Retrievability: An Evaluation Measure for Higher Order Information Access Tasks

Leif Azzopardi
Department of Computing Science
University of Glasgow
Glasgow, UK
leif@dcs.gla.ac.uk

Vishwa Vinay
Microsoft Research Cambridge
7 J J Thomson Avenue
Cambridge, UK
vvinay@microsoft.com

ABSTRACT

Evaluation in Information Retrieval (IR) has long focused on effectiveness and efficiency. However, new and emerging access tasks now demand alternative evaluation measures which go beyond this traditional view. A retrieval system provides a means of gaining access to documents, therefore intuitively, our view of the collection is shaped by the retrieval system. In this paper, we outline some emerging information access related scenarios that require knowledge about how the retrieval system affects the users' ability to access information. This provides the motivation for the proposed evaluation measures and methodology where the focus is on capturing the behavior of the system, in terms of how retrievable it makes individual documents within the collection. To demonstrate the utility of the proposed methods, we perform an extensive analysis on two TREC collections showing how the measures can be applied to evaluate different information access questions. For higher order information access tasks that are inherently dependent on retrievability, our novel evaluation methodology emphasizes that effectiveness is an insufficient characterization of a retrieval system. This paper provides the foundations for the evaluation of higher order access related tasks.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Retrieval Models

General Terms

Measurement, Performance, Experimentation

1. INTRODUCTION

With search playing an increasingly crucial role in the access to information, there are growing concerns over the role of this technology [8, 11]. This is because as larger amounts of information is being made available online, Information Retrieval (IR) systems, as exemplified by the many search engines available today, are becoming the primary means of accessing this information [10]. Consequently, issues are being raised from a number of different areas questioning the influence that IR systems have on the access to information. For example, media regulators are concerned over whether search engines are biased towards particular websites over others [16, 9], while e-Government administrators in the U.S. are now legally required to ensure that all government information online is accessible through search engines [11]. Legal and patent searchers also need to ensure that they have IR systems which enable them to find all documents in the collection relevant to their information need. These situations require a way to evaluate an IR system in terms of how much access they provide into the underlying collection. To address these types of questions, we must develop new measures that indicate how easily documents or sets of documents in the collection can be accessed given the IR system. These measures can then be used to compare the influence of different IR systems on the access to information.

Given these motivations, it is the purpose of this paper to develop suitable methods to quantify the influence that a retrieval system has on the access to information. This influence of an IR system, exerted either explicitly or implicitly, is evoked at two junctions: deciding which documents are used as input into the index of the system, and when producing a ranked list of documents as output in response to a query. The ability to find a document through the retrieval process is therefore a combination of whether or not the document is indexed, and then whether or not a document can be retrieved through querying. The first factor has already been well studied [14, 6], illustrating the importance of making content discoverable by search engine crawlers to ensure inclusion in the index. The second factor is more subtle and is less understood. This is because it is generally assumed that if a document is indexed, it can be retrieved; the only impeding factors being:

- a user's ability to formulate his/her need in the form of a suitable query,
- the retrieval system's *matching function*, and,
- the user's willingness to examine documents.

However, the combination of these factors means that some documents are more easily accessible than others. This is because, search is a process of discrimination: a user deciding how to pose the query, the IR system attempting to discriminate between relevant and non-relevant content, and

CIKM'08, October 26–30, 2008, Napa Valley, California, USA.

the user wading through portions of the results retrieved by the system trying to find content relevant to his/her need. The main objective of the retrieval system is to maximize effectiveness, which implicitly leads to favoring certain types of documents over others; and so it can be said that the system is inherently biased[1]. Applying or removing particular retrieval biases have been shown to significantly improve or degrade effectiveness. For example, using inlinks to favor more popular documents [5], or accounting for document length normalization [13] can lead to significant improvements to retrieval effectiveness. The dual nature of discrimination, e.g. "favor longer over shorter", means that some documents can become more *retrievable* at the expense of others. The interest of this paper lies in understanding what influence the pursuit of increased effectiveness, and the associated accumulation of biases, has on all documents in the collection, not just the set of relevant documents.

The contribution in this paper is two-fold. Firstly, we propose a retrieval-oriented methodology for evaluating the influence of an IR system on information access. Secondly, we provide a quantitative measure of a *document's retrievability* and demonstrate how it can be used in the evaluation of a number of different information access scenarios. A document's retrievability captures the ease with which the document can be retrieved given a particular IR system, and enables for instance, a way to assess the extent to which information within e-Government websites is accessible. Examples to this effect are shown in Section 4, where we perform an extensive analysis of several standard retrieval models on two different TREC collections, the AQUAINT (News) collection and the .GOV (Web) collection. We first calibrate and test the proposed measure of retrievability and verify that the measure is working as expected, before conducting an analysis on how the retrievability imposed by the standard IR models affects the access into the documents indexed by the system. Then, we conduct a final series of experiments where we relate the influence a retrieval system exerts on the access to the collection and the retrieval effectiveness it achieves. Our main findings are:

- Well known retrieval systems do exhibit retrieval bias, and the presence and influence of retrieval bias is not fully captured by traditional relevance-based evaluation,
- In extreme cases, up to 80% of the least retrievable documents in the collection can be removed without significantly degrading the system effectiveness,
- The least retrievable documents, those that the retrieval system least favors, are significantly more difficult to find, and,
- The extent to which this is a problem in terms of accessing information in the collection depends on the degree of bias imposed by the retrieval system. The more bias imposed, the greater the impedance to the least retrievable documents in the collection.

Finally, in Section 5, we conclude that retrievability provides a useful indicator of the interaction between the collection and the retrieval system, before providing an outline of directions for future work.

[1]In this paper, the term '*bias*' is used according to the definition in [9], to refer to situations when a group of documents is favored or preferred over others.

2. RELATED WORK

Information Retrieval is the area that deals with the storage, organization and access of information [15]. The purpose of an IR system is to deliver relevant content to the user's request and it should do this effectively and efficiently. There is an extensive amount of literature that aims to address these facets of IR. Largely, research is concerned with effectiveness. That is, retrieving relevant information in response to a query. The necessary pre-cursor to relevance is therefore retrieval, i.e., an indexed document must be retrieved, before it can be judged relevant or not[3]. This condition determines whether or not a document can be accessed through the system, and how easily the document can be accessed.

First of all, the document must be present in the index of the IR system; otherwise there is no possibility for the document to ever be presented in a ranked list of results. Providing effective responses to queries therefore relies on possessing the right content in the collection. In the context of web search, it is vitally important for web-site owners to increase the visibility of their web pages to search engine crawlers in order to increase the likelihood of their pages being indexed by the system. Dasgupta et al [6] refer to this as the *discoverability* of pages on the web, and Upstill et al [14] refer to this as the *crawlability*. In the current paper, we assume that the documents of concern have been indexed, and turn our attention to the second junction of retrieval: the ranking function.

