Algorithmic Bias: Do Good Systems Make Relevant Documents More Retrievable?

Colin Wilkie
School of Computing Science
University of Glasgow
Glasgow, Scotland
c.wilkie.3@research.gla.ac.uk

Leif Azzopardi
Computer & Information Sciences
University of Strathclyde
Glasgow, Scotland
leif.azzopardi@acm.org

ABSTRACT

Algorithmic bias presents a difficult challenge within Information Retrieval. Long has it been known that certain algorithms favour particular documents due to attributes of these documents that are not directly related to relevance. The evaluation of bias has recently been made possible through the use of retrievability, a quantifiable measure of bias. While evaluating bias is relatively novel, the evaluation of performance has been common since the dawn of the Cranfield approach and TREC. To evaluate performance, a pool of documents to be judged by human assessors is created from the collection. This pooling approach has faced accusations of bias due to the fact that the state of the art algorithms were used to create it, thus the inclusion of biases associated with these algorithms may be included in the pool. The introduction of retrievability has provided a mechanism to evaluate the bias of these pools. This work evaluates the varying degrees of bias present in the groups of relevant and non-relevant documents for topics. The differentiating power of a system is also evaluated by examining the documents from the pool that are retrieved for each topic. The analysis finds that the systems that perform better, tend to have a higher chance of retrieving a relevant document rather than a non-relevant document for a topic prior to retrieval, indicating that retrieval systems which perform better at TREC are already predisposed to agree with the judgements regardless of the query posed.

1 INTRODUCTION

Algorithmic bias presents numerous challenges, in particular, within the domain of Information Retrieval [6]. For many years, researchers have been aware that performance issues are often related to algorithmic bias. For example, TF.IDF was renown for it's bias towards longer documents, spurring researchers to investigate ways to mitigate against this length bias eventually leading to Singhal *et al's* Pivoted TF.IDF [9]. On the other hand, the introduction of PageRank meant that new pages were less likely to be ranked due to the bias towards older more linked pages [4]. Many retrieval algorithms, including the state of the art, contain various biases towards particular documents. Sometimes this is beneficial to performance (or certain groups) but other times it is not.

It has been hypothesised that fairer retrieval systems are better performing systems [11]. However, this has only been shown in particular circumstances and has not been generalised. Instead of making such a broad claim, it is perhaps more realistic to pose the hypothesis that retrieval systems that contain little unwanted biases, thus being fairer, are more likely to improve performance by allowing documents to be judged purely on a query by query basis. In this work, a related hypothesis is proposed; that better performing systems actually exhibit a bias towards the relevant documents for a query, prior to retrieval. A system that performs well in terms of a TREC style performance evaluation will be more likely to retrieve relevant documents than non-relevant documents, *a priori*.

2 RELATED WORK

Retrievability was introduced as a document centric evaluation measure by Azzopardi and Vinay with the intention of evaluating the access to the collection provided by the retrieval mechanism [1]. Retrievability evaluates the likelihood that a document will be retrieved from the collection when given some arbitrary query without considering relevance. A document d has a retrievabilty score r as defined by the following equation:

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

where q is a query from the universe of queries Q, meaning O_q is the probability of a query being chosen. Then $k_d q$ is the rank at which d is retrieved given q. and $f(k_{dq}, \{c\})$ 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 $r(d)$ of a document across all q's in the query set Q. Obviously, it is impractical to launch ever 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 [1]. 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 models however it is most common to use a cumulative scoring model. In the cumulative measure, the access function $f(k_{dq}, c)$ evaluates to 1 if d is retrieved in the top c documents given q, otherwise it evaluates to 0. Intuitively, the measure is a count of number of times the document is retrieved in the top c.

