

Studying Retrievability of Publications and Datasets in an Integrated Retrieval System

Dwaipayan Roy
Indian Institute of Science Education
and Research, Kolkata
India
dwaipayan.roy@iiserkol.ac.in

Zeljko Carevic
GESIS – Leibniz Institute for the
Social Sciences, Cologne
Germany
Zeljko.Carevic@gesis.org

Philipp Mayr
GESIS – Leibniz Institute for the
Social Sciences, Cologne
Germany
Philipp.Mayr@gesis.org

ABSTRACT

In this paper, we investigate the retrievability of datasets and publications in a real-life Digital Library (DL). The measure of *retrievability* was originally developed to quantify the influence that a retrieval system has on the access to information. Retrievability can also enable DL engineers to evaluate their search engine to determine the ease with which the content in the collection can be accessed. Following this methodology, in our study, we propose a system-oriented approach for studying dataset and publication retrieval. A speciality of this paper is the focus on measuring the accessibility biases of various types of DL items and including a metric of *usefulness*. Among other metrics, we use Lorenz curves and Gini coefficients to visualize the differences of the two retrievable document types (specifically datasets and publications). Empirical results reported in the paper show a distinguishable diversity in the retrievability scores among the documents of different types.

CCS CONCEPTS

• **Information systems** → *Retrieval models and ranking*.

KEYWORDS

Retrievability, Dataset Retrieval, Interactive IR, Diversity

ACM Reference Format:

Dwaipayan Roy, Zeljko Carevic, and Philipp Mayr. 2022. Studying Retrievability of Publications and Datasets in an Integrated Retrieval System. In *The ACM/IEEE Joint Conference on Digital Libraries in 2022 (JCDL '22)*, June 20–24, 2022, Cologne, Germany. ACM, New York, NY, USA, 9 pages. <https://doi.org/10.1145/3529372.3530931>

1 INTRODUCTION

With the availability of various types of information from different sources, sometimes with different modalities, the necessity of federated or integrated search systems [1, 2] has become a prominent research topic in information retrieval and digital library. Textual

data still remains the predominant type among them and significant research has been conducted in the domain of textual document retrieval. Among the rest, recent research on dataset retrieval [25] has become increasingly important in the (interactive) information retrieval and digital library community. One of the reasons is undoubtedly the massive amount of research datasets available. However, the underlying characteristics of dataset retrieval also contribute to the attention in this area. One often mentioned characteristic is the increased complexity of datasets over traditional document retrieval. While the latter is well known and adequately studied, datasets often include more extensive material and structures that are relevant for retrieval. This may involve the raw data, descriptions of how the data was collected, taxonomic information, questionnaires, codebooks, etc. Recently, numerous studies have been conducted to further identify the characteristics of dataset retrieval. These studies include the observation of data retrieval practices [24], interviews and online questionnaires [15, 23] and transaction log analysis [9, 22].

In this paper, we follow a system-oriented approach for studying dataset retrieval. By employing the measure of *retrievability* [3], we aim to gain insights into the particularities of dataset retrieval in comparison to traditional document retrieval. The measure of retrievability was originally developed to quantify the influence that a retrieval system has on the access to information. In a simplified way, retrievability represents the ease with which a document can be retrieved given a particular IR system [3]. The measure of retrievability can be utilised for several use cases. Retrievability can, for instance, be employed to identify potential biases in the underlying corpus. Retrievability can also enable digital library (DL) engineers to evaluate their search engine to determine the ease with which the content in the collection can be accessed.

In this paper, we investigate the retrievability of various types of documents in the integrated DL *GESIS Search* (see Section 4), focusing on datasets and publications in the same system. The assumption followed here is that in an ideal ranking system¹, the retrievability of each indexed item (dataset or other publication) is equally distributed. Likewise, a discrepancy to this assumption may reveal an inequality between the items in a collection caused by the system. By employing a measure of retrievability, we expect to gain further insight into the characteristics of dataset retrieval compared to traditional document retrieval.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

JCDL '22, June 20–24, 2022, Cologne, Germany

© 2022 Association for Computing Machinery.

ACM ISBN 978-1-4503-9345-4/22/06...\$15.00

<https://doi.org/10.1145/3529372.3530931>

¹In this paper, by *ranking system* or *IR system*, we refer to a system containing a corpus together with the retrieval model to be used to search on that corpus.

Research questions

Considering the real-life digital library *GESIS Search*, in this paper, we address the following research questions:

- **RQ1:** In the integrated search system with various types of items, can we observe any prior bias of accessibility of documents from a particular type?
- **RQ2:** Can we formalize this type-accessibility bias utilizing the concept of document retrievability?
- **RQ3:** How diverse are the retrievability score distributions in the different categories of documents in our integrated search system?

In sum, our contributions are as follows: 1) we utilize the retrievability measure to better understand the diversity of accessing datasets in comparison to publications with real-life queries from a search log; 2) building on retrievability, we propose the measurement of *usefulness*, which represents implicit relevance signals observed for datasets and publications. Our understanding of bias follows the argumentation provided in [32] where bias denotes the inequality between documents in terms of their retrievability within the collection. Bias can be observed when a document is overly or unduly favoured due to some document features (e.g. length, term distribution etc.) [31].

The rest of the paper is organized as follows. We first present related work in Section 2. A formal introduction of the concept of retrievability is presented in Section 3. The integrated search system *GESIS Search* along with the motivation of our retrievability study is presented in Section 4. Section 5 discusses the empirical results and analysis of the outcome of the experimentation before introducing the novel concept of usefulness in Section 6 along with the experimental study of usefulness. We conclude the paper in Section 7 highlighting the contributions and findings of the paper.