An IR system is evaluated in terms of being able to identify those documents in the collection that match the user's request. The inability of a retrieval system to do so leads to analysis of the result sets so that the ranking algorithm can be subsequently improved. For example, in [13] a bias towards length is corrected, and in [10] a method to remove bias due to popularity is suggested. Length and (link-based) popularity are just two features of a document that might influence its position in ranked results produced by a retrieval function. Consistently favoring the retrieval of documents based on such features will invariably lead to a persistent retrieval bias, where documents with such features are in general more likely to be retrieved.

Because of this, a number of studies have been conducted which attempt to determine whether search engines are biased. Such studies have considered a range of possible biases, for example, if one site has more coverage than another [8], whether sites in particular geographical locations are favored [16], or whether search engines are biased given a particular topic [9].

However, the studies performed have used crude measures based on coverage to determine the existance of bias, and the later studies were performed using only a handful of samples. In this paper, we propose a robust measure for quantifying the level of access afforded to individual documents, which can be used to determine whether there is any relative retrieval bias. It should be pointed out that search is a process of discrimination; and as such, retrieval systems will be naturally biased in some way; so simply detecting bias is not enough. So as part of this work, we examine whether the retrieval bias actually impedes one's access to information in the collection, and if so, to what extent. This has not been considered in prior work where the assumption has been that "bias is bad". This latter stance might be justified in certain situations. For instance, perceived bias in web

search rankings has led to legal action being taken against a well known search engine company[2]. While, in the area of e-Government, ensuring that online content is accessible is a very important concern, because citizens of a democratic country have a right to access the information. If the information is hidden from the public then this could jeopardize the integrity of the government. The importance of e-Government content being made accessible through search technology was highlighted in a recent report[3]. This resulted in changes to U.S. legislation[4], which requires that government websites be monitored and assessed in terms of how "searchable" they are, to ensure that all government information is accessible through search engines.

Our main objective is to more precisely understand how the retrieval system affects one's ability to access the content housed within the index/collection. In order to do so, we introduce a measure of retrievability in the following section. This measure will enable the evaluation of a number of different higher order information access tasks, such as search engine bias and e-Governement accessibility.

3. A MEASURE OF RETRIEVABILITY

Given a collection $\mathbf{D}$, an IR System accepts a user query $\mathbf{q}$ and returns a ranking of documents $\mathbf{R_q}$ which are deemed to be relevant to $\mathbf{q}$ from within $\mathbf{D}$. We can consider the retrievability of a document as a system dependant factor that measures how likely the document is to be returned to the user, with respect to the collection $\mathbf{D}$ and the ranking function used by the system. Consider $\mathbf{Q}$, the universe of all possible queries. Each $\mathbf{q} \in \mathbf{Q}$ is associated with a weight o_q which indicates how likely it is that a user will issue this query to the IR system. Such a weight can be used to capture query popularity, for example, to associate $\mathbf{q}$="celebrity gossip" with a higher weight than $\mathbf{q}$="information retrieval".

Intuitively, the retrievability of a document $\mathbf{d}$ should be high if:

1. there are many queries in $\mathbf{Q}$ which can be expressed in order to retrieve $\mathbf{d}$, or a few queries with a very high o_q which can be expressed in order to retrieve $\mathbf{d}$, and
2. when retrieved, the rank k_{dq} of the document $\mathbf{d}$ is as low as possible (the top-most rank being 1), and certainly lower than the rank c; the point at which the user would stop examining the ranked list.

Thus we formulate the following measure of the retrievability of $\mathbf{d}$:

$$r(\mathbf{d}) = \sum_{\mathbf{q} \in \mathbf{Q}} o_q \cdot f(k_{dq}, c) \quad (1)$$

$f(k_{dq}, c)$ is a generalized utility/cost function where k_{dq} is the rank of $\mathbf{d}$ in the result for $\mathbf{q}$, and c denotes the maximum rank that a user is willing to proceed down the ranked list. The function $f(k_{dq}, c)$ returns a value of 1 if $k_{dq} \leq c$, and 0 otherwise. Note that different functions could be used to reflect the likelihood of a user examining documents at a particular rank k (e.g. rank dependant top-heavy functions for web-search)[5].

Defined in this way, the retrievability of a document is essentially a cumulative score that is proportional to the number of times the document can be retrieved within that cutoff c over the set $\mathbf{Q}$. A document that is never returned in the top-c results for any query will have an $r(\mathbf{d})$ value of zero. While, normalizing the $r(\mathbf{d})$ values would then provide an indication of the likelihood of a document being retrieved by the system. However, even though we have a measure that satisfies our intuitions, we require a method for approximating the retrievability of a document in an operational setting.

3.1 Estimating Document Retrievability

Clearly, it is impractical to calculate the absolute $r(\mathbf{d})$ scores because the set $\mathbf{Q}$ would be extremely large and require a significant amount of computation time as each query would have to be issued against the index for a given retrieval system. So, in order to perform the measurements in a practical way, we need to obtain a reasonable approximation of the relative document retrievability and arrive at some estimate of retrievabilty $\hat{r}(\mathbf{d})$. Our approach in this paper is to use a subset of all possible queries that is sufficiently large and which contains relatively probable or possible queries. For instance, we could use a historical set of queries that have been received by the system in the past, i.e., a query log. Alternatively, we could adopt a simulation based methodology by using an approach like Query Based Sampling [4]. The latter is the approach we take here. Since, we are using a subset of queries, it is worth noting that the estimate of $\hat{r}(\mathbf{d})$ provides a relative measurement of the retrievability of a document. This enables the ready comparison between two different retrieval systems, when the same set of queries has been used to draw the samples which are used in the estimation. For instance, for a given subset of queries, if the retrievability of document $\mathbf{d}$ under system A, is $r_A(\mathbf{d}) = 40$, and for system B, it is $r_B(\mathbf{d}) = 10$, then system A makes $\mathbf{d}$ four times more retrievable than system B.

4. EMPIRICAL ANALYSIS

In this section, we first outline the experimental setup, before calibrating the measurements taken. We then introduce a global measure of retrievability bias that provides a single measure to quantify the inequality between documents in the collection. Following this, an analysis is performed in a controlled environment on two standard IR test collections in order to simulate a number of different possible information access evaluations. We consider the following different scenarios:

- **Media Regulator or Watchdog**: Determine whether a search engine favors the documents of certain news providers over other news providers. On a collection of news articles, we assume that we want to investigate what parts of the collection the retrieval algorithms favor and whether they have any particular biases that we should be aware of.