Collections	AP	T45	AQ
# of Docs.	242,919	528,156	1,024,324
# Bigrams	510,019	453,722	618,964
Topics	51-200	351-400	303-689

Table 1: Collection Information and the number of bigrams issued to produce $r(d)scores$

	Rel/ NonRel			Ret.Rel/ Ret.NonRel			NotRet.Rel/ NotRet.NonRel		
	AP	AQ	T45	AP	AQ	T45	AP	AQ	T45
BM25	0.91*	0.71*	0.88*	0.95*	0.57	0.83*	-0.77*	0.44	0.63
PL2	-0.44	0.70*	0.55	-0.19	0.61	0.66	0.67*	0.15	-0.19
LMD	0.83*	0.85*	0.95*	0.25	0.89*	0.91*	-0.76*	0.26	0.84

Table 2: Table of Pearson's correlations between MAP and odds of retrieving relevant over non-relevant for the different groups. * represents a statistically significant correlation where $p < 0.05$

Retrievability Bias

The bias that systems impose on the document collections 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. In Economics and the Social Sciences, the Lorenz Curve is used to visualise the inequality in 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 we would expect this cumulative distribution to 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 [5] is used.

In the context of retrievability, if all documents were equally retrievable 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, the Gini coefficient would be one (denoting total inequality). Many factors affect the retrievability bias (denoted by the Gini coefficient). These include: the retrieval model, 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, denoted by the c parameter).

The relationship between retrievability bias and performance has been examined in various contexts (e.g. web, news, patents, archives, etc. [1–3, 8, 10–12]) and across number of different factors (query length, document length and document features [1, 12], query expansion [2], retrieval algorithms [11], over time [8], etc.) Within these works, the retrievability bias (summarised by Gini) has been correlated with performance to better understand the relationship between bias and performance. For example, in [12], Wilkie and Azzopardi explored how length normalisation parameters changed the bias of the system and how it related to various performance measures. They found a moderate correlation with bias for P10, MAP and NDCG measures and a strong correlation with bias for TBG and U-Measure - such that reducing bias lead to better performance. Similarly, in [3], Bashir and Rauber found a strong correlation between bias and recall. In a comparison across algorithms, Wilkie and Azzopardi, hypothesised that fairer systems may lead to better performance - again they showed that there was a strong correlation such that selecting a system based on the lowest bias would tend to correspond to good performing system. Rather than examining bias at the system level, in this work, we consider the bias exhibited by systems towards the set of relevant and non-relevant documents and consider at the document level the relationship with performance.

3 EXPERIMENTAL METHOD

The purpose of the experiments performed in this work is to generate a set of average retrievability scores for subsets of the collection. Namely, for each topic, the average retrievability is calculated for the Retrieved Relevant documents (Ret.Rel), Retrieved Non-Relevant documents (Ret.NonRel), Not Retrieved Relevant documents (NotRet.Rel) and the Not Retrieved Non-Relevant documents (NotRet.NonRel).

3.1 Research Questions

Given the hypothesis that better performing systems exhibit a bias towards the relevant documents, the following research question was derived: Do systems with better performance also make the relevant documents more retrievable than the non-relevant documents? This question is investigated across three different aspects: (1) Rel vs NonRel, (2) Ret.Rel vs Ret.NonRel and (3) NotRet.Rel vs NotRet.NotRel

3.2 Data and Materials

For our analysis we used three TREC collections using three parameterised retrieval algorithms. The four collections employed are Associated Press 88-90 (AP), Aquaint1 (AQ), and TREC disks 4 and 5 (T45). Details of these collections can be found in Table 1. The three retrieval algorithms featured are BM25, PL2 and Language Modelling with Dirichlet Smoothing (LMD), all implemented in the lucene4ir [1] search package, based on Lucene. For tuning the parameters for BM25, PL2 and LMD, a parameter sweep is performed