2 RELATED WORK

Retrievability as a measure was proposed in [3]. In the work, the authors demonstrate the utility of retrievability on two TREC collections using query based sampling [7]. Retrievability has been primarily used to detect bias in ranking systems. For instance, [28] employ retrievability to research the effect of bias across time for different document versions (treated as independent documents) in a web archive. Their results show a ranking bias for different versions of the same document. Furthermore, the study confirms a relationship between retrievability and findability measured by Mean Reciprocal Rank (MRR). They follow the assumption that the lower a document's retrievability score the more difficult it is to find the document. Another application of the retrievability measure can be found in patent or legal document retrieval which provides a unique use case due to its recall oriented application. In both studies, [4, 5] look at document retrievability measurements and argue that a single retrievability measure has several limitations in terms of interpretability. In [4] they try to improve accessibility measurement by considering sets of relevant and irrelevant queries for each document. In this way, they try to simulate recall oriented users. In addition, they plot different retrievability curves to better spot the gaps between an optimal retrievable system and the tested system. In [5] they analyze the bias impact of different retrieval models and query expansion strategies. Their

experiments show that clustering based document selection for pseudo-relevance feedback is an effective approach for increasing the findability of individual documents and decreasing the bias of a retrieval system.

A study on the query list generation phase for determining the measure of retrievability is presented in [6]. The study addresses two central problems when determining retrievability: 1) query selection and 2) query characteristics identification. It is argued that the query selection phase is usually performed individually without well-accepted criteria for query generation. Hence their goal is to evaluate how far the selection of query subsets provides an accurate approximation of retrieval bias. The second shortcoming is addressed by determining retrievability bias considering different query characteristics. In their experiments they recognise that query characteristics influence the increase or decrease of retrievability scores. As recognised in [6], the majority of retrievability experiments employ simulated queries to determine retrievability. To study the ability of the retrieval measure in detecting a potential retrievability bias using real queries issued by users, [30] conducted an experiment on a newspaper corpus. Their study confirms the ability to expose retrievability bias within a more realistic setting using real world queries. A comparison of simulated and real queries with regard to retrievability scores further shows considerable differences which indicates a need for an improved construction of simulated queries.

The concept of usefulness was first introduced in [13] as a criterion to determine how well a system is able to solve a user's information need. In [20], it has been operationalised within a log-based evaluation approach to determine the usefulness of a search term suggestion service. In [10], usefulness has been operationalised to determine the effects of contextualised stratagem browsing on the success of a search session.

Recently, a considerable amount of research has been carried out concerning the characteristics of dataset retrieval. A comprehensive literature review on dataset retrieval practices is provided in [17] focusing on dataset retrieval practices in different disciplines. Research in this area covers, for instance, the analysis of information-seeking behaviour during dataset retrieval through observations [24], questionnaires and interviews [15, 23], and transaction-log studies [9, 22]. In their work in [23], the authors investigated the requirements that users have for a dataset retrieval system. Their findings on dataset retrieval practices suggest that users invest greater effort during relevance assessment of a dataset. They conclude that the selection of a dataset is a much more important decision compared to the selection of a piece of literature. This results in high demands on metadata quality during the dataset retrieval. The complexity of assessing the relevance of a dataset is also highlighted in [24]. Besides topical relevance, access to metadata as well as documentation about the dataset plays a crucial role. A query log analysis from four open data portals is presented in [22]. Their study indicates differences between queries issued towards a dataset retrieval system and queries in web search. In a subsequent study [21], the extracted queries are further compared to queries generated from a crowdsourcing task. The intuition and focus of this work is to determine whether queries issued towards

a data portal differ from those collected in a less constrained environment (crowdsourcing).

3 RETRIEVABILITY

The *retrievability of a document* in a collection, as defined by Azopardi and Vinay, is the concept of systematically measuring the possibility of retrieval of the document by *any query*, or in other words, it indicates of how easily the information within the collection can be accessed with a given retrieval model [3]. Formally, given a collection C of documents, and Q , a set of all possible queries for which the answers are expected to be present in C , the retrievability of a document d ($d \in C$) is defined as:

$$r(d) = \sum_{q \in Q} w_q \cdot f(k(d, q), c) \quad (1)$$

In Equation 1, w_q is the weight of the query q , and $k(d, q)$ specifies the rank of document d after querying with q with some retrieval model. The function $f(k(d, q), c)$ is a boolean function indicative of whether $k(d, q)$ (the rank of document d) is within a rank cut-off c . Specifically, the function $f(\cdot)$ returns 1 if $k(d, q) \leq c$, and 0 otherwise, as presented in Equation 2.

$$f(k(d, q), c) = \begin{cases} 1, & \text{if } k(d, q) \leq c. \\ 0, & \text{otherwise.} \end{cases} \quad (2)$$

Informally, the retrievability score $r(d)$ of a document d can also be realized as the number of queries for which d is ranked within rank-cut-off c .

The query weight (w_q in Equation 1) may depend on the *popularity* of the query which can be treated as a uniform constant for all the queries in Q to sidestep the effect of query bias (as presented in [3]); that is, all queries are considered to have equal weights. The approximated retrievability score of document d will then be a discrete value x indicating the number of queries for which d is retrieved within rank c . Certainly, this is a simplifying assumption and the queries submitted to a search system in practice vary vastly both in terms of *popularity* as well as *difficulty* [12]. Constructing Q , the set of all queries, expects the utilization of a query log in absence of which a query-based sampling method [8] can be applied to randomly populate Q . Considering a fairly sized collection of documents, there can be infinitely many distinct queries that can be answered by various documents in the collection. Hence, the construction of Q based on either query log or random sampling of terms from the collection are only some practical approximations that we can adapt in order to realize the concept of retrievability of documents in a collection.

