Efficiently Estimating Retrievability Bias

Colin Wilkie and Leif Azzopardi

University of Glasgow, 18 Lilybank Gardens, G12 8QQ, Glasgow, UK
{colin.wilkie, leif.azzopardi}@glasgow.ac.uk

Abstract. Retrievability is the measure of how easily a document can be retrieved using a particular retrieval system. The extent to which a retrieval system favours certain documents over others (as expressed by their retrievability scores) determines the level of bias the system imposes on a collection. Recently it has been shown that it is possible to tune a retrieval system by minimising the retrievability bias. However, to perform such a retrievability analysis often requires posing millions upon millions of queries. In this paper, we examine how many queries are needed to obtain a reliable and useful approximation of the retrievability bias imposed by the system, and an estimate of the individual retrievability of documents in the collection. We find that a reliable estimate of retrievability bias can be obtained, in some cases, with 90% less queries than are typically used while estimating document retrievability can be done with up to 60% less queries.

1 Introduction

Retrievability is a document centric evaluation measure which generates an objective score that describes the likelihood for any given document to be retrieved by a particular retrieval system. It has been applied to a variety of different contexts. A wide range of applications exist including improving recall in retrieval systems [5, 6], improving the effectiveness of pseudo-relevance feedback [4], detecting bias towards particular organisations by search engines [1] and even relating retrievability to performance [8]. In this paper we choose not to focus on the application of retrievability but rather, the investigation of how to accurately estimate retrievability. We focus on estimating retrievability because in trying to establish a complete picture of retrievability we must issue millions of queries. This obviously, is a very time consuming and computationally intensive process.

While it is likely that a large number of queries is necessary to gain an accurate measure of the retrievability of individual documents, it may be possible to use a reduced set of queries in order to estimate the relative retrievability bias that the system imposes across the collection. In this paper we shall investigate the number of queries required to produce a reasonable/useful estimate of retrievability bias and of document retrievability.

2 Background

The concept of the document-centric evaluation measure was first introduced by Azzopardi and Vinay which they branded Retrievability [2]. This measure evaluates how likely a document is to be retrieved by a particular configuration of an IR system given the universe of potential queries. The retrievability r of a document d with respect to an IR system can be defined as:

$$r(d) \propto \sum_{q \in Q} O_q . f(k_{dq}, \{c, g\}) \tag{1}$$

where q is a query from the universe of queries Q, meaning O_q is the probability of a query being chosen. $k_d q$ is the rank at which d is retrieved given q and $f(k_{dq}, \{c, g\})$ is an access function denoting how retrievable d is given q at rank cut-off c with discount factor g. To calculate retrievability, we sum the $r(d)$ across all q's in the query set Q. Obviously, it is impractical to launch every query in the universe of possible queries, as such, it is common to use a very large set of queries instead. This query set is often automatically generated bigrams [2, 8]. The intuition is then the more queries that can retrieve d before the rank cut-off, the more retrievable d is. Calculating retrievability can then be performed using a number of different user models. The simplest model being a cumulative scoring model. In the cumulative model, we employ the access function $f(k_{dq}, c)$ such that $f(k_{dq}, c) = 1$ if d is retrieved in the top c documents given q otherwise $f(k_{dq}, c) = 0$. In other words, this model accumulates a score for a document so long as it is retrieved before the specified cut-off, while documents appearing after the cut-off are ignored completely which simulates a user who is willing to rigorously look at all documents up until a set point. A more complex and realistic retrievability model exists, this gravity based scoring model applies a weighting to the documents position in the ranked list, meaning as the rank approaches the cut-off, the documents contribute less score. Formally defined as $f(k_{dq}, g) = 1/k_{dq}^{g})$ where g is a discount factor that defines the magnitude of the penalty applied to a document given its rank position, increasing as we traverse down the ranked list. The intuition behind this model being that a document further down a ranked list is less retrievable as the users interest or attention diminishes. Therefore, a document appearing at rank 1 contributes substantially more $r(d)$ than a document at rank 10 when $c > 10$. Again, if a document appears after cut-off c, that document contributes $r(d) = 0$.