[1]Code is available at: https://github.com/lucene4ir/lucene4ir

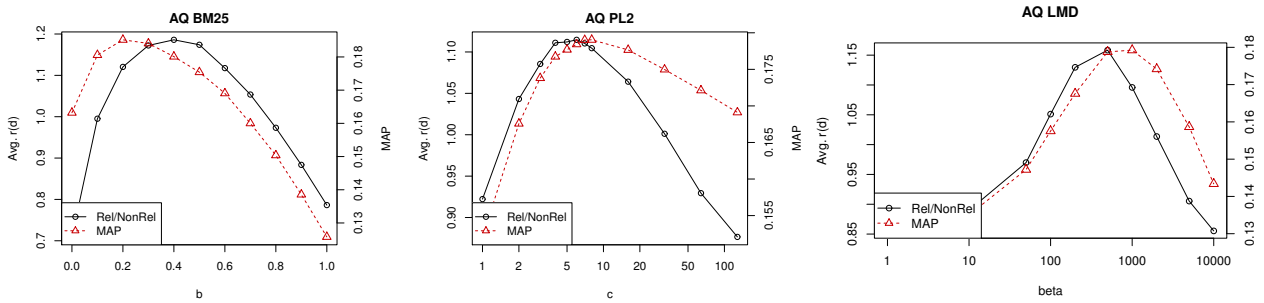

Figure 1: The MAP and Odds of retrieving a Rel over a NonRel given the model parameters (left: BM25, middle: PL2, right: LMD) for the TREC Aquaint collection. As the Odds increases, performance also tend to increase.

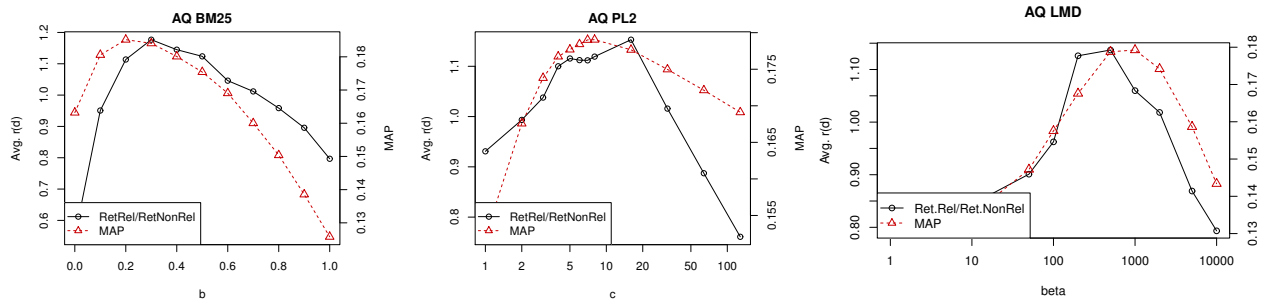

Figure 2: The MAP and Odds of retrieving a Ret.Rel over a Ret.NonRel given the model parameters (left: BM25, middle: PL2, right: LMD) for the TREC Aquaint collection.

across their b, c and β parameters, respectively, to allow insights into the effects these parameters are having on document retrievability. MAP is calculated for each topic on each collection using each model and used in the analysis stage as an indicator of system performance.

3.3 Retrievability Analysis

To compute the retrievability scores for documents, we first generated queries from the collection and then issued the queries to each of the different configurations (collection, retrieval model, parameter setting). The method used for generating queries was as follows. The collections were indexed in the lucene4ir framework. Documents were tokenised using a shingle tokeniser which creates shingles of 2 terms to index. This tokeniser removed stop words, applied porter stemming and only accepted terms longer than 3 characters long before stemming. A list of bigrams was then generated from the index by returning the shingles indexed along with their document frequencies and collection frequencies. Bigrams that occur 4 or more times were taken, returning a sizeable list of bigrams to be used in the retrievability estimation (see Table 1).

Each index was then queried with the bigram queries using the chosen retrieval models and parameter settings, generating results list of up to 100 documents for each query issued. Following this, the results lists were used to compute the retrievability of each document using the cumulative measure, given Equation 1 where $c = 100$.