The second factor of the per-query component in Equation 1 is defined to be a boolean function the output of which depends solely on the rank at which document d is retrieved. Increasing the value of the rank cut-off (c) broadens the domain of documents retrieved which will positively influence the retrievability scores of more documents. Note that, being selected by a retrieval model for some queries does not ensure the relevance of the document which can only be comprehended by human judgements.

4 RETRIEVABILITY IN AN INTEGRATED RETRIEVAL SYSTEM

We define an *integrated search system* as a system that searches multiple sources of different types and integrates the output in a unified framework². The retrieval in such a system requires sophisticated decision making considering the various modalities in documents in the collection of data.

Following Equation 1, the retrievability score of documents is dependent on the other documents in the collection³: considering a rank-cut c , the rank of a document under consideration can be greater than c ($> c$) due to the documents, taking the top c positions, being more relevant or duplicate [27]. Another factor that can influence the retrievability score of a document is the popularity; a trendy document will be retrieved multiple times by the users over time. In case of an integrated search engine, where the documents belong to various categories, some particular types could be having higher chances than the others in terms of being retrieved. In general, there can be some disparity in the number of documents of various categories being retrieved which can be a result of popularity bias in the collection. This type of popularity bias can impede the satisfaction of the information need of a user, and in turn, can affect the performance of the system. Traditional methods of query performance prediction [14, 18], in this scenario, can not be effectively applied.

In this paper, we are going to study the diversity in retrievability scores for different categories of document in the integrated search system *GESIS Search*⁴ [19]. To the best of our knowledge, this is the first attempt to formalize this type of diversity utilizing the concepts of retrievability.

5 EXPERIMENTAL STUDY

As presented in Section 4, we use the integrated search system with various categories of documents in this work. In this section, we start with describing the data that we have used in the work along with different statistics of the data; this will be followed by the experimental evaluation of the study.

5.1 Datasets

To identify possible differences between publication and dataset retrieval concerning their retrievability, we use the integrated search system *GESIS Search* containing a total of 830K indexed records. A screenshot showing the interface of *GESIS search* is presented in Figure 1. The indexed records are divided into six categories based on their types, covering more than 118K *publications* and 64K *research datasets* (also referred to as *datasets*). Given a query, the system returns six search result pages (SERP) corresponding to each of the categories. The segregation of the SERP enables us to study the retrievability of the different types. In this study, we concentrated on the types *publications* and *datasets* independently.

In the integrated search system, the interaction of the users with the system is logged and stored in a database. A total of more than 40 different interaction types are stored covering, for instance, searches (queries), record views and export interactions etc. [19].

²This is similar to the concepts of aggregated search [26] or federated search [2].

³Here, we are considering the employed retrieval function as constant.

⁴<https://search.gesis.org>



Figure 1: Screenshot of GESIS Search showing result sets for research data and publications.

The export of a record belongs to an umbrella of categories including various interactions such as bookmarking, downloading or citing. These interactions are specifically useful for the application of implicit relevance feedback as they indicate a relevance of a record that goes beyond a simple record view. The interaction log of the search system provides the basis for our analysis in Section 5.4 (and later in Section 6.2). The queries contained in the log (more than 1.2 M) form the basis for determining the retrievability of documents. This ensures realistic queries in Q of Equation 1 as opposed to the simulated queries used in [3] or [30]. The data used in this study was extracted from the interaction log of the integrated search system between July 2017 and December 2021. Statistics regarding the extracted interactions utilized in our study can be found in Table 1.

5.2 Measuring retrievability in a collection

In a traditional relational databases, there can be a one-to-one correspondence between a query and a record. In other words, each record stored in R-DBMS can be distinctly selected by a specific query raised to the system. Also the different queries that can be

Record type	Size	#queries (unique)	avg. query length	#exports
Publication	113,000	626,335 (276,553)	15.4	8,146
Dataset	64,000	602,581 (187,644)	20.5	24,310

Table 1: Statistics of the extracted information.

submitted and addressed by such systems are finitely tractable which depends on the number of records stored. However, considering the nature of *free-text query*, infinitely many queries can be answered by a traditional information retrieval system through the selection of documents containing the responses. Also due to the notion of relevance grade (extremely relevant, loosely relevant etc.), and the necessity of having a ranked list, certain documents have more chances of being retrieved than the others.

Considering a traditional document collection C, all the documents are not equally important to a query, hence paving the need

Rank cutoff	Publication				Dataset			
	μ	geo- μ	σ^2	σ	μ	geo- μ	σ^2	σ
10	54.86	15603.2	124.91	24.26	103.27	144018.77	379.5	20.63
20	103.95	38350.5	195.83	50.8	176.63	317920.25	563.84	38.21
30	152.49	66770.93	258.4	79.17	247.01	540999.45	735.53	56.93
40	200.36	100471.29	316.97	108.51	315.71	794970.82	891.61	76.73
50	247.66	134928.78	367.33	138.47	383.59	1080000.55	1039.23	97.57
100	477.78	352835.83	594	291.22	708.76	2724762.92	1650.69	214.74

Table 2: The mean (both arithmetic and geometric), variance and standard deviation of the retrievability values when the rank-cutoff is varied.

to have a ranked retrieval. Now given a set of all possible queries Q , some documents in C will be relevant to more number of queries (depending on the topical coverage of the document) than others which can be measured by the concept of retrievability (see Section 3). Formally with the notion of retrievability, some documents will be having higher $r(d)$ in a collection, resulting in an unequal distribution of retrievability scores.

