A Topical Approach to Retrievability Bias Estimation

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

Leif Azzopardi
University of Strathclyde
Glasgow, United Kingdom
leifos@acm.org

ABSTRACT

Retrievability is an independent evaluation measure that offers insights to an aspect of retrieval systems that performance and efficiency measures do not. Retrievability is often used to calculate the retrievability bias, an indication of how accessible a system makes all the documents in a collection. Generally, computing the retrievability bias of a system requires a colossal number of queries to be issued for the system to gain an accurate estimate of the bias. However, it is often the case that the accuracy of the estimate is not of importance, but the relationship between the estimate of bias and performance when tuning a systems parameters. As such, reaching a stable estimation of bias for the system is more important than getting very accurate retrievability scores for individual documents. This work explores the idea of using topical subsets of the collection for query generation and bias estimation to form a local estimate of bias which correlates with the global estimate of retrievability bias. By using topical subsets, it would be possible to reduce the volume of queries required to reach an accurate estimate of retrievability bias, reducing the time and resources required to perform a retrievability analysis. Findings suggest that this is a viable approach to estimating retrievability bias and that the number of queries required can be reduced to less than a quarter of what was previously thought necessary.

1. INTRODUCTION

Retrievability is an estimate of how easily documents in a collection can be found using a particular retrieval system [2]. Retrievability offers an alternative view of a retrieval system by investigating the influence the system exerts on the collection by which documents it provides access to. Performing a retrievability analysis is primarily done to evaluate whether a system is biased towards a particular set of documents for inappropriate reasons, such as a system favouring documents that are incredibly long. This is particularly useful when assessing new systems for any unknown biases which may influence performance. Retrievability can also be used to identify individual documents that are hard to find which can help identify problem areas where important documents are difficult to access in enterprise settings. Retrievability has also been shown to correlate with some performance measures and so, researchers have suggested that it could be used to tune a system depending on what aspect of performance is most important to the user [14]. Doing so would avoid recourse to relevancy judgements when the standard method of computing retrievability is performed. These aspects of retrievability make it a valuable tool in the evaluation of systems, however, retrievability has not yet seen widespread use due to some limitations in how it is computed.

The biggest limitation when performing a retrievability analysis concerns the resources required, especially in terms of time as the current method of computing retrievability requires a brute force approach, launching large sets of queries to cover a range of potential queries that could be issued to the system. Typically, these large query sets are generated automatically, often by extracting very large numbers of bigrams from the whole collection to attempt to cover all the documents in the collection. This means that a retrievability analysis simply cannot be performed on large data sets as the compute time would be huge. This analysis also cannot be performed on continually streaming data sets as each new document would require the retrievability to be recomputed for the full collection. This is a huge problem as it limits the usefulness of retrievability in its current state. This work provides an investigation into providing an alternate, more efficient, method of computing the overall retrievability bias of a system. This method, however, will not allow for an accurate retrievability score to be generated for every document in the collection. This novel method of computing retrievability is based on the idea that there are particular topics within a collection that can be deemed important and as such should be the focus of such an analysis. In focussing on these documents, the space is reduced in such a way that important documents are assessed in a much more efficient way.

The remainder of this paper will introduce the relevant background material covering both retrievability and its place in research literature so far. Following this, the method used to garner results for this analysis is described before an analysis of these results is performed, detailing the key findings of this study. Finally, future work is listed to confirm the generalisability of these findings as well as how to expand on these findings and further applications of this work.

2. BACKGROUND

Retrievability is a document centric evaluation method [2], that provides an alternate view on how a retrieval system interacts with a collection. Retrievability measures 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, 5, 7, 12]. 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.

The typical approach to compute retrievability has been to extract a large query set from the collection in the form of either bigrams, unigrams, n-grams or a combination of all. This set is often very large and so authors have put in artificial limits that dictate how many times a query must appear in the collection before it will be added to the set [11]. Once the query set has been created, all the queries are issued to the system in question and the top 100 (dependant on the $\mathbf{c}$ selected) results are recorded. Once all queries have been issued, each result is taken in turn and the scores for each document are applied (based on the chosen utility function.) From all the results, a list of the $\mathbf{r}(\mathbf{d})$ for every document in the collection will be generated. This allows the user to investigate individual document scores and see which documents are particularly difficult or easy to find.

To convert the $\mathbf{r}(\mathbf{d})$ for each document into a single value describing bias, an inequality metric is used to assess the distribution of wealth in a population. However, the retrievability of a document in a collection fits this paradigm as the retrievability can be considered the documents wealth and the collection is the population that retrievability is distributed amongst. The Gini Coefficient can be used to calculate the level of inequality in a population by comparing the distribution to the Lorenz Curve. This estimate indicates the retrievability bias, with a score near 1 indicating total inequality while a score approaching 0 denotes total equality.

