# An Initial Investigation of Query Expansion Bias

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

Query expansion is a useful retrieval mechanism for creating more verbose queries from the users initial key word search. Query expansion generally have multiple parameters that allow the user to define how many terms and where those terms come from are introduced to the expanded query. However, the idea that query expansion may be introducing biases into the system by selecting terms from overly retrievable documents has never been formally evaluated. In this work, the relationship between performance and retrievability bias is explored when various query expansion methods are employed to aide retrieval. Several parameters are altered, independently, to identify those that have an impact on bias. Parameters altered include; Rocchio's beta, length normalisation parameters, the number of terms added and the number of documents those terms are extracted from. The evaluation performed here identifies a strong correlation between performance and retrievability bias, suggesting that performance is increased by making the system more biased thus more likely to pick terms from a set of overly retrievable documents.

## 1 INTRODUCTION

Information Retrieval's (IR) primary objective is to return all of the most relevant information ordered in an useful way to a user, given their information need. One function of an IR system is to infer a users information need when provided with a few key words that a user believes reflects their search goal. However, these key words often form a very vague query that could cover a massive range of documents, thus methods have been developed to bolster the users key words with additional content in the hopes that this new query will identify the most relevant documents. A Query Expansion (QE) mechanism must identify documents that are relevant to the original query and extract new terms from these documents to expand the user query with meaningful content. This new query should then identify more relevant documents that satisfy the users information need. Obviously, the documents selected are of vital importance to the success of the method as terms extracted from non-relevant documents may cause the query to drift from the users intended information need. Ideally, the user would deem which documents are relevant to them and allow the QE mechanism to select the terms from them, however this is often not possible and instead the QE must rely on some pseudo relevance feedback (PRF) [8]. Pseudo relevance feedback assumes that the top $n$ documents are relevant and as such, important terms from these documents will improve the users query. This technique can lead to improved performance but it is pivotal that the terms are extracted from the appropriate documents. To this end, QE can be adjusted in various ways to alter the weightings terms receive as well as how many terms are extracted and which documents are taken into consideration.

This method of query improvement clearly opens doors for biases to creep into the system, for example, overly retrievable documents may be given the opportunity to contribute terms which then increases their relevance when the document may not be relevant to the original query at all. Ideally, QE would select a relevant document and extract important terms that steer the query towards a specialised set of relevant documents that may not be easy to retrieve with particular terms.

This work investigates the impact of QE on the relationship between retrievability bias and performance [10]. The main question is whether the gain in performance that is often found is the result of an increase in bias. Identifying the effect of QE on retrievability bias, we can then correlate the bias with the performance metrics. By doing so, we can infer whether the introduction (or reduction) of bias in the system increases (or decreases) the performance of the system and whether or not this is beneficial to the user.

## 2 BACKGROUND

Retrievability, a document centric evaluation method, has become a popular method of evaluation in domains where system bias influences retrieval. Retrievability, proposed by Azzopardi and Vinay [2], provides an alternative view on how an IR system interacts with a collection by evaluating how *likely* a document is to be retrieved by a particular configuration of an IR system. The retrievability $\mathbf{r}$ of a document $\mathbf{d}$ with respect to the configuration of an IR system is defines as:

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

where $\mathbf{q}$ is a query from the large query set $\mathbf{Q}$. $\mathbf{k_{dq}}$ is the rank at which $\mathbf{d}$ is retrieved given $\mathbf{q}$, therefore the utility function $\mathbf{f}(\mathbf{k_{dq}}, \mathbf{c})$ determines the score that document $\mathbf{d}$ attains for query $\mathbf{q}$ given the rank cutoff $\mathbf{c}$. $\mathbf{r}(\mathbf{d})$ 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 $\mathbf{Q}$ is very large set of queries [1, 2, 4, 5, 10]. The standard measure of retrievability used employs the utility function $\mathbf{f}(\mathbf{k_{dq}}, \mathbf{c})$, such that if a document, $\mathbf{d}$, is retrieved in the top

$\mathbf{c}$ documents given $\mathbf{q}$, then $\mathbf{f}(\mathbf{k_{dq}}, \mathbf{c}) = 1$, otherwise $\mathbf{f}(\mathbf{k_{dq}}, \mathbf{c}) = 0$. This measure provides an intuitive value for each document as it is simply the number of times that the document is retrieved in the top $\mathbf{c}$ documents. Documents falling outside the the top $\mathbf{c}$ attain no scores.