Similar types of inequalities are observed in economics and social sciences, and they are traditionally measured using the Gini coefficient or Lorenz curve [16] which measures the statistical dispersion in a distribution⁵.

Mathematically, the Gini coefficient (G) of a certain value v in a population \mathcal{P} can be defined as:

$$G = \frac{\sum_{i=1}^N (2 * i - N - 1) * v(i)}{N \sum_{j=1}^N v(j)} \quad (3)$$

where N is the size of the population and $v(i)$ specifies the value of i^{th} item in \mathcal{P} . The value of Gini coefficient in the population will be between 0 and 1 and is proportional to the inequality inherent in the population: higher value of G indicates greater disparity and vice versa. In other words, a value of G equal to 0 in Equation 3 indicates that all the items in the population are equally probable to be selected where as higher values of G specify a bias implying that only certain items will be selected.

5.3 Experimentation

As explained in Section 3, the retrievability of a document is a measurement of how likely the document will be retrieved by *any* query submitted to the system⁶. Hence, the study of retrievability in a collection of documents requires rigorous retrieval with a set of diversified queries to cover all topics discussed in the collection. In other words, the retrievability of the documents should be calculated considering all sorts of queries submitted to the system. However, an infinite number of queries are possible to be answered by a collection for free-text queries. To cover all the topics, a traditional approximation is to simulate a set of queries randomly, accepting the risk of erratic queries not aligned with the real scenario [3, 30]. With the availability of a query log, the process of query generation can be made more formalized and streamlined to consider the

actual queries submitted by real users. For the study reported in this article, we utilize the query log presented in Section 5.1.

As reported in the earlier study, the retrievability distribution in a collection depends on the employed retrieval model [3]. Following the findings by Azzopardi and Vinay, we use BM25 as the retrieval model [29]. Particularly, we use the implementation available in Elasticsearch⁷ which use Lucene⁸ as the background retrieval model. Following Equation 1, the retrievability of a document depends on the selection of the rank cutoff value (c) - a rank threshold to indicate how deep in the ranked list are we going to explore before finding that document. Considering the model employed for retrieval and the set of all queries Q as fixed, c is the only parameter in calculating the retrievability. For a query q , setting a lower value to c will restrict less number of documents being considered retrievable because $f(k(d, q), c)$ will be 1 only if $k(d, q) \leq c$ (see Equation 2). Having a higher value of c will allow more documents to be considered retrievable reducing the overall inequality. In this study, we have varied the value of c in the range 10 to 100 in steps of 10 and have analyzed the observations which is reported in the next section.

5.4 Observation and analysis

We start this section with describing different statistical properties of the retrievability distribution of documents (both publication and dataset) when the value of c is varied. The mean (μ), geometric mean (geo- μ), variance (σ^2), and standard deviation (σ) of the retrievability score distribution on different types (publication and dataset) are given in Table 2. On varying the value of c from 10 to 100, we observe a drastic 770% change in the mean retrievability score in case of publication while for datasets, the change in the mean is around 580%. Similar observations are recorded for variance and standard deviation as well when computed using the distribution of $r(d)$ on publication and dataset with different c values. The variations are presented graphically in Figure 2. From the table, we can conclude that most of the statistical measurements (specifically mean, variance, and standard deviation) are higher for the datasets than publications. One contrasting observation can be made from Figure 2b where the variation in geometric mean (geo- μ) is presented. Unlike the other three statistics, the geometric mean is seen to be higher for publication than dataset across all

⁵Lorenz curve and Gini coefficient are popular in economics to measure of wealth disparity in a community/country.

⁶By a system, we are referring to the organization of the collection, along with a retrieval model to be used for retrieval for a given query.

⁷<https://www.elastic.co/>

⁸<https://www.lucene.apache.org/>

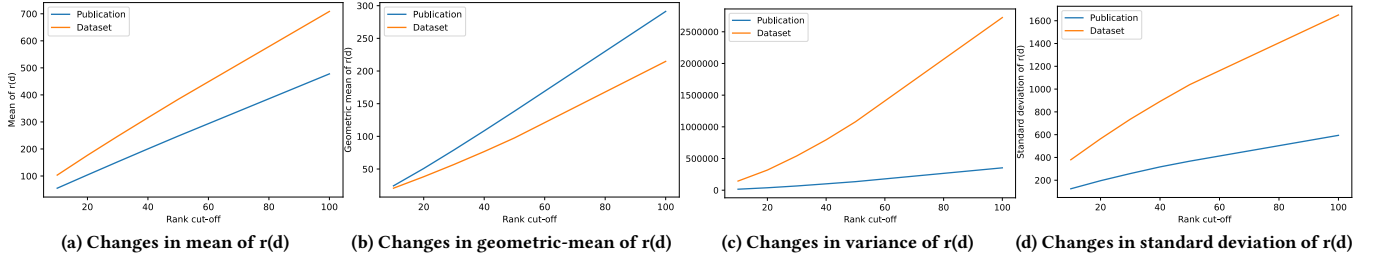


Figure 2: Graphical representation of the change in various statistical measures of the observed distribution of retrievability scores. The mean, geometric-mean, variance and standard deviation of the distribution of retrievability scores of publication (in blue) and dataset (in orange) are presented.

