# Query Length, Retrievability Bias and Performance

Colin Wilkie
School of Computing Science, University of Glasgow
Glasgow, United Kingdom
Colin.Wilkie@glasgow.ac.uk

Leif Azzopardi
School of Computing Science, University of Glasgow
Glasgow, United Kingdom
Leif.Azzopardi@glasgow.ac.uk

## ABSTRACT

Past work has shown that longer queries tend to lead to better retrieval performance. However, this comes at the cost of increased user effort effort and additional system processing. In this paper, we examine whether there are benefits of longer queries beyond performance. We posit that increasing the query length will also lead to a reduction in the retrievability bias. Additionally, we speculate that to minimise retrievability bias as queries become longer, more length normalisation must be applied to account for the increase in the length of documents retrieved. To this end, we perform a retrievability analysis on two TREC collections using three standard retrieval models and various lengths of queries (one to five terms). From this investigation we find that increasing the length of queries reduces the overall retrievability bias but at a decreasing rate. Moreover, once the query length exceeds three terms the bias can begin to increase (and the performance can start to drop). We also observe that more document length normalisation is typically required as query length increases, in order to minimise bias. Finally, we show that there is a strong correlation between performance and retrieval bias. This work raises some interesting questions regarding query length and its affect on performance and bias. Further work will be directed towards examining longer and more verbose queries, including those generated via query expansion methods, to obtain a more comprehensive understanding of the relationship between query length, performance and retrievability bias.

## Categories and Subject Descriptors

H.3.4 [**Information Storage and Retrieval**]: Systems and Software:Performance Evaluation

## General Terms

Theory, Experimentation

## Keywords

Retrievability, Performance, Evaluation, Query Length


*CIKM'15,* October 19–23, 2015, Melbourne, Australia.

DOI: http://dx.doi.org/10.1145/2806416.2806604.

## 1. INTRODUCTION

The retrievability of a document is essentially how easily a document can be found given a large set of queries and a retrieval system [3]. Consequently, retrievability is considered a precursor to relevance as it is a measure that predicts the likelihood that a document will be retrieved, and a document must be retrieved before it can be judged relevant. Recently, a number of studies have examined the relationship between retrievability and performance [12, 14]. The working hypothesis is, if a system reduces the retrieval bias (i.e. makes all documents equally retrievable) it will lead to better performance [12]. The intuition being if a system provides some chance to each document in the collection (assuming the document is worthy of retrieval) to be retrieved if that document is relevant to an information need, then there would be a number of queries that would retrieve the document at a rank sufficiently high enough that a user would encounter it. In these papers, they have shown that this hypothesis tends to hold for system ranking, parameter tuning, and document length normalisation Here we investigate the relationship between retrievability bias, query length and retrieval performance. Specifically, we examine whether increasing query length decreases the retrievability bias. The intuition being that as queries become more specific (by increasing their length), they focus on the a smaller set of documents and so provide those documents some level of retrievability. On the other hand, if one term is issued, then a large set of documents will be returned, so the documents at the top will amass retrievability. Consequently, if the system tends to favour certain documents over others then shorter queries are more likely to be subject to the system bias. We further speculate that longer queries will introduce other biases i.e. tend to favour the retrieval of longer documents because they have a higher likelihood of matching more query terms, and so could inadvertently introduce length bias thus may require more document length normalisation. To this end, we perform a comprehensive retrievability analysis to explore the relationship between queries containing one to five terms, on two TREC Test collections, using three standard retrieval model (using a range of parameter settings). The main contribution here is the first investigation into the impact of query length on retrievability bias.