Retrievability Bias

The bias that a system imposes on the document collection can be determined by examining the distribution of $r(d)$ scores. Here, bias denotes the inequality between documents in terms of their retrievability within the collection. It is possible to visually assess the inequality / bias by plotting a Lorenz Curve [7]. In Economics and the Social Sciences, the Lorenz Curve is used to visualise the inequality in a population given their incomes.

To estimate bias we treat the retrievability of a document as its wealth. Therefore, using the Gini coefficient, we can see how skewed the distribution of retrievability is towards a particular set of documents.

In the context of retrievability, if all documents were equally retrievable then the Gini coefficient would be zero (denoting equality within the population). On the other hand if only one document was retrievable and the rest were not then the Gini coefficient would be one (denoting total inequality). Usually, documents have some level of retrievability and thus the Gini coefficient is somewhere between one and zero. Many factors affect the retrievability bias, these include: the retrieval model/system, the parameter settings, the indexing process, the

documents and collection representations/statistics, as well as how the system is used by the user (i.e. the types of queries and the number of documents that they are willing to examine). Obviously, the Gini coefficient is only a overview of the presence of bias and as such it is important to generate accurate$r(d)$ scores to examine *where* the bias lies.

Estimating Retrievability

For a given retrieval system configuration (i.e. retrieval model and parameter setting), the estimation process consists of three parts: (1) the generation and selection of a sufficiently large query set (2) issuing the queries to system and (3) calculating the retrievability of the documents, and the retrievability bias of the system.

In [2], it was pointed out that in order to obtain an exact estimate of retrievability the universe of all queries would need to be issued. However, because that is not feasible, typically, a very large set of queries is used instead (either single term queries or a subset of two word queries). Most papers that perform a retrievability analysis use two or three word queries and issue anywhere between 200,000 to 2,000,000 queries [2, 6, 8].

This, is obviously a very time consuming process requiring a huge amount of resources to complete. Additionally, for each configuration this would have to be repeated.

With the increasing usage of the retrievability evaluation measure, it is important to look to optimise the process to reduce the amount of time and resources required. Doing so makes retrievability a more accessbile measure and improves its viability.

One attempt has been made to improve the efficiency of retrievability. Rather than directly estimating retrievability scores for documents Bashir proposed a method for the estimation of retrievability employing techniques from machine learning [3]. Bashir extracted document features, such as normalised average term frequency, number of frequent terms, average document frequency, etc., to then estimate how likely that document is to be retrieved without resorting to posing any queries to a retrieval system. This method of determining retrievability bias is obviously far more efficient as it goes on document rankings rather than document retrievability scores. The disadvantage is that we can only gain an insight to bias but cannot understand exactly where the bias lies in terms of which individual documents are more or less favoured than others.

3 Experimental Method

We propose a set of experiments to answer some important research questions given the hypothesis; a lower bound exists, such that increasing the number of queries issued to the system provides no additional insights in terms of bias or document retrievability. Our research questions are thus:

1. Do similar trends occur when smaller sets of queries are used?
2. How many queries are needed to gain a comparable approximation of Gini?
3. How correlated are the $r(d)$ scores between varying numbers of queries?

We are ultimately interested in minimizing the amount of processing required to calculate retrievability. While previous work has managed to achieve relatively

AQ									
	10%	**20%**	**30%**	**40%**	**50%**	**60%**	**70%**	**80%**	**90%**
BM25	0.68	0.76	0.84	0.88	0.93	0.95	0.97	0.98	0.99
PL2	0.70	0.78	0.85	0.89	0.93	0.95	0.97	0.98	0.99
DPH	0.68	0.76	0.84	0.88	0.93	0.95	0.97	0.98	0.99
TF-IDF	0.76	0.87	0.92	0.94	0.96	0.98	0.98	0.99	1.00

