Chapter 1
Retrieval Models versus Retrievability

Shariq Bashir, Andreas Rauber

Abstract

Retrievability is an important measure in information retrieval that can be used to analyze retrieval models and document collections. Rather than just focusing on a set of few documents that are given in the form of relevance judgments, retrievability examines what is retrieved, how frequently it is retrieved, and how much effort is needed to retrieve it. Such a measure is of particular interest within the recall oriented retrieval systems (e.g. patent or legal retrieval), because in this context a document needs to be retrieved before it can be judged for relevance. If a retrieval model makes some patents hard to find, patent searchers could miss relevant documents just because of the bias of the retrieval model. In this chapter we explain the concept of retrievability in information retrieval. We also explain how it can be estimated and how it can be used for analysing a retrieval bias of retrieval models. We also show how retrievability relates to effectiveness by analysing the relationship between retrievability and effectiveness measures and how the retrievability measure can be used to improve effectiveness.

Shariq Bashir
Department of Computer Science
Mohammad Ali Jinnah University, Islamabad
e-mail: shariq.bashir@jinnah.edu.pk

Andreas Rauber
Institute of Software Technology and Interactive Systems
TU Wien
e-mail: rauber@ifs.tuwien.ac.at

1.1 Introduction

Access to information from the web and internet is playing an important part in a society relying increasingly on information looked-up ad-hoc from an ever growing pool of information in databases and the web in general. Information retrieval (IR) systems are essential components of this process. IR systems deal with the storage (indexing), organization, management and retrieval of information [7, 10, 11]. Besides indexing one important factor that shapes the access to information is the role of the retrieval strategy (or model) [36]. It acts as a middleware between the users' required information and the users' effort to access the information. The main role of a retrieval model is to first discriminate between relevant and irrelevant information, and then to display the relevant results to the users. Additionally such models determine the order of results by relevance. Users should be able to view the most relevant information at the top ranked positions. In the last few years, a large number of retrieval models have been proposed for various kinds of retrieval tasks. One of the main problems is therefore how to choose the right model for a given retrieval task. It is a tedious task, and falls under the research domain of evaluation of retrieval models [29, 22, 23, 32]. Historically, research on the evaluation of retrieval models has always focused on either the effectiveness or the efficiency (speed/memory). These are still the most common two measures that are determining the quality of retrieval models. The main limitation of these measures is that they focus only on a few documents that are the (most) relevant ones being returned at the top of a ranked list. This constitutes the primary criterion in most standard retrieval tasks (web retrieval, question answering [21], opinion retrieval [28], etc). With evaluation measures such as recall and F_β, aspects of the completeness of information are being brought into consideration. As a complement to these, a so-called higher order evaluation has been proposed based on the accessibility (retrievability, how easily the information can be accessed). Instead of analyzing how well the system performs in terms of speed or effectiveness, the retrievability measure provides an indication of how easily the information within the collection can be reached or accessed with a given retrieval models [5]. This offers a higher and more abstract level for understanding the influence of the given IR systems or retrieval models. Such models are effecting the access to all relevant information in the collection, not just the set of information that is given in the form of judged relevant documents by a group of few people. This is particularly important for recall-oriented retrieval domains like patent or legal retrieval, where it is necessary to ensure that everything relevant has been found and furthermore, the non-existence of a document has to be proven (e.g. a document which invalidates a new patent application that does not exist) [1, 24]. Moreover, retrievability specifically examines whether the lack of access to information actually impedes one's ability to access the required information within the collection.

1.2 Analogy of Retrievability in Information Retrieval

The approach for analyzing the effectiveness of retrieval models in terms of retrievability has been translated from logistics and transportation planning [5]. (Hansen et al.)[16] draws the definition of accessibility in transportation planning as follows.

"a measurement of the spatial distribution of activities about a point adjusted for the ability and the desire of people or firms to overcome spatial separation. More specifically, the formulation states that the accessibility at point 1 to a particular type of activity at area 2 (say employment) is directly proportional to the size of the activity at area 2 (number of jobs) and inversely proportional to some function of the distance separating point 1 from area 2. The total accessibility to employment at point 1 is the summation of the accessibility to each of the individual areas around point 1. Therefore, as more and more jobs are located nearer to point 1, the accessibility to employment at point 1 will increase."

Based on this definition, accessibility (or just access) refers to the effort of completing different activities of daily life from given opportunities (buses, trains, cares, metro, paths, roads and airports) of a transportation system [16, 18, 15, 19]. This approach is different to the study of efficiency of transportation systems, where the focus of analysis is based on certain positions or procedures, for example the travel time between particular (important) locations. While being related to efficiency accessibility considers access in a more general sense. It provides a high level view on an evaluation of transportation systems.

(Azzopardi et al.)[4] draw the analogy of accessibility between transportation planning and IR systems as follows: in the context of transportation planning, accessibility is the idea that models the opportunities (train, car, metro, bus) of the transportation systems with the objective of completing daily life activities (to reach a location e.g. going to office, dining, shopping at a supermarket). Completing a particular activity is subject to a certain associated cost function such as traveling distance, number of stops, changes of buses, trains etc. Similarly, in the context of information retrieval, accessibility (retrievability) is the idea that models the searching or retrieving of a particular information. This is subject to a cost function based on the simplicity in using the system for retrieval of certain information, and the amount of retrieved documents to examine for reaching the desired information. Here, documents replace locations, and queries replace the opportunities for the completion of a particular activity. However, in IR there is no concept of physical space, therefore there is no constraint on the user's current location (i.e., at a particular document). The IR system models a stop or a location in such way that every possible opportunity can be available (i.e., the universe of all possible queries), and users can select any route desired (he/she can query the IR system with any type of search terms) at any time regardless of the locations. While this makes every document in the collection potentially

equally retrievable, however, the choice of selecting a suitable opportunity (in the form of a query) to reach a particular location where the user wants to travel (retrieving the desired information), and the bias in the ranking methodology of IR systems affect just how documents are retrievable in the information space with the given IR system.

Based on the analogy described above, the accessibility of a document, therefore depends upon the following three factors,

1. user's ability to formulate his/her need in the form of a suitable query,
2. the bias of the retrieval model, and,
3. the user's willingness to go through and down the ranked result list of the query.

In the context of this research we are considering the accessibility of documents as a retrieval task where documents are accessed by querying the IR system.

1.3 Analyzing Retrievability of Retrieval Systems

Web coverage and documents' retrievability are two well-known measures for discovering IR systems accessibility. The body of literature on web coverage based measures contains a range of possible biases [26, 39, 40], for example, if one website has more coverage than another, whether sites in a particular geographical location are favored over others [39], or whether search engines are biased given a particular topic. These studies are usually motivated by the view that search engines may be providing biased content, and these measures are aiming at providing guideline for regulation. Opposed to the web coverage based measures, retrievability focuses on the individual document retrievability scores, which can be also used for analyzing the accessibility of IR systems.

Estimating Retrievability Given a collection D, an IR system accepts a user query q and returns a ranked list of documents, which are deemed to be relevant to q. We can consider the retrievability of a document based on two system-dependent factors, (a) how likely the documents are returned to the user with respect to the collection D, and (b) the effectiveness of the ranking strategy (retrieval model) of the IR system. In order to derive an estimate of this quantity [5] used query set based sampling for approximating the retrievability (retrieval bias) [9]. Q query set could either be a historical sample of queries or an artificially simulated substitute similar to users' queries. Then, each $q \in Q$ is issued to the IR system, and the retrieved documents along with their positions in the ranked list are recorded. Intuitively, retrievability of a document d is high when:

1. There are many probable queries in Q, which can be expressed in order to retrieve d, and
2. when retrieved, the rank r of the document d is lower than a rank cutoff (threshold) c. This is the point at which the user would stop examining the ranked list. This is a user dependent factor and reflects a particular retrieval scenario in order to obtain a more accurate estimate of this measure. For instance, in web-search scenarios a low c would be more accurate as users are unlikely to go beyond the first page of the results, while in the context of recall-oriented retrieval settings (for instance, legal or patent retrieval), a high c would be more accurate.