Research by a number of authors has began to address the issue of resources required to compute effective estimates of retrievability. The most naive approach, by Wilkie and Azzopardi [13] investigated the impact of simply reducing the number of bigrams issued to the collection. In this work, the authors followed their previous methodology of extracting bigrams from the whole collection and issuing them to a system and estimating retrievability however, they created multiple sets of bigrams for each collection which were all subsets of the original bigram extraction. The authors ordered the bigrams by their log likelihood scores and then removed percentages of the lowest scoring bigrams to create new sets. This way, they created 10 sets of bigrams ranging from 10% of the highest scoring bigrams up to all of the bigrams extracted in intervals of %10 (i.e. 10%, 20%, 30%... etc). The findings of this work revealed that, on very biased systems (i.e. systems that have very high Gini Coefficients, like TF.IDF) large amounts of the bigrams can be removed, using sets as small as 40% of all bigrams, still produced a very reasonable estimate of the Gini Coefficient for the collection. On the other hand, systems that were know to be less biased, such as BM25 and PL2, had very little leeway in the number of bigrams necessary and that removing as few as 10% on a well tuned system lead to statistically significant differences in the estimation of the Gini Coefficient. Therefore, this work demonstrated that reducing the number of bigrams was not a viable method of reducing the required resources to compute retrievability bias accurately. More recent work by Lipani and Lupu [8] laid the groundwork for the creation of an analytical method of computing retrievability bias. This work, essentially, creates an upper bound to how retrievable a document can be by employing a boolean model. The work can be broken down to two main findings, when using bigrams with an *AND* between terms, the document containing both terms accumulate retrievability while all other documents receive no retrievability. When using an *OR*, any document that contains either term will receive a share of retrievability. The disadvantage of this technique is that there is no document ranking in place, therefore every document must be considered equally retrievable, thus cumulative and gravity scoring models have no place here. The method, in its current state, is therefore too naive to effectively compute an accurate estimate of retrievability but is useful for defining this upper bound. Finally, work by Bashir [4] also attempted to take an alternate, analytical approach to calculating retrievability bias. In this work, Bashir employed features of the documents in the collection (such as length, unique vocabulary, etc) to estimate retrievability. The key findings of the work were that this method could actually be used to estimate the Gini Coefficient reasonably well but, like the other work mentioned here, was unable to efficiently and accurately estimate the individual retrievability score of the documents in the collection. Therefore, further work is still required to find a method of efficient estimation that can accurately determine both the Gini Coefficient of the system as a whole and the individual retrievability scores of the documents in the collection.

Another important aspect of research regarding retrievability focuses on the relationship between retrievability bias and retrieval performance [2, 3, 5, 6, 11, 12, 14]. Work by Wilkie and Azzopardi has investigated this relationship in a variety of ways including the relationship between performance and bias when selecting which system to employ [12] as well as the relationship when tuning a particular systems length normalisation settings [11, 14]. The authors found the relationship between performance and bias to be non-linear when TREC performance measures are used. However, when the authors investigated system tuning using metrics like Time Biased Gain [10] and U-Measure [9], the parameter setting that minimised bias also maximised performance. This was a very important finding as it suggested that it could be possible to tune a system well using the results of a retrievability analysis, thus removing the need for resorting to relevancy judgments.

