Retrievability in an integrated retrieval system: an extended study

Roy, Dwaipayan; Carevic, Zeljko; Mayr, Philipp

Funded by the German Research Foundation (DFG) - Grant MA 3964/10-1, the "Establishing Contextual Dataset Retrieval - transferring concepts from document to dataset retrieval" (ConDATA) project at GESIS

International Journal on Digital Libraries (2024) 25:287–301
https://doi.org/10.1007/s00799-023-00363-4

Retrievability in an integrated retrieval system: an extended study

Dwaipayan Roy[1] · **Zeljko Carevic**[2] · **Philipp Mayr**[2]

Received: 6 December 2022 / Revised: 30 March 2023 / Accepted: 31 March 2023 / Published online: 28 April 2023

Abstract
Retrievability measures the influence a retrieval system has on the access to information in a given collection of items. This measure can help in making an evaluation of the search system based on which insights can be drawn. In this paper, we investigate the retrievability in an integrated search system consisting of items from various categories, particularly focussing on datasets, publications and variables in a real-life digital library. The traditional metrics, that is, the Lorenz curve and Gini coefficient, are employed to visualise the diversity in retrievability scores of the three retrievable document types (specifically datasets, publications, and variables). Our results show a significant popularity bias with certain items being retrieved more often than others. Particularly, it has been shown that certain datasets are more likely to be retrieved than other datasets in the same category. In contrast, the retrievability scores of items from the variable or publication category are more evenly distributed. We have observed that the distribution of document retrievability is more diverse for datasets as compared to publications and variables.

Keywords Retrievability · Dataset retrieval · Interactive IR · Diversity

1 Introduction

In the present era of information, we are generating a colossal amount of data that needs to be handled and processed efficiently for quick look-ups. The expeditious advancement in technologies has made data generation even more complex with a diversified form of information coming from divergent sources. This necessitates the need to have a federated or integrated system [1, 2] that searches and assimilates results from assorted sources. Textual data still remain 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 [24] has become increasingly important in the (interactive) information retrieval and digital library communities. One of the reasons is undoubtedly the enormous number 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 [23], interviews and online questionnaires [15, 22] and transaction log analysis [11, 20].

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 with traditional document retrieval. The measure of retrievability was initially developed to quantify the influence that a retrieval system has on 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.

✉ Dwaipayan Roy
Dwaipayan.Roy@iiserkol.ac.in

✉ Philipp Mayr
Philipp.Mayr@gesis.org

Zeljko Carevic
Zeljko.Carevic@gesis.org

[1] Indian Institute of Science Education and Research, Kolkata 741246, India

[2] GESIS – Leibniz Institute for the Social Sciences, Unter Sachsenhausen 6-8, 50667 Cologne, Germany

As an extension of our prior work [27], we investigate the retrievability of various types of documents in an integrated digital library *GESIS Search* (see Sect. 3), focusing on various types of data, particularly datasets, publications, and variables. The assumption followed here is that in an ideal ranking system[1], 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.

1.1 Research questions

We verify the research questions put forward and discussed by [27] in the updated system with a variety of item types tested with more queries (see Sect. 4). Similar to the previous work, we substantiate the following research questions on the integrated search system *GESIS Search* focusing on an additional type of item: *Variables* together with *Publication* and *Dataset*:

- **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 formalise this type-accessibility bias utilising the concept of document retrievability?
- **RQ3** How diverse are the retrievability score distributions in the different categories of documents in our integrated search system?

Our previous study [27] was designed to take all queries in the query log into account. This had the benefit of being as close to the real search behaviour as possible. At the same time, this design choice introduced a popularity bias caused by reoccurring queries that positively influence the retrievability score of documents in the corresponding result set. Additionally, the popularity bias of queries has been ignored in this work. Thus, contrasting with the previously reported results, we address the following research question:

- **RQ4** In a real-life search system, does popularity bias of queries influence the inequality in any way?

In sum, our contributions are as follows: 1) we utilise the retrievability measure to better understand the diversity of accessing datasets in comparison with publications with real-life queries from a search log; 2) building on retrievability, we propose to employ the measurement of *usefulness*, which represents implicit relevance signals observed for datasets and publications. Our understanding of bias follows the argumentation provided in [36] 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.) [33].

The rest of the paper is organised as follows: We first present background and related work in Sect. 2 together with formally introducing the concept of retrievability. The integrated search system *GESIS Search* along with the motivation of our retrievability study is presented in Sect. 3. Section 4 discusses the empirical results and analysis of the outcome of the experimentation before introducing the novel concept of usefulness in Sect. 5 along with the experimental study of usefulness. We conclude the paper in Sect. 6 highlighting the contributions and findings of the paper with directions to extend the work.

2 Background and related work

Considering a collection of items, the retrievability of items can be defined as how accessible or findable the items are by some searching techniques. In context of document retrieval, the concept was developed and proposed in [3]. Informally, the retrievability of a document in a collection indicates the expectation of selection of the document by some retrieval model within a rank cut-off. Mathematically, the retrievability of any document d in a collection C is defined as:

$$r(d) = \sum_{q \in Q} w_{\mathrm{q}} \cdot f(\mathrm{rank}(d, q, M), c) \tag{1}$$

where

- Q - the set of all queries which are answerable by the collection;
- w_{q} - weight of the query q;
- $\mathrm{rank}(d, q, M)$ - rank of the document d when retrieval is performed with query q using retrieval model M;
- c - the rank cut-off.

The function $f(\mathrm{rank}(d, q, M), c)$ is an *indicator function* that returns either 1 or 0 depending on whether the rank $(\mathrm{rank}(d, q, M))$ of document d is within the rank cut-off c or not. The indicator function can be mathematically defined as the following:

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

[1] In 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.

In Eq. 1, the retrievability of a document is computed based on retrieval performed with all sets of queries Q addressable by the document collection. Considering a sizeable collection of documents, there can be infinitely many distinct queries that can be answered by various documents in the collection. One of the practical approaches to get this set of all queries Q is to use a query log; however, acquiring such a log is not always feasible. In the absence, a query-based sampling method [9] can be applied to randomly populate Q. In [3], the authors considered generating queries with unigrams and bigrams based on the collection frequency of them above a threshold in the collection. This approach may result in an enormous number of queries if a large collection of documents is considered. To keep the experimental setup tractable, one approach here is to truncate the list again based on a certain threshold value (e.g. 2 million as selected by Azzopardi and Vinay). Hence, the construction of Q based on either query log or random sampling of terms from the collection are some practical approximations that we can adapt in order to realise the concept of retrievability of documents in a collection.

The query weight w_q in Eq. 1 may be used for incorporating a bias (such as popularity, importance, etc.) in the retrievability computation. Ignoring these biases, this weight is considered as uniform for all queries in earlier works [3, 5, 7]. The approximated retrievability score ($\hat{r}(d)$) of document d will then be a discrete value x indicating the number of queries for which d is retrieved within top rank c. Certainly, this is a simplifying assumption and the queries submitted to a search system in practice vary vastly in terms of both *popularity* and *difficulty* [13].

The second factor of the per-query component in Eq. 1 is a Boolean function that depends solely on the rank at which document d is retrieved. Increasing the value of the rank cutoff (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 assessed by human judgements.

