Analyzing the Influence of Bigrams on Retrieval Bias and Effectiveness

Abdulaziz AlQatan
abdulaziz.alqatan@strath.ac.uk
University Of Strathclyde
Glasgow, Scotland

Leif Azzopardi
leif.azzopardi@strath.ac.uk
University Of Strathclyde
Glasgow, Scotland

Yashar Moshfeghi
yashar.moshfeghi@strath.ac.uk
University Of Strathclyde
Glasgow, Scotland

ABSTRACT

Prior work on using retrievability measures in the evaluation of information retrieval (IR) systems has laid out the foundations for investigating the relationship between retrieval effectiveness and retrieval bias. While various factors influencing bias have been examined, there has been no work examining the impact of using bigram within the index on retrieval bias. Intuitively, how the documents are represented, and what terms they contain, will influence whether they are retrievable or not. In this paper, we investigate how the bias of a system changes depending on how the documents are represented using unigrams, bigrams or both. Our analysis of three different retrieval models on three TREC collections, shows that using a bigram only representation results in the lowest bias compared to unigram only representation, but at the expense of retrieval effectiveness. However, when both representations are combined it results in reducing the overall bias, as well as increasing effectiveness. These findings suggest that when configuring and indexing the collection, that the bag-of-words approach (unigrams), should be augmented with bigrams to create better and fairer retrieval systems.

ACM Reference Format:
Abdulaziz AlQatan, Leif Azzopardi, and Yashar Moshfeghi. 2020. Analyzing the Influence of Bigrams on Retrieval Bias and Effectiveness. In *Proceedings of the 2020 ACM SIGIR International Conference on the Theory of Information Retrieval (ICTIR '20), September 14–17, 2020, Virtual Event, Norway.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3409256.3409831

1 INTRODUCTION

Traditionally, research has been focused on evaluating the efficiency and effectiveness of Information Retrieval systems [10]. However, Investigating and analyzing the bias of Information Retrieval (IR) Systems and how fairly such systems retrieve items has become an increasingly important area of research [3]. This is because the IR system may be unfairly treating items (and groups represented) within the collections. As such, if an item (or group of items) are unfairly treated and thus hard to retrieve because of the system – then the ensuing retrieval effectiveness for such requests for such items, would be poor. This line of argument led to the idea of the "*fairness hypothesis*" [20]: that fairer systems would result in better retrieval effectiveness – because if, all documents in the collection would be afforded a similar chance of being retrieved, then, if the item was ever relevant to a request, then it could be retrieved. This notion, called retrievability is fundamental within IR because retrieval precedes relevancy [2]. That is, an item can not be judged relevant, if it is not retrieved. Consequently, a number of works have attempted to explore the different factors that influence the retrievability within IR systems – and how retrievability bias relates to retrieval effectiveness in various contexts (e.g. web, news, patents, archives, etc. [5–7, 12, 15, 20]) and across a number of different factors (e.g. query length, document length, fielding [2, 21, 22], query expansion [4], retrieval algorithms [20], over time [15, 18], etc.) From these works it has been shown that retrievability bias tends to correlated with effectiveness – for example – in [21], they found that different document length normalisation settings, lead to different levels of bias, but minimizing bias lead to greater retrieval effectiveness. While in [22], they examined the influence of fielded document representations and showed that when fields were index separately it resulted in greater bias, and lower effectiveness. In these studies, and other prior works, they have focused mainly on bag-of-words representations. However, if bigrams are also indexed, then it will affect the document length statistics if combined within the same field as the unigrams (bag-of-words), or require separate fields to represent the unigrams and bigrams changing the scoring function. Both alternatives, are likely to influence and affect the bias of the system – but it is largely unknown in what way – nor is it clear how best to configure a system (if bigrams are going to be included as part of the document representation). Furthermore, while past work has investigated and shown that including bigrams tends to lead to small increases in effectiveness [11, 16, 17], we wonder if they also provide additional benefit (or not) with respect to bias. Thus, in this paper, we set out to investigate the relationship between effectiveness and retrieval bias when bigrams are included as part of the indexing and retrieval processes. Specifically,

- How do bigrams influence the bias of retrieval systems?
- What index representation (unigram, bigram, or combination) provides the lowest retrieval bias?
- And, does minimizing the retrieval bias, given the different representations, result in higher retrieval effectiveness?