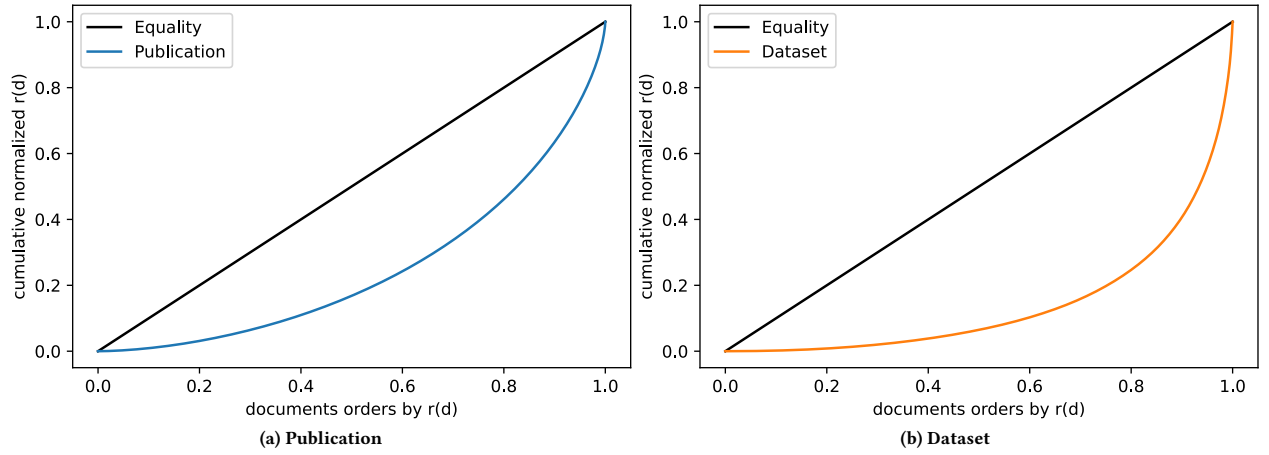


Figure 3: The Lorenz curve with the retrievability (rank cutoff set to 100). The straight line going through the origin (in black) indicate the equality, that is, when all the documents are equally retrievable.

values of c . Comparing geometric-mean values with the other statistics, we can perceive that for some dataset items, the retrievability values are extensive (for the popular datasets); at the same time, there are number of datasets items with poor $r(d)$ values (for those which are rarely cited/used). The first category of datasets are contributing to the high average $r(d)$, which is consistent across different c values, while the datasets of the second category cause the geometric-mean to fall. This is also in line with the observed higher values of variance (and standard deviation) for datasets that indicates diverse $r(d)$ scores. As opposed to that, we witness a similar behaviour of the publications in terms of both mean and its geometric counterpart.

To examine the imbalance in the $r(d)$ scores across the documents of a category, in Figure 3 we plot the Lorenz curve with the $r(d)$ scores computed separately for publications and datasets. The Lorenz curve shows that retrievability of datasets is more imbalanced than publications. The value of c is set to 100 in this plot however similar trends are observed with the lower values of c as well. As discussed in Section 3, the retrievability score of documents escalate with higher values of c ; consequently, the overall retrievability-balance of the collection also changes positively. To

empirically see the variations in terms of imbalance, Gini coefficients (G) attained at different values of c are presented for both the category of documents in Table 3 which is also graphically displayed in Figure 4. From the table, it can be noticed that the fall of G for publications is more than 19% while there is a 10% change in the G value for datasets. This confirms that the inequality of documents being searched for is higher in case of datasets than publications.

Additionally, we report the percentage of total documents (respectively of type, publication and dataset) being retrieved while changing c in Table 3. Note that, more than 95% of documents of type publication are retrieved within the top 10 retrieved documents; in contrast, just above 76% datasets are retrievable within the same rank cut-off. Increasing the value of c , it is noticed that more than 98% documents are retrievable within the top 100 ranked document by all the queries for both publication and dataset. The significant change in the percentage of retrieved documents of type dataset indicates that searching for datasets is more complex than publications; a deeper ranked list traversal might be essential to find a relevant dataset.

Rank cutoff	Gini coefficient		Retrieved			
	Publication	Dataset	Publication	%	Dataset	%
10	0.6230	0.7975	105795	95.02	48934	76.19
20	0.5863	0.7793	109483	98.33	55528	86.45
30	0.5639	0.7661	110413	99.17	58429	90.97
40	0.5481	0.7562	110796	99.51	60025	93.45
50	0.5359	0.7479	110987	99.68	60991	94.96
100	0.5008	0.7159	111269	99.93	62966	98.03

Table 3: Change in Gini coefficient when the rank cut off is increased. Also the number and percentage of documents retrieved of type Publication and Dataset is presented.

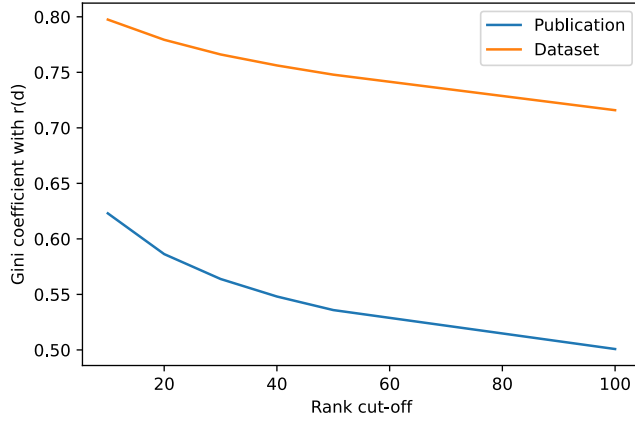


Figure 4: The change in Gini coefficient when the rank cut-off is varied in the range from 10 to 100. The blue line indicates the publication while dataset is specified by the orange curve.

