The Impact of Fielding on Retrieval Performance and Bias

Colin Wilkie
University of Glasgow, United Kingdom.
c.wilkie.3@research.gla.ac.uk

Leif Azzopardi
University of Strathclyde, United Kingdom.
leif.azzopardi@strath.ac.uk

ABSTRACT

Within many domains, such as news, medicine and patent, documents contain a variety of fields such as title, author, body, source, etc. As such fielded retrieval models that query across fields are often employed. It is largely presumed that fielding provides a better representation of the document and offers more control when querying, and that this will lead to improved retrieval performance. However, depending on how the fields are weighted and if the fields are populated, the retrieval algorithm may unduly favour certain documents over others. This is known as algorithmic bias and it may be detrimental to retrieval systems performance. In this paper, we explore the impact of fielding on retrieval bias and performance across a variety of TREC News Test Collections. We perform an extensive large-scale analysis on two types of fielded retrieval model variations that are based on the popular **BM25** retrieval algorithm where either: fields are scored independently and then combined (**Model 1)**, or fields are first combined and then scored (**Model 2)**. Our findings show that for **Model 1** fielding, a strong correlation exists between retrieval bias and performance such that as title fields are weighted more heavily, bias increases, while retrieval performance decreases. When weighting is applied to content-based fields, performance increases as bias decreases, showing that relying more on content may be favourable in terms of fairness and performance. On the other hand, for **Model 2** fielding, the relationship between retrieval bias and performance is more complex. But, crucially we show that **Model 2** fielding results in lower retrieval bias and greater performance than **Model 1** fielding. And, we observed that under **Model 1,** news articles without titles are substantially less retrievable (i.e. more susceptible to algorithmic bias). These findings have serious ramifications as many popular Open Source Information Retrieval frameworks, commonly used by professional searchers, use the default implementation of **Model 1** for their fielded search capability. This research shows the importance of analysing retrieval algorithms with respect to both bias and performance to ensure they minimize any unwanted or unintended biases when maximising performance. Further work is required to examine this phenomenon in more detail and to design fielded retrieval models that have the advantages of control and performance without detrimental biases.

KEYWORDS

Algorithmic Bias, Information Retrieval, Search Engine Bias

INTRODUCTION

In the field of Information Retrieval (IR) the focus has typically been on trying to make search as effective and efficient as possible. However, depending on the algorithmic choices, the data and the configuration of the information retrieval system, algorithmic biases can creep in that can adversely affect the performance of the system by treating particular users or groups of documents unfairly (Kirkpatrick, 2016). Researchers have argued for the need of legislative controls to be applied to retrieval systems to reduce the level of bias present in search results (Goldman, 2005). Thus, it is also important to analyze and evaluate the bias within a system and how it relates to the performance of the system. In this work we aim to do so in the context of fielded retrieval. Treating a document as a bag-of-words greatly simplifies how documents are modeled and represented. However, in many domains, documents contain structured meta-data and various fields that characterize the contents of the document (Robertson, Zaragoza and Taylor, 2004). For example, email documents contain subjects, body, to and from, while news articles contain title, source, contents, author, etc. Whether it be patent, medical, academic, email or news, documents typically have a title and contents field that can be used to identify relevant material. The premise is that having a more structured representation of the document provides searchers with more control when querying a collection (Kim, Xue and Croft, 2009). Often query terms are related to specific fields within a document (Azzopardi, De Rijke and others, 2006) so query terms are mapped either explicitly or implicitly to fields and/or the importance of fields are weighted in order to improve retrieval performance.

In terms of ranking based on fielded documents, various retrieval algorithms have been developed including **BM25F** (Robertson, Zaragoza and Taylor, 2004) and variants of (Itakura and Clarke, 2010; Blanco and Boldi, 2012), Fielded Language and Relevance Models (Ogilvie and Callan, 2003; Azzopardi, De Rijke and others, 2006; Kim and Croft, 2012), Fielded Multinomial Randomness Models (Plachouras and Ounis, 2007), etc. These retrieval algorithms are used to score the document given the set of fields ***f***, by either: (**1**) a weighted combination of field level scores (Kim and Croft, 2012), or (**2**) a combination of weighted fields, which is then scored (Robertson, Zaragoza and Taylor, 2004). While both have merits, the first approach to fielding has been criticized because it can lead to an imbalance due to how the term frequencies are handled on a per field basis and that terms missing from particular fields may lead to a greater disparity in scores. This could lead to retrieval biases creeping in that adversely affect the performance of the system. In a recent study on the second approach to fielding, it was shown that field weights can have a major impact on performance and that any improvements are very much dependent on the task, the domain and are sensitive to parameter tuning (Jimmy, Zuccon and Koopman, 2016). It is currently an open question whether fielding (and the different approaches to fielding) affect retrieval bias and consequently retrieval performance. Thus, we posit that fielding may introduce systematic algorithmic biases that are detrimental to retrieval performance when fielding is incorrectly applied. To explore whether this is the case we perform the first investigation on how fielded retrieval affects retrieval performance and retrieval bias given the two types of approaches to fielding. Given the concerns over the first approach, we contend that the second approach will be fairer i.e. it will exhibit less bias across the population of documents, and lead to greater retrieval performance. Our working hypothesis is that a fairer system will lead to better retrieval performance (Wilkie and Azzopardi, 2014a). Given that the most widely used Open Source IR systems, such as Elastic, Solr and Lucene, currently provide fielding out of the box using the first approach by default, it is important to study the implications of this choice. Thus, this work is both relevant and timely and will provide much needed guidance on how algorithmic biases stemming from fielding can be avoided or minimized.