[2]see `http://www.searchenginewatch.com`

[3]Hiding in Plain Sight: Why Important Government Information Cannot be Found Through Commercial Search Engines, Center for Democracy and Technology, `http://www.ombwatch.org/info/searchability.pdf`

[4]U.S. Legislation: E-Government Act 2002, and the e-Government Reauthorization Act 2007

[5]For an alternative derivation of the retrievability measure based on the concept of "Accessibility" from Transportation Planning we refer the reader to [2], which also describes other more sophisticated utility/cost functions.

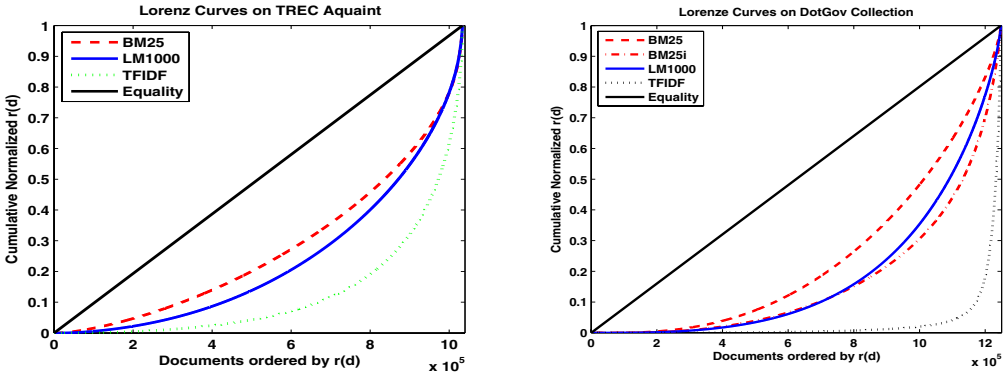

Figure 1: The Lorenz Curve visually depicts the inequality between the population of documents

- **E-Gov Site Administrator**: Assess how easily the content in the collection can be accessed and identify areas of the collection which could be improved or enriched to facilitate their retrieval. Assuming the role of the .GOV collection administrator, according to legislation it is required that the sites be monitored to determine how accessible they are through search engines.
- **IR Practitioner/Researcher**: Detect any untoward bias detrimental to performance and/or study the influence of retrieval algorithms on particular collections and understand more precisely the benefits and limitations of such algorithms.
- **Search Engine Optimizer**: Detect any favoritism of a search engine so that content can be optimized to increase the chances of retrieval.

These different scenarios follow from our initial motivations for considering this work, as well as being diverse enough to help identify other areas of application. Our analysis reveals that the retrieval systems evaluated exhibit different retrieval biases. To determine whether such retrieval biases impede or restrict one's ability to access the documents in an adverse manner, we conduct a further analysis. In these follow-up experiments, we examine the relationship between effectiveness and retrievability.

4.1 Experimental Setup

Datasets We used two different TREC Collections: (1) The .GOV dataset is a collection of just over 1.2 million web documents that consists of a crawl over 643 sub-domains of the U.S. Government website. The collection comes with a set of 225 topics from the Web Track 2004, which we shall use during the analysis. We also use the associated links file, which contains a list of all links between documents in the domain; (2) The AQUAINT dataset is a collection of just over 1 million newswire articles from three different news providers (APW, NYT and XIE). We also use the fifty TREC topics from the ROBUST Track '05.

Retrieval Models For the purposes of this paper, we used three standard IR models to ensure that we evaluate the measure using known functions. This is so we can check that the measure is performing as expected, reflecting known behavior of these algorithms. The models we employed were TFIDF, a Language Modelling approach using Bayes Smoothing with $\mu = 1000$ denoted by LM1000, and the OKAPI retrieval function (BM25). On the .GOV dataset, we also used BM25 combined with an inlink prior using the method and parameters suggested in [5], which we shall denote as BM25i. By design, BM25i favors documents with more inlinks, while TFIDF is known to be biased towards longer documents. LM1000 and BM25 are provided to contrast these methods and show the differences between systems more clearly.

AQUAINT	TFIDF	LM1000	BM25	BM25i
MAP	0.1743	0.1956	0.1740	-
MRR	0.4603	0.6134	0.4923	-
P@5	0.4735	0.4640	0.3444	-
P@20	0.3413	0.3821	0.333	-
.GOV	**TFIDF**	**LM1000**	**BM25**	**BM25i**
MAP	0.0914	0.2244	0.2405	0.3350
MRR	0.1026	0.2645	0.3006	0.4182
P@5	0.0329	0.0818	0.0924	0.1440
P@20	0.0184	0.0371	0.0493	0.0638

Table 1: Performance of each algorithm on AQUAINT and .GOV: Mean Average Precision (MAP), Mean Reciprocal Rank (MRR), Precision at 5 and 10 documents (P@5, P@20) with respect to their query sets

System Effectiveness Table 1 provides an overview of the effectiveness of each retrieval model over a number of performance metrics with their respective TREC topics. Significance testing[6] of the MAP values reveals that LM1000 is significantly better than TFIDF and BM25 on AQUAINT. On .GOV, BM25i is significantly better than the other retrieval systems, and BM25 and LM1000 were both significantly better than TFIDF. Notably the difference in performance on AQUAINT between the systems is small, while the difference on .GOV is quite large. It appears that on the AQUAINT collection, any of the three retrieval methods considered would provide equivalent performance. The following sections show that despite similar effectiveness, the

[6]Conducted using a paired t-test, $p < 0.05$ [12]

three system vary in terms of what areas of the collection they favor, and to what extent.

All experiments reported were conducted using the LEMUR toolkit [7], which was used to index both the collections, with Porter stemming and the removal of stopwords.

4.2 Large Scale Retrieval Simulation

In order to estimate the $r(\mathbf{d})$ values for each document, a number of approximation choices need to be made. For this initial set of experiments, we would like to ensure that the choice of the parameters is appropriate and that our prior knowledge of known retrieval biases in the chosen set of retrieval systems are brought out. For the experiments reported here, we assume the simple binary function for $f(.)$ (indicating just presence or absence in result sets of size c) and set the weight of each query to a constant (we use $o_q = 1$). The latter assumption indicates that we wish that each $\mathbf{q} \in \mathbf{Q}$ contributes equally to the retrievability score $r(\mathbf{d})$ of all $\mathbf{d}$, situations where the query distributions are skewed (e.g. head and tail queries for web-search) might be better handled by alternative choices of o_q. While this is a simple configuration, it provides a very intuitive retrievability measurement, from which to obtain a good idea about how the measure behaves and what it means.