Thus based on the Q, r and c, we formulate the following measure for the retrievability of d

$$r(d) = \sum_{q \in Q} p(q) \cdot \hat{f}(k_{dq}, c) \tag{1.1}$$

$f(k_{dq}, c)$ is a generalized utility/cost function, where k_{dq} is the rank of d in the result list of query q, c denotes the maximum rank that a user is willing to proceed down in the ranked list. The function $\hat{f}(k_{dq}, c)$ returns a value of 1, if $k_{dq} \leq c$, and 0 otherwise. $p(q)$ denotes the likeliness that a user actually issues query q. This probability is hard to determine explicitly, and is set to 1, i.e. to give all queries equal probabilities [5]. More complex heuristics integrating the length of the query, the specificity of the vocabulary etc. may be considered. 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 Q. This fulfils our aim, in that the value of $r(d)$ will be high when there is a large number of highly probable queries that can retrieve the document d at the rank less than c, and the value of $r(d)$ will be low when only a few queries retrieve the document. Furthermore, if a document is never returned at the top ranked c positions, possibly because it is difficult to retrieve by the system, then the $r(d)$ is zero.

This $r(d)$ function is an intuitive way to show the relative bias between both: sets of retrieval models and sets of documents. For example, if system A returns $r(d1) = 30$ and system B returns $r(d1) = 1$ then relatively, the system A will retrieve document $d1$ 30 times more than the system B. This gives a more concise view of the relation of system A to system B with respect to document $d1$. Alternatively if for system A $r(d1) = 30$ and $r(d2) = 1$ then system A favors and retrieves $d1$ 30 times more than it does $d2$.

The cumulative measure of the retrievability score of a document on the basis of the binary $f(k_{dq}, c)$ function ignores the ranking position of a document in a ranked result list, i.e. how accessible the document is in the ranking. A gravity based measure can be used for this purpose by setting the function to reflect the effort of going further down in the ranked result list, and it is defined as

$$\hat{f}(k_{dq}, \beta) = \frac{1}{(k_{dq})^{\beta}} \tag{1.2}$$

The rank cutoff factor is changed to β which is a dampening factor that adjusts how accessible the document is in the ranking. In our experiments we calculate the retrievability score of documents only on the basis of cumulative measure.

Retrievability inequality can be further analyzed using the *Lorenz Curve* [14]. In Economics and Social Sciences, a Lorenz Curve is used to visualize the inequality of 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 wealth was distributed equally in the population then we would expect this cumulative distribution to be linear. The extent to which a given distribution deviates from the equality is reflected by the amount of skewness in the distribution. (Azzopardi et al.)[5] used a similar idea in the context of a population of documents, where the wealth of documents is represented by $r(d)$ function. The more skewed the plot, the greater the amount of inequality, or bias within the population. The *Gini-Coefficient* [14] G is used to summarize the amount of retrieval bias in the Lorenz Curve and provides a bird's eye view. It is computed as follows:

$$G = \frac{\sum_{i=1}^{|D|}(2 \cdot i - |D| - 1) \cdot r(d_i)}{(|D| - 1)\sum_{j=1}^{|D|} r(d_j)} \tag{1.3}$$

D represents the set of documents in the collection. If $G = 0$, then no bias is present because all documents are equally retrievable. If $G = 1$, then only one document is retrievable and all other documents have $r(d) = 0$. By comparing the Gini-Coefficients of different retrieval methods, we can analyze the retrieval bias imposed by the underlying retrieval systems on a given document collection.

1.4 Applications of Retrievability Analysis

Retrievability analysis can be used for a range of applications, such as

- **Media Regulator or Watchdog:** The retrievability measure can act as watchdog. It can be used to determine whether a retrieval system favors documents of certain news providers over others [3]. It can also be used to investigate what parts of the collection are favored over others, and whether they have any particular biases that the users should be aware of.
- **Bias Detection for IR Practitioner/Researcher:** The retrievability measure can provide help to the IR practitioner/researcher for detecting any untoward bias detrimental to the effectiveness of retrieval models. It

can further examine the influence of different retrieval models on a particular collection to understand more precisely the benefits and limitations of different models.

- **Automatic Ranking Retrieval Models:** One important area of IR research is to evaluate the effectiveness of retrieval models using test collections [29, 22, 23, 32]. A test collection consists of a set of documents, a set of queries, and (relevance judgments i.e. a list that describes which documents are relevant to which query topic). While documents and queries are relatively easy to gather, creating relevance judgments requires significant effort and resources. In recent years there has been increased research interest in methods for automatically ranking retrieval models without human generated relevance judgments [29, 22, 23, 32, 27, 37, 12, 34, 33, 31, 17, 8, 41]. One important aspect of the retrievability measure is that it can be analyzed without relevance judgment, and this provides an attractive alternative for producing the automatic ranks of retrieval models.
- **E-Gov Site Administration:** Ensuring that online contents are accessible is very important in the area of e-Government, because citizens of a democratic country have a right to information [38]. If such information is non-accessible to the public this can jeopardize the integrity of the government. Currently, there exists no quantitative measure or methodology that can be employed for ensuring access to relevant content. The retrievability measure can be used for this task in order to determine whether a sufficient amount of access is accorded to the information housed within the e-Government websites.
- **Search Engine Optimization:** The retrievability measure can be used to detect any favoritism in the search engine so that the content can be optimized to increase the chance of retrieval.

1.5 What Retrievability Cannot Examine

We cannot use retrievability for examining the following factors.

1. Retrievability cannot predict the search effectiveness of professional users (e.g. patent examiners). An experienced user can easily bypass the retrieval bias by recalling most relevant terms of his/her information need and applying bitwise *"AND"* operator between query terms. A long bitwise *"AND"* query decreases the size of query's result list (i.e. total number of retrieved documents) and the user can easily access his/her required information from a few number of top ranked documents.
2. On the basis of retrievability scores, low retrievability scores documents are more difficult to find than the high retrievability scores documents. However, it is wrong to say that low retrievability scores documents are unlikely to be found by any type of query. As we explained earlier, along

Dataset	Total Docs.	Seed Docs.	Rank cutoff factors
TREC-CRT	1.2 million	34,205	50,100,250
ChemAppPat	36,998	36,998	5,10,15,20,25
DentPat	27,988	27,988	5,10,15,20,25
ATNews	47,693	47,693	5,10,15,20,25

Seed Docs: = *This is the set of documents that are used for query generation and retrievability analysis.*

Table 1.1 The properties of document collections used for the retrieval bias analysis.

with the users' ability to formulate the queries and the retrieval bias of retrieval models, a third factor affecting the retrievability of documents is the difference between the size of the query result list and the users' ability or willingness to proceed down in the ranked result list. This difference is controlled with the help of a rank cutoff factor. If this difference is large, this means that the users will go through only a small portion of the total number of retrieved documents, and from this we can expect a large retrievability inequality between the documents of collection. On the other hand, if this difference is small or the sizes of query result lists are less than the rank cutoff factor, then the users will go through a large portion of the retrieved documents and thus there will be less retrievability inequality between the documents of collection. Ideally, queries used for retrievability analysis should have large result list sizes. Because, this allows us to precisely understand the role of the retrieval bias of retrieval models in the retrievability; however, this ignores many specific, focused queries.

1.6 Retrievability Experiments

In this section, we examine the retrieval bias of different retrieval models for different collections. The collections that we use for experiments contain patent and news documents. For these collections, we first determine the retrievability of documents with different retrieval models, and then we analyze to what extent these retrieval models are differing in terms of retrieval bias that they imposed on the documents of collections. The overall retrievability of documents provides an indication of how easily the documents are accessible with different retrieval models. The overall retrievability inequality between the documents of collection shows the retrieval bias of retrieval model.

1.6.1 Document Collections

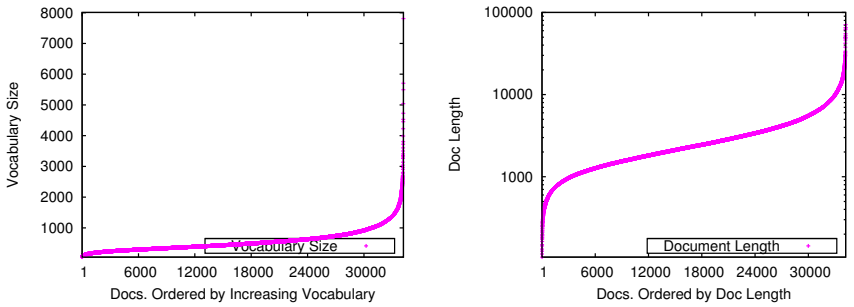
We use the following four collections (Table 1.1) for retrieval bias analysis. Table 1.1 presents the basic properties of these collections. Seed documents represent the set of those documents that are used for query generation and retrievability analysis.