T123									
	10%	**20%**	**30%**	**40%**	**50%**	**60%**	**70%**	**80%**	**90%**
BM25	0.69	0.79	0.86	0.90	0.92	0.95	0.96	0.98	0.99
PL2	0.74	0.82	0.88	0.91	0.93	0.95	0.96	0.98	0.99
DPH	0.92	0.95	0.97	0.98	0.99	0.99	0.99	1.00	1.00
TF-IDF	0.97	0.99	0.99	1.00	1.00	1.00	1.00	1.00	1.00

Table 1. AQ (Top) & T123 (Bottom): Tables of correlations between all queries and the percentage of queries stated. All correlations were found to be significant at $p < 0.05$

good rankings of document retrievability, the only method for effectively estimating individual document $r(d)$ remains to be in launching these huge query sets. We investigate whether we can achieve similar Gini coefficients on different sized query sets. However, it is entirely possible for similar Gini coefficients to be calculated from very differently biased sets. Therefore, we need to explore how correlated the individual document r(d) scores are to fully understand if and where this lower bound exists.

Data and Materials

Two TREC test collections: Aquaint (AQ) and Trec123 (T123) were used in these initial experiments. Each collection contains approximately one million documents and the query sets used were created by extracting bigrams from the collection and selecting those bigrams which occurred at least 20 times, giving about a quarter of a million queries for each collection. We employed four retrieval models: BM25, PL2, DPH and TF-IDF. We considered various parameter settings for BM25 and PL2, where we performed a parameter sweep ($b = [0, 1]$ and $c = [0.1, 1, 2, \ldots, 10, 100]$) to determine which parameter value minimises the Gini coefficient as done in [8].

In [2, 8] it has been observed that BM25 is the least biased model, while TF-IDF is the most biased (overly favouring long documents). We included these models in our experiments to compare how many queries are needed to find a stable estimation for different quantities of bias in models (i.e. does a more bias model require more or less queries). We also choose PL2 to compliment BM25 as these both have adjustable parameters for length normalisation. We can then investigate whether the parameter that exhibits minimum Gini changes with the number of queries issued.

4 Results and Analysis

Do similar trends occur when smaller sets of queries are used? The plots in Figure 1 show how Gini changes as the b and c parameters are varied for BM25 (left) and PL2 (right), respectively. This is shown for different numbers of queries. It is clear that the same trend is present regardless of the number of queries used. These results shows that it is possible to use substantially less

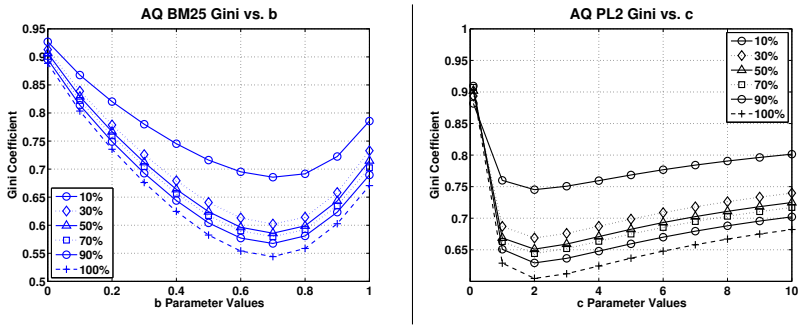

Fig. 1. AQ BM25 (Left) & PL2 (Right): Gini vs. parameter setting on BM25 and PL2.

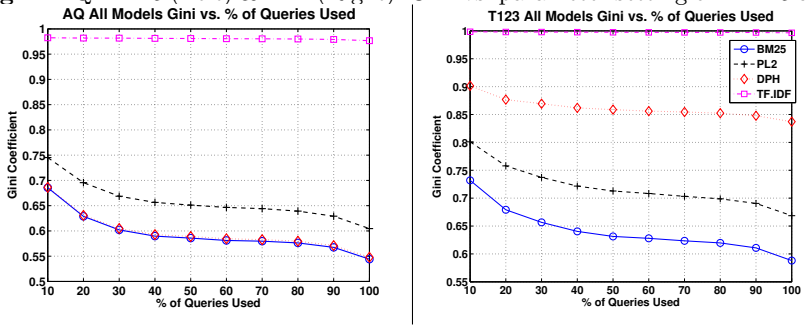

Fig. 2. AQ (Left) & T123 (Right): Gini vs. Percentage of queries used.

queries to find the parameter value that minimise Gini. In [8], this point was correlated with high retrieval performance.