## 2. BACKGROUND

In this section, we will give a brief overview of query length studies, before formally defining the retrievability measures and how retrievability relates to performance. Past research has shown that query length and retrieval performance is

| Collection | | AQ | | | | | DG | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | # Terms | One | Two | Three | Four | Five | One | Two | Three | Four | Five |
| **BM25** | $b$ | 0.70 | 0.75 | 0.80 | 0.85 | 0.85 | 0.90† | 0.90† | 1.00 | 1.00 | 1.00 |
| | **Min Gini** | 0.69 | 0.44 | 0.38 | 0.36 | 0.35 | 0.76 | 0.58 | 0.56 | 0.56 | 0.56 |
| | **Max S@10** | 0.06 | 0.24 | 0.43 | 0.56 | 0.66 | 0.04 | 0.16 | 0.22 | 0.25 | 0.26 |
| **LM** | $\beta$ | 200† | 100† | 100† | 100† | 200† | 100† | 50† | 1† | 50† | 50† |
| | **Min Gini** | 0.69 | 0.40 | 0.35 | 0.34 | 0.34 | 0.76 | 0.58 | 0.56 | 0.57 | 0.58 |
| | **Max S@10** | 0.06 | 0.26 | 0.45 | 0.58 | 0.68 | 0.05 | 0.15 | 0.21 | 0.24 | 0.25 |
| **PL2** | $c$ | 2.0 | 3.0 | 2.0 | 2.0 | 1.0 | 0.1† | 1.0† | 1.0† | 1.0† | 1.0† |
| | **Min Gini** | 0.69 | 0.46 | 0.40 | 0.38 | 0.37 | 0.76 | 0.64 | 0.63 | 0.63 | 0.64 |
| | **Max S@10** | 0.06 | 0.26 | 0.44 | 0.57 | 0.67 | 0.05 | 0.14 | 0.20 | 0.22 | 0.23 |

**Table 1: Shows the parameter setting at which minimum bias occurs (†indicates the parameter setting also achieved maximum Success@10) followed by the value of the minimum Gini and the value of the maximum Success@10 for each model on both collections.**

related, such that longer queries tend to result in better performance. In [7], they show that as the length of TREC queries increases the performance also increases. This has lead researchers in interactive IR to try and extract longer queries from users [1, 10]. In a more systematic study, Azzopardi [2], generated queries of length 1 to 30, issued them to three retrieval models and showed that both precision and MAP increased as query length increased. However, once the length exceeded 2-3 terms, smaller and smaller increases in performance were observed (i.e. the law of diminishing returns). The author argues that this is why query lengths tend to be short in practice. Performance does not always increase as the length of a query increases. Invariably, as more terms are added, the relevance of the terms with respect to the information need decreases, and so performance can drop because more noise is being added.

Little work has directly studied the relationship between query length and parameter estimation. When studying document length normalisation in BM25, Cummins [8] found that as the query length increased, more normalisation was required (i.e. $b$ has to increase) to achieve optimal performance. They hypothesised that this is due to longer queries giving longer documents more chances to match terms and thus, score higher. This potentially leads to a bias towards longer documents and so should be regulated. This result was also observed in a study by He and Ounis [9]. They noted that for the divergence from randomness model PL2, more normalisation was required for longer queries (i.e. a decrease in the $c$ parameter) in order to achieve best performance. In [15], Zhai and Lafferty performed a study of smoothing parameters on various language models, they found that longer queries required more smoothing (i.e. an increase in $\beta$) to extract the best performance from the model. To explore the notion in more detail, we will be investigating the bias that results from different query lengths across the parameter space of BM25, PL2, and Language Modelling. To estimate the bias stemming from query length, we shall use the novel evaluation measure, retrievability.

## Retrievability

In [3], Azzopardi and Vinay introduced the concept of retrievability, a measure that defined how *easily* a document could be retrieved by a particular configuration of an IR system. Formally, retrievability $\mathbf{r}$ of a document $\mathbf{d}$ with respect to an IR system is defined as:

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

where $\mathbf{q}$ is a query from the very large query set $\mathbf{Q}$, meaning $\mathbf{O_q}$ is the opportunity of the query being chosen from this set. $\mathbf{k_{dq}}$ is the rank at which $\mathbf{d}$ is retrieved given $\mathbf{q}$, and $\mathbf{f}(\mathbf{k_{dq}}, \mathbf{c})$ is a utility function which denotes how retrievable the document $\mathbf{d}$ is for the query $\mathbf{q}$ given the rank cutoff $\mathbf{c}$. Retrievability is then 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 [3, 4, 6, 12]. The standard measure of retrievability used is a cumulative based approximation, which employs an 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 $\mathbf{c}$ are completely ignored, simulating a user who is only willing to pursue the first $\mathbf{c}$ results. Essentially, the more queries that retrieve a document, the more retrievable a document is. Using the $r(d)$ scores of each of the documents, the Gini Coefficient is used to then estimate bias in a single value.