To this end, we performed an empirical analysis using three TREC Collections and three standard baseline retrieval models, where the indexing structure was varied (unigram only, bigram only, combined, and fielded) – exploring how different parameterizations influence bias and effectiveness. Our findings indicate the bigram only index results in poor retrieval effectiveness but provide

considerably lower bias than the unigram only index. However, the different combinations of unigrams and bigrams, both lead to a synergistic effect, increasing effectiveness, lowering bias, and, crucially, increasing the number of items that are retrievable.

2 BACKGROUND

In Information Retrieval, many different retrieval models have been proposed over the years [1, 13, 14, 16]. Yet, the default, baseline approach to IR has typically relied upon the simple bag-of-words representation (i.e. unigrams) combined with BM25, TF.IDF, or some other weighting function. While there are many sophisticated extensions, one of the most obvious extensions, is to include term dependencies [13, 14, 16], such as bigrams in order to improved effectiveness. However, while including such dependencies tends to increase effectiveness, it is seldom, if ever employed. In this paper, we wonder if using bigrams can provide any additional benefits, such as reducing the over system bias. To measure the bias of a system, we can draw upon Retrievability theory.

Retrievability provides a way to quantify the influence of a system on a collection. It measures how *likely* a document is to be retrieved given an IR system configuration [2]. The retrievability $\mathbf{r}$ of a document $\mathbf{d}$ is defined as:

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

where $\mathbf{q}$ is a query from a large query set $\mathbf{Q}$, and $\mathbf{k_{dq}}$ is the rank at which $\mathbf{d}$ is retrieved given $\mathbf{q}$. The utility function $\mathbf{f}(\mathbf{k_{dq}}, \mathbf{c}, \beta)$ determines the score that document $\mathbf{d}$ attains for query $\mathbf{q}$ given the rank cutoff $\mathbf{c}$ and a discount β. $\mathbf{r}(\mathbf{d})$ is 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 it is commonly approximated with a large set of queries [2, 5]. The standard measure of retrievability used employs the utility function $\mathbf{f}(\mathbf{k_{dq}}, \mathbf{c}, \beta)$, such that if a document $\mathbf{d}$ is retrieved in the top $\mathbf{c}$ documents given $\mathbf{q}$, then $\mathbf{f}(\mathbf{k_{dq}}, \mathbf{c}, \beta) = 1/\mathbf{k}_{\mathbf{dq}}^{\beta}$, otherwise $\mathbf{f}(\mathbf{k_{dq}}, \mathbf{c}, \beta) = 0$. When $\beta = 0$, the measure is essentially cumulative i.e. the number of times that the document is retrieved in the top $\mathbf{c}$ documents, whereas when $\beta > 0$, documents further down the ranked list are assigned less utility (this is referred to as a gravity-based measure by Azzopardi and Vinay [2]).

To measure the retrieval bias given the retrievability scores, the Gini Coefficient is used to calculate the level of inequality in a population [8]. Intuitively, if all the documents have the same level of $\mathbf{r}(\mathbf{d})$, then there is no inequality within the population of documents, and so Gini = 0.0. However, if all the documents have an $\mathbf{r}(\mathbf{d}) = 0$, except one document, then there is high inequality within the population (i.e. a King and all the peasants), so Gini = 1.0 denoting total inequality.

Given this measure of retrieval bias, a number of studies have been undertaken examining the relationship between bias and effectiveness [4–7, 15, 19, 20]. By and large, these works have shown that different retrieval models result in different levels of bias, and this is affected by document length normalization, query length, query expansion, fielding, pruning, etc.. In general, it has been shown that optimizing the IR system, such that it minimizes retrieval bias, tends to lead to good effectiveness on standard retrieval measures such as P@10 and MAP, and for more recent measures, such as Time Biased Gain and the U-measure, there is a much higher correlation [21]. In this work, we explore the influence of bigrams on retrieval bias and effectiveness.