How many queries are needed to gain a comparable approximation of Gini: Figure 1 shows that regardless of how many queries are issued, the parameter setting that achieves the minimum point of Gini is found to be the same (0.7 for AQ and 0.8 for T123 on BM25). This result indicates that we can significantly reduce the amount of queries used (by up to 90% in this case) and the parameter setting at which minimum Gini occurs does not change. This means we can estimate the parameter for minimum Gini reliably with 90% less queries. These results hold for both collections used but also for PL2 where the minimum point of Gini occurs at 2 for AQ and T123 for all query sets.

Examining the plots of Figure 2 we see how Gini changes as more queries are issued on the 4 models (BM25, PL2 DPH and TF-IDF). In these plots, the minimum point of Gini discovered for BM25 and PL2 in the previous plots are used here. The first point we note is, the extremely biased model (TF-IDF) shows very little change to the Gini as the number of queries increase. This tells us that a heavily biased model reaches a stable estimation of bias in very few queries and using extremely large query sets is not necessary. The information in Table 1 backs up this claim showing that on T123, the results can converge within as few as 40% of the query set.

How correlated are the $r(d)$ scores between varying numbers of queries: Looking at the less biased models presented here we see a trend develop. If too few queries are issued (Less than 40%) the estimation of bias is not particularly accurate and provides a more biased picture than larger numbers. However between 40% and 80%, the estimation is fairly stable, giving the impression that

the results have converged to a stable point but if we continue to issue more queries we see another substantial drop ($> 80\%$) in bias. This makes estimation of bias difficult as there is often an area where it appears enough queries have been launched but an accurate estimate has not yet been reached. Table 1 shows the correlations between the volumes of queries and when all queries have been sent. We can see that high correlations exist between 40% of the queries upwards suggesting this final dip in bias is not substantially different from the stable points of previous amounts of queries.

5 Summary and Conclusions

In this paper we investigated whether or not there existed a minimum amount of queries to estimate retrievability to a highly accurate degree. We found that this minimum amount is largely dependant on the retrieval model employed. Mainly the more biased a model is, fewer queries need to be issued to reach a stable estimate of retrievability. When a model is known to be fairer we see that estimating this lower bound is very difficult as plateaus exist where it appears the results have converged and become stable when there is another drop to come.

Further investigation of these findings is required to determine whether their is some link between collection size or type and how many queries must be issued. We must also investigate whether the ordering of queries plays a major impact in the estimation; for example, if we reverse or randomise the query list does this positively or negatively affects the estimation of bias.

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

References

1. Azzopardi, L., Owens, C.: Search engine predilection towards news media providers. In: Proc. of the 32nd ACM SIGIR. pp. 774–775 (2009)
2. 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)
3. Bashir, S.: Estimating retrievability ranks of documents using document features. Neurocomputing (2013)
4. Bashir, S., Rauber, A.: Improving retrievability of patents with cluster-based pseudo-relevance feedback documents selection. In: Proc. of the 18th ACM CIKM. pp. 1863–1866 (2009)
5. Bashir, S., Rauber, A.: Improving retrievability & recall by automatic corpus partitioning. In: Trans. on large-scale data & knowledge-centered sys. II, pp. 122–140 (2010)
6. Bashir, S., Rauber, A.: Improving retrievability of patents in prior-art search. In: Proc. of the 32nd ECIR. pp. 457–470 (2010)
7. Gastwirth, J.L.: The estimation of the lorenz curve and gini index. The Review of Economics and Statistics 54, 306–316 (1972)
8. Wilkie, C., Azzopardi, L.: Relating retrievability, performance and length. In: Proc. of the 36th ACM SIGIR conference. pp. 937–940. SIGIR '13 (2013)