RELATED WORK

Retrieval algorithms have generally been developed with a particular goal in mind; to increase performance. However, unintentional biases can appear in retrieval algorithms that can unwittingly impact performance. For example, earlier variants of **TF.IDF** either overly favored short documents or overly favored long documents depending on whether the term frequency was normalized (i.e. *favoring short documents*) or not (i.e. *favoring long documents*). Thus, pivoted length normalization was proposed as a way to mitigate length bias (Singhal, Buckley and Mitra, 1996). Now length normalization is common to most retrieval algorithms. When scoring documents, retrieval algorithms tend to have two main goals; first, to estimate the query terms information content and second, to apply length normalization factor to penalize long (or short) documents. Models which focus only on one component tend to lead to exhibit higher levels of retrieval bias (Wilkie and Azzopardi, 2017a).

However, the inclusion of fields introduces a new layer of complexity when scoring a document as the weights between and within fields will have an impact on retrieval performance and may also introduce additional harmful biases. In this paper, we intend to explore whether there is a relationship between retrieval bias and performance given the different types of fielding models. First, we will define precisely what we mean by retrieval bias and how it is derived from measures of retrievability. Following this, we will outline previous work examining the relationship between retrieval bias and performance in other contexts before formally describing the two different approaches to scoring fielded documents.

Retrievability and Retrieval Bias

The concept of retrievability was first introduced by Azzopardi and Vinay (Azzopardi and Vinay, 2008b, 2008a) to describe and measure how easily a document can be retrieved when using a particular configuration of an IR system. The measure of retrievability ***r*** of a document ***d*** with respect to an IR system was formally defined as:

$$r(d) \propto \sum_{q \in Q} O_q \cdot u(k_{dq}, c, a)$$

where the retrievability ***r(d)*** is calculated by taking the sum over the universe of all possible queries ***Q***, weighted by the probability of each query ***q*** being chosen, O_q multiplied by the utility access function, $u(k_{dq}, c, a)$, which denotes how easily ***d*** can be found given the query ***q***. The utility function takes three parameters: the rank of the document given the query k_{dq}, a rank cut-off ***c*** and discount factor ***a***. Two different utility functions have been proposed, which ascribe a value to how easily a document can be retrieved for a given query. The simplest is a cut-off based scoring model (or cumulative model), where the utility access function $u(k_{dq}, c, a)$ is defined such that $u(k_{dq}, c, a) = 1$ if ***d*** is retrieved in the top ***c*** documents given the query ***q***, otherwise $u(k_{dq}, c, a) = 0$. In other words, this model accumulates score for a document so long as it is retrieved before the specified cut-off while documents appearing after the cut-off are ignored. This simulates a user who is willing to look at all documents up until a set point (e.g. all the documents on the first page of results).

An alternative utility function is the gravity-based scoring model, which simulates a user who is more likely to examine documents higher in the ranking, and less likely to examine documents further down. Formally the gravity-based utility function can be defined as:

$$u(k_{dq}, c, a) = \frac{1}{[k_{dq}]^a}$$

where a is a discount factor that defines the magnitude of the penalty applied to a document given the document's rank position, when $k_{dq} < c$, otherwise $u(k_{dq}, c, a) = 0$. The idea behind the gravity-based model is that a document further down the ranked list is less retrievable because users are less likely to visit deeper ranks. Therefore, a document appearing at rank **1** contributes substantially more to the overall retrievability, ***r(d)***, than a document at rank **10** or rank **100**. If the document appears beyond the cut-off ***c***, it does not contribute anything to the documents overall retrievability.

The intuition behind the retrievability score is that documents that are retrieved at higher ranks by many queries are more retrievable. While, documents that few queries retrieve at a sufficiently high rank to be encountered by a user, modeled by ***c*** and ***a***, are less retrievable. Assuming we can calculate the ***r(d)*** for a collection of documents, it is possible to examine the overall bias across the population of documents and determine if one set of documents is more retrievable than other. Put another way, does the retrieval system unduly favor one set of documents over another set of documents? To obtain a high-level view of the retrieval bias that a system imposes on the document collection, the *Lorenz Curve* (Gastwirth, 1972) can be used. The Lorenz Curve visually depicts the inequality within a population of documents (according to their ***r(d)*** scores). In Economics and the Social Sciences, the Lorenz Curve is used to show the inequality within a population given their income or wealth. 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 this cumulative distribution will be linear (and thus denotes equality). 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. The most extreme case being, when a monarch has all the wealth, while the peasants have none. To summarize the inequality of such distributions the *Gini Coefficient* is used (Gastwirth, 1972). The Gini Coefficient is essentially the area under the Lorenz Curve, **B**, divided by the area under the Line of Equality (**A+B**) as shown in Figure **1**.