3 EXPERIMENTAL METHOD

To explore our research questions, we conducted an empirical analysis across a number of TREC collections with standard retrieval algorithms.

Data and Materials: For the analysis we used three test collections: *TREC CommonCore 2018* (approx. 600K documents from the Washington Post with 50 topics (321–825), *TREC CommonCore 2017* (approx. 1.8 million documents of New York Times articles from 1987 to 2007, with 50 topics (301–700)), and *TREC AQUAINT 2005* Corpus (approx. one million documents from three different news wires: New York Times, AP and Xinhua News Agency, with 50 topics (303–689)). Each collection was indexed in Lucene 8[1], with the standard tokenizer and Porter stemming. Stop words were also removed. Four different indexes were created:

- **Unigram Only Index (UI)**: All unigrams were indexed into one field.
- **Bigram Only Index (BI)**: All bigrams were indexed into one field.
- **Combined Index (CI)**: All unigrams and bigrams were indexed into one field. Note that this is the default option in Lucene when applying n-grams (called Shingles).
- **Fielded Index (FI)**: All unigrams and bigrams were index, but into two separate fields, FI-U and FI-B, respectively.

Retrieval Models and Evaluation Metrics: We used three different, but commonly used, retrieval models: *BM25* [14], a *Language Model* (LM) with Bayes Smoothing [16] and *Divergence from Randomness Model*, PL2 [1]. To examine, the influence of the length normalisation for each model: for BM25, b was varied between 0.1 to 1.0 at 0.1 steps, while k was kept constant at 1.2; for LM, μ was varied from 100 to 1000 in steps of 100, and then up to 5000; and for PL2, c was set to: 0.1, 0.5, 1, 5, 10, 15, 20, and 50.

For the Fielded Index, we calculated the final retrieval score as the sum over both fields – and thus, implemented the fielded versions of BM25, LM and PL2 [13, 14, 16], respectively. Each field was equally weighted. We leave exploring different combinations for future work. It should be noted that for Combined Index, the term statistics will be quite different, as both unigrams and bigrams are included in the same field within the Lucene index – and therefore – we hypothesize that this may have a negative impact on effectiveness and bias.

To determine the effectiveness of each retrieval model / parameter setting on each collection/index, we used the corresponding TREC topics, and calculated the *Mean Average Precision MAP*, *P10* and *Bpref*. Due to space constraints only *MAP* is reported, but, similar findings are observed for the other measures.

Query Generation and Retrievability Bias: To compute the retrievability scores for each system configuration (i.e. index / model / parameter setting), we followed a similar methodology as done in [2, 4]. First, we extracted bigrams that occurred at least five times in each collection of news articles. Given the set of bigrams,

[1] http://lucene.apache.org/

we then scored them using the *Pointwise Mutual Information Measure* [9] to identify popular collocations/phrases. This was to select common phrases which might be issued as queries. For each collection, 300,000 collocations were then taken and used as a query set. For each system configuration, the queries were then issued, and the retrievability $r(d)$ for each document was computed. We computed two retrievability scores: (*a*) cumulative based measure where $c = 100$, and (*b*) gravity based measures where $c = 100$ and $\beta = 0.5$. Given the $r(d)$ scores for each retrievability measure, we then computed the Gini co-efficient [8] as done in [2]. Due to space constraints, we only report on the gravity based measures (but note that our findings were similar with the cumulative based measures as well).

Table 1: The Mean Average Precision (MAP) and Bias (G) for the best parameter setting given the Unigram Index (UI), and the corresponding scores for the Bigram (BI), Fielded (FI) and Combined Indexes (CI). Bolded values indicated that whether the MAP or G improves over the Unigram Index.