Retrievability as a measure was proposed in [3] where the authors experiment on two TREC collections with queries generated using a query-based sampling technique [9]. Since then, 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 [5, 7], the authors look at document retrievability measurements and argue that a single retrievability measure has several limitations in terms of interpretability. In [5], 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. The other work [7] analyse 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. Further researches on patent retrieval reported in [6] and [4] identify content-based features that can be used to classify a set of documents based on their retrievability. Experiments on various patent collections show that these features can achieve more than 80% classification accuracy.

A study on the query list generation phase for determining the measure of retrievability is presented in [8]. 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. A topic-centric query generation technique, tested on the Associated Press (AP) document collection, is proposed in [35]. A significant correlation is reported between the traditional estimate of Gini and the estimate produced by this method of topic-centric query. As recognised in [8], 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 indicate a need for improved construction of simulated queries. To see whether there is any correlation between the retrievability bias and performance measurement, in another study, [34] examine the relationship between retrieval bias and ten retrieval performance measures. Experimentation of TREC

ad hoc data demonstrates that the retrievability bias hypothesis tends to hold for most of the performance measurements.

Retrievability of documents indicates the chance of selection by a retrieval model for various queries submitted. However, the selection of a document does not mean that the document is indeed *useful* in addressing the information need generating the query. This can only be realised by using document consumption signals (e.g. in the form of relevance judgements). This concept was first introduced in [14] as a criterion to determine how well a system is able to solve a user's information need. In their work, Cole et al denoted this notion as *usefulness*. In [18], it has been operationalised within a log-based evaluation approach to determine the usefulness of a search term suggestion service. The usefulness has been further operationalised in [10] 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 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 [23], questionnaires and interviews [15, 22], and transaction-log studies [11, 20]. In [22], 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 [23]. 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 [20]. 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 are to determine whether queries issued towards a data portal differ from those collected in a less constrained environment (crowdsourcing).

3 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 framework.[2] The retrieval in such a system requires sophisticated decision-making considering the various modalities in documents in the collection of data.

Following Eq. 1, the retrievability score of documents is dependent on the other documents in the collection[3]: 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 [26]. Another factor that can influence the retrievability score of a document is its popularity; a popular document will be retrieved multiple times by 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 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. The satisfaction of a user can only be realised via a direct feedback from them. In the absence of such explicit information, it is strenuous, if at all possible, to understand whether information need is fulfilled or not. In this article, we are going to present an extended study of the diversity in retrievability scores for different categories of documents in the integrated search system *GESIS Search*[4] [19].

4 Experimental study

As presented in Sect. 3, we use the integrated search system with various categories of documents in this work. In this section, we start by 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.

4.1 Datasets

We conduct our experimentation on the integrated search system *GESIS Search* containing a total of 860K indexed records (as of November 2022) in different categories such as Research Data, Publications, Variables, Instruments, etc. Social science publications that are indexed in *GESIS Search* use and reference survey datasets, containing hundreds or thousands of questions. These questions are using so-called survey variables (variables in the following). From an information retrieval perspective, variables in GESIS Search are information objects like datasets with specific metadata ele-

[2] This is similar to the concepts of aggregated search [25] or federated search [2].

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

[4] https://search.gesis.org.

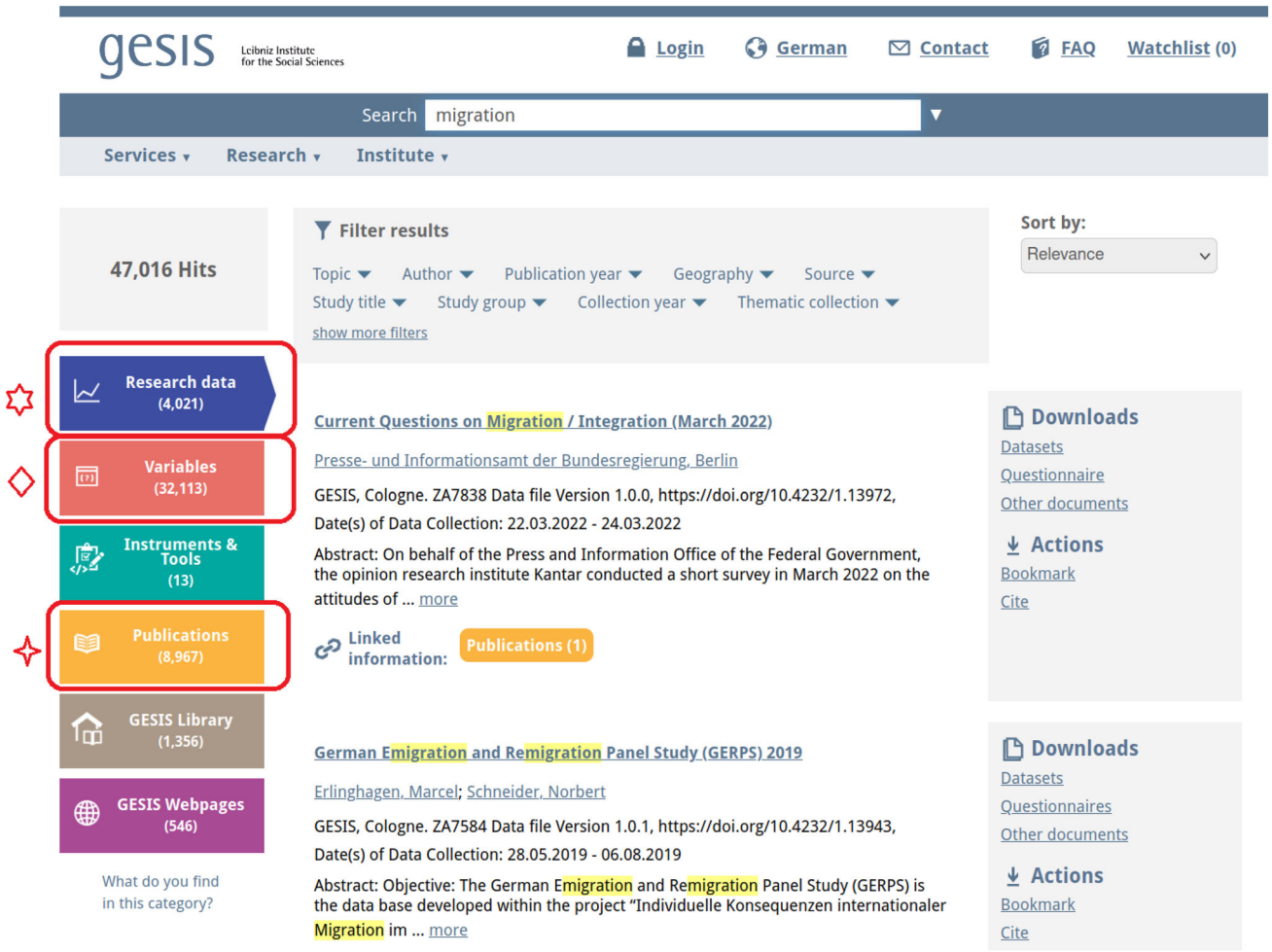


Fig. 1 Screenshot of GESIS Search showing result sets for research data, publications, and variables

ments such as question text, answer categories and frequency tables.

A screenshot showing the interface of GESIS search is presented in Fig. 1. See an example of a variable description in Fig. 2 and the according link to the variable record QD3_1[5] in GESIS Search.[6] The indexed records in GESIS Search are divided into six categories based on their types, covering more than 122K *publications*, 64K *research data* (also referred to as *datasets*), and more than 520K Variables. Given a query, the system returns six search result pages (SERP) corresponding to each of the categories (see Fig. 1). The segregation of the SERP enables us to study the retrievability of the different types. In this study, we specifically focus on the three categories having the largest number of entries, that is, *dataset*, *publications*, and *variables*.