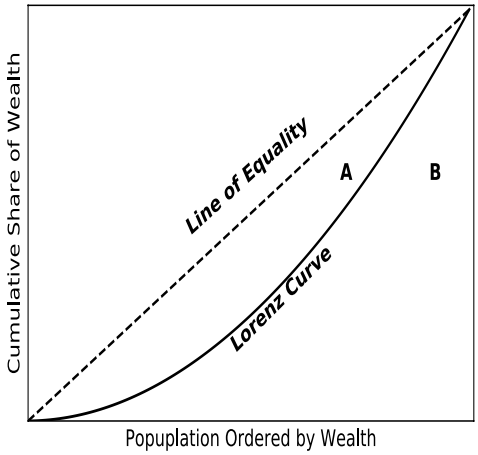

Figure 1: Lorenz curve demonstrating how inequality is depicted graphically.

In the context of retrievability, if all documents were equally retrievable then the area ***A*** in Figure **1** would be zero, and thus the Gini coefficient would be zero (denoting equality within the population). On the other hand, if the retrieval system only ever retrieved one document, and it was always the same document, while all other documents were not retrieved (and thus had a retrievability ***r(d)* = 0**), then the area **B** in Figure 1 would be equal to zero, and 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 retrieval bias (denoted by the Gini coefficient) these include:

- The indexing processes,
- The retrieval model/system and its parameter settings,
- The documents and collection representations/statistics,
- The type and length of queries, and,
- The number of documents users are willing to examine.

In this paper, we explore the impact of fielding and approaches to fielding on retrieval bias.

Retrieval Bias and Performance

The relationship between retrieval bias and performance has been examined in various contexts (e.g. web, news, patents, archives, etc. (Azzopardi and Vinay, 2008b; Azzopardi and Owens, 2009; Bashir and Rauber, 2009a) and across number of different factors (query length, document length and document features (Wilkie and Azzopardi, 2013, 2015; Chen, Azzopardi and Scholer, 2017), query expansion (Bashir and Rauber, 2009b; C. Wilkie and Azzopardi, 2017b), retrieval algorithms (Bashir and Rauber, 2014; Wilkie and Azzopardi, 2014a; Lipani *et al.*, 2015) and over time (Traub *et al.*, 2016; Samar *et al.*, 2017) etc.) From this body of work there have been several key findings: first, different retrieval systems exhibit different levels of retrieval bias but the **BM25** retrieval algorithm tends to be the fairest when tuned (i.e. lowest Gini). Second, how the retrieval model is parameterized and configured has a considerable impact on what documents are favored over others, leading to greater or lesser amounts of bias. And third, there is strong correlation between performance and retrieval bias such

that reducing bias tends to lead to better retrieval performance, leading to the *fairness hypothesis*. This hypothesis suggests that a system that affords access to all documents, and thus is fairer, will lead to better performance. The idea is that if a system makes some documents hard to retrieve i.e. ***r(d)*** either zero or close to zero, then if a user ever wanted to retrieve those documents it would very difficult for the user pose a query which would surface a document at a rank sufficiently high enough for them to encounter it. Thus, a system that affords some level of retrievability to all documents as equally as possible gives some way to all (or most) of the documents being retrieved at a rank where the user will see it. However, prior research has only considered standard non-fielded retrieval algorithms and has not investigated the impact of fielding and fields on retrieval bias nor whether the fairest hypothesis holds in this context.

Fielded Retrieval Models

Fielded retrieval became popular due to the idea that particular sections of documents may contain more information content than other areas. The simplest example is for news story collections where fields such as title, content and source are often readily available. Since the title of the document is written to attract the attention of readers, it is logical to reason that the short titles will contain keywords which are very relevant to the article. As such, when querying these articles, having a part of your query specifically target this field could yield improvements in performance. However, this has been shown to be dependent on many factors including the implementation of fielded retrieval that is employed. As previously mentioned there are two main approaches when scoring fielded documents:

Model 1: where fields are scored independently and then combined to form the overall document score.

Model 2: where the terms within fields are first combined and then scored to provide the overall document score.

The baseline approach is to treat the fields as one big bag of words – a standard retrieval model – **Model 0,** where the score of a document is:

$$s(q, d) = \sum_j s(n(q_j), d)$$

Where the query q is composed of **j** terms q_j and $s(n(q_j), d)$ is the score assigned by the retrieval model i.e. **BM25**, **TF.IDF**, **PL2**, etc, given the number of times the terms q_j appears in the document $n(q_j)$. The **Model 1** fielded retrieval approach, however, scores the fields in the document first, and the totals up the weighted field scores as follows:

$$s(q, d) = \sum_i w_i \sum_j s(n(q_j, f_i), d)$$

Where the document is composed of i fields, f_i and the $s(n(q_j, f_i), d)$ is the score assigned by the retrieval model, given the number of times the term q_j appears in field f_i in document i.e. $n(q_j, f_i)$ and w_i is the weight assigned to the field f_i.