- **TREC 2009 Chemical Retrieval Track Collection (TREC-CRT):** This dataset consists of 1.2 million patent documents from the TREC Chemical Retrieval Track (2009) (*TREC-CRT*)[1] [20]. Due to the large size of the collection, determining the retrievability for all documents in the collection requires large processing time and resources. Therefore to complete the experiments in a reasonable time, a subset of *34,205* documents (judged documents) for which the relevance assessments are available as part of *TREC-CRT* serves as seed for query generation and retrievability analysis. As compared to other three collections, the documents in this collection are very long. The distributions of document length and vocabulary size are also highly skewed (see Figure 1.1). For this collection, retrieval bias is analyzed with five rank cutoff factors *c=50*, *c=100*, *c=150*, *c=200*, and *c=250*.
- **USPTO Patent Collections (ChemAppPat, DentPat):** These collections were downloaded from the freely available US patent and trademark office website [2]. We collect all patents that are listed under the United State Patent Classification (USPC) classes *433 (Dentistry)*, and *422 (Chemical apparatus and process disinfecting, deodorizing, preserving, or sterilizing)*. These collections consist of 64,986 documents, with 36,998 documents in *USPC Class 422* and 27,988 documents in *USPC Class 433*. The *USPC Class 433* documents are called with *DentPat Collection*, and the *USPC Class 422* documents are called with *ChemAppPat Collection*. Similar to the *TREC-CRT* collection, the documents in this collection are long, however, the distributions of documents length and vocabulary size are less skewed than the *TREC-CRT* collection (see Figure 1.2 and Figure 1.3). For both collections the retrieval bias is analyzed with the rank cutoff factors *c=5*, *c=10*, *c=15*, *c=20* and *c=25*. Also, these collections are more topically focused, consisting of only two highly similar USPC classes. It is typical for a domain-specific document collection.
- **Austrian News Dataset (ATNews):** Our final collection consists of 47,693 Austrian news documents[3]. We call this collection *(ATNews Collection)*. As compared to above three collections, the documents in this collection are mostly short, however, the distributions of document length

[1] available at `http://www.ir-facility.org/research/evaluation/trec-chem-09`

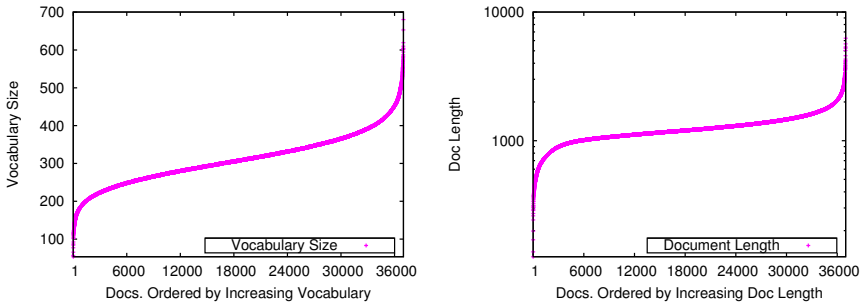
[2] available at `http://www.uspto.gov/`

[3] `http://www.ifs.tuwien.ac.at/~andi/tmp/STANDARD.tgz`



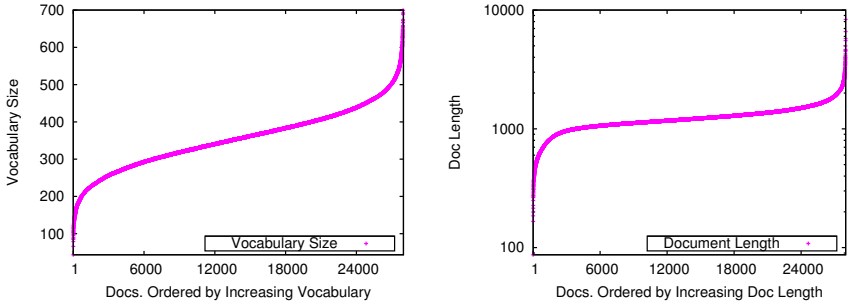
Vocabulary Distribution Doc Length Distribution

Fig. 1.1 Document vocabulary size and length distribution on the *TREC-CRT* collection.



Vocabulary Distribution Doc Length Distribution

Fig. 1.2 Document vocabulary size and length distribution on the *ChemAppPat* collection.



Vocabulary Distribution Doc Length Distribution

Fig. 1.3 Document vocabulary size and length distribution on the *DentPat* collection.

and vocabulary size are highly skewed similar to the *TREC-CRT* collection (see Figure 1.4). For this collection we use the rank cutoff factors *c=5*, *c=10*, *c=15*, *c=20* and *c=25* for the retrieval bias analysis.

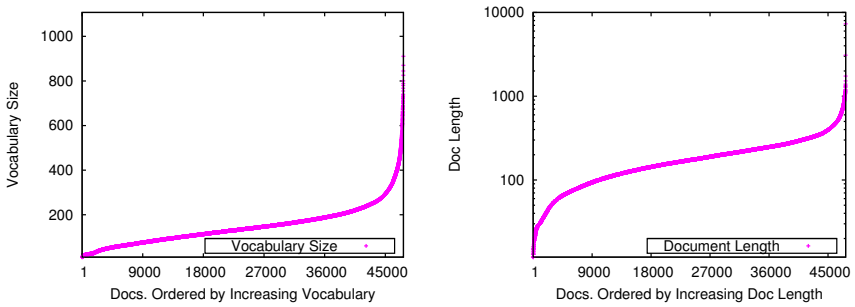


Fig. 1.4 Document vocabulary size and length distribution on the *ATNews* collection.

1.6.2 Retrieval Models

Four standard IR models and four different variations of language models with term smoothing [42] are used for the retrieval bias analysis. These are standard TFIDF, NormTFIDF, the OKAPI retrieval model BM25 [30], SMART [35], Jelinek-Mercer language model JM, Dirichlet (Bayesian) language model DirS, Absolute Discounting language model, and TwoStage language model.

1.6.3 Query Generation for Retrievability Analysis

We consider all sections (title, abstract, claims, description, background summary) of patent documents for both retrieval and query generation. Stop words are removed prior to indexing and words stemming is performed with the Porter stemming algorithm. Additionally, we do not use all those terms of the collection that have a document frequency greater than 25% of the total collection size as they are too generic. Next, queries for retrievability analysis are generated with the combinations of those terms that appear more than one time in the document. For these terms, all 3-terms and 4-terms combinations are used in the form of boolean AND queries for creating the exhaustive set of queries Q, and duplicate queries are removed from Q.

As we explained earlier, a third factor along with the user ability to formulate the query and the retrieval bias of retrieval model that effects the retrievability of documents is the difference between the result list size of the query and the user's ability that how much deeply he/she would check/read the retrieved documents of the query. In retrievability measurement this difference is controlled with a rank cutoff factor. The high difference implies that the user would go through only a small portion of the retrieved

Characteristics	TREC-CRT	ChemAppPat	DentPat	ATNews
$\|Q\|$	130.8 Million	108.1 Million	110.4 Million	57.2 Million
Minimum Query Result List Size	100	45	45	45
Avg Query Result List Size	1,943	149	154	80
Avg # of Queries/Document	1,790,604	437,969	606,879	135,395

Table 1.2 Properties of Q that are used for the retrieval bias analysis.

documents, and thus we can expect that the retrievability of documents will highly depend upon the retrieval bias of retrieval model. Less retrieval bias would make a large number of documents highly retrievable at top ranked positions. On the other hand, if this difference is small, or the size of query result lists become less than the rank cutoff factor, then the user would go through a large portions of documents and thus the bias of retrieval models will play less part on the retrievability of documents.

Therefore, in order to precisely analyze the effect of retrieval model's retrieval bias on the retrievability, the size of query result lists should not be too close to the user's rank cutoff.

