-0.95*
	WT	-0.93*†	-0.96*†	-0.98*†	-0.96*†	-0.97*†	-0.96*†	-0.93*†	-0.96*†	-0.99*	-0.99*†
	CW	-0.95*†	-0.88*	-0.01	-0.08	-0.40	-0.90*†	-0.18	-0.90*†	-0.38	-0.75*
Sig. Corrs.		14/20	9/20	9/20	9/20	8/20	8/20	9/20	10/20	17/20	18/20

Table 3: Correlations between Gini and the performance measure stated. * denotes statistical significance at $p<0.05$. †represents a perfect match-up between performance at minimum Gini and maximum performance

and the other retrieval models is the explanation for the contrasting results. For U-Measure, we can also see that all correlations are moderate to high and negative, and all but two are statistically significant.

The findings for these two recently proposed measures are in stark contrast to findings on all the other (traditional) measures. It would appear that since both TBG and the U-Measures account for document length, that a better match up is obtained. For these measures, minimising the bias tends to result in maximising the performance. The findings show that we can tune a system to perform very well in terms of U-Measure or TBG using the Gini scores achieved.

5. DISCUSSION AND FUTURE WORK

In this work, we examined the relationship between retrieval bias and ten retrieval performance measures. We found that the relationship depends on both the type of model and the performance measures, but is fairly consistent across collections. For precision-recall based performance measures, we generally found that the relationship was rather mixed and not consistent with the retrieval bias hypothesis for the three models where we adjusted the document length normalisation parameter (i.e BM25, PL2 and BS). We speculate that this is because the length of relevant documents in these test collections tends to be longer than the average document and a bias towards longer documents is required if the best performance, given these measures, is to be obtained. However, we did find that on the JM Language Model, that there was a strong negative and often significant correlation between retrieval bias and the performance measures, P@10, MRR, Recall, RET_REL and NDCG across all collections.

For Time Biased Gain and the U-measure, we found that for all the retrieval models tested there was a strong, negative and often significant relationship between retrieval bias and their performance. This was consistent across all the collections tested. These results support the retrievability bias hypothesis for these measures. Since Time Biased Gain and U-Measure both include parameters that are dependent on length, the fact that the relevant documents in the collections used are longer is offset by the additional effort required to process these long documents. Consequently, it would seem that by being less biased, and tuning the system such that the chance to retrieve documents is not disproportionate given length results in better performance on these measures. This, of course, is the notion behind the retrievability bias hypothesis. Operationally, this is a very useful finding, given these models it is possible to select the parameter settings that will give very good performance on TBG or U-Measure, without recourse to relevance judgements or usage data.

This work, however, also opens up a number of interesting directions for future work. So far the focus of research into retrievability has been on standard retrieval models. Here we have shown that the relationship between retrieval bias and performance is dependent on both models and measures. Further research is warranted to explore how generalisable these findings are to other models across the main measures including those which handle incompleteness. Furthermore, it would be interesting to examine models that include additional features other than terms within the model (i.e. document priors, fields, etc) or models that are derived through learning (i.e. SVM based retrieval models), and whether the relationship holds or not. In terms of the how the measures relate it would be interesting to undertake an analytical comparison between performance measures and retrievability measures to determine whether it is possible to infer performance measures from retrievability measurements. Also, in this work we have only considered one kind of inequality measure, i.e. the Gini Coefficient. In future work, it would be interesting to explore whether other measures of inequality such as the 20/20 ratio, Theil Index or Atkinson Index, would lead to a stronger correlation with retrieval performance measures and wether using different correlation measures change the results.

Acknowledgements: This work is supported by the EPSRC Project, *Models and Measures of Findability* (EP/K000330/1) and IAA Agreement/66375.