Collection	Model	Index	Param.	MAP	G
AQUAINT	BM25	UI	$b = 0.2$	0.152	0.637
		BI		0.109	**0.450**
		CI		**0.155**	**0.564**
		FI		**0.156**	**0.564**
	LMD	UI	$\mu = 700$	0.152	0.601
		BI		0.107	**0.423**
		CI		**0.153**	**0.506**
		FI		**0.158**	**0.520**
	PL2	UI	$c = 20$	0.151	0.591
		BI		0.108	**0.438**
		CI		**0.154**	**0.513**
		FI		**0.149**	**0.491**
CORE 2017	BM25	UI	$b = 0.4$	0.173	0.544
		BI		0.111	**0.399**
		CI		**0.183**	**0.451**
		FI		**0.183**	**0.414**
	LMD	UI	$\mu = 700$	0.171	0.538
		BI		0.110	**0.387**
		CI		**0.182**	**0.411**
		FI		**0.182**	**0.423**
	PL2	UI	$c = 15$	0.169	0.555
		BI		0.111	**0.403**
		CI		**0.177**	**0.433**
		FI		**0.179**	**0.417**
WAPO	BM25	UI	$b = 0.4$	0.153	0.503
		BI		0.050	**0.247**
		CI		**0.171**	**0.350**
		FI		0.143	**0.278**
	LMD	UI	$\mu = 1000$	0.152	0.544
		BI		0.051	**0.245**
		CI		**0.171**	**0.362**
		FI		0.144	**0.367**
	PL2	UI	$c = 15$	0.152	0.551
		BI		0.051	**0.253**
		CI		**0.167**	**0.329**
		FI		0.143	**0.311**

4 RESULTS AND DISCUSSION

Figure 1 shows the relationship between MAP and the Gravity based Gini Coefficient for the different indexes for each retrieval model. The plots show CORE 2017 (top) and WAPO (bottom), but note that AQUAINT plots were similar to CORE 2017. From these plots, we can start to see a number of different trends. Firstly, when the bigram only index is used, regardless of the collection or retrieval model, we can see that it gives very poor retrieval effectiveness (low MAP). However, it also tends to give the lowest retrieval bias (G). This suggests that retrieval is fairer when using a bigram index, but at the expense of retrieval effectiveness. This seems at odds with the fairness hypothesis, later we will re-visit this contradiction. In contrast, on the unigram index, retrieval effectiveness is substantially higher, but bias is also substantially higher. Given these two results, it is therefore surprising, that the combination of unigrams and bigrams have a synergistic effect. Generally, we can see that effectiveness increases, and the bias decreases. This is the case for both on CORE 2017. However, on WAPO we see that the Fielded Index, only leads to a reduction in bias, and not a corresponding increase in effectiveness (presumably this is due to the field weightings, further exploration is required to see if the effectiveness can be ameliorated through tuning). Table 1 shows that for the best performing parameter setting, the corresponding effectiveness and bias for each index and collection. A key finding here, is that for each model's parameter setting (i.e. b, μ and c) the corresponding effectiveness and bias on the combined index, for each retrieval model and collection, is point on point, better (higher effectiveness and lower bias). This means, however, one chooses the parameter setting, on the unigram index, improvements can be made, with the additional of bigrams – while the retrieval effectiveness differences are not substantial, the reduction in bias is considerable.

Now, turning our attention back to the Bigram index. If fairer is better, then why did it perform so poorly? We hypothesised that while the Bigram index might be fairer, per say, it may not provide as many opportunities. To investigate whether this was the case, we computed the RSum measure [6], which is the sum of $r(d)$ scores over all documents in the collection. Table 2 shows the corresponding RSum values on each index. Essentially, RSum tells us how many/much retrieval opportunities are afforded by the given index. The higher the RSum, the more opportunities for documents to be retrieved. Note that the RSum value is the same across retrieval models for a given collection, instead the retrieval models distribute those retrieval opportunities to different documents (leading to different levels of bias (G)). From Table 2, we can see that under the Bigram index the RSum scores are considerably lower than on the Unigram index. This means that while the Bigram index is fairer (lower G), it has fewer opportunities to distribute (but distributes those opportunities more equally), and we posit that this is the reason for its lower effectiveness as well. However, when unigrams and bigrams are combined, either together or through fielding, we see that there is an increase in RSum, and thus the combined and fielded indexes provide more retrieval opportunities than either the Unigram or Bigram indexes. This finding suggests that not only do we need to make fairer systems that have less bias, but also ensure the opportunities to be retrieved is maximised.