Under this principle, for the *TREC-CRT* collection, we remove all those queries from the Q that retrieve less than 100 documents. Similarly for the *ChemAppPat*, *DentPat* and *ATNews* collections we remove all those queries from the Q that retrieve less than 45 documents. Next, all these queries are used for the documents retrieval against the complete collection as boolean AND queries with subsequent ranking according to the chosen retrieval models to determine the retrievability scores of documents. Table 1.2 shows the general characteristics of Q for the different collections. Figures 1.5, 1.6, 1.7, 1.8 show the distributions of the total number of queries per document relative to the vocabulary size of documents. The *TREC-CRT* and *ATNews* collections have large differences between the documents' vocabulary size, thus for these collections this distribution is highly skewed. The *ChemAppPat* and *DentPat* collections have less differences between the documents' vocabulary size, thus for these collections the distribution of queries is less skewed.

1.6.4 Retrieval Bias Analysis

Tables 1.3, 1.4, 1.5, and 1.6 list the retrievability inequality providing Gini-Coefficients for a range of rank cutoff factors for the different collections. Note that a high bias is experienced when limiting oneself to short result lists of 5 or 50 documents. The Gini-Coefficient tends to decrease slowly for all query sets and for all retrieval models as the rank cutoff factor increases. This indicates that the retrievability inequality within the collection is mitigated by the willingness of the user to search deeper down into the result list. If a

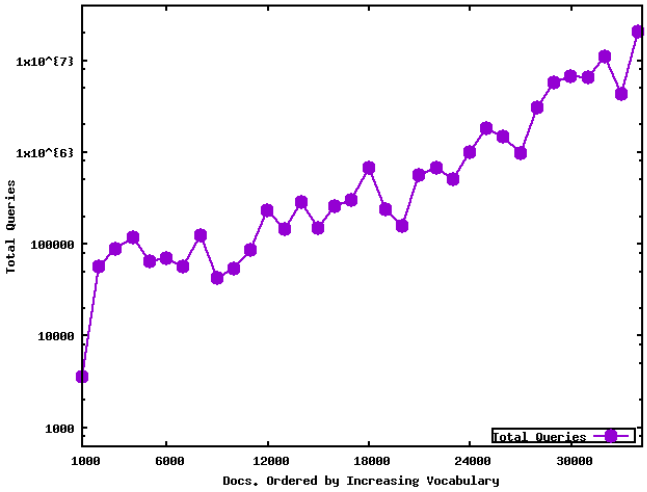


Fig. 1.5 The distribution of total number of queries generated per document for the *TREC-CRT* collection. Documents are ordered by the increasing vocabulary size.

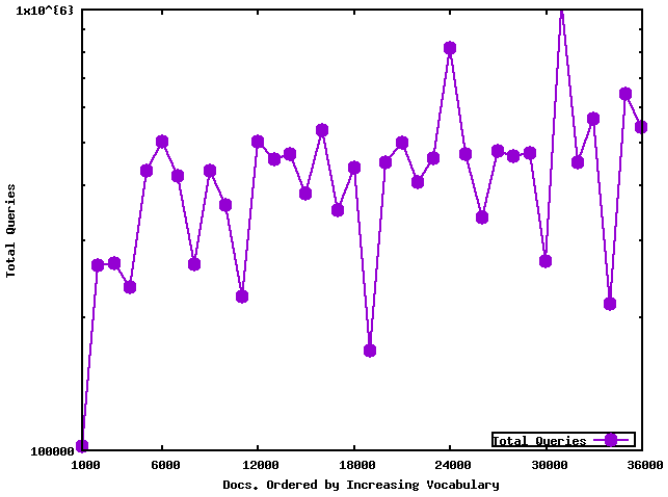


Fig. 1.6 The distribution of total number of queries generated per document for the *ChemAppPat* collection. Documents are ordered by the increasing vocabulary size.

user examines only a small portion of the result list, then he/she will face a greater degree of retrieval bias.

Overall, *BM25* on all collections exhibits lower retrieval bias than all other retrieval models. The four language modeling approaches *DirS*, *TwoStage*,

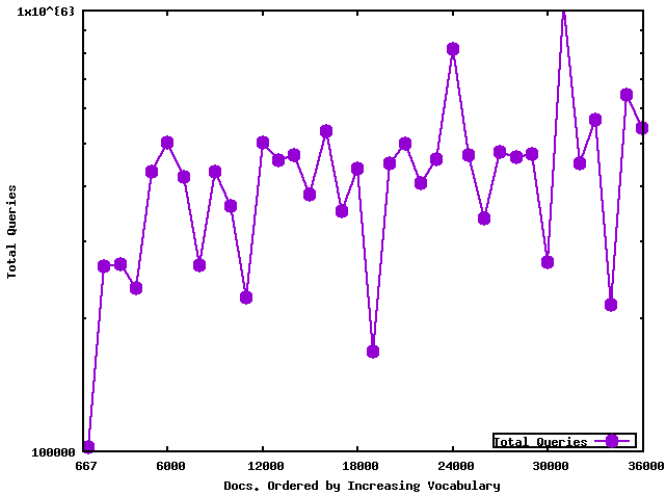


Fig. 1.7 The distribution of total number of queries generated per document for the *DentPat* collection. Documents are ordered by the increasing vocabulary size.

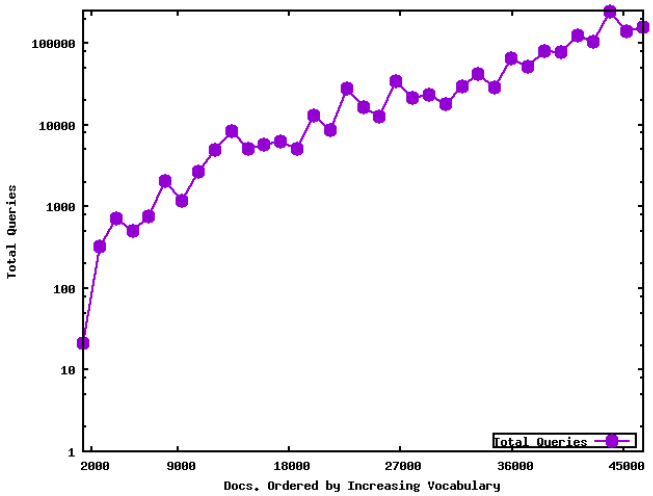


Fig. 1.8 The distribution of total number of queries generated per document for the *ATNews* collection. Documents are ordered by the increasing vocabulary size.

JM, and *AbsDis* also exhibit lower retrieval bias than *TFIDF*, *SMART* and *NormTFIDF*.

Retrieval Model	$r(d)$		
	c=50	**c=100**	**c=250**
TFIDF	0.95	0.91	0.81
NormTFIDF	0.70	0.62	0.51
BM25	0.57	0.52	0.44
SMART	0.96	0.93	0.87
DirS	0.63	0.57	0.50
JM	0.68	0.62	0.51
AbsDis	0.66	0.60	0.50
TwoStage	0.64	0.56	0.46

Table 1.3 Gini-Coefficient scores representing the retrieval bias of different retrieval models on various rank cutoff factors for the *TREC-CRT* collection. As rank cutoff factor increases, bias steadily decreases indicating that the low retrieval bias is experienced when considering the long ranked lists.

Retrieval Model	$r(d)$				
	c=5	**c=10**	**c=15**	**c=20**	**c=25**
TFIDF	0.65	0.56	0.52	0.49	0.48
NormTFIDF	0.48	0.42	0.39	0.38	0.37
BM25	0.39	0.38	0.37	0.37	0.37
SMART	0.93	0.88	0.84	0.81	0.77
DirS	0.43	0.39	0.38	0.37	0.37
JM	0.41	0.37	0.36	0.36	0.36
AbsDis	0.40	0.38	0.38	0.38	0.38
TwoStage	0.47	0.42	0.39	0.38	0.38

Table 1.4 Gini-Coefficient scores representing the retrieval bias of different retrieval models on various rank cutoff factors for the *ChemAppPat* collection. As rank cutoff factor increases, bias steadily decreases indicating that the low retrieval bias is experienced when considering the long ranked lists.