## Retrieval Bias and Query Length

The relationship between retrieval bias and performance has been investigated in a number of different ways [3, 4, 5, 11, 12, 14]. In [12] the focus was the relationship between performance and bias in system ranking, while in [11, 14] it was on how retrieval bias correlates to performance and document length normalisation. These studies have shown, that in general, a reduction in retrieval bias correlates to an increase in performance, and that systems that exhibit a lower retrieval bias tend to deliver better performance (with the strongest correlation being with Time-Biased gain and the U-Measure) [11].

However, there has not been any studies that have directly investigated the relationship between bias, performance and query length. Indeed most studies use queries of one length, typically, bigram queries [3, 11, 12, 14]. A few studies, however, do provides some insights into the relationship, as they have explored different length queries [4, 5] though for a different purpose.

In [4], Bashir and Rauber perform a retrievability analysis using query sets with two, three and four terms. There results show that the average retrievability of documents decreased with the length of the queries, suggesting that more bias may be introduced due to query length. However, they did not report the retrieval bias scores (i.e. the Gini Coefficients), so it is not possible to determine if this is the case, or whether this is an artefact of the different number of queries used in each set. In [13], Wilkie and Azzopardi

show that the retrieval bias estimate is greatly influenced by the size of the query set used. Therefore in our experiments we control and ensure that the same number of queries per length is used.

In [5], Bashir and Rauber undertake a different retrievability analysis where they examine the influence of query expansion on retrieval bias. They show that for typical query expansion methods retrieval bias increased substantially between the non-expanded (short) queries and the expanded (long) queries. This suggests that longer queries may indeed increase retrieval bias. While they devise an alternative query expansion technique that ameliorates the bias increase, they did not report the corresponding retrieval performances, so it is unclear how the bias stemming from query length relates to performance. In this paper, we specifically probe the relationship between between query length, performance and bias using a methodology designed to compared the differences across lengths.

# 3. EXPERIMENTAL METHOD

The aim of this study is to investigate how varying the length of queries to estimate retrievability alters how retrievability bias relates to performance. More succinctly, we wish to observe whether using longer queries queries leads to a decrease in the estimation of bias by a system.

## Data and Materials

To perform our experiments, we employed 2 standard TREC test collections: AQ (Aquaint) and DG (DotGov). We utilised 3 standard retrieval models: BM25, PL2 and Language Modelling with Bayes Smoothing and varied the document length normalisation parameter. With BM25 we used 11 parameter settings for $b$ between 0.0 and 1.0 increasing in steps of 0.1. We also include $b$ at 0.75 and 0.85 for further granularity. For PL2 we set parameter $c$ to values between 1 and 10 but also included 0.1 and 100 to test extreme cases. For Bayes Smoothing (LM), we set the $\beta$ to values of: 1, 5, 10, 20, 50, 100, 200, 300, 500, 1000, 2000, 3000, 5000 and 10,000.

## Query Generation Method

To perform our experiments we use a novel query generation technique that generates queries from the titles of documents. We generate queries of 1 to 5 terms by extracting terms from the title and when there is not enough terms in the title, we extract terms from the main text. This means we have 5 query sets (a set of single term queries, a set of two term queries, etc.) that contain a query for every document in the collection. A number of issues exist in the domain of query generation that we must address. One such issue that we must mitigate against is the generation of duplicate queries. For example when extracting queries from two unique documents on similar topics, it would be common to extract the same query from both. To reduce the chance of this occurring, we extract terms ordered by their IDF in the hope that more original terms will be picked out for each document that are less likely to have been seen before. Extracting queries strictly from the title was impractical as many titles are very short and many unique documents share the same title. Therefore to avoid this issue, we include the main body of the document to locate additional terms.