6 FROM RETRIEVABILITY TO USEFULNESS

Usefulness was introduced in [13] designed initially as a criterion for the evaluation of interactive search systems. The *usefulness* of a document can be defined as how often the document is retrieved and *exported* (see Section 5.1) by the end user. Of course the concept of usefulness can only reliably be recognized by relevance judgements submitted by the user for a given query and the relevance of a document may also depend on the perspective of the user which may vary across users. Without an explicit relevance judgement, the approximation of usefulness or importance of documents can not be reliably accomplished. Considering the availability of this information from the query log, we can define the usefulness of a document ($u(d)$) by the following equation:

$$u(d) = \sum_{q \in Q} w_q \cdot g(d, q) \quad (4)$$

In Equation 4, the weight of the query (w_q) can be defined in a similar way as defined in retrievability (Equation 1). The usefulness of a document may also depend on the *difficulty* of the query [11, 12]⁹. A document d should be considered more useful if

⁹A query can be considered as *difficult* if the top ranked documents are mostly non-relevant in which scenario, the user has to go deep down the ranked list to get the document addressing the query [11].

it is retrieved and consumed following a query Q than any other document, say d' with an associated query Q' which is relatively easier than Q (i.e. $\text{difficulty}(Q) > \text{difficulty}(Q')$). Hence, we extend the definition of the weight of the query taking into account a difficulty factor in Equation 5.

$$w'_q = w_q * h(q) \quad (5)$$

where the function $h(q)$ represents the difficulty of the query q . The function $g(\cdot)$ in Equation 4 indicates usefulness in terms of relevance of the document d for the query q . Mathematically, $g(\cdot)$ can be defined as follows:

$$g(d, q) = \text{rel}(d, q) \quad (6)$$

The function $\text{rel}(d, q)$ in Equation 6 indicates the relevance of d for the query q . It will work similar to the way function $f(k(d, q), c)$ is defined in Equation 2 if a binary relevance, (that is d can be either relevant - $\text{rel}(d, q) = 1$, or non-relevant - $\text{rel}(d, q) = 0$ to the query q) is considered.

The usefulness of a document cannot be realised until it is retrieved by some retrieval model even though it is relevant. Given a pair of documents d_1 and d_2 having the same relevance score, a document, say d_1 , should be considered more useful than the other (d_2) if d_1 is ranked higher than d_2 in the ranked list produced by a retrieval model. Hence, the relevance factor of Equation 6 can be extended considering the rank of the document as in the following Equation:

$$g(d, q) = \frac{1}{k(d, q)} * \text{rel}(d, q) \quad (7)$$

where the function $k(\cdot)$ indicates the rank of document d after performing a retrieval with query q . Based on Equation 7, we can assert that the usefulness and the rank of the document are inversely proportional to each other; the chance of being consumed by the end user will decrease if the document d is being retrieved at lower rank. Note that, the function $g(\cdot)$ in Equation 7 is dependent on the retrieval model that is used to retrieve d for q ; the rank-order of a pair of documents in two ranked lists can be different if the corresponding retrievals are performed using different retrieval models.

6.1 Experimentation

As presented and argued earlier in Section 6, the signal of document consumption by the user after submitting a query is essential in order to compute the usefulness of the documents. We utilize the information stored in the interaction-log of the integrated search

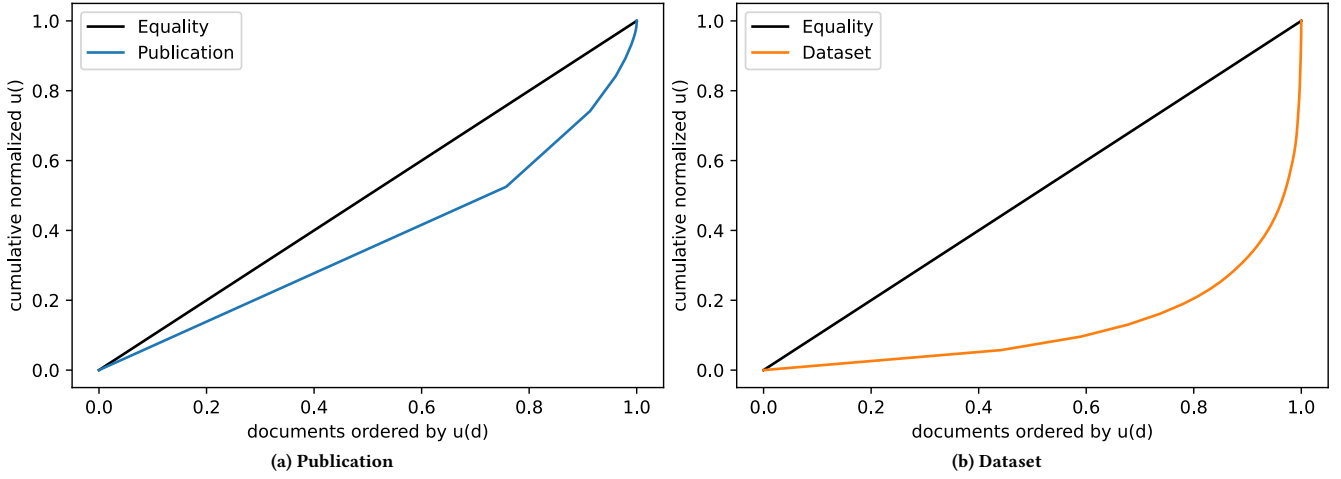


Figure 5: The Lorenz curve with usefulness. The straight line going through the origin (in black) indicate the *equality*, that is, when all the documents are equally useful. The blue line specifies (Figure 5a) the publication while dataset is indicated by the orange curve (Figure 5b).

system as the indication of document consumption by the user. Particularly, the usefulness is determined on the basis of implicit relevance feedback from the export interactions (see Section 5.1). A rank cut-off value (c) of 10 is considered; that is, the top 10 documents of the SERP are considered as viewed/consumed. Utilizing the difficulty of the query in order to compute the usefulness has been left as part of future work; the difficulty of the query is kept as constant ($h(q)$ in Equation 5 set to 1) in this study.