Retrieval Model	$r(d)$				
	c=5	**c=10**	**c=15**	**c=20**	**c=25**
TFIDF	0.65	0.58	0.53	0.50	0.48
NormTFIDF	0.51	0.44	0.41	0.40	0.39
BM25	0.41	0.39	0.38	0.38	0.38
SMART	0.93	0.89	0.85	0.81	0.78
DirS	0.46	0.42	0.40	0.39	0.38
JM	0.43	0.40	0.38	0.37	0.37
AbsDis	0.42	0.40	0.39	0.39	0.39
TwoStage	0.49	0.44	0.41	0.40	0.39

Table 1.5 Gini-Coefficient scores representing the retrieval bias of different retrieval models on various rank cutoff factors for the *DentPat* collection. As rank cutoff factor increases, bias steadily decreases indicating that the low retrieval bias is experienced when considering the long ranked lists.

1.7 Relationship between Retrievability and Effectiveness

The retrievability measure has been used in a number of different contexts (see [5] for more details and [2, 6] for examples of its usage in practice).

Retrieval Model	$r(d)$				
	c=5	**c=10**	**c=15**	**c=20**	**c=25**
TFIDF	0.95	0.92	0.90	0.88	0.87
NormTFIDF	0.54	0.53	0.54	0.56	0.57
BM25	0.52	0.52	0.53	0.55	0.57
SMART	0.87	0.83	0.79	0.76	0.73
DirS	0.77	0.73	0.72	0.70	0.69
JM	0.53	0.52	0.53	0.55	0.56
AbsDis	0.56	0.57	0.59	0.60	0.61
TwoStage	0.78	0.75	0.73	0.71	0.70

Table 1.6 Gini-Coefficient scores representing the retrieval bias of different retrieval models on various rank cutoff factors for the *ATNews* collection. As rank cutoff factor increases, bias steadily decreases indicating that the low retrieval bias is experienced when considering the long ranked lists.

However, there has been little work on analyzing the relationship between retrievability and more standard IR effectiveness measures. One important aspect of the retrievability measure is that it can be analyzed or estimated without the availability of explicit ground truth (relevance judgments). It provides an attractive alternative for the automatic ranking of retrieval models. Additionally, it can be also used for tuning a retrieval model's effectiveness by varying its parameter values or retrieval features over retrievability so that they can perform well for a given collection. However, this is possible only if there is a significant positive correlation between the retrieval bias (i.e., the summarized retrievability of all documents) and effectiveness measures. This is because high or low retrieval bias of retrieval models does not imply that the retrieval models will also perform well on the effectiveness measures. For instance - given the definition of retrievability - a retrieval model that ranks the documents by randomly selecting from the document collection would provide a better retrievability to all documents. This would result in a low retrieval bias, but very poor effectiveness for finding the relevant documents. Conversely, a retrieval model that only ranks the set of known relevant documents at the top ranked positions, regardless of given queries would provide a high inequality among documents (poor retrievability and high retrieval bias) but better effectiveness for a set of known topics. This indicates that neither extreme is desirable. However, to what extent we need to trade-off between the retrievability and the effectiveness depends upon the correlation between them. In the following experiments, we rank all retrieval models on both measures independently, and test to what extent the two rankings agree with each other, i.e. to what extent the low retrieval bias of a retrieval model leads to high effectiveness.

1.7.1 Effectiveness Analysis

We select the prior-art (PA) task of the *TREC-CRT* collection for analyzing the effectiveness of retrieval models. The PA task consisted of 1,000 topic queries that are the full-text patent documents (i.e., consisting of at least claims and abstract or description) taken from both the European Patent Office (EPO) and the US Patent Office (USPTO). The goal of searching a patent database for the prior-art search task is to find all previously published related patents on a given topic [25, 13, 20]. It is a common task for *patent examiners* and *attorneys* to decide whether a new patent application is novel or contains technical conflicts with some already patented invention. They collect all related patents and report them in a search report. We use these reports as relevance judgments. Next, we apply a standard approach for query generation in the patent retrieval domain. From each topic, we select only the claim section being the most representative piece of text, as it describes the scope of the invention. In order to build prior-art queries from the claim sections, we first sort all the term in the claim sections on the basis of their increasing term frequencies. Next, we select the top 30 terms that have highest frequencies, and use these terms in the form of a long query for searching the relevant documents. Note, that more complex query generation approaches may be used. Yet, as our primary motivation is to analyze the relationship between effectiveness and retrievability, this standard baseline is sufficient.

We performed effectiveness analysis with Precision@30 (P@30), Recall@100 (R@100), Mean Average Precision (MAP) and b-pref[4].

1.7.2 Retrieval Models

The retrieval models that we use for retrieval bias analysis include standard retrieval models (*NormTFIDF*, *BM25*, *TFIDF* and *SMART*), language modeling based retrieval models (*DirS*, *JM*, *AbsDis* and *TwoStage*), and low level retrieval features of IR.

We use following low-level features in experiments:

- document length ($|d|$),
- document vocabulary size ($|T_d|$),
- sum of absolute query term frequencies within document (Equ. 1.4),

$$tf(d,q) = \sum_{t \in q} tf_{t,d} \tag{1.4}$$

  where $tf_{t,d}$ is term frequency of document (d).

[4] `http://trec.nist.gov/pubs/trec16/appendices/measures.pdf`

- sum of normalized query term frequencies relative to document length (Equ. 1.5),

$$ntf(d,q) = \sum_{t \in q} tf_{t,d}/|d| \tag{1.5}$$

where $|d|$ is length of d.

- sum of document frequency of query terms (Equ. 1.6),

$$sdf(d,q) = \sum_{t \in q} df_t/|D| \tag{1.6}$$

- and the sum of probability of query's terms occurring in the collection (Equ. 1.7).

$$scf(d,q) = \sum_{t \in q} cf_t / \sum_{d \in D} |d| \tag{1.7}$$

where cf_t is collection frequency of t in D.

Retrieval Model	**Rank cutoff factors**		
	c=50	**c=100**	**c=250**
Standard Retrieval Models and Language Models			
TFIDF	0.95	0.91	0.81
NormTFIDF	0.70	0.62	0.51
BM25	⋆**0.57**	⋆**0.52**	⋆**0.44**
SMART	0.96	0.93	0.87
DirS	0.63	0.57	0.50
JM	0.68	0.62	0.51
AbsDis	0.66	0.60	0.50
TwoStage	0.64	0.56	0.46
Low Level Retrieval Functions			
tf(d,q)	0.95	0.92	0.83
ntf(d,q)	⋆**0.71**	⋆**0.63**	⋆**0.51**
sdf(d,q)	0.85	0.85	0.83
$\|d\|$	0.80	0.74	0.61
$\|T_d\|$	0.99	0.99	0.99
scf(d,q)	0.85	0.85	0.83

Table 1.7 Gini-Coefficient scores representing the retrieval bias of different retrieval models on the *TREC-CRT* collection.

Retrieval Model	Retrieval Bias and Effectiveness Scores				
	G@100	R@100	P@30	MAP	b-pref
BM25	0.52	0.156	0.101	0.049	0.428
TwoStage	0.56	0.174	⋆**0.110**	⋆**0.055**	0.474
DirS	0.57	⋆**0.177**	⋆**0.110**	⋆**0.055**	⋆**0.470**
AbsDis	0.60	0.170	0.108	0.052	0.440
JM	0.62	⋆**0.184**	⋆**0.113**	⋆**0.058**	⋆**0.483**
NormTFIDF	0.62	0.082	0.045	0.023	0.320
ntf(d,q)	0.63	0.107	0.061	0.028	0.470
$\|d\|$	0.74	0.001	0.000	0.000	0.256
sdf(d,q)	0.85	0.042	0.027	0.010	0.414
scf(d,q)	0.85	0.002	0.001	0.000	0.237
TFIDF	0.91	0.008	0.003	0.003	0.115
tf(d,q)	⋆**0.92**	⋆**0.016**	⋆**0.008**	⋆**0.004**	⋆**0.428**
SMART	⋆**0.93**	⋆**0.074**	⋆**0.044**	⋆**0.021**	⋆**0.276**
$\|T_d\|$	⋆**0.99**	⋆**0.001**	⋆**0.000**	⋆**0.000**	⋆**0.245**

Table 1.8 Retrieval bias (Gini-Coefficient) and effectiveness scores of different retrieval models on *TREC-CRT* Collection. Retrieval Models are ordered by increasing Gini-Coefficient scores.