The QREL file associated with each collection was then used to identify the relevant and non-relevant documents for each topics. The documents $r(d)$ scores were extracted from the full list of $r(d)$ scores and then averaged for each of the different sets: Ret.Rel, Ret.NonRel, NotRet.Rel and NotRet.NonRel for each topic.

To compute the performance of each system, we used the TREC topic titles as the query for each topic. When discussing relevant and non-relevant documents, only those included in the QREL file were considered. Un-judged documents were excluded from the analysis reported here.

4 RESULTS

Results of the experiments detailed in Section 3 are presented in the following subsections, breaking down the research question to examine the three aspects of retrievability and performance. Due to space limitations we only present the plots for the AQ collection, however, Table 2 provides the correlations across all the collections and models used. The plots presented show the MAP scores across the parameter sweeps as well as the odds of retrieving a relevant item over a non-relevant item, given the model, collection and parameter setting.

4.1 Rel vs NonRel

Figure 1 shows plots of the *p*riori Odds of retrieving relevant (Rel) vs. Non-Relevant (NonRel) documents given the pool as the model parameters change - and the corresponding change in MAP. It can

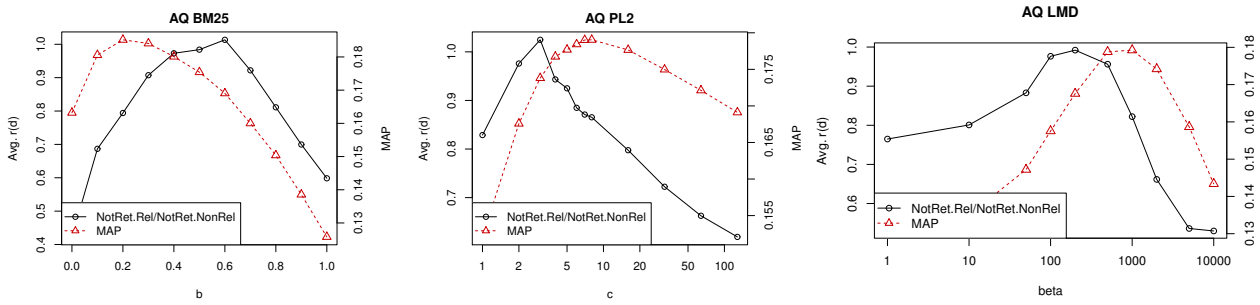

Figure 3: The MAP and Odds of retrieving a Rel.NotRet over a NonRel.NotRet given the model parameters (left: BM25, middle: PL2, right: LMD) for the TREC Aquaint collection. As expected there is less of a correlation between NotRet.Rel/NotRet.NonRel and MAP.

be seen that as Odds increases, so too does the MAP, however, for most models, there is a small offset between when the Odds peaks and when MAP peaks. These plots, however, suggest that making relevant items more retrievable than non-relevant items tends to lead to better performance. Interestingly, when the Odds of $Rel/NonRel > 1.0$ the performance is always better that when the Odds is $Rel/NonRel < 1.0$.

Table 2 reports the Pearson's correlation between the performance and the Odds showing that for three of the collections there is a strong positive (and significant correlation) for most of the models. Also apparent is that BM25 and LMD exhibit greater correlations than PL2 yet all have comparable MAP scores indicating that systems can also perform well without strongly favouring relevant over non-relevant.

4.2 Ret.Rel vs Ret.NonRel

Figure 2 shows plots of the Odds of retrieving relevant and retrieved (Rel.Ret) vs. non-relevant and retrieved (NonRel.Ret) documents. As above, we see a similar relationship to the plots in Figure 1. Given this subset of documents, i.e. the set of documents actually retrieved, we can see that there is greater agreement, and now the best performing configuration is more closely related to the Odds of relevant vs. non-relevant. While they tend to match up better, the correlation, is slightly weaker suggesting that there is greater mis-match in other areas of the space. There are also fewer significant correlations possibly meaning the relationship is not as stable here, or different way of analysing the data would be more appropriate.