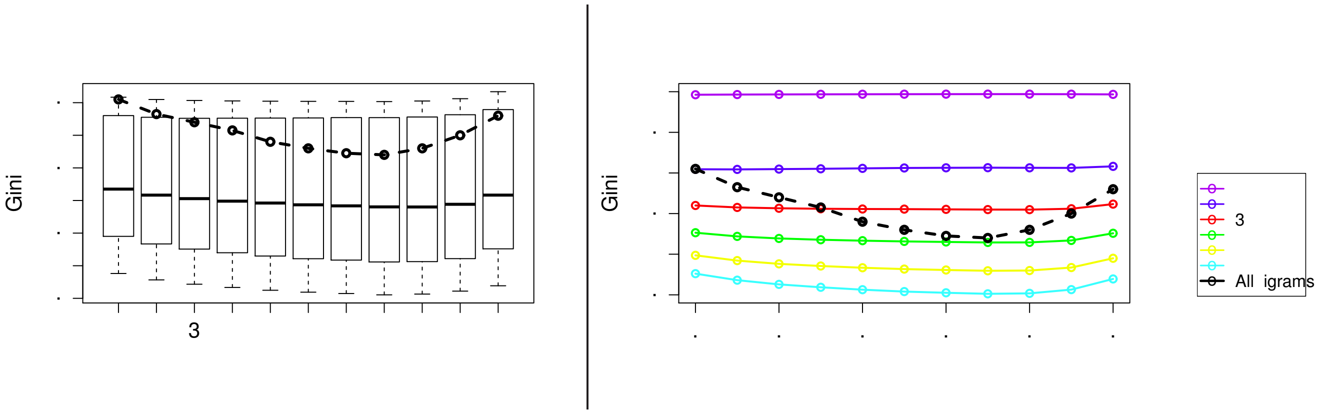

Figure 1: (Left) Box plot showing the max, min and mean Gini score across the 50 topics issued. The dashed line represents the Gini computed when using bigrams from the whole collection. (Right) Line plot showing the mean Gini Coefficient for each query set used. The lines are ordered from top to bottom by increasing volume of queries in each set.

3. METHOD

The focus of this work is to analyse whether or not retrievability bias can be accurately estimated by using a novel topic centric approach. This approach focusses on the topic pools of the TREC collections. Given that this approach is viable, the next step of the investigation will assess how few queries are needed, per topic, to arrive at a reasonable estimation of retrievability bias, that agrees with the estimate of the more traditional, high volume approach.

The new method is performed by first identifying the pool of documents that were judged for a topic. Bigrams are then extracted from these pooled documents only, thus creating a small bigram set which is, theoretically, topic centric. This is repeated for each topic, thus giving several small bigram query sets. The query sets generated were trimmed down to 600 queries as this was near the average number of queries extracted for each topic. Doing so prevented a small set of topics being overly dominant in the study as some topics were providing in excess of 20,000 queries on their own. The sets were trimmed down to the 600 queries with the highest log likelihood scores thus creating a set of 600 'realistic' bigrams. Following this, the bigrams are issued to the system and the top 100 results recorded, this is repeated for several parameter settings of the retrieval system. Next, the retrievability of each document is calculated by counting how many times each document occurs in the top 100 results for each query issued. However, unlike previous studies, retrievability is only calculated for documents that are retrieved at least once from any of the queries issued. Essentially, all documents that are never retrieved are ignored, thus removing the large set of documents with 0 score. Finally, retrievability bias is computed as normal, using the Gini Coefficient. This produces an estimate of retrievability for a set of bigrams extracted from a single topic, therefore the process is repeated for each topic in the collection. Once the process has been done for every topic, the Gini Coefficients from each topic are averaged across to give a single estimate of Gini.

The experiment detailed above was performed on the Associated Press (AP) collection as this study is an initial investigation of the viability of the ideas presented. With AP, topics numbered 151 to 200 were used as were their respective relevancy judgements. The system used was BM25 and the b parameter space was investigated. b was set to values ranging from 0.0 to 1.0 in intervals of 0.1.

4. RESULTS AND ANALYSIS

Analysing the results the method used created, a number of interesting observations are apparent. To begin with, the left plot of Figure 1 presents information regarding the Gini coefficient calculated from each of the 50 topics, using their respective (600 query) query sets. The plot presents the mean, quartiles, maximum and minimum Gini coefficients found as the b parameter of BM25 is adjusted. The plot demonstrates that the topics have a very broad range of estimations for Gini, generally covering a range of roughly 0.8 to 0.2. This obviously suggests that the Gini Coefficient of the system cannot be accurately estimated when using a random set of topics and as the range is so broad (as indicated by the quartiles) it would be difficult to effectively select a subset of topics to create an estimate with. It is also worth noting that the mean Gini at each b value follows a similar trend to the Gini found when the traditional approach was used. Upon further investigation, it was found that a very strong, significant correlation of 0.982 between the traditional approach and this new approach exists. This suggests that, for this collection and system, a user can predict the b setting at which the minimum Gini coefficient will be found as well as the consequences of either increasing or decreasing b. The benefit of this approach is the large reduction in queries that must be issued to accurately predict the b setting for minimum Gini. The traditional approach requires 81,000 queries to be issued to the system, the new method only requires 30,000 queries issued. This is a huge saving in terms of resources however, it must be noted that although it is possible to predict the setting for minimum Gini, this method gives a very different estimate of Gini due to the fact that it is only performed on a subset of the documents in the collection (i.e. only documents that were retrieved for at least 1 of the 30,000 queries).