6.2 Observation and analysis

The experimental results on the integrated system are graphically presented in Figure 5 where a pair of Lorenz curves are displayed with the usefulness of the documents of type publication and dataset. In Figure 5a, the usefulness distribution of publications is presented indicating that publications are close to being equally distributed. In comparison, the similar distribution of datasets (presented in Figure 5b) is observed to be more skewed with an evident inclination towards certain items being more useful. The corresponding Gini coefficient of the two distributions (for publications and datasets) is presented in Table 4. The value of G for the usefulness of dataset distribution is seen to be almost three times greater than the publication. This observation clearly highlights that a few datasets are more useful than the rest, whereas the usefulness distribution of the publications is considerably close to being uniform.

	Publication	Dataset
Gini coefficient	0.2594	0.7466

Table 4: The Gini coefficient computed with the distribution of usefulness of the publication and dataset. A higher Gini coefficient (upper bound 1.0) indicates an uneven distribution of usefulness.

7 CONCLUSION

In this paper, we have studied the retrievability of documents from different types belonging to an integrated, real-life digital library, and have observed noticeable differences between documents from the publication and dataset categories. In response to **RQ1**, we have seen significant popularity-bias of certain items being retrieved than the others. Particularly it has been shown that certain items from the dataset category are more likely to be retrieved than the other items of the same category. In contrast, the retrievability scores of documents from the publication type are more evenly distributed. For the **RQ2**, the intra-document selection bias is formalized using the common measures of Lorenz curve and Gini coefficient. In response to **RQ3**, we have observed that the distribution of document retrievability is more diverse for the dataset as compared to publications. This can be again attributed to the popularity bias of certain items in the dataset category.

Additionally, we have proposed a measurement of *usefulness* of documents based on the signal of document consumption by the users after submitting a query to the system. From the experimentation, we have noticed that the usefulness of documents from the publication category, in general, is substantially greater than the usefulness of documents from the datasets. Comparing the intra-category usefulness, a vast diversity is reported for datasets which signifies that certain documents are considerably more *trendy* in terms of being retrieved and satisfying the information need in case of searching in datasets than publications.

Acknowledgement

This work was funded by DFG under grant MA 3964/10-1, the Establishing Contextual Dataset Retrieval - transferring concepts from document to dataset retrieval (ConDATA) project, <http://bit.ly/Condata>.