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]. 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 Sect. 4.4 (and later in Sect. 5.2). These real-user queries form the basis of determining the retrievability of documents. This ensures realistic queries in Q of Eq. 1 as opposed to the simulated queries used in [3] or [30]. The data used in this study are an extended version of our previous work [27]; in this log, all the interactions of real users with the search system were recorded for a period of more than five years, specifically between July 2017 and July 2022. The log records more than 2.3 million queries submitted to the integrated search sys-

[5] https://search.gesis.org/variables/exploredata-ZA5876_Varqd3_1.

[6] Further explanation and examples of Social Science variables and its utilisation for information retrieval can be found in [31].

qd3_1 - EU CITIZENSHIP: FEEL TO BE EU CITIZEN

Actions
Bookmark

Item: QD3_1 - You feel you are a citizen of the EU

Study: ZA5876 - Eurobarometer 80.1 (2013)

Question number: QD3

Pre question text: ASK QD ONLY IN EU28 – OTHERS GO TO QE1

Question text: For each of the following statements, please tell me to what extent it corresponds or not to your own opinion.

Interviewer instructions: SHOW CARD WITH SCALE - ONE ANSWER PER LINE - READ OUT

Topics: SOCIETY AND CULTURE | Cultural and national identity | POLITICS | International politics and organisations | ECONOMICS | Economic conditions and indicators | Economic systems and development | MEDIA, COMMUNICATION AND LANGUAGE | Media

Item categories:

Wert	Wertelabel
QD3_1	**You feel you are a citizen of the EU**
QD3_2	You know what your rights are as a citizen of the EU
QD3_3	You would like to know more about your rights as a citizen of the EU

Answer categories:

Value	Value label
1	Yes, definitely
2	Yes, to some extent
3	No, not really

Fig. 2 Screenshot of the variable description of variable QD3_1 in the GESIS Search

tem. Detailed statistics regarding the extracted interactions utilised in our study can be found in Table 1. Together with the previous observations for record type Publication and Dataset, we report the results on another category, the Variables.

Repeated queries can influence the retrievability score of a document. Formally, the set of all queries Q in Eq. 1 may contain the same queries more than once. For synthetically generated queries (used by [3] and [7]), this can be avoided by keeping track of the already generated queries. However, the query log of a real-life search system records all such instances where the same queries are given multiple times by the users. This factor additionally introduces popularity bias into the reproducibility of documents in the form of query popularity. The results and observations reported in our earlier study [27] were based on this type of interaction log. In order to exclusively understand the reproducibility without the query popularity factor, we have only considered unique queries in this work.

4.2 Measuring retrievability in a collection

One way of quantifying the information coverage of a collection is by the count of queries that can be addressed (or answered) by the items in the collection. From the traditional point of view of a web search, the most sought-after way of composing the queries is using free text where vocabulary terms are used to represent an information need. In a moderate-sized document collection, an intractable number of queries formed using a free-text format are possible. Also due to the significant number of documents that can match a free text query, a Boolean matching algorithm is not sufficient; this leads to the development of ranked retrieval that returns an ordered list of items sorted based on their relevance.

Table 1 Statistics of the extracted information belonging to the three selected record types

Record type	Size	# Queries (unique)	Avg. query length	# Exports
Publication	113K	1,028,485 (345,144)	2.6	63,577
Dataset	64K	1,208,108 (268,208)	2.3	142,184
Variables	523K	79,221 (23,909)	2.1	18,832

Considering a traditional document collection C, all the documents are not equally important to a query, hence paving the need to have a ranked retrieval. Now given a set of all possible queries Q, some documents in C will be relevant to more queries (depending on the topical coverage of the document) than others, which can be measured by the concept of retrievability (see Sect. 2). 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

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

Rank	Publication				Research data				Variables			
cut-off	μ	g-μ	σ^2	σ	μ	gμ	σ^2	σ	μ	g-μ	σ^2	σ
10	27.46	7.20	6554.97	80.96	28.16	6.45	12582.64	112.17	2.52	1.77	12.57	3.55
20	37.56	10.49	9983.99	99.92	39.28	9.11	20022.23	141.50	2.77	1.91	15.05	3.88
30	46.13	13.65	12666.31	112.54	48.49	11.63	27404.71	165.54	2.97	2.03	16.98	4.12
40	53.34	16.88	14975.97	122.38	56.3	14.17	33835.99	183.95	3.13	2.12	18.47	4.30
50	59.66	20.15	16821.35	129.70	63.52	16.97	40087.10	200.22	3.25	2.20	19.52	4.42
100	66.80	26.09	17923.59	133.88	90.81	32.88	63517.06	252.03	3.67	2.48	22.68	4.76

the Gini coefficient or Lorenz curve [16], which measures the statistical dispersion in a distribution.[7]

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

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

where N is the size of the population and $v(i)$ specifies the value of ith item in $\mathcal{P}$. The 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 Eq. 3 indicates that all the items in the population are equally probable to be selected, whereas higher values of G specify a bias, implying that only certain items will be selected.

4.3 Experimentation

As explained in Sect. 2, the retrievability of a document is a measurement of how likely the document will be retrieved by *any* query submitted to the system.[8] 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 of 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 formalised and streamlined to consider the actual queries submitted by real users. For the study reported in this article, we utilise the query log presented in Sect. 4.1.

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 Elasticsearch[9] which uses Lucene[10] as the background retrieval model. Following Eq. 1, the retrievability of a document depends on the selection of the rank cut-off 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 reduce the number of documents being considered retrievable because $f(k(d, q), c)$ will be 1 only if $k(d, q) \leq c$ (see Eq. 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 10–100 in steps of 10 and have analysed the observations which are reported in the next section [11].

4.4 Observation and analysis

We start this section with describing different statistical properties of the retrievability distribution of items (from all the three different document types that we experimented with) when the value of c is varied. The mean (μ), geometric mean (g-μ), variance (σ^2), and standard deviation (σ) of the retrievability score distributions on different types (publication, dataset and variable) are given in Table 2. In general, it can be noticed that all the statistical measures for datasets are far more diverse than the other categories. On varying the value of c from 10 to 100, we observe a change of more than 140% and 220% in mean retrievability scores in case of publication and dataset, respectively, while only 45%

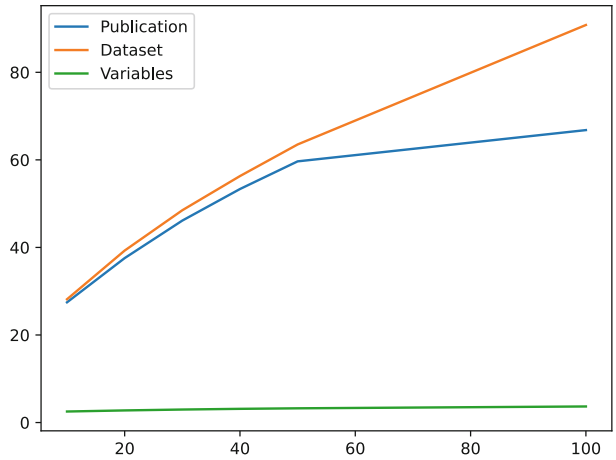
[7] Lorenz curve and Gini coefficient are popular in economics to measure of wealth disparity in a community/country.