To convert the $\mathbf{r}(\mathbf{d})$ for each document into a single value describing bias, inequality metrics that assess the distribution of wealth in a population are used. However, the retrievability fits this paradigm as the documents retrievability can be considered its wealth wealth and the collection is the population that the retrievability is distributed amongst. The Gini Coefficient [6] is one inequality metric that can be used to to calculate the level of inequality in a population by comparing the distribution to the Lorenz Curve.

An interesting line of research emerging from the theory of retrievability is how performance metrics relate to retrievability bias. Wilkie and Azzopardi have conducted several studies investigating this relationship [9–11]. The first of these works investigated how retrievability bias related to both performance and document lengths [9]. In this study, Wilkie and Azzopardi investigated how altering the length normalisation parameters of the BM25 and PL2 retrieval models impacted both the performance and the bias that the systems exerted upon the collection. Their findings indicated that the relationship between bias and performance was non-linear and that TREC performance metrics had a poor match up with bias (i.e. the parameter where performance was at its highest was not the parameter where the least amount of bias was found). In a follow up study by Wilkie and Azzopardi, system ranking based on performance and retrievability bias was performed [10]. The results demonstrated a much stronger correlation between bias and performance when a range of retrieval models were utilised. They found that choosing a less biased retrieval system would often improve performance while also reducing bias. From these two studies it could be concluded that when choosing a retrieval model, models which exert less bias are generally better performers. However, when tuning the length normalisation parameter of a retrieval model selecting the least bias setting for that parameter will often not lead to optimal performance. In many cases, the difference in the best performance and performance at the least biased setting was not significant however and so it was stated that the least biased point of length normalisation is a good starting point when tuning a system. A final study by Wilkie and Azzopardi on the relationship between bias and performance employed the use of a large number of evaluation metrics including some recently proposed metrics that were more user centric than TREC metrics. Again, they investigated how altering the length normalisation parameter of retrieval models impacted the relationship between bias and performance. They found that when employing metrics like Time Biased Gain and U-Measure in almost all cases, the parameter setting that minimised bias also maximised performance. Again suspecting that the poor match up for TREC metrics occurred due to length biases in the relevancy pools, the authors investigated these pools and found that there was a strong length bias towards longer documents in the relevancy pools of the TREC collections.

While the studies performed by Wilkie and Azzopardi were conducted in the domain of ad hoc web and news search, Bashir and Rauber have performed studies investigating retrievability bias in the patent retrieval domain [3, 5]. In these studies, the authors find a stronger link between recall and retrievability bias than what was observed by Wilkie and Azzopardi in ad-hoc search. These results lead Bashir and Rauber to investigate methods of QE which account for retrievability bias [4]. In this study, Bashir and Rauber compare the bias of a number of competing retrieval models and QE methods in a study to demonstrate how their new cluster based QE method provides results that are less biased than all other approaches compared. The authors work was motivated by the finding that current QE methods would often increase retrievability bias by extracting terms from the highly retrievable documents, thus making them more retrievable. However, in their study the QE methods investigated were only investigated on their default settings and the performance of the methods was never evaluated. Therefore, the relationship between bias and performance when QE is performed remains completely unknown.

Converse to Bashir and Rauber, Pickens *et al* evaluated the performance of the traditional QE methods against their new QE method, the reverted index [7]. In this work, Pickens *et al* explored the parameter space of multiple QE methods and evaluated performance at each setting to find what the impact of altering each parameter was.

Between the work of Bashir and Rauber, and the work of Pickens *et al* there is an interesting gap in the literature about the relationship between bias and performance when QE is employed to improve performance. No previous work has explored the parameter space available in QE and quantified both bias and performance to determine whether QE improves performance by employing a more biased retrieval model or if its success is due to a reduction in bias.