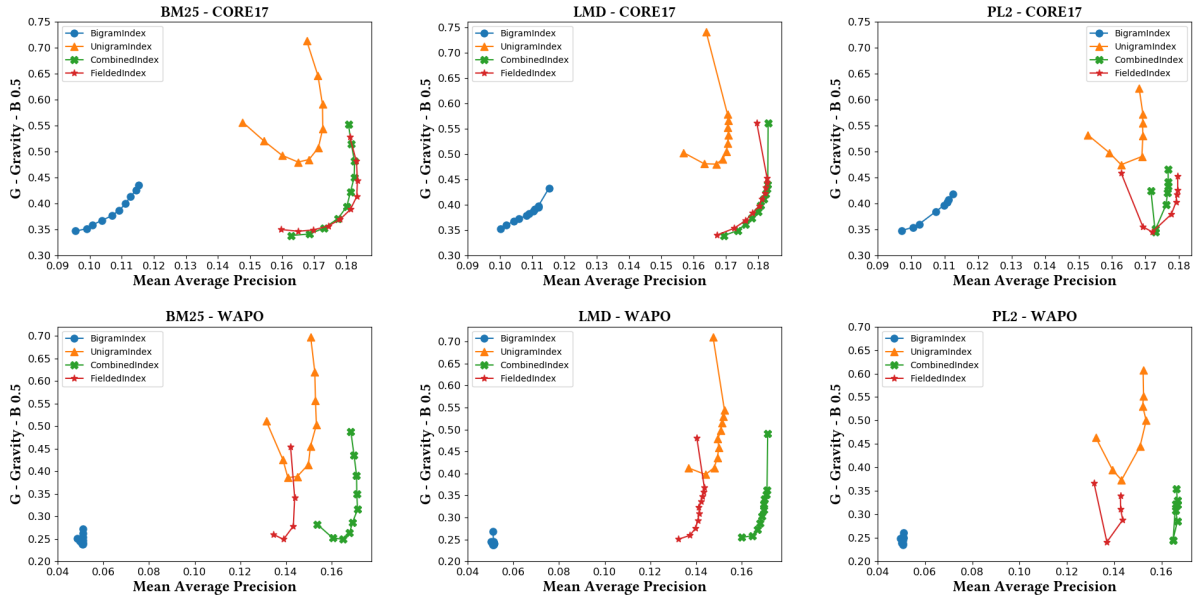

Figure 1: Effectiveness (MAP) versus Bias (G). Top: Common Core 2017 and Bottom: Common Core 2018 (Washington Post) for each index for each of the different retrieval models (Left: BM25, Mid: LM, and Right: PL2). The Bigram Index provides low bias, but poor effectiveness, while the Combined Index gives better effectiveness and lower bias.

Table 2: RSums values for the Gravity based Retrievability measure (when $b = 0.5$).

Index	AQUAINT	CORE 2017	WAPO
UI	4814557	4765623	4869681
BI	2902082	3855955	4151477
CI	**5376883**	**5390534**	**5387239**
FI	**5376883**	**5390534**	**5387239**

5 SUMMARY AND FUTURE WORK

In summary, we have shown that the inclusion of bigrams has a positive impact, creating systems that are not only more effective, but that are also fairer, and also provide more retrieval opportunities. These findings provide deeper insights into developing fairer systems – and motivates future work investigating alternative retrieval models and methods to combine document representations as well as exploring the relationship between bias and effectiveness in different collections and in different contexts.

Acknowledgements. The first author would to thank the Embassy of the State of Kuwait PhD Scholarship (grant no. *19CS0089*).