[8] By a system, we are referring to the organisation of the collection, along with a retrieval model to be used for retrieval for a given query.

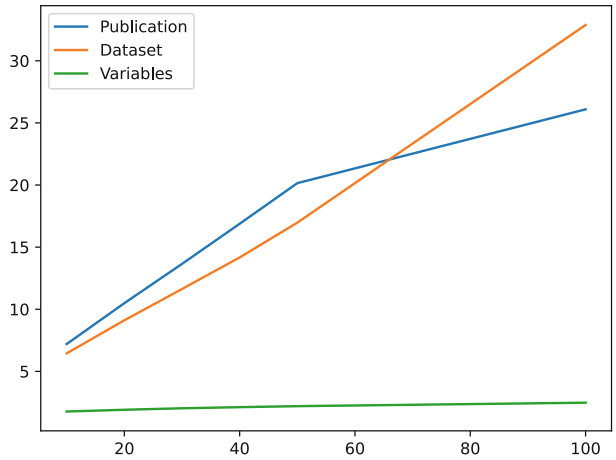
[9] https://www.elastic.co/.

[10] https://www.lucene.apache.org/.

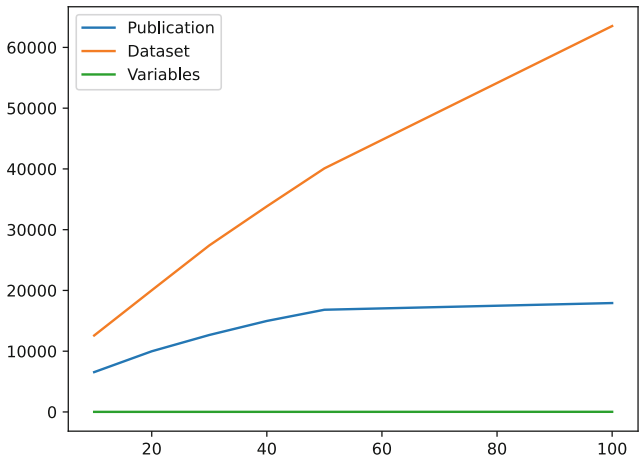
[11] All codes are available here: http://u.pc.cd/vzKctalK.



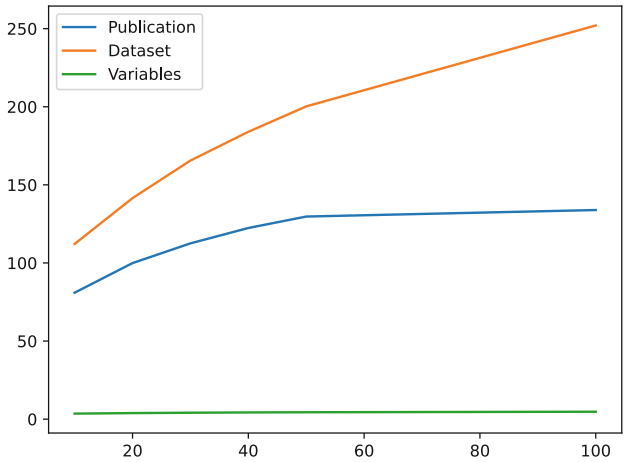
(a) Changes in mean of r(d)



(b) Changes in geometric-mean of r(d)



(c) Changes in variance of r(d)



(d) Changes in standard deviation of r(d)

Fig. 3 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), dataset (in orange), and variables (in green) are presented (colour figure online)

change is noticed in case of variables. In comparison with our earlier work [27], we can see these changes in the retrievability scores are moderate and are not as substantial as seen before. Note that we have excluded repeated queries from the interaction log in this work, which were considered in [27]. This indicates that there is a significant number of repeated queries submitted into the system that had contributed to the momentous change reported earlier resulting in a vast diversity in retrievability scores (see [27], Table 2). Similar trends are recorded for variance and standard deviation as well when computed using the distribution of $r(d)$ on all three categories with different c values. From Table 2, we can conclude that most of the statistical measurements (specifically mean, variance, and standard deviation) are higher for the datasets than publications. In comparison, the geometric mean (**g**-μ in Table 2) is seen to be higher for publications than datasets at the lower rank cut-offs. However, the geometric mean of retrievability of datasets surpasses that of publications at the rank cut-off 100. Combining the observation that can be drawn from geometric-mean values together with the other statistics, we can perceive that for some dataset items, the retrievability values are extensive (popular datasets retrievable by a number of queries); at the same time, there are datasets with poor $r(d)$ values that are rarely retrieved through the submitted queries. The first category of datasets are contributing to the high mean of $r(d)$, which is con-