## 3 EXPERIMENTAL METHOD

### 3.1 Research Questions

The hypothesis of this study states that performing QE during retrieval leads to increases in retrievability bias. This work seeks to answer several research questions in the following experiments. The first question investigates how altering length normalisation in retrieval model impacts both bias and performance. Previous work by Wilkie and Azzopardi [11] has shown that employing the length normalisation setting that minimises bias does not maximise performance. As such, in QE we expect that due to there being two rounds of retrieval under the one system that the bias may be compounded and biased systems will introduce larger biases. The next question concerns how much weight should be applied to the new expansion terms compared with the original query terms. Does weighting the original query terms lower than the expansion terms introduces more bias to the system as the expansion terms are being extracted from documents which may be highly retrievable, thus bolstering the retrievability of these already retrievable documents? The next research question investigates the length of the queries that are issues to the retrieval system. As QE adds new terms to the original query, we investigate whether generating longer queries using QE actually increases biases as longer documents have more chance to match more terms. Finally, the work explores the impact that the number of documents the QE terms are extracted from has on the performance and bias relationship. It is posited that

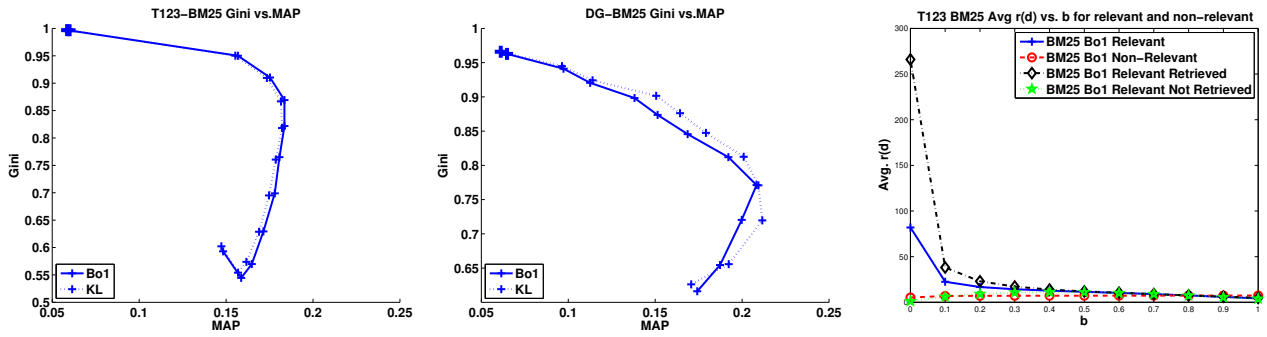


**Figure 1: Plots of Gini vs. MAP Altering the b parameter of BM25 (Left & Middle). The larger points indicate $b = 0.0$. And plots of Average r(d) vs. b for BM25 using Bo1 expansion on T123 (Right).**

using more documents from the rankings should help lower bias by utilising a larger sample of documents which would allow for more diverse terms to be generated for the expanded query. Following is the experimental methodology employed to achieve the results required for this investigation.

## 3.2 Data and Materials

**Collections** The experiments were performed on three TREC test collections: two news collections, TREC123 (T123) and Associated Press (AP), and one web collection; .Gov (DG). Each collection was indexed on Terrier[1] with stop words removed and Porter Stemming.
**Retrieval Models** Experiments were performed using 3 common retrieval models implemented in Terrier: BM25, DPH and TF.IDF. For the parameterised BM25 model, only the **b** parameter was altered as it influences length normalisation, leaving the remaining parameters of BM25 to their default values ($k_1 = 1.2$ and $k_3 = 8$). These models were selected to fit a range of profiles of performance and bias.
**Query Expansion Models** Experiments were performed using two common query expansion models, KullbackâĂŞLeibler divergence (KL) and Bose-Einstein (Bo1). These models both feature a suite of tuneable parameters. The first of these parameters is the Rocchio's beta, a parameter that is used to alter the weighting applied to the original query terms and the new expanded terms. Parameters specifying how many terms should be extracted and from how many documents these terms are to be selected from were also altered for the corresponding research questions.
**Performance and Retrievability Bias** To measure the performance in each of the experiments Mean Average Precision (MAP) was used. The retrievability bias was quantified using the Gini Coefficient similar to previous studies [2, 4, 11]. To quantify the bias of the system these steps are followed: first, generate a very large set of bigrams for each collection using automatic extraction where bigrams which occur at least 20 times were selected. Next, launch this query set for the corresponding collection and compute the document retrievability scores. Following this, Gini is used to compute the bias of the system using the **r(d)** scores. This outputs a decimal between 0 and 1 that represents the level of bias the system exerts on the collection, this can then be used to compare systems.