Robertson *et. al.* proposed a revision of this model, that altered how the fielded scores were combined (Robertson, Zaragoza and Taylor, 2004; Robertson and Zaragoza, 2009). The authors argued that simple linear combinations of the scores could interrupt the saturation of term frequencies across the fields of the collections thus negatively impacting performance. The authors performed experiments on news collections that featured title and content, and found that their alternative combination term frequencies from fields greatly improved upon the weighted linear combination. The **Model 2** approach scores the document as follows:

$$s(q, d) = \sum_j s(n_j, d)$$

Where:

$$n_j = \sum_i w_i \cdot n(q_j, f_i)$$

Where $n(q_j, f_i)$ is the number of times the term q_j appears in the field f_i in the document. However, **Model 2** fielding is not without its problems either. In Jimmy *et al.* they explored the influence of boosting the weights of different fields and found that significant reductions in performance can result depending on the configuration (Jimmy, Zuccon and Koopman, 2016). They posit that weighting title over content can have a negative impact on retrieval for exploratory queries whereas for navigational queries, title boosting often improves performance. Neither of these investigated the impact fielding has on retrieval bias. In this work, like the aforementioned studies, we will used **BM25** to provide the score between query terms and fields and/or documents and explore the influence of field weighting and the fielding approach on retrieval performance and bias.

EXPERIMENTAL METHOD

Research Questions

The primary research question of this study is: what is the relationship between the different approaches to fielding, retrieval performance and retrieval bias? More specifically:

- How does boosting the weights of fields affect retrieval bias and retrieval performance?
- How retrievable are documents when they have missing fields?
- Which fielding approach is more robust?

While exploring these questions, we will also examine whether the *fairness hypothesis* holds in this context and uncover if the fielding approach that is fairer (i.e. lower retrieval bias) also result in better retrieval performance?

Data and Materials

For this analysis, three TREC News test collections were employed as they contained a title field and a content field: Associated Press 88-89 (**AP**), TREC volumes **4&5** (**T45**) and Aquaint (**AQ**) (see Table **1** for the collection statistics). The collections were indexed using a single term tokenizer with Porter stemming. Additionally, all stop words were removed as were terms less than 3 characters long. Several indexes were created, one to represent **Model 0** that contained one field, "all", which housed both title and content i.e. the baseline bag of words approach). Another was created that contained, "title" and "content" fields separately which was used for **Model 1** experiments, where each field could be independently weighted. Then additional indexes were created which contained one field to house the weighted combination of "title" and "content" fields to support **Model 2** experiments. For the purposes of our experiments we employed a number of different weightings: ($w_{title} = 1$ and $w_{content} = 0, 1, 2, 4, 8$) and ($w_{content} = 1$ and $w_{title} = 0, 1, 2, 4, 8$) to explore how first, increasing the content weighting affects performance and bias, and second how increasing the title weighting affects performance and bias. The retrieval algorithm used across all fielding models was **BM25** (Robertson and Zaragoza, 2009) that has a length normalization parameter b which was set to 0**.75**, unless stated otherwise. All experiments were constructed using the Lucene4IR[1] add-on for experimental evaluations in IR, based on Apache Lucene.

	AP	T45	AQ
Documents	164146	527559	1032673
Topics	51-200	351-400	303-689 (50 topics)
Missing Titles	7541	55733	55080

Table 1. Collection statistics for each of the TREC test collections used – along with the number of missing titles.

Retrievability and Retrieval Bias Estimation

To estimate the retrievability of the documents in each collection, we followed the methodology of Azzopardi and Vinay (2008). Since it is impractical to issue every query in the universe of all possible queries an approximation of the r(d) scores is obtained by using a very large set of queries instead. In prior work, automatically generated bigrams are extracted from the collection (Azzopardi and Vinay, 2008b; Bashir and Rauber, 2011; Wilkie and Azzopardi, 2014b; Samar *et al.*, 2017). For each collection, we extracted the top **100,000** collocations (bigrams) given their point wise information score i.e. common phrases (mostly people, places and events) that are prevalent in the collection and are likely to be issued as queries. This formed the query set Q. These queries were then issued to each of the different indexes, where we recorded the documents and rank of the documents for each query. To calculate the $r(d)$ scores, we used the cumulative based scoring function, where $c=100$. Note that other cut-offs were used as well that led to similar findings as those reported below. Given the $r(d)$ values for a given fielding approach and configuration, the Gini coefficient was then calculated to provide an estimate of the retrievability bias.

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

Performance Estimation

To estimate the retrieval performance for a given fielding approach (and configuration i.e. fielding weighting) we issued the corresponding TREC topics[2] and their relevance judgements. We then calculated the Mean Average Precision (**MAP**) and the Rank Biased Precision (**RBP**). However, due to space constraints we only report **MAP** scores to accommodate for comparison with prior work but note that we obtained similar findings for **RBP**.