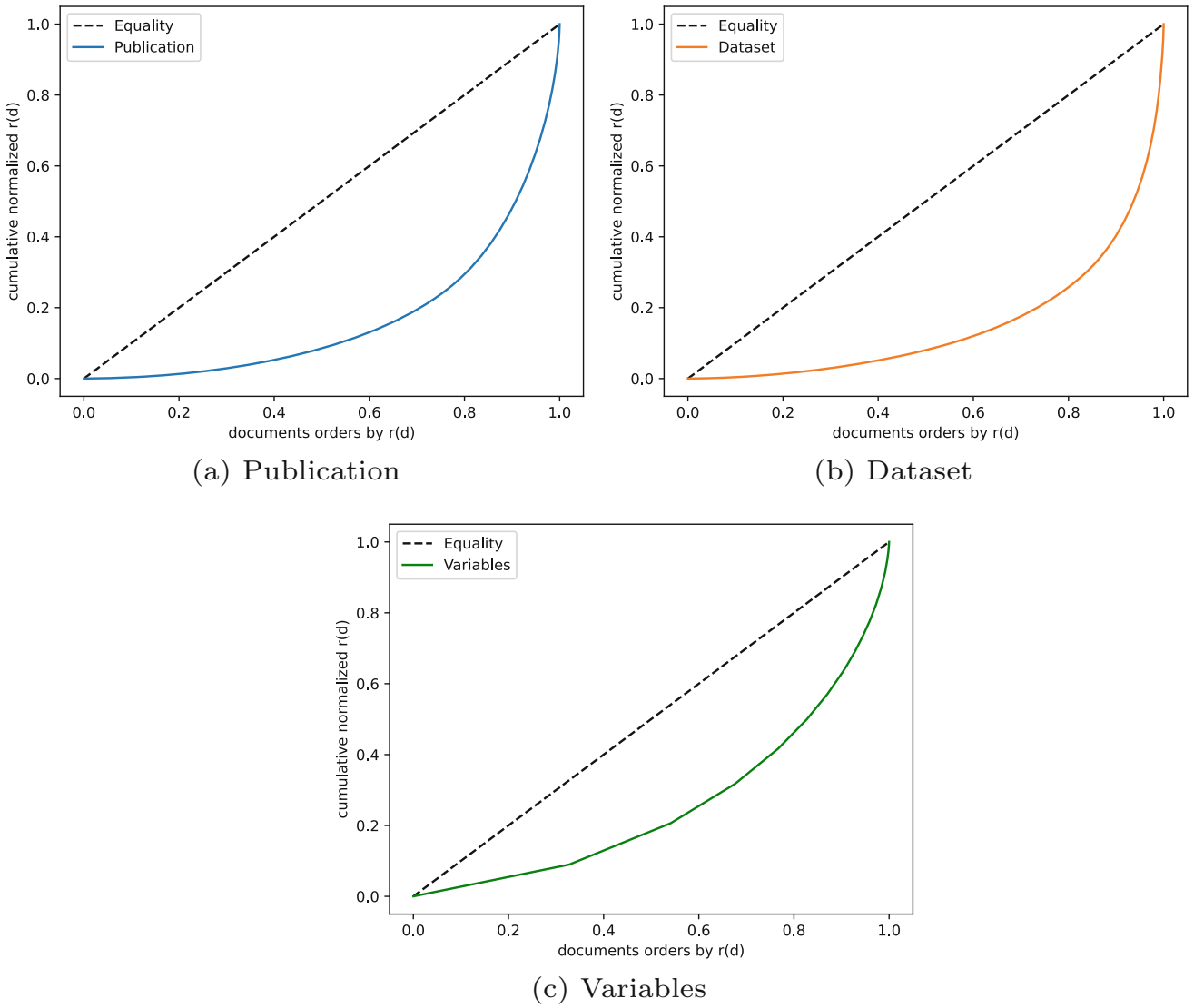


Fig. 4 The Lorenz curve with the retrievability (rank cut-off set to 100). The straight line going through the origin (in black) indicates the equality, that is, when all the documents are equally retrievable (colour figure online)

sistent across different c values, while the datasets of the second category cause the geometric-mean to fall. For the variables, we report all these measures are noticeably smaller than for publications and dataset. The reason behind this is the relatively small number of queries of the variable category compared to the other types; as a result, the variables in general are selected for less number of queries in comparison with other categories. These variations are presented graphically in Fig. 3.

As proposed in [3] and used in our earlier work [27], we utilise the Gini coefficient (G) to quantify the variation in retrievability scores, and Lorenz curve to graphically represent the disparity in retrievability among the items in different categories. Figure 4 plots the Lorenz curve with the $r(d)$ scores computed separately for publications, datasets and variables. To consider the highest coverage, we set the rank cut-off c to 100 while plotting the $r(d)$ values.[12] From Fig. 4, it is seen that retrievability of datasets (presented in Fig. 4b) is more imbalanced than the other two types with Gini coefficient 0.7000. Also, variables are seen to be the closest to the equality (in Fig. 4b) attaining a Gini coefficient of 0.4806.

As discussed in Sect. 2, the retrievability score of documents escalates with higher values of c; consequently, the overall retrievability-balance of the collection also changes positively bringing in the curve close to the equality. To empirically see this variation, Gini coefficients attained at different rank cut-offs are presented in Table 3, which is also graphically displayed in Fig. 5. From the table, it can be noticed that the fall in G for variables (green curve in

[12] Similar trends are observed with c set to lower values.

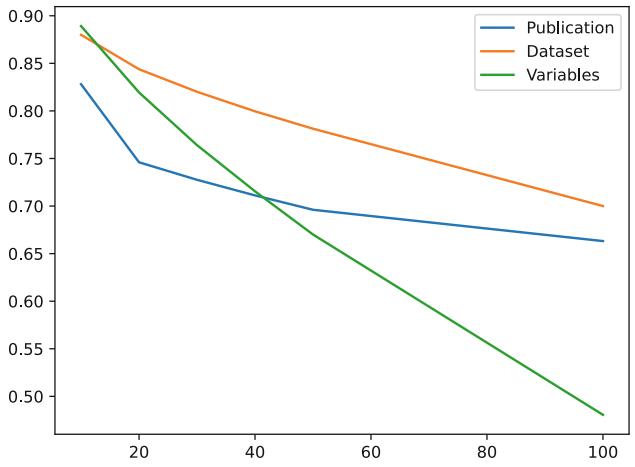


Fig. 5 The change in Gini coefficient when the rank cut-off is varied in the range from 10 to 100. The blue line indicates the publication, while dataset is specified by the orange curve (colour figure online)

Fig. 5) is more than 45%. From a severe unequal distribution with G having 0.8281 till rank 10 (highest among all the categories), the Gini value falls sharply at 0.4806 when the rank cut-off is set to 100. This indicates that more variables are discernible if the ranked list is explored beyond the top position.

Additionally, we report the percentage of total items retrieved while changing c in Table 3. Note that more than 92% of publication are retrieved within the top 10 positions, while only 58% and 10% items, respectively, from the category dataset and variables are retrieved 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 documents 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. Note that only half of the items from variables category (specifically 50.43%) are retrieved within the top 100 positions although the Gini value indicates more balance in retrievability ($G = 0.4806$). This leads to an interesting observation: as reported in Table 2, the average retrievability scores for variables are significantly smaller ($r(d) = 3.67$ at cut-off 100), the difference in not being retrieved (having $r(d) = 0$) and retrieved with average retrievability score is merely a small value. Due to this seemingly inconsequential difference in $r(d)$ score, the Gini is not affected significantly. However, these variables, which are not retrieved at all, lower the percentage of retrieved items.

4.5 Comparing influence of query popularity bias

Considering a real-life query log, there is an obvious possibility of having more than one entry for the popular queries. While computing the retrievability, the items retrieved by those repeated queries get a boost in the retrievability score due to the popularity bias of the queries. To understand the influence of this query popularity bias, in this section, we report relationship between the retrievability scores of the items computed with i) Q_{r} - the interaction log containing *repeated* queries, and ii) Q_{u} - the query log with only the *unique* queries.[13] Particularly, we report how disjoint the documents with the highest retrievability scores are when the retrievabilities are computed with the two types of queries separately. If the documents are ordered by their retrievability scores, we get two individual ranked lists of documents each when Q_r and Q_u are employed. In order to compare and contrast the lists produced by the two types of query lists, we adapt three ways to quantify the difference:

- **Set-based** We compute the Jaccard's coefficient between the two lists ranked by their retrievability scores till different rank cut-offs. Particularly, the first 1K, 5K, 10K, 20K and 50K top-ranked items are considered and their set-based overlap is computed. The results are reported in Table 4. From the results, we can see that overlap in items having the highest 1K retrievability scores are 10% and 12%, respectively, for the categories publication and dataset. However, around 31% overlap is observed for the variable category among top 1K items. The Jaccard's coefficient changes swiftly for all the categories when higher number of items are considered. This indicates that the diversity between the two types of ranked lists are significant for all the three categories of items.
- **Correlation-based** Further, we compare the two ranked lists in terms of their correlations. Based on the discordant and concordant pairs, we compute the Kendall's τ correlation coefficient. Additionally, the Spearman's rank correlation is also assessed and reported in Table 5 for all three categories. Considering these measures, we note that the rank correlations indicate an imperceptible relation between the two lists for all of the types, while the most diverse results are observed in the case of publication category. For variables, the correlations are noted to be higher as compared to the other types, whereas it is too inconsiderable for the other types.
- **Rank overlap-based** The correlation-based measures suffer from certain limitations such as the lists needing to be conjoint and the measurement does not consider the position where the disagreements are happening; that is, the measure does not discriminate between mismatch at

[13] Note that as the system may evolve with new documents being added into the index, the exact ranked list produced for the same query submitted at two different times may differ. However, we have ignored the evolving nature of the index and have considered the latest snapshot of the index to perform the retrieval.