[1] http://terrier.org/

## 3.3 Experiments

**Experiment 1: Altering the Parameters of the Retrieval Model** The first experiment altered the $b$ parameter between the bounds of 0 and 1, traversing the space in steps of 0.1 (i.e 0, 0.1, 0.2, ..., 0.9, 1). Additionally, in these experiments Bo1 and KL were used with their parameters at default values of: $Rocchio's\beta = 0.4$, extracting 10 terms for expansion from the top 3 documents.
**Experiment 2: Altering Rocchio's Beta in the QE Method** The second experiment consisted of altering the Rocchio's Beta for the employed QE method on each of the retrieval models. As BM25 has the adjustable $b$ parameter, for the remaining experiments it was fixed to $b = 0.7$. For QE, 10 terms were extracted from the top 3 documents.
**Experiment 3: Altering the Number of Terms for Expansion** The third experiment again utilised both QE methods on the 3 retrieval models similar to the previous experiment. In this experiment, Rocchio's Beta was set to 0.4 while the number of terms extracted was explored. In each run, a different number of expansion terms were selected from the top 3 documents. The number of terms extracted were: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 12, 15, 20, 25 and 50. This approach replicated some of the experiments performed by Pickens (et al) [7] where performance was evaluated. We will also evaluate retrievability bias, an evaluation that has not previously been performed.
**Experiment 4: Altering the Number of Documents for Expansion** The final experiment follows the same method as Experiment 3 however, the number of documents that terms are extracted from are altered. 10 terms are extracted from a varying number from the top $x$ documents (1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 12, 15, 20, 25, 50).

# 4 RESULTS AND ANALYSIS

The results of the experiments are presented in this section, however for brevity, focus is placed on the results of T123 as similar patterns are observed on the other collections. When increasing the $b$ parameter for length normalisation in BM25 using the Bo1 and KL QE methods, results similar to those observed by Wilkie

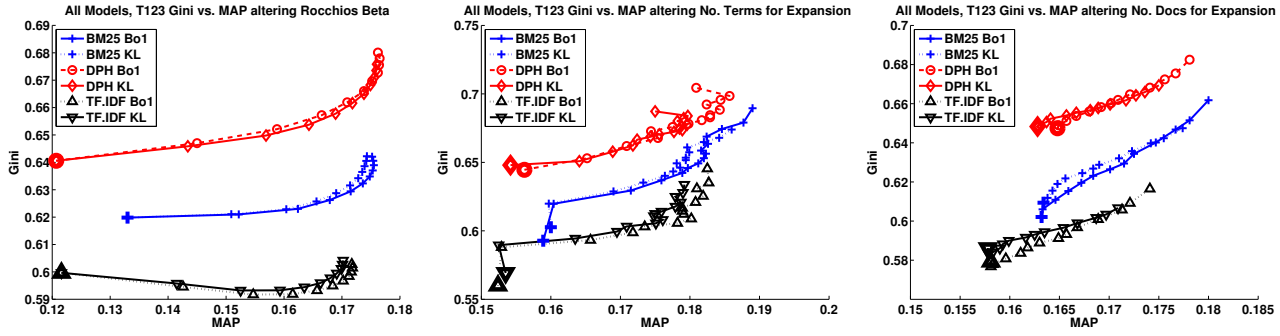


**Figure 2: Plots of Gini vs. MAP for BM25, DPH and TF.IDF when varying Rocchio's Beta (Left), the number of documents for expansion (Middle) or the number of terms for expansion (Right). The larger points towards the left of each graph indicate the smallest settings.**