The next choice is the set of queries to be used. For each collection (i.e., .GOV and AQUAINT), we created the reference set of queries $\mathbf{Q}$ by extending, and exploiting, the idea behind Query Based Sampling [4]. The idea was to extract a sufficiently large sample of documents on which to base our estimate of retrievability by issuing millions of queries to probe the collection via the IR system. To do so, a reference set of queries for each collection was created containing queries of one or two terms. The single term queries were constructed by taking each term in the vocabulary that occurred 5 times or more and posing the term as a query. The bi-term queries were constructed by taking each bigram in the collection (i.e., every pair of consecutively occurring terms) that occurred at least 20 times, and ranking them by number of occurrences before truncating the list at 2 million. Each bigram in the set was then posed as a query. Table 2 provides details of the two datasets as well as the query sets used to obtain a reasonable approximation of $r(\mathbf{d})$.

	AQUAINT	.GOV
No. of docs.	1,033,461	1,247,753
Vocabulary size	663,158	5,895,123
No. of single term queries	663,158	881,230
No. of bi-term queries	1,134,362	2,000,000
Total	1,797,520	2,881,230
Total no. of docs retrieved	100,147,410	213,829,937
Expected ($\hat{r}(d)$ @ $c = 100$)	96.9	171.4

Table 2: Details of the TREC datasets and large scale retrieval simulations.

A large scale retrieval simulation was then conducted, where we took each $\mathbf{q} \in \mathbf{Q}$ and issued it to a given retrieval system, collecting up to 100 results for each query. This was performed for each system and on each collection. The total number of documents retrieved during the course of the simulations for a given run is also shown in Table 2, along with the number of times we would expect to retrieve any given document in the top 100. Since our approximation of $r(\mathbf{d})$ is essentially a cumulative measure which reflects how many times document $\mathbf{d}$ is going to be retrieved, it is a function of the size of the result set that is returned to the user, characterized by the parameter c. For each collection and retrieval model, we computed the $r(\mathbf{d})$ values over the all the documents in the collection for $c = 10, 20, 30, 50$ & 100.

In order to choose a suitable c parameter, we investigated the correlation between retrievability measurements given the different values of c. In a series of pairwise comparisons, with $c = 10$, we found that a significant correlation exists between the measurements at different values of c, regardless of the retrieval system. This suggests that changing the c parameter will not dramatically alter the relative estimate of $r(\mathbf{d})$, but only affects the magnitude of the approximation. Ideally, the c parameter should be chosen to reflect the particular scenario in order to obtain a more accurate estimate. For instance, in web-search a low c would be more accurate because users are unlikely to go beyond the first page of results; whereas a high c would be more appropriate for a more thorough legal or patent searcher. For the purposes of the analysis shown in the remainder of this paper, we only report experiments using $c = 100$ as this provides the largest sample[8].

The first interesting observation from examining the retrievability scores of documents, is that there are many documents in the collection that attract a very low retrievability score, while there are few documents that attract a very high retrievability score. When comparing different retrieval models, we witnessed that this trend was more pronounced in TFIDF and BM25i than LM1000 and BM25. In order to quantify the difference in retrievability amongst the population of documents, we require a global measure of the retrievability bias.

4.3 A Global Measure for Retrievability Bias

By examining the distribution of $r(\mathbf{d})$ scores of all the documents, it is possible to assess the inequality between documents within a collection by using the Lorenz Curve [7]. In Economics and the Social Sciences, a Lorenz Curve is used to visualize the inequality of the wealth in a population. This is performed by first sorting the individuals in the population in ascending order of their wealth and then plotting a cumulative wealth distribution. If the wealth in the population was distributed equally then we would expect this cumulative distribution to be linear. The extent to which a given distribution deviates from equality is reflected by the skew in the distribution. We employ the same idea in the context of a population of documents, where their wealth is represented by $r(\mathbf{d})$ and plot the result in Figure 1. The more skewed the plot, the greater the amount of inequality, or bias within the population. In the example, TFIDF displays the most inequality, whereas BM25 results in the least. The Gini Coefficient G was proposed as a way to summarize the amount of bias in the Lorenz curve [7], and is computed as follows:

$$G = \frac{\sum_{i=1}^{N} (2 * i - N - 1) * r(\mathbf{d_i})}{N \sum_{j=1}^{N} r(\mathbf{d_j})}$$

[7]http://www.lemurproject.org

[8]The same analysis at different values of c yielded similar findings to those reported here.

Col.	Ret. Mod.	c				
		10	20	30	50	100
AQ	TFIDF G	0.78	0.77	0.77	0.76	0.74
		ρ	0.95*	0.92*	0.87*	0.81*
	LM1000 G	0.63	0.60	0.59	0.57	0.54
		ρ	0.89*	0.82*	0.71*	0.65*
	BM25 G	0.56	0.52	0.51	0.49	0.46
		ρ	0.86*	0.75*	0.64*	0.55*
.GOV	TFIDF G	0.96	0.95	0.95	0.95	0.94
		ρ	0.91*	0.84*	0.77*	0.67*
	LM1000 G	0.75	0.72	0.70	0.67	0.64
		ρ	0.97*	0.93*	0.89*	0.81*
	BM25 G	0.63	0.60	0.57	0.55	0.52
		ρ	0.95*	0.91*	0.86*	0.77*
	BM25i G	0.77	0.74	0.72	0.70	0.67
		ρ	0.97*	0.95*	0.91*	0.83*

Table 3: Gini coefficient values for all the retrieval models considered with different values of c. As c increases, G steadily decreases indicating that lesser bias is experienced when considering deeper ranked list. Also shown is the Pearson's correlation coefficient between retrievability values calculated with $c = 10$ and other values of c. A statistically significant relationship exists between all the pairs, indicating that the measure is relatively stable with respect to choice of c.

where the retrievability values, $r(\mathbf{d_i})$, have been sorted in ascending order and N is the number of documents in the collection. If $G = 0$ then no bias is present because all the documents are equally retrievable (i.e., $r(\mathbf{d_i}) = r(\mathbf{d_j})$ for all i, j), whereas if $G = 1$ then only one document is retrievable and all other documents have $r(\mathbf{d}) = 0$. By comparing the Gini coefficient of different retrieval methods, we can obtain a bird's eye view of the retrievability bias imposed on the collection of documents by different retrieval systems, given the reference set of queries.

We can see in Table 3 that as c is increased, the Gini coefficient tends to slowly decrease. This suggests that the amount of inequality within the population is mitigated by the willingness of the user to search deeper into the ranking. Consequently, if a user is only willing to examine the top documents, they will be subject to a greater degree of retrieval bias.