Table 3 Change in Gini coefficient when the rank cut-off is increased

Rank	Gini coefficient			Retrieved					
cut-off	Publication	Dataset	Variable	Publication	%	Dataset	%	Variable	%
10	0.8281	0.8800	0.8892	110666	92.15	37554	58.31	53799	10.28
20	0.7460	0.8438	0.8194	116322	96.86	46437	72.10	89959	17.20
30	0.7276	0.8201	0.7640	118050	98.30	51160	79.44	118961	22.74
40	0.7112	0.7996	0.7155	118819	98.94	54503	84.63	144393	27.60
50	0.6961	0.7813	0.6701	119259	99.31	56761	88.13	167691	32.06
100	0.6632	0.7000	0.4806	119847	99.80	63735	98.96	263801	50.43

Also the number and percentage of documents retrieved of type Publication Dataset and Variable are presented

Table 4 The Jaccard's coefficient (set-based similarity) between the ranked lists of items obtained with different query sets Q_r and Q_u are reported

Top items	Jaccard's coefficient		
considered	Publication	Dataset	Variable
1000	0.1025	0.1287	0.3199
5000	0.2917	0.2606	0.4319
10000	0.3896	0.3546	0.5473
20000	0.4584	0.4821	0.6353
50000	0.5756	0.8383	0.8897

The first column indicates the number of top retrievable items considered to compute the similarity

top position and at later positions. As an alternative, [32] proposed a ranked-biased overlap measure (RBO) that weights the difference considering the position at which they are occurring. Mathematically, the RBO between two ranked lists S and T is computed as:

$$\mathrm{RBO}(S, T, p) = (1 - p) \sum_{d=1}^{\infty} p^{d-1} \cdot A_d \tag{4}$$

In the Equation, d is the depth of the list, p is a weighting factor (between 0 and 1) and A_d is the common items at depth d divided by the depth d itself.
Following Webber et al, we have set the weight parameter p to 0.9. The RBO-based similarity between the two types of results is reported in Table 5. Again, it is prominent from the results that the dissimilarities between the rank of the items based on their retrievability scores are noteworthy, particularly for the publication and dataset categories.

From the dissimilarities between the two ranked items of all three categories, it can be concluded that the popularity bias of queries affects the retrievability irrespective of the type. Out of the three categories, comparatively the least influence by this bias is observed for items belonging to the variable categories. The retrievability of items from the publication and dataset categories is noted to be the most impacted with less than 13% common items being observed among the top 1K.

5 From retrievability to usefulness

Usefulness was introduced in [14] and 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 Sect. 4.1) by the end user. Of course, the concept of usefulness can only reliably be recognised 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 and different points in time. Without an explicit relevance judgement, the approximation of usefulness of documents cannot be reliably accomplished. Considering the availability of the export and utilisation information from the query log, we can define the usefulness of a document ($u(d)$) by the following equation:

$$u(d) = \sum_{q \in \mathsf{Q}} w_{\mathrm{q}} \cdot g(d, q) \tag{5}$$

In Eq. 5, the weight of the query (w_{q}) can be defined in a similar way as defined in retrievability (Eq. 1). The usefulness of a document may also depend on the *difficulty* of the query [12, 13][14]. A document d should be considered more useful if 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. $difficulty(Q) > difficulty(Q')$). Hence, we extend the definition of the weight of the query taking into account a difficulty factor in Eq. 6.

$$w'_q = w_q * h(q) \tag{6}$$

[14] A 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 [12].

Table 5 The rank correlation-based (Kendall's τ and Spearman's r) and rank-bias-based (RBO) similarities between the ranked lists of items obtained with different query sets Q_r and Q_u are reported

Measure	Correlation coefficient		
	Publication	Dataset	Variable
Kendall's τ	0.0279	0.0789	0.1275
Spearman's r	0.0390	0.1179	0.2267
RBO	0.4594	0.6211	0.7119

where the function $h(q)$ represents the difficulty of the query q. The function $g(\cdot)$ in Eq. 5 indicates usefulness in terms of relevance of the document d for the query q. Mathematically, $g(\cdot)$ can be defined as follows:

$$g(d, q) = \text{rel}(d, q) \tag{7}$$

The function $\text{rel}(d, q)$ in Eq. 7 indicates the relevance of d for the query q. It works, in the same way, $f(k(d, q), c)$ is defined in Eq. 2 considering 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).

Informally speaking, the usefulness of a document can be generally stated as the number of queries for which, it is exported (i.e. consumed) by the user. Considering a SERP without any duplicate documents, the usefulness can be further simplified as the count of exportation of the document.

5.1 Experimentation

As presented and argued earlier in Sect. 5, the signal of document consumption by the user is essential in order to compute the usefulness of documents. We utilise the information stored in the interaction log of the integrated search system *GESIS Search* 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 Sect. 4.1). The difficulty of the query is kept as constant ($h(q)$ in Eq. 6 set to 1) in this study and further study in this regard has been left as part of future work.

5.2 Observation and analysis

The experimental results on usefulness are graphically presented in Fig. 6 where a pair of Lorenz curves are displayed with the usefulness of the documents of type publication, dataset and variable. From Fig. 6c, we can observe that the usefulness distribution of variables is close to being equally distributed as compared to the other types. In comparison, the similar distribution of datasets (presented in Fig. 6b) is observed to be more skewed with an evident inclination towards certain items being more useful. The corresponding Gini coefficient of the distributions is presented in Table 6 where the value of G for the usefulness of dataset distribution is seen to be almost three times greater than the variables. The difference in publications and datasets is also evident. This observation clearly highlights that a few datasets are more useful than the rest, whereas the usefulness distribution of the variables is considerably close to being uniform. In the case of publications, the distributions are also observed to be similar to that of variables which are close to uniformity.

6 Conclusion and future work

In [27], we have reported a significant difference in retrievability of items belonging to various categories in the integrated search system *GESIS Search*. We particularly focused on the types *publications* and *datasets* and concluded that there is a significant difference in the retrievability scores if the item belonged to the category of publication or dataset. As an extension to that work, we have included another category to study the retrievability which is *variables*. Along with that, we have used a newer and larger version of interaction logs for our experimentation. A noticeable difference in the experimental setup from our earlier work is that we have used a deduplicated version of the log. That is, only the unique queries from the interaction log are considered excluding any repeated entries. This ensures bypassing any query popularity bias, which may influence the retrievability of the items.

In this extended study, we observe similar phenomena on the newer data as well as on the variable type. In response to **RQ1**, we have seen a significant popularity-bias with certain items being retrieved more often than others. Particularly, it has been shown that certain items from the dataset category are more likely to be retrieved than the other items in the same category. In contrast, the retrievability scores of items from variable or publication types are more evenly distributed. For the **RQ2**, the intra-document selection bias is formalised 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 datasets as compared to publications. This can be attributed again to the popularity bias of certain items in the dataset category. The earlier study used an interaction log not employing any deduplication of queries; as a result, the items retrieved for those popular queries (occurring frequently in the log) gain a boost in the computed retrievability scores. In this paper, we have further included an explicit discussion and comparison of the retrievability scores of items in different categories when the query popularity bias is factored out by the deduplication of the queries. In this connection, as a response to **RQ4**, we showed that there can be a positive