Retrieval Model	Correlation Analysis				
	G@100(Rank)	R@100(Rank)	P@30(Rank)	MAP(Rank)	b-pref (Rank)
BM25	1	5	5	5	6
TwoStage	2	3	2	2	2
DirS	3	2	3	3	3
AbsDis	4	4	4	4	5
JM	5	1	1	1	1
NormTFIDF	6	7	7	7	9
ntf(d,q)	7	6	6	6	4
$\|d\|$	8	13	13	12	11
sdf(d,q)	9	9	9	9	8
scf(d,q)	10	12	12	13	13
TFIDF	11	11	11	11	14
tf(d,q)	12	10	10	10	7
SMART	13	8	8	8	10
$\|T_d\|$	14	14	14	14	12
Correlation with G@100		0.79	0.80	0.81	0.72

Table 1.9 Relationship between retrieval bias and effectiveness on *TREC-CRT* Collection. Retrieval Models are ordered by increasing Gini-Coefficient ranks. Last column shows correlation of effectiveness measures with retrieval bias. We calculate correlation using Pearson product-moment correlation coefficient.

1.7.3 Relationship between Two Measures on the basis of Retrieval Models Ranks

So far, we examined the retrieval bias of different retrieval models. Our results show that the retrieval models differ substantially in terms of the retrieval bi-

ases that they impose on the population of documents. However, the question still remains, what is the relationship between minimizing the retrieval bias and maximizing a retrieval model's effectiveness? In a TREC-style definition of effectiveness, it is important to ensure that all relevant documents of topic queries should have high retrievability scores. But given the recall oriented retrieval domains such as patents, legal, or government administration, it is necessary to ensure that all documents should be high retrievable. In this section, we will now specifically examine to what extent the low or high retrieval bias of retrieval models correlates with their effectiveness. That is, if a retrieval model has less retrieval bias than other models, then does it also mean that it is more effective than the others models? If this holds true then the retrievability will provide a valuable alternative for the automatic ranking of retrieval models in the case when there are no resource to relevance judgments available for a given collection. In order to examine these premises we perform the following experiment.

In this experiment, we compare the relationship between the two measures on the basis of retrieval model ranks. Hereby we want to examine to what extent the low retrieval bias of retrieval models leads to high effectiveness. In order to analyze this, we test and rank all retrieval models independently on both measures. Table 1.8 shows the retrieval bias and effectiveness scores and Table 1.9 shows retrieval model ranks on both measures and the relationship between them. Although the relationship between two rank lists is not perfect, it can be observed from the results that the best retrieval models are consistently ranked in at least the top half of the ranking. This indicates a systematic relationship between retrievability and effectiveness measures. When comparing only the standard and the language modeling based retrieval models on the basis of these rankings, then *BM25* and the four language modeling approaches (*JM*, *DirS*, *AbsDis*, and *TwoStage*) have higher effectiveness than other models possibly due to their lower retrieval bias. *NormTFIDF* has lower retrieval bias than *TFIDF* and *SMART*, plus it has higher effectiveness. However, *NormTFIDF* has a higher retrieval bias than *BM25*, *JM*, *DirS*, *AbsDis*, and *TwoStage*, and also lower effectiveness than these models. If we focus only on the low level retrieval features in the bottom half of the table, then *ntf(d,q)* has the lowest retrieval of these, while also having the highest effectiveness. The main reason behind the systematic relationship between a high retrieval bias and a low effectiveness may be the level of retrievability inequality between the documents. When relevant documents show low retrievability, then these are less likely to retrieve at top ranked positions due to the presence of high retrievable documents. This high level of retrievability inequality between documents decreases the overall effectiveness of retrieval models.

1.7.4 Improving Effectiveness of Retrieval Models by Tuning Parameters using Retrieval Bias

Commonly retrieval models are tuned with the help of parameter values. These parameters either control the query terms normalization relative to document length or smooth the document relevance scores in case of unseen query terms. In this experiment we tune the parameter values of different retrieval models over specified ranges and examine their sensitivity and change with both measures (effectiveness and retrieval bias). Four language modeling approaches with term smoothing (*JM*, *DirS*, *AbsDis TwoStage*) along with *BM25* are used for this purpose. In case of *BM25*, *JM* and *TwoStage*, the parameters b and λ are varied from 0.1 to 1.0 in steps of 0.1, while the parameter μ in case of *DirS* is varied from 500 to 10000 in steps of 1000. Figures 1.9, 1.10, 1.11, 1.12, and 1.13 are showing the effect of parameter values on both measures. We can observe that all those parameter value settings that exhibit high retrieval bias do not correspond to the maximum effectiveness. The maximum effectiveness is gained only when the parameter values result in low retrieval bias. Along with the parameter values that exhibit low retrieval bias and high effectiveness, there exist also some parameter values that - while achieving low retrieval bias - also hurt the effectiveness by a small fraction. This decrease in effectiveness occurs due to the relevance bias on long documents, while the tuned retrieval models aim at providing equal access to all documents. These findings again indicate the presence of a strong relationship between the Gini-Coefficient and P@30, R@100, MAP, and b-pref, representing effectiveness of retrieval models.

1.8 Conclusion

Retrievability measures to what extent a retrieval model provides theoretically equal access to all documents, i.e., returns all documents with equal likelihood if all possible queries are posed against a specific document collection. The studies presented in this chapter reveal that at least for recall-oriented application domains, where users are willing to proceed down to 50, 100 or 250 items in a ranked search result list, there is a high correlation with the effectiveness of a retrieval models. This indicates that retrieval models may be tuned using retrievability as a guiding measure, rather than more conventional approaches of tuning through effectiveness measures. The comparisons clearly show for all retrieval models that there is a very strong relationship between a low retrievability bias and the effectiveness measures. Nonetheless, the assumption that the setting with the lowest retrievability bias always leads to a top effectiveness value cannot be confirmed in all cases.

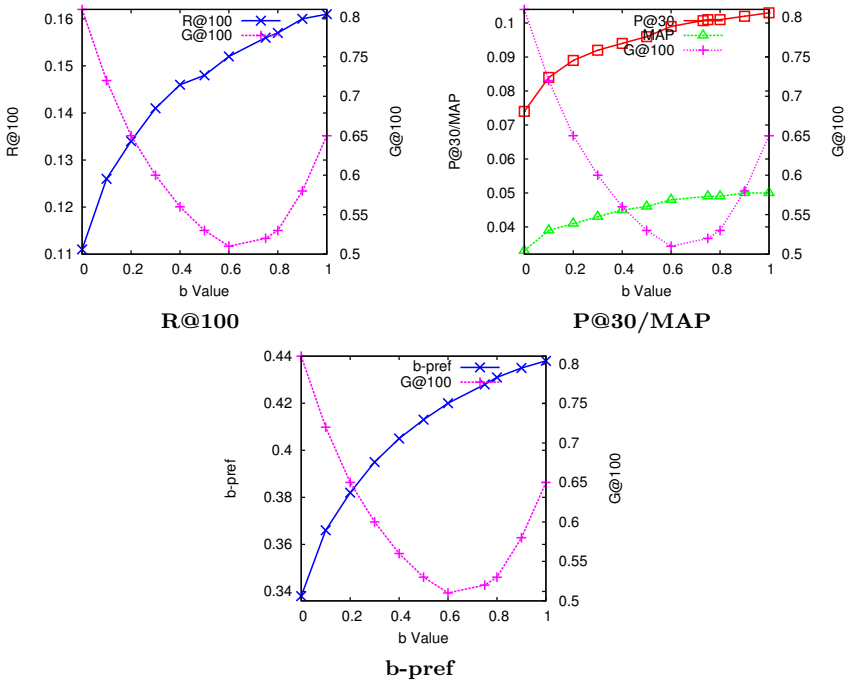


Fig. 1.9 Graphical relationship between the retrieval bias and the effectiveness across various parameter (b) values of *BM25*.

While the parameter settings can be tuned this way and often a good result is obtained, sometimes the results can also turn out slightly below average.