RESULTS

To provide an overview of our main findings, we will focus reporting the results for collection **AQ** though our results for the other collections are similar (see Table **2** for an overview). Figure **2** provides an overview of the Lorenz curves for **Model 1** & **2** where we show how the inequality across the population of documents changes when the fielding boosting is varied: **T1-C1**, **T1-C8** and **T8-C1**. Of note is that for **Model 1**, changes in field boosts lead to greater changes in inequality – such that boosting the title leads to the greatest inequality – while boosting the content leads to lower inequality. On **Model 2** we can see that there is less variation in terms of inequality and that boosting either field leads to slightly less inequality. Note that **Model 2**, **T1-C1** is equivalent to **Model 0** (the baseline) and so under **Model 2** fielding appears to help in reducing bias. However, as both boost settings (**T1-C8** and **T8-C1**) lead to reductions in bias – this seems counter to intuition – how could they both reduce overall bias? We hypothesize that this is because bias can manifest from both the fielding and the length of fields, so it is possible that boosting one field might mitigate one and boosting the other might mitigate the other.

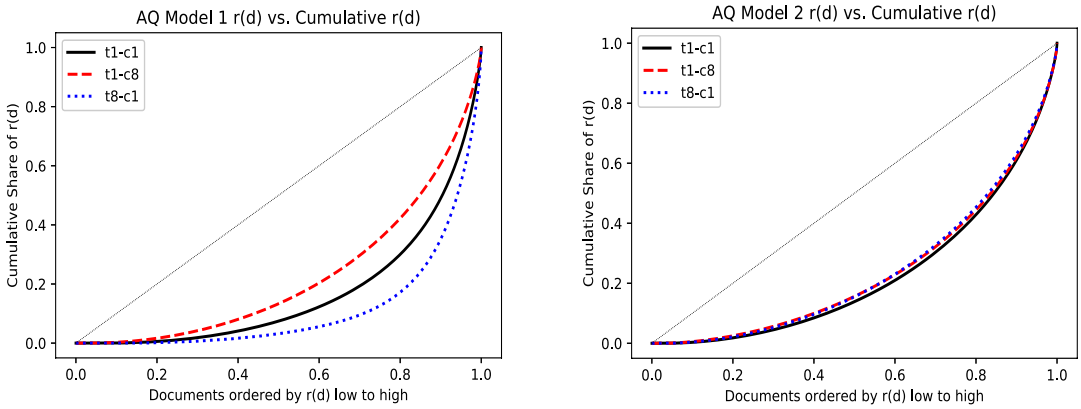

Figure 2. Lorenz curves for Model 1 (left) and Model 2 (right). For Model 1 the field boosting substantially changes the inequality between documents, while field boosting in Model 2 has less of an impact.

Figure 3 shows how boosting affects performance and bias. In the left plot, we can see how boosting the content tends to decrease the bias while increasing the performance (the changes are more dramatic for **Model 1** compared to **Model 2)**. In the right plot, we can see the influence of boosting the title. For **Model 1** we see that boosting the title is clearly detrimental, decreasing the overall performance, and inducing more bias across the collection. For **Model 2**, boosting the title field leads to reductions in bias but at the expense of performance. These plots indicate that a fairer system (less biased) does indeed lead to better performance when **Model 1** is employed. And thus **Model 1** tends to uphold the fairness hypothesis. **Model 2** on the other hand, shows that including both fields substantially reduces the bias and increases performance, but then the field boosts lead to a trade-off between performance and bias.

Figure **4** depicts the average retrievability of documents that have titles and documents that do not have titles. These plots show how the different fielding boostings can substantially affect the retrievability of documents. The left plots show the influence on **Model 1**, where increasing the title boosting (the two righthand boxplots), leads to even lower retrievability afforded to documents without titles whereas increasing the content boosting (the two middle boxplots) leads to increasing their retrievability instead. On **Model 2**, there is less difference between the two groups. Recall that when **Model 2** is set to **T1-C1** it represents the default bag of words model (i.e. no fielding, **Model 0**). We can see that under **Model 0** (**Model 2 T1-C1**) there is not much difference between the groups. However, when content boosting is applied (**Model 2 T1-C8**) the disparity is lessened, whereas when title boosting is applied (**Model 2 T8-C1**) the disparity increases slightly and is statistically significant.