Finding that the b parameter can be estimated using this new approach, the next step in the investigation was to find how few queries were needed from the 50 topics to gain a reasonable estimate of Gini. The right plot of Figure 1 presents the results of the retrievability analysis when query sets containing various amounts of queries (between 100 and 600

No. Queries	100	200	300	400	500	600	All
Min. Gini	0.99	0.81	0.71	0.63	0.56	0.50	0.64
b	0.0	0.1	0.8	0.7	0.7	0.7	0.7
Corr. w/ All	*-0.96	-0.41	*0.83	*0.95	*0.97	* 0.98	1

Table 1: Table presenting the actual minimum Gini values, along with the *b* parameter setting it was found at, and the correlation with the traditional method of calculating Gini. * denotes statistical significance at $p<0.05$.

queries) are issued to the system. It is immediately evident that lower volumes of queries (100 - 300) are not producing any notable differences when the *b* parameter is adjusted. However, we see that the Gini estimates for the traditional method (dashed black line) sits in the region between the estimates produced from 200 to 400 queries being issued. However, as noted before, this traditional method is calculating Gini based on the retrievability of all the documents in the collection, while this new method only computes Gini based on documents retrieved. Therefore, the Gini estimate for the traditional method is expected to remain high as not all documents have the chance to be retrieved. From the plot it can be established that the lowest Gini estimate in the traditional method appears when *b* is set to 0.7. This plot also shows that the point of minimum Gini for 400, 500 and 600 query set is also $b = 0.7$ however, launching less than 400 queries results in different minimum settings being found, also demonstrated in Table 1. This table also shows that with as few as 400 queries issued for the 50 topics, a strong, significant correlation with the traditional method is still found. This suggests that as few as 20,000 queries can be issued to the system and still produce an accurate estimation of Gini, less than a quarter of the queries originally needed.

5. CONCLUSIONS AND FUTURE WORK

The investigation performed in this work was an initial exploration of an alternative, cheaper way to calculate retrievability by extracting the queries from the topic pools, rather than the full collection, and computing the Gini coefficient across documents which were retrieved for at least one query, instead of all documents. The key findings were, for this collection and configuration of the system, a very strong significant correlation is observed between the traditional estimate of Gini and the estimate produced by the method introduced in this paper. Further to this, this work expanded on the efficiency aspect by finding that a user can further reduce the number of queries extracted from each topic pool an still reach a reasonable estimate of Gini for identifying the *b* parameter setting which minimises Gini. These findings suggest that this is a viable method for computing retrievability bias cheaper than the existing approaches by utilising less than a quarter of the volume of queries originally required. However, as this is an initial study, performed on only one small collection, it is vital that these findings are explored on a wide range of both test collections and retrieval models before any wider conclusions can be drawn. It would also be pertinent to investigate how the Gini estimates relate to performance on a topic level to find if particular topics suggest that minimising Gini also maximises performance.

6. REFERENCES

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

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

[3] Bache, R.: Measuring and improving access to the corpus. In: Current Challenges in Patent Information Retrieval, The Information Retrieval Series, vol. 29, pp. 147–165 (2011)

[4] Bashir, S.: Estimating retrievability ranks of documents using document features. Neurocomput. 123, 216–232 (Jan 2014), http://dx.doi.org/10.1016/j.neucom.2013.07.011

[5] Bashir, S., Rauber, A.: Improving retrievability of patents with cluster-based pseudo-relevance feedback documents selection. In: Proc. of the 18th ACM CIKM. pp. 1863–1866 (2009)

[6] Bashir, S., Rauber, A.: Improving retrievability & recall by automatic corpus partitioning. In: Trans. on large-scale data & knowledge-centered sys. II, pp. 122–140 (2010)

[7] Bashir, S., Rauber, A.: Improving retrievability of patents in prior-art search. In: Proc. of the 32nd ECIR. pp. 457–470 (2010)

[8] Lipani, A., Lupu, M., Aizawa, A., Hanbury, A.: An initial analytical exploration of retrievability. In: Proc. of the 2015 ICTIR. pp. 329–332. ICTIR '15, ACM (2015)

[9] Sakai, T., Dou, Z.: Summaries, ranked retrieval and sessions: a unified framework for information access evaluation. In: Proc. of the 36th ACM SIGIR. pp. 473–482. SIGIR '13 (2013)

[10] Smucker, M.D., Clarke, C.L.: Time-based calibration of effectiveness measures. In: Proc. of the 35th ACM SIGIR. pp. 95–104 (2012)

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

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

[13] Wilkie, C., Azzopardi, L.: Efficiently estimating retrievability bias. In: Advances in Information Retrieval. pp. 720–726 (2014)

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