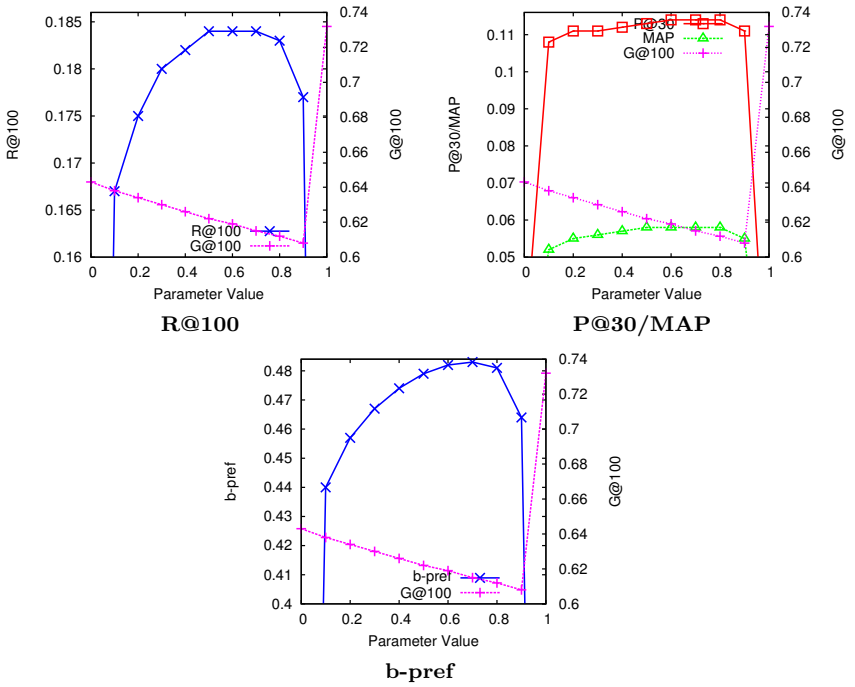


Fig. 1.10 Graphical relationship between the retrieval bias and the effectiveness across various parameter (λ) values of *JM*.

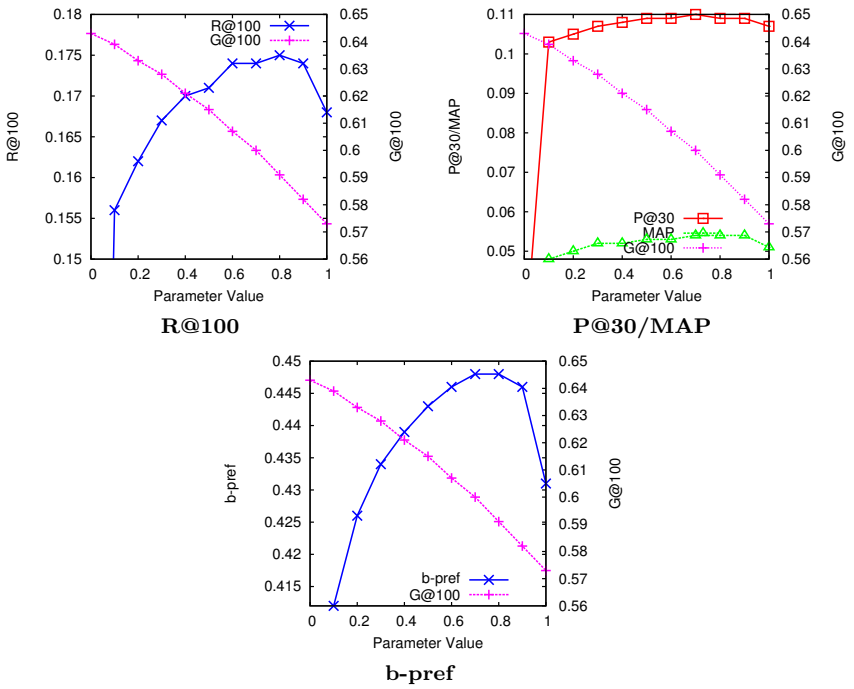


Fig. 1.11 Graphical relationship between the retrieval bias and the effectiveness across various parameter (δ) values of *AbsDis*.

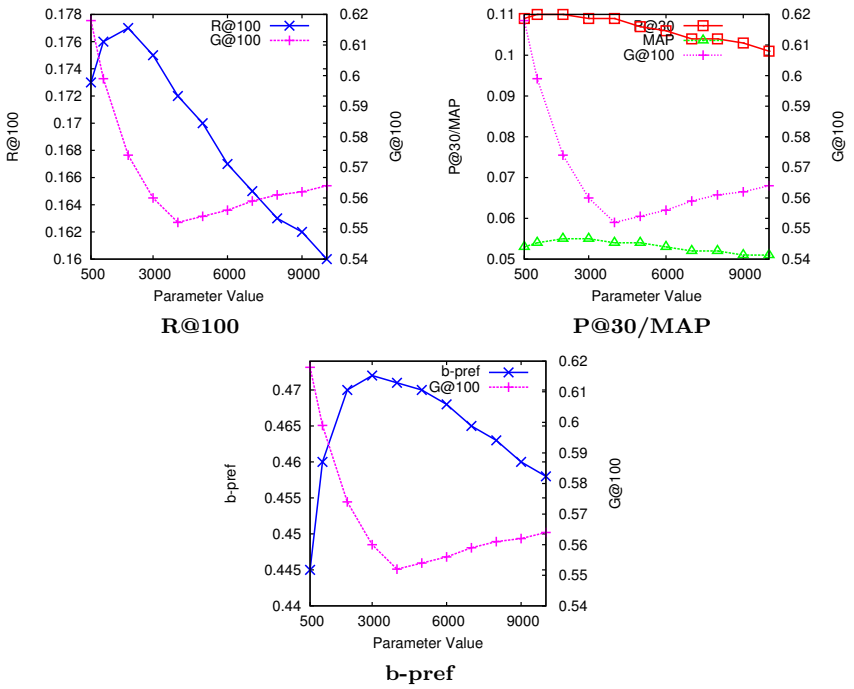


Fig. 1.12 Graphical relationship between the retrieval bias and the effectiveness across various parameter (μ) values of *DirS*.

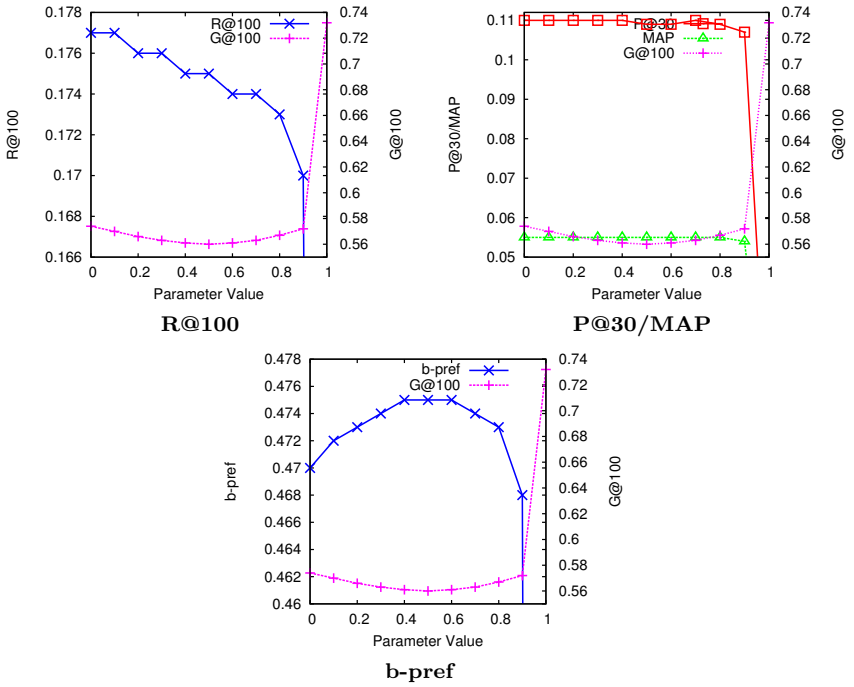


Fig. 1.13 Graphical relationship between the retrieval bias and the effectiveness across various parameter (λ) values of *TwoStage*.