[2] Topics available at: https://trec.nist.gov/data/topics_eng/index.html

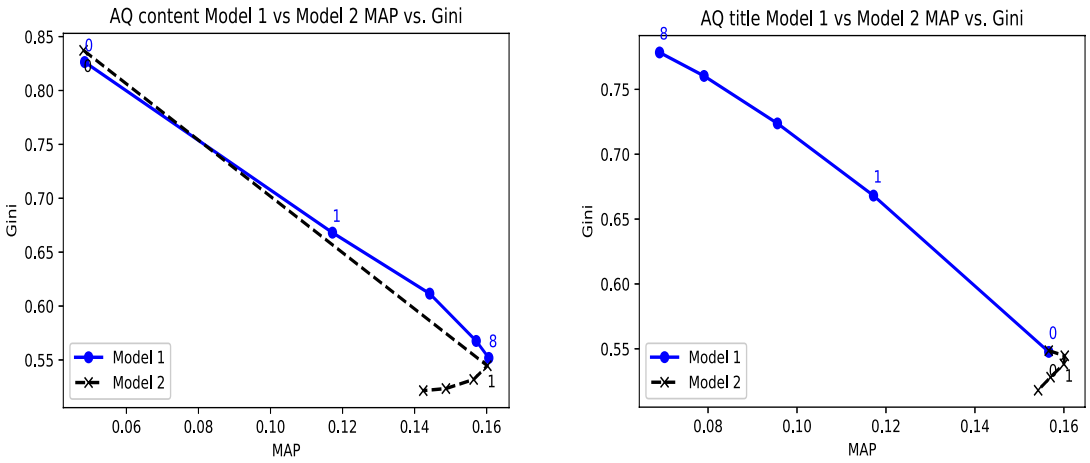

Figure 3. Retrieval Bias (Gini) vs Performance (MAP) for Content Boosting (left) and Title Boosting (right) for Model 1 and 2. When the content is boosted, it tends to increase performance and decrease bias. When title is boosted, on Model 1 performance decreases, while bias increases.

DISCUSSION AND FUTURE WORK

In this work we explored the impact that fielding has on both retrieval performance and retrieval bias and produced three key findings. The first finding highlighted is that **Model 2** was more robust to field boosting in terms of inequality. While **Model 1** was highly sensitive to changes in the degree of boost applied, **Model 2** was able to keep the distribution of ***r(d)*** scores very similar, even when large boosts were applied. Next, we found that **Model 1** exhibits a linear relationship between retrieval performance and retrieval bias such that boosting titles degrades performance and increases bias while boosting content improves performance and decreases bias. This suggests that if **Model 1** is to be employed, minimizing bias will lead to maximizing the performance. On the other hand, **Model 2** demonstrates a more complex relationship between performance and bias: *(a)* including both fields dramatically reduced the bias and increases the performance, however, *(b)* boosting one field leads to a trade-off between performance and bias meaning that a fairer system can slightly degrade performance on the collections used. Therefore, it will be important to examine if the fairer system generalizes better than the optimized system when using held out queries. Finally, we found that the amount of retrievability ***r(d)*** assigned to documents with titles was far higher than those without when **Model 1** was employed. **Model 2** had less bias toward documents with titles than without titles and thus tended to be fairer overall. However, there were still significant differences in terms of retrievability between these two groups. These findings suggest that **Model 2** is fairer than **Model 1**, and that it is more robust and less sensitive to field boosting and can better handle documents with missing fields. This is an important finding given that **Model 1** fielding is the

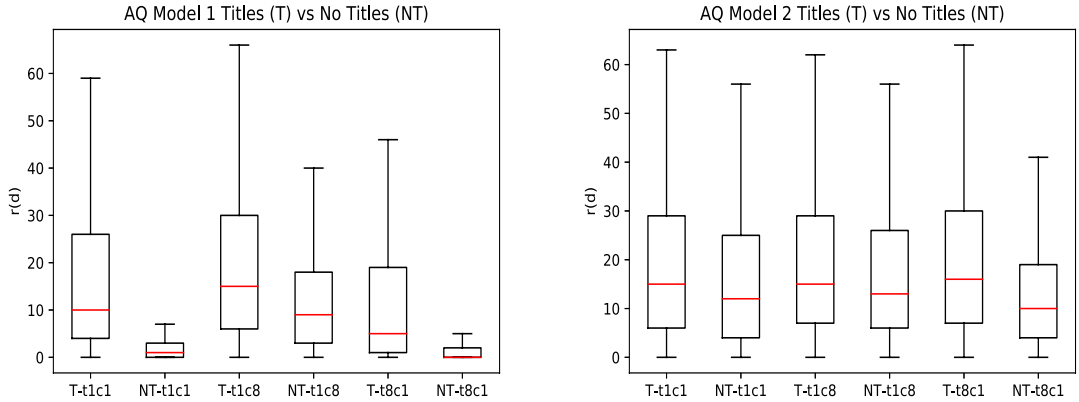

Figure 4. Boxplots of r(d) for with (T-) and without titles (NT-) for Model 1 (left) and Model 2 (right). On Model 1 there is greater disparity between document with titles and without, than on Model 2. When content is boosted on Model 2 (and to a lesser extent Model 1), there is less inequality between groups. Note the red line indicates the median.

default fielding algorithm employed in the most widely used Open Source IR toolkits (e.g. Apache Lucene, Solr and Elastic). Our work suggests that if such an approach is to be used then **Model 1** has to be very carefully tuned, whereas a switch to **Model 2** type fielding is likely to lead to immediate benefits (i.e. reduced bias and increased performance).

Finally, this work presents multiple avenues for future work, the first concerning how to manage missing fields such as title fields in news stories. It may be the case that imputing a title for these documents could lead to significant reductions in bias on **Model 1** due to how the scoring handles empty fields. If a title could be imputed from the content or other features, then the bias towards documents with missing fields in **Model 1** might be redressed, leading to an alternative way of mitigating the algorithmic bias. In this work, we observed that **Model 2** reduced the bias of the system when boosting titles or content which seems counter to intuition and we hypothesize that this may be because the boosting tackles different biases, not just fielding but also length bias. As such, we plan to also examine the influence of the length normalization parameter within **BM25** to try and isolate the cause of the different biases.