In terms of the bias induced by the tested retrieval models, we note that TFIDF has the greatest inequality between documents over both datasets while BM25 appears to provide the least inequality. The introduction of a prior (a favoritism towards documents with more inlinks) is reflected in the Gini coefficient for BM25i being more than that for BM25, i.e., it is more biased. In it interesting to compare the overall effectiveness of the system to the retrievability bias (summarized by the Gini coefficient). We can clearly see that global retrieval bias can be very harmful to effectiveness, in the case of TFIDF on .GOV, but conversely retrieval bias can be beneficial to effectiveness as in the case of BM25i versus BM25. At this level, there appears to be no relationship between global retrieval bias and overall effectiveness. Intuitively, this is correct, because bias can be both damaging and supportive in the process of searching for relevant content. In section 4.6, we examine at a lower level the relationship between effectiveness and retrievability and show a lower level relationship between the two types of measures.

4.4 Subset Analysis

In this subsection, we consider how we can examine the influence of an IR system on different logical divisions of the collection. Such analysis would be useful in order to identify whether certain sites are favored over others, or whether there are sections of the collection that need to be improved. Both AQUAINT and .GOV, can be logically sub-divided into meaningful divisions. For the e-Government scenario, the curator might be interested in noting which parts of the .GOV domain are more readily accessible through the search engine; and if it is found that some sections of the domain are noticeably less retrievable then remedial action that more effectively exposes the content in the domain can be taken. Alternatively, when comparing retrieval systems, preference may be given to one that makes a larger fraction of the domain visible; for instance, this may be the case with the AQUAINT collection where all the systems deliver comparable retrieval effectiveness.

.GOV: For the current set of experiments, we associate every document in .GOV with a sub-domain based on their URL strings. We can then consider each sub-domain in turn, calculating the average $r(\mathbf{d})$ value for all the documents in a given sub-domain for a given retrieval algorithm. Similarly for the AQUAINT collection of newswire articles, the three different sources (APW, NYT and XIE) provide three parts to the collection and the average retrievability of the documents from a particular source can be calculated. Table 4 shows the top 10 and bottom 10 sub-domains of .GOV, which contained more than 100 documents. Most of the domains which had less than 100 documents, tended to have lower mean $r(\mathbf{d})$ than those shown. Interestingly, we found that no correlation existed between the size of the domain and the mean retrievability of the domain. However, we did find a significant positive correlation between the total retrievability of each domain and the size domain, which is to be expected because a larger domain will have more documents which could be retrieved.

The relative difference between the top 10 domains and bottom 10 domains, is by orders of magnitude, where documents in some domains are, on average, 200 times more likely to be retrieved than documents in other domains. This is quite a disparity; for smaller domains the problem tends to be substantially worse. Some of the least retrievable subdomains consist of considerable numbers of documents. If search is the primary information access mechanism on this collection, then this represents a problem for the site administrator because there are parts of the collection which are relatively inaccessible.

AQUAINT: Table 5 shows how differently the three retrieval methods retrieve different parts of the AQUAINT collection. In particular, the mean $r(\mathbf{d})$ value for NYT documents is almost 8 times larger than the equivalent number for XIE when using TFIDF retrieval. What this means is that even if relevant content existed within XIE, the TFIDF retrieval algorithm will be unlikely to retrieve it such that it is accessible within the top 100 results. BM25 on the other hand is the *least biased* of the three algorithms considered providing the least imbalance amongst the sub-divisions of the collection. A one-way ANOVA test on each of the

TFIDF	BM25	BM25i	LM1000
ngisc.gov 1032.87 (238)	irs.gov 902.82 (136)	↓rstgov.gov 1195.15 (117)	irs.gov 1226.62 (136)
stat-usa.gov 750.89 (464)	nyc.gov 879.72 (1097)	nyc.gov 970.77 (1097)	nyc.gov 807.32 (1097)
mhc.gov 739.50 (301)	↓rstgov.gov 772.44 (117)	irs.gov 963.62 (136)	fedworld.gov 717.49 (1600)
bop.gov 713.32 (177)	fedworld.gov 600.40 (1600)	businesslaw.gov 923.76 (180)	↓rstgov.gov 683.87 (117)
↓tness.gov 644.37 (163)	bop.gov 594.15 (177)	medicare.gov 785.05 (164)	medicare.gov 680.18 (164)
ncr.gov 628.18 (346)	medicare.gov 555.01 (164)	fedworld.gov 777.48 (1600)	bop.gov 585.11 (177)
negp.gov 598.84 (271)	ustreas.gov 504.74 (3089)	consumer.gov 703.19 (104)	mhc.gov 533.80 (301)
fedworld.gov 597.38 (1600)	businesslaw.gov 480.42 (180)	mass.gov 590.36 (138)	ustreas.gov 530.33 (3089)
nyc.gov 580.86 (1097)	4woman.gov 456.18 (1235)	peacecorps.gov 556.82 (106)	consumer.gov 469.10 (104)
tigta.gov 559.32 (283)	panynj.gov 441.23 (207)	4woman.gov 514.09 (1235)	↓tness.gov 462.48 (163)
...	...	...	...
...	...	...	...
...	...	...	...
arserrc.gov 9.19 (103)	itrd.gov 32.97 (331)	nro.gov 26.28 (121)	fgdc.gov 32.44 (605)
ngi.gov 9.13 (179)	arm.gov 32.04 (2012)	fnal.gov 23.45 (15314)	fedcirc.gov 31.63 (113)
disabilitydirect.gov 9.09 (197)	bnl.gov 31.34 (17872)	ncs.gov 22.89 (335)	dnfsb.gov 31.46 (259)
export.gov 8.10 (157)	ngi.gov 29.50 (179)	xml.gov 21.80 (197)	bnl.gov 29.60 (17872)
ojp.gov 8.02 (243)	fgdc.gov 29.15 (605)	itrd.gov 21.74 (331)	fnal.gov 29.14 (15314)
disabilities.gov 7.92 (320)	fnal.gov 29.03 (15314)	fgdc.gov 19.21 (605)	nersc.gov 27.22 (2487)
businesslaw.gov 7.33 (180)	nersc.gov 28.72 (2487)	nhm'.gov 17.55 (126)	arm.gov 25.75 (2012)
nhm'.gov 6.59 (126)	dnfsb.gov 28.44 (259)	cwc.gov 16.70 (122)	ngi.gov 25.57 (179)
csce.gov 6.46 (296)	xml.gov 22.17 (197)	fedcirc.gov 16.09 (113)	xml.gov 17.97 (197)
orau.gov 5.51 (150)	nhm'.gov 9.83 (126)	nclis.gov 11.91 (486)	nhm'.gov 10.74 (126)