References

1. Avi Arampatzis, Jaap Kamps, Martijn Kooken, and Nir Nussbaum. Access to legal documents: Exact match, best match, and combinations. In *Proceedings of The Sixteenth Text Retrieval Conference (TREC'07)*, 2007.
2. Leif Azzopardi and Richard Bache. On the relationship between effectiveness and accessibility. In *SIGIR '10: Proceeding of the 33rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 889–890, Geneva, Switzerland, 2010.
3. Leif Azzopardi and Ciaran Owens. Search engine predilection towards news media providers. In *SIGIR '09: Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 774–775, Boston, MA, USA, 2009.
4. Leif Azzopardi and Vishwa Vinay. Accessibility in information retrieval. In *ECIR'08: Proceedings of the 30th European Conference on IR Research*, pages 482–489, 2008.
5. Leif Azzopardi and Vishwa Vinay. Retrievability: an evaluation measure for higher order information access tasks. In *CIKM '08: Proceeding of the 17th ACM Conference on Information and Knowledge Management*, pages 561–570, Napa Valley, California, USA, 2008.
6. Richard Bache and Leif Azzopardi. Improving access to large patent corpora. In *Transactions on Large-Scale Data- and Knowledge-Centered Systems II*, volume 2, pages 103–121. Springer, 2010.
7. Ricardo Baeza-Yates and Berthier Ribeiro-Neto. Modern information retrieval. In *ACM Press*, 1999.
8. Shariq Bashir and Andreas Rauber. Automatic ranking of retrieval models using retrievability measure. *Knowl. Inf. Syst.*, 41(1):189–221, 2014.
9. Jamie Callan and Margaret Connell. Query-based sampling of text databases. In *ACM Transactions on Information Systems (TOIS) Journal*, volume 19, Issue 2, pages 97–130, 2001.
10. G. G. Chowdhury. Introduction to modern information retrieval. In *Second Edition, Facet Publishing, London*, 2004.
11. Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schutze. Introduction to information retrieval. In *Cambridge University Press*, 2008.
12. Miles Efron. Using multiple query aspects to build test collections without human relevance judgments. In *Advances in Information Retrieval, 31th European Conference on IR Research, ECIR 2009, Toulouse, France, April 6-9, 2009. Proceedings*, pages 276–287, 2009.
13. Atsushi Fujii, Makoto Iwayama, and Noriko Kando. Introduction to the special issue on patent processing. In *Information Processing and Management Journal*, volume 43, number 5, pages 1149–1153, 2007.
14. Joseph L. Gastwirth. The estimation of the Lorenz curve and Gini index. In *The Review of Economics and Statistics*, volume 54, number 3, August, pages 306–16, August 1972.
15. Karst T. Geurs and Bert van Wee. Accessibility evaluation of land-use and transport strategies: Review and research directions. In *Journal of Transport Geography*, volume 12, pages 127–140, 2004.
16. Walter G. Hansen. How accessibility shape land use. In *Journal of the American Institute of Planners*, volume 25, pages 73–76, 1959.
17. Claudia Hauff, Djoerd Hiemstra, Leif Azzopardi, and Franciska de Jong. A case for automatic system evaluation. In *Advances in Information Retrieval, 32nd European Conference on IR Research, ECIR 2010, Milton Keynes, UK, March 28-31, 2010. Proceedings*, pages 153–165, 2010.

18. P. L. Dumble J. M. Morris and M. R. Wigan. Accessibility indicators for transport planning. In *Transportation Research Part A: General*, volume 13, pages 91–109, 1979.
19. Todd Litman. Evaluating accessibility for transportation planning. In *Victoria Transport Policy Institute*, 2008.
20. Mihai Lupu, Jimmy Huang, Jianhan Zhu, and John Tait. TREC-CHEM: large scale chemical information retrieval evaluation at trec. In *SIGIR Forum*, volume 43, number 2, pages 63–70. ACM, 2009.
21. Ellen M. Voorhees. Overview of the trec 2001 question answering track. In *Proc. of the Text Retrieval Conference, TREC'01*, pages 42–51, 2001.
22. Ellen M. Voorhees. The philosophy of information retrieval evaluation. In *CLEF'01, Springer-Verlag*, pages 355–370, 2002.
23. Ellen M. Voorhees and Donna K. Harman. Trec experiment and evaluation in information retrieval. In *Cambridge, Massachusetts: MIT Press*, 2005.
24. Walid Magdy and Gareth J. F. Jones. Pres: A score metric for evaluating recall-oriented information retrieval applications. In *SIGIR'10: ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 611–618. ACM, 2010.
25. Hisao Mase, Tadataka Matsubayashi, Yuichi Ogawa, Makoto Iwayama, and Tadaaki Oshio. Proposal of two-stage patent retrieval method considering the claim structure. In *ACM Transactions on Asian Language Information Processing (TALIP)*, volume 4, number 2, pages 190–206, 2005.
26. Abbe Mowshowitz and Akira Kawaguchi. Bias on the web. In *Communications of the ACM*, volume 45, number 9, pages 56–60, New York, NY, USA, 2002. ACM.
27. Rabia Nuray and Fazli Can. Automatic ranking of information retrieval systems using data fusion. *Inf. Process. Manage.*, 42(3):595–614, May 2006.
28. Iadh Ounis, Maarten De Rijke, Craig Macdonald, Gilad Mishne, and Ian Soboroff. Overview of the trec 2006 blog track. In *Proc. of the Text Retrieval Conference, TREC'06*, 2006.
29. Stephen P. Harter and Carol A. Hert. Evaluation of information retrieval systems: Approaches, issues, and methods. In *Annual Review of Information Science and Technology (ARIST)*, volume 32, pages 3–94, 1997.
30. S. E. Robertson and S. Walker. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *SIGIR '94: Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 232–241, Dublin, Ireland, 1994.
31. Tetsuya Sakai and Chin-Yew Lin. Ranking retrieval systems without relevance assessments: Revisited. In *Proceedings of the 3rd International Workshop on Evaluating Information Access, EVIA 2010, National Center of Sciences, Tokyo, Japan, June 15, 2010*, pages 25–33, 2010.
32. Mark Sanderson and Justin Zobel. Information retrieval system evaluation: Effort, sensitivity, and reliability. In *SIGIR'05: ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 162–169. ACM, 2005.
33. Zhiwei Shi, Peng Li, and Bin Wang. Using clustering to improve retrieval evaluation without relevance judgments. In *COLING 2010, 23rd International Conference on Computational Linguistics, Posters Volume, 23-27 August 2010, Beijing, China*, pages 1131–1139, 2010.
34. Zhiwei Shi, Bin Wang, Peng Li, and Zhongzhi Shi. Using global statistics to rank retrieval systems without relevance judgments. In Zhongzhi Shi, Sunil Vadera, Agnar Aamodt, and David B. Leake, editors, *Intelligent Information Processing*, volume 340 of *IFIP Advances in Information and Communication Technology*, pages 183–192. Springer, 2010.
35. Amit Singhal. At&t at trec-6. In *The 6th Text Retrieval Conference (TREC6)*, pages 227–232, 1997.
36. Amit Singhal. Modern information retrieval: A brief overview. In *IEEE Data Engineering Bulletin*, volume 24, pages 34–43, 2001.

37. Anselm Spoerri. Using the structure of overlap between search results to rank retrieval systems without relevance judgments. *Inf. Process. Manage.*, 43(4):1059–1070, 2007.
38. Ingemar J. Cox Vaclav Petricek, Tobias Escher and Helen Margetts. The web structure of e-government - developing a methodology for quantitative evaluation. In *WWW '06 Proceedings of the 15th international conference on World Wide Web*, pages 669–678, 2006.
39. Liwen Vaughan and Mike Thelwall. Search engine coverage bias: evidence and possible causes. In *Information Processing and Management Journal*, volume 40, issue 4, May, pages 693–707, May 2004.
40. Handy W. Lauw, Ee-Peng Lim, and Ke Wang. Bias and controversy: Beyond the statistical deviation. In *In Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 625–630, Philadelphia, PA, USA, 2006.
41. Colin Wilkie and Leif Azzopardi. A retrievability analysis: Exploring the relationship between retrieval bias and retrieval performance. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management, CIKM 2014, Shanghai, China, November 3-7, 2014*, pages 81–90, 2014.
42. ChengXiang Zhai. Risk minimization and language modeling in text retrieval. In *PhD Thesis, Carnegie Mellon University*, 2002.