More generally future work should examine how well our results generalize to other document genres (i.e. web, patent, medical, etc.) where fielding is extensively used and to explore how the number of fields and types of fields influence the performance and bias of the system. We also must explore other retrieval algorithms and other fielded approaches in such contexts as well as examine how fielded queries – that is where the query has structure specifying which field(s) should be used – influences performance and bias.

REFERENCES

Azzopardi, L. and Owens, C. (2009) 'Search engine predilection towards news media providers', in *Proceedings - 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2009*. doi: 10.1145/1571941.1572122.

Azzopardi, L., De Rijke, M. and others (2006) 'Query intention acquisition: A case study on automatically inferring structured queries', in *Proceedings of the 6th Dutch-Belgian Information Retrieval Workshop*, pp. 3–10.

Azzopardi, L. and Vinay, V. (2008a) *Accessibility in information retrieval*, *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. doi: 10.1007/978-3-540-78646-7_46.

Azzopardi, L. and Vinay, V. (2008b) 'Retrievability: An evaluation measure for higher order information access tasks', in *International Conference on Information and Knowledge Management, Proceedings*. doi: 10.1145/1458082.1458157.

Bashir, S. and Rauber, A. (2009a) 'Identification of Low/High Retrievable Patents Using Content-based Features', in *Proceedings of the 2Nd International Workshop on Patent Information Retrieval*. New York, NY, USA: ACM (PaIR '09), pp. 9–16. doi: 10.1145/1651343.1651346.

Bashir, S. and Rauber, A. (2009b) 'Improving retrievability of patents with cluster-based pseudo-relevance feedback documents selection', in *Proc. of the 18th ACM CIKM*, pp. 1863–1866.

Bashir, S. and Rauber, A. (2011) 'On the relationship b\/w query characteristics and IR functions retrieval bias', *J. Am. Soc. Inf. Sci. Technol.*, 62(8), pp. 1515–1532.

Bashir, S. and Rauber, A. (2014) 'Automatic ranking of retrieval models using retrievability measure', *Knowledge and Information Systems*. Springer, 41(1), pp. 189–221.

Blanco, R. and Boldi, P. (2012) 'Extending BM25 with Multiple Query Operators', in *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*. New York, NY, USA: ACM (SIGIR '12), pp. 921–930. doi: 10.1145/2348283.2348406.

Chen, R.-C., Azzopardi, L. and Scholer, F. (2017) 'An empirical analysis of pruning techniques performance, retrievability and bias', in *International Conference on Information and Knowledge Management, Proceedings*. doi: 10.1145/3132847.3133151.

Gastwirth, J. L. (1972) 'The estimation of the Lorenz curve and Gini index', *The Review of Economics and Statistics*, 54, pp. 306–316.

Itakura, K. Y. and Clarke, C. L. A. (2010) 'A Framework for BM25F-based XML Retrieval', in *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. New York, NY, USA: ACM (SIGIR '10), pp. 843–844. doi: 10.1145/1835449.1835644.

Jimmy, Zuccon, G. and Koopman, B. (2016) 'Boosting Titles Does Not Generally Improve Retrieval Effectiveness', in *Proceedings of the 21st Australasian Document Computing Symposium*. New York, NY, USA: ACM (ADCS '16), pp. 25–32. doi: 10.1145/3015022.3015028.

Kim, J., Xue, X. and Croft, W. B. (2009) 'A Probabilistic Retrieval Model for Semistructured Data', in *Proceedings of the 31th European Conference on IR Research on Advances in Information Retrieval*. Berlin, Heidelberg: Springer-Verlag (ECIR '09), pp. 228–239. doi: 10.1007/978-3-642-00958-7_22.

Kim, J. Y. and Croft, W. B. (2012) 'A Field Relevance Model for Structured Document Retrieval', in *Proceedings of the 34th European Conference on Advances in Information Retrieval*. Berlin, Heidelberg: Springer-Verlag (ECIR'12), pp. 97–108. doi: 10.1007/978-3-642-28997-2_9.

Kirkpatrick, K. (2016) 'Battling Algorithmic Bias: How Do We Ensure Algorithms Treat Us Fairly?', *Commun. ACM*, 59(10), pp. 16–17.

Lipani, A. *et al.* (2015) 'An Initial Analytical Exploration of Retrievability', in *Proc. of the 2015 ICTIR*. ACM (ICTIR '15), pp. 329–332.

Ogilvie, P. and Callan, J. (2003) 'Combining Document Representations for Known-item Search', in *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval*. New York, NY, USA: ACM (SIGIR '03), pp. 143–150. doi: 10.1145/860435.860463.

Plachouras, V. and Ounis, I. (2007) 'Multinomial Randomness Models for Retrieval with Document Fields', in *Proceedings of the 29th European Conference on IR Research*. Berlin, Heidelberg: Springer-Verlag (ECIR'07), pp. 28–39.

Robertson, S. and Zaragoza, H. (2009) 'The Probabilistic Relevance Framework: BM25 and Beyond', *Foundations and Trends in Information Retrieval*, 3(4), pp. 333–389.