## Performance Evaluation

Given the query generation method, for each query there is a corresponding relevant document, creating a series of known-item query pairs which result in a relevance judgements file. Using these judgements we computed a number of measures (i.e. P@10, MAP/MRR, Success@10), but report Success@10. Similarly correlations and findings were observed on these other measures.

# 4. RESULTS AND ANALYSIS

**Query Length and Length Normalisation:** Table 1 demonstrates the relationship between query length and the length normalisation parameter associated with a model. From this table it is apparent, that to minimise the bias of a system, issuing short queries requires less length normalisation than issuing longer queries. For example, when using BM25 on AQ to minimise bias, $b$ must increase as we issue more terms (0.7, 0.75, 0.8, 0.85 and 0.85 in that order for one to five terms). This trend can also be observed on PL2 although it is not as pronounced. For PL2 on DG $c$ must be set to 0.1 for minimal bias using one term and $c = 1$ for two or more terms. The need for different parameter settings for two to five terms may have been evident if the space between 0.1 and 1 was further explored. The increase in the amount of length normalisation required to minimise bias may be attributed to longer documents being returned when longer queries are issued, as was suggested by Cummins [8]. Interestingly, LM performs very differently from BM25 and PL2 and we see a dip in the amount of length normalisation required. This finding suggests that an ideal length of query for LM may exist but would require further exploration of queries with more than five terms and the $\beta$ parameter space between 0 and 100.

**Query Length and Retrievability Bias:** The effect of query length and retrievability bias is shown to be complex as we see conflicting evidence in Table 1. For most models we see that increasing the number of terms results in a decrease in bias but this trend is subject to diminishing returns (particularly evident when using BM25 on AQ). However, when employing some models on DG (LM and PL2) we see that adding more than three terms actually results in an increase in bias. We hypothesise this increase is due to the large differences in average document length in each of the collections. As we previously observed more terms requires more length normalisation, it is plausible that given the much larger average document length present in DG causes the system to be unable to apply enough length normalisation once too many terms are used to continue to reduce bias. We argue this explanation is more probable than that adding more terms adds more noise [5], due to the fact our method of query generation extracts discriminative terms from a single document rather than from a several of documents like typical query expansion techniques.

**Query Length, Retrievability Bias and Performance:** The plots of Figure 1 in conjunction with Table 2 summarise the relationship between query length, retrievability bias and performance. Table 2 shows that there is strong negative and significant correlations between performance and bias in all but one case. This indicates that decreasing bias tends to improve performance, agreeing with the findings reported in [12, 14]. Tables 1 and 2 also shows that in over half of the cases, minimising bias leads to the

| Collection | | AQ | | | | | DG | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| # Terms | | One | Two | Three | Four | Five | One | Two | Three | Four | Five |
| Model | BM25 | -0.84* | -0.91* | -0.89* | -0.88* | -0.91* | -0.99* | -1.00* | -0.99* | -0.99* | -0.99* |
| | LM | -0.74* | -1.00* | -0.98* | -0.99* | -1.00* | -0.99* | -1.00* | -1.00* | -1.00* | -1.00* |
| | PL2 | -0.57 | -0.99* | -0.91* | -0.80* | -0.71* | -0.99* | -0.97* | -1.00* | -0.99* | -0.98* |