and Azzopardi when no QE was performed [9] were evident in the left and middle plots of Figure 1. In terms of MAP, from $b = 0.0$ (the poorest performing point on the graph) there is a steady increase in MAP until the maximum MAP is found at $b = 0.3$ for T123 and $b = 0.7$ for DG for both Bo1 and KL. The difference here can be attributed to the variance in length between the collections and that DG has an average document length of 1108 compared with T123 of 439 meaning more length normalisation must be applied to DG to improve performance (thus a higher value of $b$). In terms of bias there is a steady decrease from $b = 0.0$, the most biased point, until the minimum point of bias of $b = 0.9$ for T123 and $b = 1.0$ on DG for both QE methods. Obviously, the point of minimum bias does not coincide with the point of maximum performance, also observed by Wilkie and Azzopardi. The rightmost plot of Figure 1 demonstrates how at low settings of $b$, the models apply very high retrievability to relevant documents making the rest of the collection unretrievable. As these low $b$ settings do not correspond with high performance scores, it is evident that a small set of the relevant documents are receiving huge $r(d)$ scores, dragging the average up (supported by the standard deviation at $b = 0.0$ of 2600 compared to 8 at $b = 1.0$) as this is a small set of the whole collection.

For the second experiment where the Rocchio's Beta was altered, results are shown in the left plot of Figure 2. In all cases, an increase in Rocchio's $\beta$ leads to improvements in performance however, gains are diminishing. In terms of bias for this experiment, for BM25 and DPH a direct correlation between performance and retrievability bias is observed, signifying that as performance improves, bias also increases.

The central plot of Figure 2 shows how the findings agree with Pickens *et al* [7] in that there is a constant increase in MAP as we add more terms for expansion. Similar to the finding in the previous experiment, there is a corresponding rise in bias as performance increases. Again, suggesting that the terms extracted are relevant but are causing the retrieval system to focus on a smaller set of documents than when done without QE. The Rocchio's Beta appears to have most impact on performance while the number of terms has the biggest impact on bias.

The final experiment yields very similar findings to the previous two experiments where a steady rise in MAP and a corresponding rise in bias is observed as more documents become available to select terms from.

## 5 CONCLUSION

Given the results presented, the following conclusions about QE's effect on retrievability bias and performance can be drawn. Increases in the number of terms and number of top documents used to extract the terms leads to increases in MAP as well as in bias, meaning the system is selecting useful terms for expansion but in doing so, is narrowing the set of documents that can be retrieved to a subset of the collection. Altering Rocchio's beta on each model, it was again evident that applying higher weight to the expansion terms lead to improvements in MAP and increases in bias. Finally, altering the $b$ parameter for BM25 provided results that reflect previous findings when no QE is performed where no match up between maximum performance and minimum bias appears. These findings suggest that the effectiveness of QE is in part, linked with an increase in bias associated with the system.

## REFERENCES


[1] L. Azzopardi and R. Bache. On the relationship between effectiveness and accessibility. In *Proc. of the 33rd ACM SIGIR*, pages 889–890, 2010.

[2] L. Azzopardi and V. Vinay. Retrievability: An evaluation measure for higher order information access tasks. In *Proc. of the 17th ACM CIKM*, pages 561–570, 2008.

[3] S. Bashir and A. Rauber. Analyzing document retrievability in patent retrieval settings. In *Database and Expert Systems Applications*, pages 753–760. 2009.

[4] S. Bashir and A. Rauber. Improving retrievability of patents with cluster-based pseudo-relevance feedback documents selection. In *Proc. of the 18th ACM CIKM*, pages 1863–1866, 2009.

[5] S. Bashir and A. Rauber. Improving retrievability of patents in prior-art search. In *Proc. of the 32nd ECIR*, pages 457–470, 2010.

[6] J. Gastwirth. The estimation of the lorenz curve and gini index. *The Review of Economics and Statistics*, 54:306–316, 1972.

[7] J. Pickens, M. Cooper, and G. Golovchinsky. Reverted indexing for feedback and expansion. In *Proc. of the 19th ACM CIKM*, pages 1049–1058, 2010.

[8] J. J. Rocchio. Relevance feedback in information retrieval. 1971.

[9] C. Wilkie and L. Azzopardi. Relating retrievability, performance and length. In *Proc. of the 36th ACM SIGIR conference*, pages 937–940, 2013.

[10] C. Wilkie and L. Azzopardi. Best and fairest: An empirical analysis of retrieval system bias. *Advances in Information Retrieval*, pages 13–25, 2014.

[11] C. Wilkie and L. Azzopardi. A retrievability analysis: Exploring the relationship between retrieval bias and retrieval performance. In *Proc. of the 23rd ACM CIKM*, pages 81–90, 2014.