Robertson, S., Zaragoza, H. and Taylor, M. (2004) 'Simple BM25 extension to multiple weighted fields', in *Proceedings of the 13th ACM CIKM*, pp. 42–49.

Samar, T. *et al.* (2017) 'Quantifying retrieval bias in Web archive search', *International Journal on Digital Libraries*. Springer, pp. 1–19.

Singhal, A., Buckley, C. and Mitra, M. (1996) 'Pivoted document length normalization', in *Proce. of the 19th ACM SIGIR conference*. (SIGIR '96), pp. 21–29.

Traub, M. C. *et al.* (2016) 'Querylog-based assessment of retrievability bias in a large newspaper corpus', in *2016 IEEE/ACM Joint Conference on Digital Libraries (JCDL)*, pp. 7–16.

Wilkie, C. and Azzopardi, L. (2013) 'Relating retrievability, performance and length', in *SIGIR 2013 - Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*. doi: 10.1145/2484028.2484145.

Wilkie, C. and Azzopardi, L. (2014a) *Best and fairest: An empirical analysis of retrieval system bias*, *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. doi: 10.1007/978-3-319-06028-6_2.

Wilkie, C. and Azzopardi, L. (2014b) *Efficiently estimating retrievability bias*, *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. doi: 10.1007/978-3-319-06028-6_82.

Wilkie, C. and Azzopardi, L. (2015) 'Query length, retrievability bias and performance', in *International Conference on Information and Knowledge Management, Proceedings*. doi: 10.1145/2806416.2806604.

Wilkie, C. and Azzopardi, L. (2017) 'Algorithmic Bias: Do Good Systems Make Relevant Documents More Retrievable', in *Proceedings of the International ACM CIKM*. (CIKM '17).

Wilkie, C. and Azzopardi, L. (2017) 'An initial investigation of query expansion bias', in *ICTIR 2017 - Proceedings of the 2017 ACM SIGIR International Conference on the Theory of Information Retrieval*. doi: 10.1145/3121050.3121097.

ACKNOWLEDGMENTS

We would like to thank the EPSRC for supporting this research through grants EP/M508056/1 and EP/K000330/1.

COPYRIGHT

		AP			T45			AQ		
		Gini	MAP	T vs NT r(d) Diff	Gini	MAP	T vs NT r(d) Diff	Gini	MAP	T vs NT r(d) Diff
Model 0	**Base**	0.344	0.261	4.000*	0.49	0.159	-37.000*	0.545	0.16	3.000*
Model 1	**T1C0**	0.611	0.076	9.000*	0.795	0.058	2.000*	0.826	0.048	**3.000***
	T1C1	0.448	0.197	13.000*	0.627	0.12	**-1.000***	0.668	0.117	9.000*
	T1C2	0.389	0.24	12.000*	0.527	0.149	-12.000*	0.612	0.144	10.000*
	T1C4	0.358	0.256	10.000*	0.484	**0.16**	-21.000*	0.568	0.157	8.000*
	T1C8	0.349	**0.259**	7.000*	**0.482**	0.159	-26.000*	0.552	**0.161**	6.000*
	T0C1	**0.348**	0.257	**3.000***	0.494	0.155	-31.000*	**0.548**	0.157	**3.000***
	T1C1	0.448	0.197	13.000*	0.627	0.12	**-1.000***	0.668	0.117	9.000*
	T2C1	0.507	0.145	12.000*	0.715	0.093	4.000*	0.724	0.096	8.000*
	T4C1	0.549	0.121	11.000*	0.754	0.081	4.000*	0.76	0.079	6.000*
	T8C1	0.572	0.107	10.000*	0.771	0.075	3.000*	0.778	0.069	5.000*
Model 2	**T1C0**	0.62	0.069	8.000*	0.812	0.039	**3.000***	0.837	0.048	3.000*
	T1C1	0.344	**0.261**	4.000*	0.49	0.159	-37.000*	0.545	**0.16**	3.000*
	T1C2	0.342	0.253	3.000*	0.472	**0.161**	-35.000*	0.532	0.156	3.000*
	T1C4	0.329	0.243	**2.000***	0.465	**0.161**	-36.000*	0.523	0.149	**1.000***
	T1C8	0.335	0.232	**2.000***	**0.46**	**0.161**	-35.000*	0.522	0.142	2.000*
	T0C1	0.348	0.258	3.000*	0.493	0.155	-38.000*	0.549	0.157	3.000*
	T1C1	0.344	**0.261**	4.000*	0.49	0.159	-37.000*	0.545	**0.16**	3.000*
	T2C1	0.338	**0.261**	4.000*	0.484	0.16	-36.000*	0.538	**0.16**	4.000*
	T4C1	0.324	0.258	6.000*	0.472	**0.161**	-34.000*	0.528	0.157	5.000*
	T8C1	**0.306**	0.253	6.000*	0.461	**0.161**	-33.000*	**0.518**	0.154	6.000*

Table 2. Table of results from each collection featuring the Gini Coefficient, MAP and the Difference in average r(d) score between documents with and without titles. * denotes a statistically significant difference at $p < 0.02$