Table 4: The top 10 and bottom 10 domains in .GOV in terms of retrievability, restricted to domains which have over 100 documents. The mean $r(\mathbf{d})$ for each domain is shown along with the size of the domain in brackets. Notice the similarity of sites favored between retrieval methods and their relative differences

	Source		
	APW	**NYT**	**XIE**
Avg. doc length	426 (385)	798(745)	205(186)
No. of docs.	239,576	314,452	479,433
Ret. method	Mean $r(\mathbf{d})$		
TFIDF	109.6 (44)	197.3 (84)	24.5 (10)
LM1000	123.0 (78)	128.0 (98)	63.0 (35)
BM25	120.9 (81)	83.7 (58)	93.5 (69)

Table 5: Mean $r(\mathbf{d})$ values for each document in AQUAINT grouped by source (median values shown in brackets).

groups, and follow up significance test shows that there is a significant difference between the retrievability of documents from each of the different sources (regardless of retrieval algorithm). Consequently, we find that none of these retrieval models provide unbiased access across the different sources.

4.5 Features Analysis

An alternative study is to analyse the document features to determine if there is a relationship with retrievability. While one must have an idea of the features in mind *a priori*, it is possible to build up a picture of what features may affect the retrievability of documents. On one hand, this provides a diagnostic tool for IR practitioners and researchers, and on the other hand it provides a investigative tool for search engine optimizers. Such an analysis can be seen as a first step towards taking remedial action - if we know that documents with particular characteristics are being penalised by the retrieval algorithm, as the administrator of the collection, we might want to ensure that the documents in the collection possess the positive characteristics of a highly retrievable document.

For the purposes of demonstration, we consider three document features that could be used to characterize documents in the .GOV collection and examine how they relate with $r(\mathbf{d})$. The features we consider, are document length $n(\mathbf{d})$, the number of inlinks to a document, $n(i, \mathbf{d})$ and the number of outlinks from a document, $n(o, \mathbf{d})$. However, any feature that is available could be used, and if we had access to the retrieval function instead then we could precisely evaluate the influence of each feature. To perform the analysis: for each of the three features, we sorted the documents in increasing order of their value for that feature. We then placed documents into bins of size 2000 leading to 613 groups over the .GOV collection. For each bin, we then calculated the mean and median $r(\mathbf{d})$ for documents in that bin. This is plotted on the Y-axes in Figure 2, with the particular feature on the X-axis. For comparison, we provide a dotted line representing the situation where the documents have been randomly assigned to bins. The differences between the retrieval systems can be clearly seen from the plots shown in Figure 2.

As one would expect, TFIDF tends to favor longer documents, and BM25i tends to favor documents with more inlinks. But perhaps surprisingly, TFIDF, LM1000 and to a lesser extent BM25 tend to favor documents with fewer outlinks. For instance, TFIDF tends to retrieve documents with fewer outlinks on average three times more than documents with many outgoing links. On the other hand, BM25i appears to favor documents with more outlinks. The observed increased positive correlation of $r(\mathbf{d})$ with inlinks when moving from BM25 to BM25i, as well as dependance of TFIDF on length, again provides a sanity check for our measure. However, the observation on outlinks illustrates non-obvious behavior of the retrieval algorithms that is elucidated using our measure. There are of course two problems with this naive treatment:

1. We have not considered correlations between the features. That is to say, the observed dependence of increased $r(\mathbf{d})$ with respect to outlinks may simply be because pages in .GOV with many inlinks also tend to have many outlinks.

2. We have had to decide beforehand what features of the documents will constitute the Y-axes of the plots. Consequently, the feature analysis only enables correlations to be found between features and retrievability as opposed to indicating a causal link.

This second point means that a search engine optimizer who only observes a bias in retrievability but does not have any knowledge of the underlying retrieval mechanism cannot

definitively say whether the feature is the cause, or whether it was just a correlation. Even taking this into account, the feature analysis serves as very helpful diagnostic and investigative tool to aid in determining how the retrieval system behaves given a particular feature.

4.6 Retrievability and Effectiveness

So far we have examined what levels of retrievability different retrieval systems provide to (individual and groups of) documents in a collection. We have seen that the algorithms differ significantly and substantially in terms of the retrieval biases that they impose on the population of documents. In this subsection, we specifically examine whether such retrieval bias actually impedes one's ability to access content within the collection. That is, given that the IR system favors certain documents over others, does it mean that less retrievable documents will be significantly harder to find. Or conversely, if some documents are less retrievable, then removing them from the collection is unlikely to have an impact on effectiveness because they are unlikely to be retrieved (by definition of $r(\mathbf{d})$). In order to examine these two premises, we construct two separate experiments.

Experiment 1 We replicate a standard IR evaluation environment by considering each collection with its corresponding set of TREC topics. We speculate that the less retrievable documents are unlikely to be retrieved in response to these topics for a given algorithm; and so if these documents are removed, no significant degradation to effectiveness will be witnessed. The extent to which we can remove documents, will depend on two things: (1) the quality of the documents within the collection, and (2) the extent to which the retrieval system is biased. Such that if a system is more biased, in terms of its Gini coefficient, it is likely that more documents can be removed, because a greater proportion of the documents are unlikely to be retrieved.

Given the set of TREC Topics, we calculated traditional measures of retrieval effectiveness on the collection of documents. Using the hypothesis that the removal of documents with low $r(\mathbf{d})$ is unlikely to significantly affect effectiveness, we successively removed a fraction f of each of our two collections. To pick the documents that get removed, we first arranged the documents from the index in decreasing order of their $r(\mathbf{d})$ values calculated using each retrieval method in turn. We then progressively removed documents from the lower end, and measured the retrieval effectiveness (using Mean Average Precision calculated on the top 1000 documents returned for each query on this reduced collection. A plot showing the percentage drop in MAP versus fraction of the collection removed is shown in Figure 3.

On AQUAINT, we found that for any particular model, there was significant degradation in performance once 10% or more of the collection was removed, except for TFIDF where over 30% of the collection was removed before there was a significant degradation in performance. While, on .GOV, we see quite a different picture. Up to 50% of the collection could be removed before there was a significant degradation in performance for BM25 and LM1000. Whereas for BM25i and TFIDF, up to 70% and 80% of the collection could be removed, respectively. Surprisingly, for BM25i we found that when 30% to 50% of the collection was removed, retrieval effectiveness on the reduced collection was actually better than on the complete collection, and this difference was statistically significant. This experiment provides evidence to suggest that the more biased the algorithm, the more of the collection can be removed without a significant drop in performance.