REFERENCES

[1] Gianni Amati and Cornelis Joost Van Rijsbergen. 2002. Probabilistic models of information retrieval based on measuring the divergence from randomness. *ACM Transactions on Information Systems (TOIS)* 20, 4 (2002), 357–389.
[2] Leif Azzopardi and Vishwa Vinay. 2008. Retrievability: An Evaluation Measure for Higher Order Information Access Tasks. In *Proc. of CIKM '08*. ACM, 561–570.
[3] Ricardo Baeza-Yates. 2018. Bias on the Web. *Comm. ACM* 61, 6 (2018), 54–61.
[4] Shariq Bashir and Andreas Rauber. 2009. Improving retrievability of patents with cluster-based pseudo-relevance feedback documents selection. In *Proc. of CIKM '09*. 1863–1866.
[5] Shariq Bashir and Andreas Rauber. 2010. Improving retrievability of patents in prior-art search. In *Proc. of ECIR '10*. 457–470.
[6] Ruey-Cheng Chen, Leif Azzopardi, and Falk Scholer. 2017. An Empirical Analysis of Pruning Techniques: Performance, Retrievability and Bias. In *Proc. of the 2017 ACM on Conference on Information and Knowledge Management*. 2023–2026.
[7] Debasis Ganguly, Ayan Bandyopadhyay, Mandar Mitra, and Gareth J.F. Jones. 2016. Retrievability of Code Mixed Microblogs. In *Proc. of the 39th International ACM SIGIR Conference* (Pisa, Italy) *(SIGIR '16)*. 973–976.
[8] J Gastwirth. 1972. The Estimation of the Lorenz Curve and Gini Index. *The Review of Economics and Statistics* 54 (1972), 306–316. Issue 3.
[9] Christopher D Manning, Christopher D Manning, and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. MIT press.
[10] Christopher D Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to information retrieval*. Cambridge university press.
[11] Seung-Hoon Na, Jungi Kim, In-Su Kang, and Jong-Hyeok Lee. 2008. Exploiting proximity feature in bigram language model for information retrieval. 821–822.
[12] Jiaul H. Paik and Jimmy Lin. 2016. Retrievability in API-Based "Evaluation as a Service". In *Proc. of the 2016 ACM International Conference on the Theory of Information Retrieval* (Newark, Delaware, USA) *(ICTIR '16)*. 91–94.
[13] Vassilis Plachouras and Iadh Ounis. 2007. Multinomial Randomness Models for Retrieval with Document Fields. In *ECIR*.
[14] Stephen Robertson and Hugo Zaragoza. 2009. The Probabilistic Relevance Framework: BM25 and Beyond. *FNTIR* 3, 4 (2009), 333–389.
[15] Thaer Samar, Myriam C. Traub, Jacco Ossenbruggen, Lynda Hardman, and Arjen P. Vries. 2018. Quantifying Retrieval Bias in Web Archive Search. *Int. J. Digit. Libr.* 19, 1 (March 2018), 57–75.
[16] Fei Song and W. Bruce Croft. 1999. A General Language Model for Information Retrieval. In *Proc. of the Eighth International Conference on Information and Knowledge Management* (Kansas City, Missouri, USA) *(CIKM '99)*. Association for Computing Machinery, New York, NY, USA, 316–321.
[17] Chade-Meng Tan, Yuan-Fang Wang, and Chan-Do Lee. 2002. The use of bigrams to enhance text categorization. *IPM* 38, 4 (2002), 529–546.
[18] Myriam C. Traub, Thaer Samar, Jacco van Ossenbruggen, Jiyin He, Arjen de Vries, and Lynda Hardman. 2016. Querylog-Based Assessment of Retrievability Bias in a Large Newspaper Corpus. In *Proc. of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries* (Newark, New Jersey, USA) *(JCDL '16)*. Association for Computing Machinery, New York, NY, USA, 7–16.
[19] Colin Wilkie and Leif Azzopardi. 2013. Relating retrievability, performance and length. In *Proc. of SIGIR '13* (Dublin, Ireland). 937–940.
[20] Colin Wilkie and Leif Azzopardi. 2014. Best and Fairest: An Empirical Analysis of Retrieval System Bias. *Advances in Information Retrieval* (2014), 13–25.
[21] Colin Wilkie and Leif Azzopardi. 2014. A Retrievability Analysis: Exploring the Relationship Between Retrieval Bias and Retrieval Performance. In *Proc. of CIKM '14* (Shanghai, China). 81–90.
[22] Colin Wilkie and Leif Azzopardi. 2018. The impact of fielding on retrieval performance and bias. *Proc. of the Association for Information Science and Technology* 55, 1 (2018), 564–572.