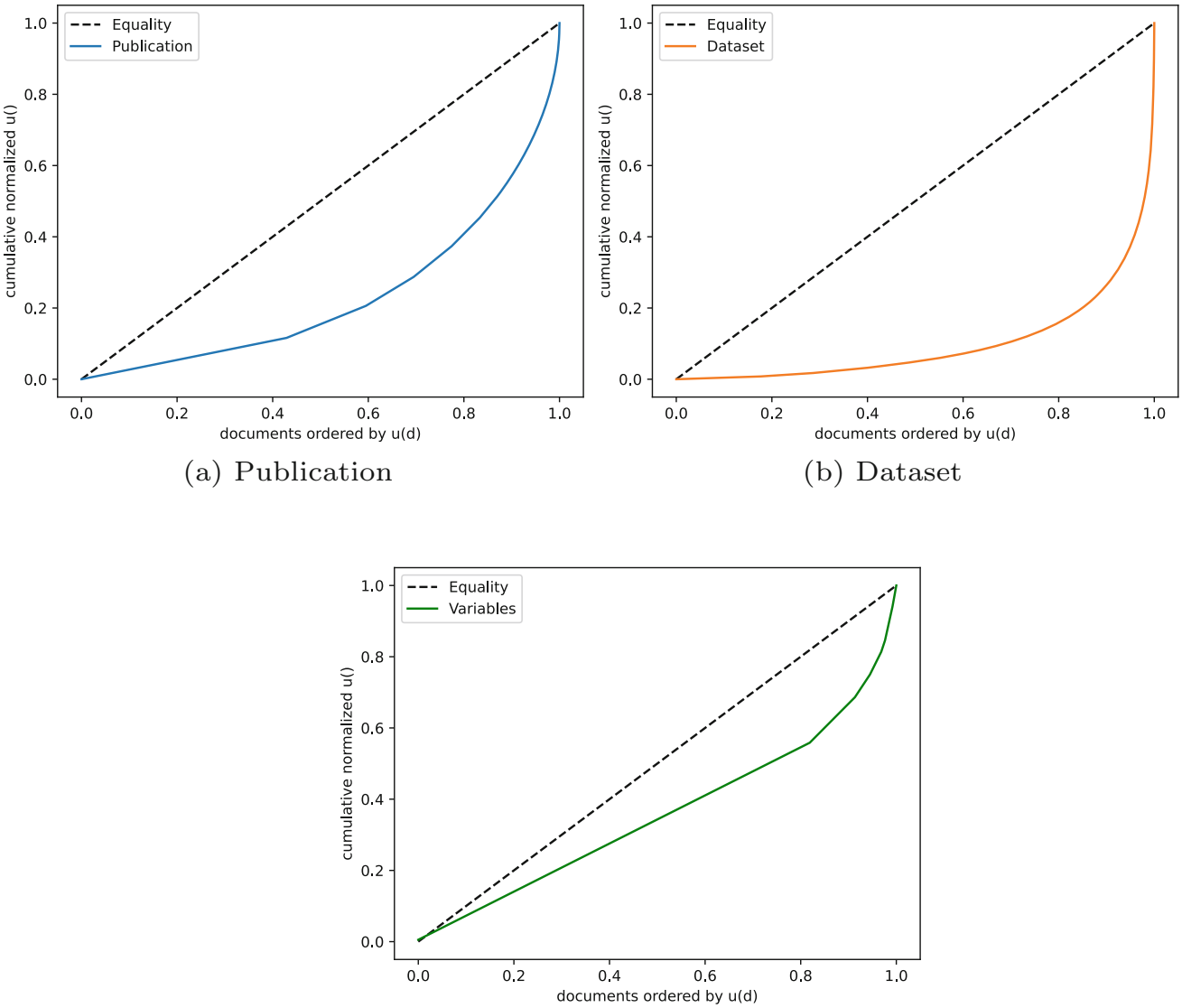


Fig. 6 Plotting Lorenz curves for usefulness values. The straight line going through the origin (in black) indicates the *equality*, that is, when all the documents are equally useful. The blue (**a**) and orange (**b**) curves, respectively, specify the publication and dataset, while variable is indicated by the green curve (**c**) (colour figure online)

Table 6 The Gini coefficient computed with the distribution of usefulness of the publication, dataset and variables

	Publication	Dataset	Variables
Gini coefficient	0.3160	0.8031	0.2876

A higher Gini coefficient (upper bound 1.0) indicates an uneven distribution of usefulness

influence of the query popularity bias on the distribution of the retrievability scores.

Further study on the measurement of usefulness (proposed in our earlier work [27]) reveals a prominent diversity in the nature of consumption of items among the different types. We notice that variables are close to having an equality in usefulness which is significantly disparate in publication and dataset categories. 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. Experimenting with the variables, we observe that the usefulness of items in this category is closer to equality than items in the other categories.

The proposed usefulness metric indicates its popularity in terms of being consumed by the users. Hence, one possible extension of this work will be to test the applicability of usefulness to improve retrieval performance. Incorporating the usefulness of documents as a feature in the learning to rank framework could actually boost the retrieval effectiveness. In terms of presenting the results (SERP) to end users, usefulness can be used as a sorting measure to organise the retrieved items based on popularity. Specifically, together with the provision of presenting the results sorted based on the recency or relevance, it can also be extended to provide an ordering based on how popularly the document is viewed by the users.

Acknowledgements 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 at GESIS. Dwaipayan Roy wants to acknowledge a research grant provided by the GESIS Research Gateway EUROLAB in summer 2022.

Funding Open Access funding enabled and organized by Projekt DEAL.

Declarations

Conflict of interest Philipp Mayr is on the Editorial Board of the "International Journal on Digital Libraries" and guest co-editor of the special issue "JCDL 2022". In this case, the co-editors are handling the review process.