The above experiment also indicates that the different retrieval algorithms make certain parts of the collection virtually inaccessible, to the point that the documents are *expendable* (i.e. can be removed without a significant loss to effectiveness). This influence of a retrieval system over the collection is not reflected in standard effectiveness based evaluations and highlights how retrievability provides a distinctly different dimension to IR system evaluation.

Experiment 2 For the next experiment, we consider the hypothesis that if we were trying to retrieve documents of varying retrievability, we would expect that it would be more difficult to formulate a query which would retrieve documents of low retrievability than those of high retrievability; even if we pose a query which is specifically crafted to retrieve that specific document. In order to test this hypothesis, we divide the collection of documents into four bins, according to their retrievability values. The first bin contains the 25% of the documents with the lowest retrievability, while the fourth bin contains the 25% of documents with the highest retrievability for a given retrieval method. From each bin, we simulated known-item queries using the method proposed in [1] with the suggested parameter settings[9]. For each of the quartiles, a document was chosen at random and query terms were randomly selected from this document according to the probability of the term being present in the document. A total of 1000 queries were simulated for each quartile. These queries were then issued against the collection, and a set of results were generated using one of the retrieval methods and the position of the target document in the returned result list was used to calculate the Mean Reciprocal Rank (MRR). The MRR of known-item searches for documents in each quartile represent the effectiveness of future users' searches for the documents in that group. The results are provided in Table 6.

		Quartile			
Col.	**Ret. Mod.**	1st	2nd	3rd	4th
AQUAINT	**TFIDF**	0.09	0.16	0.19	0.21
	LM1000	0.21	0.26	0.28*	0.28
	BM25	0.20	0.26	0.34*	0.35
.GOV	**TFIDF**	0.03	0.05	0.08	0.16
	LM1000	0.11	0.22	0.27	0. 30
	BM25	0.16	0.24	0.30*	0.29
	BM25i	0.09	0.19	0.25	0.32

Table 6: Effectiveness of known-item searches measured by MRR. An '*' indicates that the effectiveness of the queries in this set was not significantly different to the performance of the queries in the 4th quartile using the Kolmogorov-Smirnov test between the two distributions ($p > 0.05$). For all other results, there was a significant difference between the performance of the other quartiles and the 4th quartile.

[9]The specific parameters used were: the term-frequency (or popular) term selection strategy, the mean length of queries was set to four, and the probability of a noisy query term was set to 0.2.

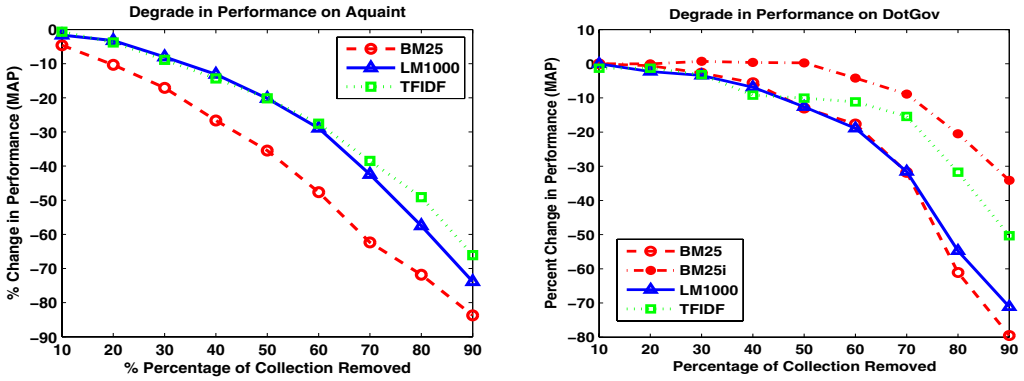

Figure 3: Reduction in MAP when the least retrievable parts of the collections are progressively removed.

It can be seen from the table that documents with lower $r(\mathbf{d})$ are more difficult to find when compared to document that are more retrievable. This is despite the fact that the queries were specifically designed to bring back those documents. While the results are dependant on the method used to generate the simulated topics, they show that across retrieval algorithms, the documents in the fourth bin are significantly easier to find (through a search) than documents in first bin. Note that TFIDF on the .GOV collection provides the most extreme example where the effectiveness for documents in the fourth bin is over five times greater than in the first bin. For BM25i, even though it is very effective at retrieving the documents in the fourth bin, it is four times worse at retrieving documents in the first bin. The disparity between bins for BM25 is less than a factor of two. When we consider these results against the global measure of retrievability bias, it indicates that if a system is highly biased, then the disparity in retrieval effectiveness over the collection is likely to be greater than if the system is less biased.

Finally, these results show that the bias of the retrieval system imposed upon a collection of documents can seriously affect the retrieval effectiveness of attempting to access documents which are less retrievable. The more biased the system the greater the disparity in retrieval effectiveness measured on known-item searches.

5. CONCLUSION AND FUTURE WORK

In this paper, we have proposed a methodology to evaluate retrieval systems based on the access they provide into a collection of documents. This required a measure to capture the *retrievability* of documents. This measure was designed to reflect the ease with which the document can be found through the retrieval system. The motivation for such a measure stems from the concern over biases of search engines and retrieval systems, and the need to ensure that content is accessible through such systems. This is because of the growing reliance of users to engage such systems in order to find content. Since effective retrieval necessarily involves a preference for one set of documents (i.e., the relevant ones) over another (the non-relevant), the existence of some bias is inevitable. We have demonstrated that the proposed measure of retrievability provides a useful way in which to quantify the retrieval bias of a system and how it can be used in the evaluation of higher order information access tasks.

Given the presence of retrieval bias, it was also important to determine whether such bias had any impact on one's ability to access the content through the system. In this paper, we have tested to see if the imposition of such bias actually has an impact, negative or otherwise, on system effectiveness. Experiments conducted on AQUAINT and .GOV yielded two main findings:

- In a TREC-style evaluation, a proportion of the least retrievable documents could be removed without significantly degrading performance. In highly biased retrieval systems up to 80% of the collection could be removed. This is because the retrieval system is unlikely to ever retrieve these documents due to the bias it exhibits over the collection of documents.

- The least retrievable documents within the collection are significantly harder to find than the rest of the collection; the difference in MRR/MAP on a known-item search can be up to a factor of five compared to the most retrievable documents. The extent of this disparity in retrieval performance appears to be exacerbated by the amount of bias imposed by the system on collection access.

Finally, measuring retrievability provides a novel way in which to assess a retrieval system's influence on the access to documents in a collection. This paper has provided a methodology that can be used to assess the impact of IR systems on collections; so that the collection and/or retrieval system can be improved to facilitate better access. While the main goal of retrieval is to maximize the system performance for specific sets of information needs, our findings suggest that it is also important to consider the impact that the retrieval system's bias has on the access to the entire collection. Determining the trade off between effectiveness and retrievability poses an interesting direction for future work, along with (i) exploring alternatives to estimate and approximate the retrievability values; (ii) improving the accuracy of the estimate by considering different cost functions and query sets, and (iii) examining different application areas (e.g. measuring search engine bias, or the accessibility of e-Government information).