**Table 2: Correlations between Gini and Success@10. * denotes statistical significance at p<0.05.**

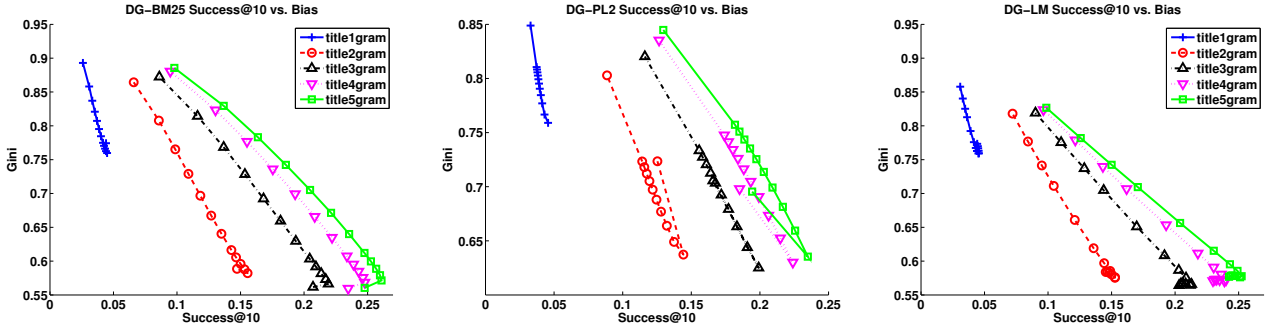


**Figure 1: Plots showcasing the relationship between bias, performance and query length. We see a trend that decreasing bias leads to better performance and that increasing the length of queries does not necessarily lower bias.**

best performance and the plots of Figure 1 reinforce this. In terms of query length and performance, we see that as query length increases, then the performance increases but with diminishing returns (as was observed in [2]).

## 5. FUTURE WORK AND CONCLUSIONS

This work has helped further discern the nature of the relationship between bias and performance by investigating the effects of query length, another variable in the retrieval process that impacts on retrieval bias. The findings have highlighted the relationship between performance and bias where reducing bias tends to improve performance, conforming with the findings of previous studies [12, 14]. In addition to this, we have found that increasing the length of queries generally reduces bias, but at a decreasing rate (i.e. diminishing return). However, issuing queries longer than five terms may actually lead to increases in bias due to noise being added which motivates examining queries with more terms. Query expansion techniques often generate very long queries and so is an ideal domain in which to explore the relationship in more detail. Finally, during this study it became evident that the parameter space we explored for certain models (mainly PL2 and LM) was not fine grained enough, so it would be interesting follow up this study probing more deeply into the parameter space.

## 6. REFERENCES


[1] E. Agapie, G. Golovchinsky, and P. Qvarfordt. Leading people to longer queries. In *Proc. of ACM SIGCHI*, pages 3019–3022, 2013.

[2] L. Azzopardi. Query side evaluation: an empirical analysis of effectiveness and effort. In *Proc. of the 32nd ACM SIGIR*, pages 556–563, 2009.

[3] 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.

[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 & recall by automatic corpus partitioning. In *Trans. on large-scale data & knowledge-centered sys. II*, pages 122–140. 2010.

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

[7] N. J. Belkin, D. Kelly, G. Kim, J.-Y. Kim, H.-J. Lee, G. Muresan, M.-C. Tang, X.-J. Yuan, and C. Cool. Query length in interactive information retrieval. In *Proc. of the 26th ACM SIGIR*, pages 205–212, 2003.

[8] R. Cummins and C. O'Riordan. Learning in a pairwise term-term proximity framework for information retrieval. In *Proc of the 32nd ACM SIGIR*, 2009.

[9] B. He and I. Ounis. Parameter sensitivity in the probabilistic model for ad-hoc retrieval. In *Proc. of the 16th ACM CIKM*, pages 263–272, 2007.

[10] D. Kelly, V. D. Dollu, and X. Fu. The loquacious user: A document-independent source of terms for query expansion. In *Proc. of the 28th ACM SIGIR*, pages 457–464, 2005.

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

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

[13] C. Wilkie and L. Azzopardi. Efficiently estimating retrievability bias. In *Advances in Information Retrieval*, pages 720–726, 2014.

[14] 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.

[15] C. Zhai and J. Lafferty. A study of smoothing methods for language models applied to ad hoc information retrieval. In *Proc. of the 24th ACM SIGIR*, pages 334–342, 2001.