REFERENCES

- [1] Sibel Adali and Ross Emery. 1995. A Uniform Framework for Integrating Knowledge in Heterogeneous Knowledge Systems. In *Proceedings of the Eleventh International Conference on Data Engineering, March 6–10, 1995, Taipei, Taiwan*. IEEE Computer Society, 513–520. <https://doi.org/10.1109/ICDE.1995.380362>
- [2] Jaime Arguello. 2012. Federated Search in Heterogeneous Environments. 46, 1 (may 2012), 78–79. <https://doi.org/10.1145/2215676.2215686>
- [3] Leif Azzopardi and Vishwa Vinay. 2008. Retrievability: an evaluation measure for higher order information access tasks. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management, CIKM 2008, Napa Valley, California, USA, October 26–30, 2008*, James G. Shanahan, Sihem Amer-Yahia, Ioana Manolescu, Yi Zhang, David A. Evans, Aleksander Kolcz, Key-Sun Choi, and Abdur Chowdhury (Eds.). ACM, 561–570. <https://doi.org/10.1145/1458082.1458157>
- [4] Shariq Bashir and Andreas Rauber. 2009. Analyzing document retrievability in patent retrieval settings. In *International Conference on Database and Expert Systems Applications*. Springer, 753–760.
- [5] Shariq Bashir and Andreas Rauber. 2009. Improving Retrievability of Patents with Cluster-Based Pseudo-Relevance Feedback Documents Selection. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management (Hong Kong, China) (CIKM '09)*. Association for Computing Machinery, New York, NY, USA, 1863–1866. <https://doi.org/10.1145/1645953.1646250>
- [6] Shariq Bashir and Andreas Rauber. 2011. On the relationship between query characteristics and IR functions retrieval bias. *Journal of the American Society for Information Science and Technology* 62, 8 (2011), 1515–1532.
- [7] Jamie Callan and Margaret Connell. 2001. Query-based sampling of text databases. *ACM Transactions on Information Systems (TOIS)* 19, 2 (2001), 97–130.
- [8] Jamie Callan and Margaret Connell. 2001. Query-Based Sampling of Text Databases. *ACM Trans. Inf. Syst.* 19, 2 (apr 2001), 97–130. <https://doi.org/10.1145/382979.383040>
- [9] Zeljko Carevic, Dwaipayan Roy, and Philipp Mayr. 2020. Characteristics of Dataset Retrieval Sessions: Experiences from a Real-life Digital Library. In *International Conference on Theory and Practice of Digital Libraries*. Springer, 185–193. https://doi.org/10.1007/978-3-030-54956-5_14
- [10] Zeljko Carevic, Sascha Schüller, Philipp Mayr, and Norbert Fuhr. 2018. Contextualised browsing in a digital library's living lab. In *Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries*. 89–98. <https://doi.org/10.1145/3197026.3197054>
- [11] David Carmel and Elad Yom-Tov. 2010. *Estimating the Query Difficulty for Information Retrieval*. Morgan & Claypool Publishers. <https://doi.org/10.2200/S00235ED1V01Y201004ICR015>
- [12] David Carmel, Elad Yom-Tov, Adam Darlow, and Dan Pelleg. 2006. What Makes a Query Difficult?. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (Seattle, Washington, USA) (SIGIR '06)*. Association for Computing Machinery, New York, NY, USA, 390–397. <https://doi.org/10.1145/1148170.1148238>
- [13] Michael Cole, Jingjing Liu, Nicholas Belkin, Ralf Bierig, Jacek Gwizdzka, C Liu, Jin Zhang, and X Zhang. 2009. Usefulness as the criterion for evaluation of interactive information retrieval. *Proc. HCIR* (2009), 1–4.
- [14] Stephen Cronen-Townsend, Yun Zhou, and W. Bruce Croft. 2002. Predicting query performance. In *SIGIR 2002: Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, August 11–15, 2002, Tampere, Finland*. ACM, 299–306. <https://doi.org/10.1145/564376.564429>
- [15] Tanja Friedrich. 2020. Looking for data: Information seeking behaviour of survey data users. (2020). <https://doi.org/10.18452/22173>
- [16] Joseph L. Gastwirth. 1972. The Estimation of the Lorenz Curve and Gini Index. *The Review of Economics and Statistics* 54, 3 (1972), 306–316. <http://www.jstor.org/stable/1937992>
- [17] Kathleen Gregory, Paul Groth, Helena Cousijn, Andrea Scharnhorst, and Sally Wyatt. 2019. Searching data: a review of observational data retrieval practices in selected disciplines. *Journal of the Association for Information Science and Technology* 70, 5 (2019), 419–432.
- [18] Claudia Hauff, Leif Azzopardi, Djoerd Hiemstra, and Franciska de Jong. 2010. Query Performance Prediction: Evaluation Contrasted with Effectiveness. In *Advances in Information Retrieval, 32nd European Conference on IR Research, ECIR 2010, Milton Keynes, UK, March 28–31, 2010. Proceedings (Lecture Notes in Computer Science, Vol. 5993)*. Springer, 204–216. https://doi.org/10.1007/978-3-642-12275-0_20
- [19] Daniel Hienert, Dagmar Kern, Katarina Boland, Benjamin Zapilko, and Peter Mutschke. 2019. A Digital Library for Research Data and Related Information in the Social Sciences. In *2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*. IEEE, Champaign, IL, USA, 148–157. <https://doi.org/10.1109/JCDL.2019.00030>
- [20] Daniel Hienert and Peter Mutschke. 2016. A usefulness-based approach for measuring the local and global effect of IIR services. In *Proceedings of the 2016 ACM on Conference on Human Information Interaction and Retrieval*. 153–162.
- [21] Emilia Kacprzak, Laura Koesten, Jeni Tennison, and Elena Simperl. 2018. Characterising dataset search queries. In *Companion Proceedings of the The Web Conference 2018*. 1485–1488.
- [22] Emilia Kacprzak, Laura M Koesten, Luis-Daniel Ibáñez, Elena Simperl, and Jeni Tennison. 2017. A query log analysis of dataset search. In *International Conference on Web Engineering*. Springer, 429–436.
- [23] Dagmar Kern and Brigitte Mathiak. 2015. Are there any differences in data set retrieval compared to well-known literature retrieval?. In *International Conference on Theory and Practice of Digital Libraries*. Springer, 197–208.
- [24] Thomas Krämer, Andrea Papenmeier, Zeljko Carevic, Dagmar Kern, and Brigitte Mathiak. 2021. Data-Seeking Behaviour in the Social Sciences. *International Journal on Digital Libraries* 22, 2 (2021), 175–195. <https://doi.org/10.1007/s00799-021-00303-0>
- [25] Sven R. Kunze and Sören Auer. 2013. Dataset Retrieval. In *2013 IEEE Seventh International Conference on Semantic Computing, Irvine, CA, USA, September 16–18, 2013*. IEEE Computer Society, 1–8. <https://doi.org/10.1109/ICSC.2013.12>
- [26] Mounia Lalmas. 2011. Aggregated Search. In *Advanced Topics in Information Retrieval*. The Information Retrieval Series, Vol. 33. Springer, 109–123. https://doi.org/10.1007/978-3-642-20946-8_5
- [27] Hani Khoshdel Nikkhoo. 2011. The Impact of Near-Duplicate Documents on Information Retrieval Evaluation.
- [28] Thaeer Samar, Myriam C Traub, Jacco van Ossensbruggen, Lynda Hardman, and Arjen P de Vries. 2018. Quantifying retrieval bias in Web archive search. *International Journal on Digital Libraries* 19, 1 (2018), 57–75.
- [29] K. Sparck Jones, S. Walker, and S.E. Robertson. 2000. A probabilistic model of information retrieval: development and comparative experiments: Part 1. *Information Processing & Management* 36, 6 (2000), 779–808. [https://doi.org/10.1016/S0306-4573\(00\)00015-7](https://doi.org/10.1016/S0306-4573(00)00015-7)
- [30] Myriam C. Traub, Thaeer Samar, Jacco van Ossensbruggen, Jiyin He, Arjen P. de Vries, and Lynda Hardman. 2016. Querylog-based Assessment of Retrievability Bias in a Large Newspaper Corpus. In *Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries, JCDL 2016, Newark, NJ, USA, June 19 - 23, 2016*. ACM, 7–16. <https://doi.org/10.1145/2910896.2910907>
- [31] Colin Wilkie and Leif Azzopardi. 2014. Best and fairest: An empirical analysis of retrieval system bias. In *European Conference on Information Retrieval*. Springer, 13–25.
- [32] Colin Wilkie and Leif Azzopardi. 2017. Algorithmic bias: do good systems make relevant documents more retrievable?. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. 2375–2378.