6. ACKNOWLEDGMENTS

The first author would like to thank the Information Retrieval Facility (`www.ir-facility.org`) for the use of their computational resources.

7. REFERENCES

[1] L. Azzopardi and M. de Rijke. Automatic construction of known-item finding test beds. In *Proceedings of SIGIR '06*, pages 603–604, 2006.

[2] L. Azzopardi and V. Vinay. Accessibility in information retrieval. In *Proceedings of ECIR'08*, pages 482–489, Glasgow, Scotland, 2008.

[3] M. Baillie, L. Azzopardi, and I. Ruthven. Evaluating epistemic uncertainty under incomplete assessments. *Inf. Process. Manage.*, 2(44):811–837, 2008.

[4] J. Callan and M. Connell. Query-based sampling of text databases. *ACM Trans. Inf. Syst.*, 19(2):97–130, 2001.

[5] N. Craswell, S. Robertson, H. Zaragoza, and M. Taylor. Relevance weighting for query independent evidence. In *Proceedings of SIGIR '05*, pages 416–423, 2005.

[6] A. Dasgupta, A. Ghosh, R. Kumar, C. Olston, S. Pandey, and A. Tomkins. The discoverability of the web. In *Proceedings of WWW '07*, pages 421–430, 2007.

[7] J. L. Gastwirth. The estimation of the lorenz curve and gini index. *The Review of Economics and Statistics*, 54(3):306–316, 1972.

[8] S. Lawrence and C. L. Giles. Accessibility of information on the web. *Nature*, 400(6740):107–107, 1999.

[9] A. Mowshowitz and A. Kawaguchi. Assessing bias in search engines. *Inf. Process. Manage.*, 38(1):141–156, 2002.

[10] S. Pandey, K. Dhamdhere, and C. Olston. Wic: A general-purpose algorithm for monitoring web information sources. In *Proceedings of VLDB '04*, 2004.

[11] V. Petricek, T. Escher, I. J. Cox, and H. Margetts. The web structure of e-government - developing a methodology for quantitative evaluation. In *Proceedings of WWW '06*, pages 669–678, 2006.

[12] M. Sanderson and J. Zobel. Information retrieval system evaluation: effort, sensitivity, and reliability. In *Proceedings of the SIGIR '05*, pages 162–169, 2005.

[13] A. Singhal, C. Buckley, and M. Mitra. Pivoted document length normalization. In *Proceedings of SIGIR '96*, pages 21–29, 1996.

[14] T. Upstill, N. Craswell, and D. Hawking. Buying bestsellers online: A case study in search & searchability. In *7th Australasian Document Computing Symposium*, Sydney, Australia, 2002.

[15] C. J. van Rijsbergen. *Information Retrieval*. Butterworths, London, second edition edition, 1979.

[16] L. Vaughan and M. Thelwall. Search engine coverage bias: evidence and possible causes. *Inf. Process. Manage.*, 40(4):693–707, 2004.

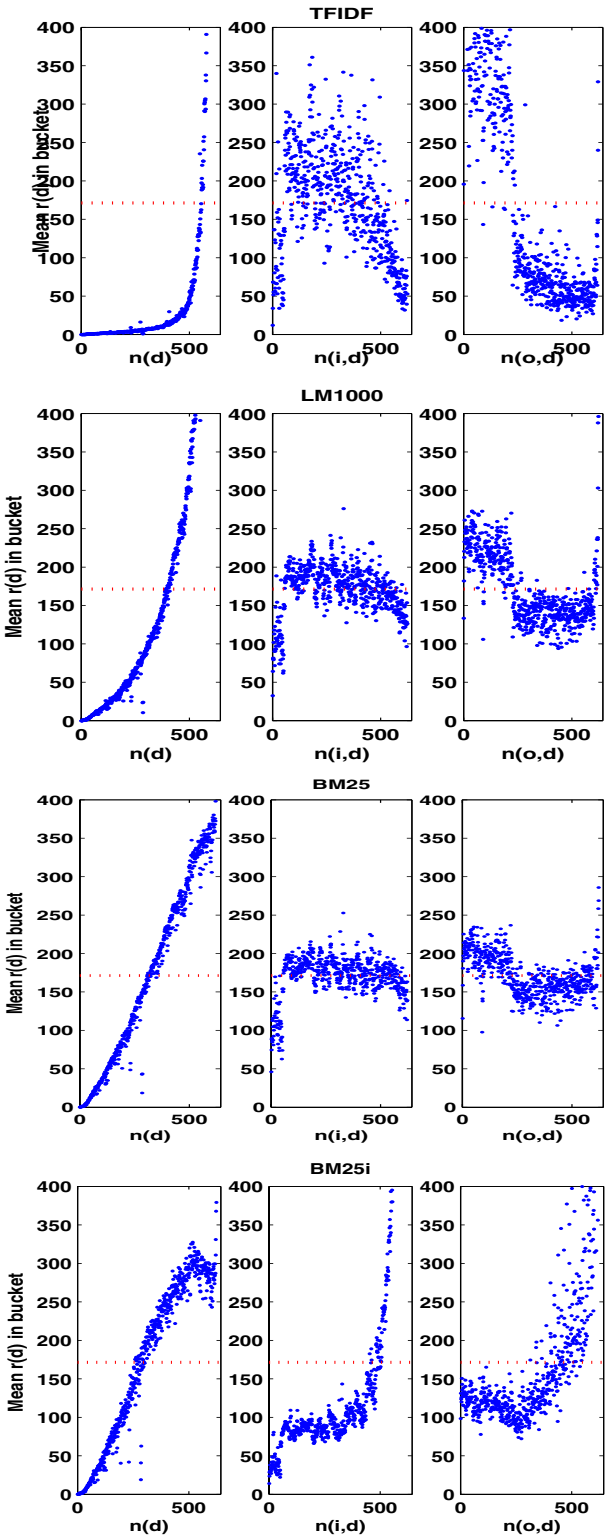

Figure 2: Retrievability plots across document dependent factors. Top to Bottom: TFIDF, LM1000, BM25, BM25i. Subplots left to right: document length $n(\mathrm{d})$, number of inlinks $n(i, \mathrm{d})$ and number of outlinks. $n(o, \mathrm{d})$. The Y-axis denotes how retrievable a document is, according to our estimate $r(\mathrm{d})$