References

1. Adali, S., Emery, R.: A uniform framework for integrating knowledge in heterogeneous knowledge systems. In: Proceedings of the Eleventh International Conference on Data Engineering, Taipei, Taiwan, 6–10 March 1995. IEEE Computer Society, pp. 513–520 (1995). https://doi.org/10.1109/ICDE.1995.380362
2. Arguello, J.: Federated search in heterogeneous environments. SIGIR Forum **46**(1), 78–79 (2012). https://doi.org/10.1145/2215676.2215686
3. Azzopardi, L., Vinay, V.: Retrievability: an evaluation measure for higher order information access tasks. In: Shanahan JG., Amer-Yahia S., Manolescu I., et al. (eds) Proceedings of the 17th ACM Conference on Information and Knowledge Management, CIKM 2008, Napa Valley, California, USA, 26–30 Oct 2008. ACM, pp. 561–570 (2008). https://doi.org/10.1145/1458082.1458157
4. Bache, R., Azzopardi, L.: Improving Access to Large Patent Corpora, pp. 103–121. Springer-Verlag, Berlin, Heidelberg (2010). https://doi.org/10.1007/978-3-642-16175-9_4
5. Bashir, S., Rauber, A.: Analyzing document retrievability in patent retrieval settings. In: International Conference on Database and Expert Systems Applications, pp. 753–760. Springer (2009a). https://doi.org/10.1007/978-3-642-03573-9_63
6. Bashir, S., Rauber, A.: Identification of low/high retrievable patents using content-based features. In: Proceedings of the 2nd International Workshop on Patent Information Retrieval. Association for Computing Machinery, New York, NY, USA, PaIR '09, pp. 9–16 (2009b). https://doi.org/10.1145/1651343.1651346
7. Bashir, S., Rauber, A.: Improving retrievability of patents with cluster-based pseudo-relevance feedback documents selection. In: Proceedings of the 18th ACM Conference on Information and Knowledge Management. Association for Computing Machinery, New York, NY, USA, CIKM '09, pp. 1863–1866 (2009c). https://doi.org/10.1145/1645953.1646250
8. Bashir, S., Rauber, A.: On the relationship between query characteristics and ir functions retrieval bias. J. Am. Soc. Inf. Sci. Technol. **62**(8), 1515–1532 (2011). https://doi.org/10.1002/asi.21549
9. Callan, J., Connell, M.: Query-based sampling of text databases. ACM Trans. Inf. Syst. (TOIS) **19**(2), 97–130 (2001). https://doi.org/10.1145/382979.383040
10. Carevic, Z., Schüller, S., Mayr, P., et al.: Contextualised browsing in a digital library's living lab. In: Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries, pp. 89–98 (2018). https://doi.org/10.1145/3197026.3197054
11. Carevic, Z., Roy, D., Mayr, P.: Characteristics of dataset retrieval sessions: experiences from a real-life digital library. In: International Conference on Theory and Practice of Digital Libraries, pp. 185–193. Springer (2020). https://doi.org/10.1007/978-3-030-54956-5_14
12. Carmel, D., Yom-Tov, E.: Estimating the Query Difficulty for Information Retrieval. Synthesis Lectures on Information Concepts, Retrieval, and Services. Morgan & Claypool Publishers (2010). https://doi.org/10.2200/S00235ED1V01Y201004ICR015
13. Carmel, D., Yom-Tov, E., Darlow, A., et al.: What makes a query difficult? In: Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Association for Computing Machinery, New York, NY, USA, SIGIR '06, pp. 390–397 (2006). https://doi.org/10.1145/1148170.1148238
14. Cole, M., Liu, J., Belkin, N., et al.: Usefulness as the criterion for evaluation of interactive information retrieval. in: Proc HCIR, pp. 1–4 (2009)
15. Friedrich, T.: Looking for data. PhD thesis, Humboldt-Universität zu Berlin, Philosophische Fakultät (2020). https://doi.org/10.18452/22173
16. Gastwirth, J.L.: The estimation of the Lorenz curve and Gini index. Rev. Econ. Stat. **54**(3), 306–316 (1972). (http://www.jstor.org/stable/1937992)
17. Gregory, K., Groth, P., Cousijn, H., et al.: Searching data: a review of observational data retrieval practices in selected disciplines. J. Assoc. Inf. Sci. Technol. **70**(5), 419–432 (2019). https://doi.org/10.1002/asi.24165
18. Hienert, D., Mutschke, P.: 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, CHIIR '16, pp. 153–162 (2016). https://doi.org/10.1145/2854946.2854962
19. Hienert, D., Kern, D., Boland, K., et al.: A digital library for research data and related information in the social sciences. In: 2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL), pp. 148–157. IEEE, Champaign, IL, USA (2019). https://doi.org/10.1109/JCDL.2019.00030

20. Kacprzak, E., Koesten, L.M., Ibáñez, L.D., et al.: A query log analysis of dataset search. In: International Conference on Web Engineering, pp. 429–436. Springer (2017). https://doi.org/10.1007/978-3-319-60131-1_29
21. Kacprzak, E., Koesten, L., Tennison, J., et al.: Characterising dataset search queries. In: Companion Proceedings of the The Web Conference 2018. International World Wide Web Conferences Steering Committee, WWW '18, pp. 1485–1488 (2018). https://doi.org/10.1145/3184558.3191597
22. Kern, D., Mathiak, B.: Are there any differences in data set retrieval compared to well-known literature retrieval? In: International Conference on Theory and Practice of Digital Libraries, pp. 197–208. Springer (2015). https://doi.org/10.1007/978-3-319-24592-8_15
23. Krämer, T., Papenmeier, A., Carevic, Z., et al.: Data-seeking behaviour in the social sciences. Int. J. Digit. Libr. **22**(2), 175–195 (2021). https://doi.org/10.1007/s00799-021-00303-0
24. Kunze, S.R., Auer, S.: Dataset retrieval. In: 2013 IEEE Seventh International Conference on Semantic Computing, Irvine, CA, USA, 16–18 Sep 2013. IEEE Computer Society, pp. 1–8 (2013). https://doi.org/10.1109/ICSC.2013.12
25. Lalmas, M.: Aggregated search. In: Advanced Topics in Information Retrieval, The Information Retrieval Series, vol. 33, pp. 109–123. Springer (2011). https://doi.org/10.1007/978-3-642-20946-8_5
26. Nikkhoo, H.K.: The impact of near-duplicate documents on information retrieval evaluation. In: Masters thesis. University of Waterloo (2011). http://hdl.handle.net/10012/5750
27. Roy, D., Carevic, Z., Mayr, P.: Studying retrievability of publications and datasets in an integrated retrieval system. In: JCDL '22: The ACM/IEEE Joint Conference on Digital Libraries in 2022, Cologne, Germany, 20– 24 June 2022. ACM, p. 8 (2022). https://doi.org/10.1145/3529372.3530931
28. Samar, T., Traub, M.C., Ossenbruggen, J., et al.: Quantifying retrieval bias in web archive search. Int. J. Digit. Libr. **19**(1), 57–75 (2018). https://doi.org/10.1007/s00799-017-0215-9
29. Sparck Jones, K., Walker, S., Robertson, S.: A probabilistic model of information retrieval: development and comparative experiments: part 1. Inf. Process. Manag. **36**(6), 779–808 (2000). https://doi.org/10.1016/S0306-4573(00)00015-7
30. Traub, M.C., Samar, T., van Ossenbruggen, J., et al.: 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, 19–23 June 2016. ACM, pp. 7–16 (2016). https://doi.org/10.1145/2910896.2910907
31. Tsereteli, T., Kartal, Y.S., Ponzetto, S.P., et al.: Overview of the SVident 2022 shared task on survey variable identification in social science publications. In: Proceedings of the Third Workshop on Scholarly Document Processing. Association for Computational Linguistics, Gyeongju, Republic of Korea, pp. 229–246 (2022). https://aclanthology.org/2022.sdp-1.29
32. Webber, W., Moffat, A., Zobel, J.: A similarity measure for indefinite rankings. ACM Trans. Inf. Syst. (2010). https://doi.org/10.1145/1852102.1852106
33. Wilkie, C., Azzopardi, L.: Best and fairest: an empirical analysis of retrieval system bias. In: Proceedings of the 36th European Conference on IR Research on Advances in Information Retrieval, vol. 8416, pp. 13–25. Springer-Verlag, Berlin, Heidelberg, ECIR 2014 (2014a). https://doi.org/10.1007/978-3-319-06028-6_2
34. Wilkie, C., Azzopardi, L.: 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. Association for Computing Machinery, New York, NY, USA, CIKM '14, pp. 81–90 (2014b). https://doi.org/10.1145/2661829.2661948
35. Wilkie, C., Azzopardi, L.: A topical approach to retrievability bias estimation. In: Proceedings of the 2016 ACM International Conference on the Theory of Information Retrieval. Association for Computing Machinery, New York, NY, USA, ICTIR '16, pp. 119–122 (2016). https://doi.org/10.1145/2970398.2970437
36. Wilkie, C., Azzopardi, L.: Algorithmic bias: do good systems make relevant documents more retrievable? In: Proceedings of the 2017 ACM on Conference on Information and Knowledge Management. Association for Computing Machinery, New York, NY, USA, CIKM '17, pp. 2375–2378 (2017). https://doi.org/10.1145/3132847.3133135