4.3 NotRet.Rel vs NotRet.NonRel

Finally, Figure 3 shows plots of the Odds of retrieving the relevant and not retrieved (NonRet.Rel) vs non-relevant and not retrieve (NotRet.NonRel). Here we see, that the there is greater disparity between the Odds and MAP. This is perhaps to be expected, because these relevant items are not contributing to the MAP score. Interestingly, the Odds tends to be below one across each model (where as for the other aspects the Odds exceeded one, and corresponded to good performance). This suggests that these subset of relevant items at best had an equal chance of being retrieved.

5 DISCUSSION AND FUTURE WORK

The results presented provide some new insights into how the retrievability of relevant and non-relevant documents across three aspects relates to performance. The findings suggest that good systems do tend to make relevant documents more retrievable. Intuitively, this makes sense, if we tune our system, such that the relevant documents are more likely to be retrieved, then the system should perform better. However, doing so, is likely to increase the overall bias, as expressed by Gini, for instance. And so, may have dire consequences on the retrieval performance of other sets of topics. In future work, it will be of interest to explore this relationship further with respect to the overall system bias and with respect to other performance measures, collections and across individual topics.

Acknowledgments Supported by EPSRC, grant №. EP/L50497X/1.

REFERENCES

[1] Leif Azzopardi and Vishwa Vinay. 2008. Retrievability: An Evaluation Measure for Higher Order Information Access Tasks. In *Proc. of the 17th ACM CIKM*. 561–570.
[2] Shariq Bashir and Andreas Rauber. 2009. Improving retrievability of patents with cluster-based pseudo-relevance feedback documents selection. In *Proc. of the 18th ACM CIKM*. 1863–1866.
[3] Shariq Bashir and Andreas Rauber. 2010. Improving retrievability of patents in prior-art search. In *Proc. of the 32nd ECIR*. 457–470.
[4] Junghoo Cho and Sourashis Roy. 2004. Impact of Search Engines on Page Popularity. In *Proceedings of the 13th International Conference on World Wide Web (WWW '04)*. ACM, New York, NY, USA, 20–29. http://doi.acm.org/10.1145/988672.988676
[5] J Gastwirth. 1972. The Estimation of the Lorenz Curve and Gini Index. *The Review of Economics and Statistics* 54 (1972), 306–316. Issue 3.
[6] Keith Kirkpatrick. 2016. Battling Algorithmic Bias: How Do We Ensure Algorithms Treat Us Fairly? *Commun. ACM* 59, 10 (2016), 16–17.
[7] David E. Losada, Leif Azzopardi, and Mark Baillie. 2008. Revisiting the relationship between doc. length and relevance. In *Proc. of the 17th ACM CIKM'08*. 419–428.
[8] Thaer Samar, Myriam C Traub, Jacco van Ossenbruggen, Lynda Hardman, and Arjen P de Vries. 2017. Quantifying retrieval bias in Web archive search. *International Journal on Digital Libraries* (2017), 1–19.
[9] Amit Singhal, Chris Buckley, and Mandar Mitra. 1996. Pivoted document length normalization. In *Proce. of the 19th ACM SIGIR conference (SIGIR '96)*. 21–29.
[10] Colin Wilkie and Leif Azzopardi. 2013. Relating retrievability, performance and length. In *Proc. of the 36th ACM SIGIR conference*. 937–940.
[11] Colin Wilkie and Leif Azzopardi. 2014. Best and Fairest: An Empirical Analysis of Retrieval System Bias. *Advances in Information Retrieval* (2014), 13–25.
[12] Colin Wilkie and Leif Azzopardi. 2014. A Retrievability Analysis: Exploring the Relationship Between Retrieval Bias and Retrieval Performance. In *Proc. of the 23rd ACM CIKM*. 81–90.