Querylog-based Assessment of Retrievability Bias in a Large Newspaper Corpus

Myriam C. Traub
Centrum Wiskunde & Informatica

Thaer Samar
Centrum Wiskunde & Informatica

Jacco van Ossenbruggen
Centrum Wiskunde & Informatica

Jiyin He
Centrum Wiskunde & Informatica

Arjen de Vries
Radboud University

Lynda Hardman
Centrum Wiskunde & Informatica
Utrecht University

ABSTRACT

Bias in the retrieval of documents can directly influence the information access of a digital library. In the worst case, systematic favoritism for a certain type of document can render other parts of the collection invisible to users. This potential bias can be evaluated by measuring the *retrievability* for all documents in a collection. Previous evaluations have been performed on TREC collections using simulated query sets. The question remains, however, how representative this approach is of more realistic settings. To address this question, we investigate the effectiveness of the retrievability measure using a large digitized newspaper corpus, featuring two characteristics that distinguishes our experiments from previous studies: (1) compared to TREC collections, our collection contains noise originating from OCR processing, historical spelling and use of language; and (2) instead of simulated queries, the collection comes with real user query logs including click data.

First, we assess the retrievability bias imposed on the newspaper collection by different IR models. We assess the retrievability measure and confirm its ability to capture the retrievability bias in our setup. Second, we show how simulated queries differ from real user queries regarding term frequency and prevalence of named entities, and how this affects the retrievability results.

CCS Concepts

•Information systems → Evaluation of retrieval results;

Keywords

Retrievability Bias, User Query Logs, Digital Library, Digital Humanities

JCDL '16, June 19 - 23, 2016, Newark, NJ, USA

ISBN 978-1-4503-4229-2/16/06…$15.00

DOI: http://dx.doi.org/10.1145/2910896.2910907

1. INTRODUCTION

For many digital libraries and archives, users are limited to the retrieval system offered by the data custodian. It is important for users that all relevant documents are equally likely to be retrieved, i.e. that retrieved results are not biased by hidden technological artefacts. If, however, the bias in the search technology impacts the findings of research tasks in a way that it renders relevant documents inaccessible or over-represents specific types of documents, this can lead to a skewed perception of the archive's contents. It is therefore important to provide data custodians and users with a measure to quantify the degree to which the retrieval system provides a neutral way of giving access to a document collection.

In the domain of Information Retrieval (IR), Azzopardi et al. introduced a way to measure how retrieval systems influence the accessibility of documents in a collection [1]. The *retrievability score* of a document d, $r(d)$, measures how *accessible* a document is. It is determined by several factors, including the matching function of the retrieval system and the number of documents a user is willing to evaluate. The retrievability score is the result of a cumulative scoring function, defined as:

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

where c defines the number of documents a user is willing to examine in a ranked list. The coefficient o_q weights the importance of a query. The function $f(k_{dq}, c)$ is a generalized utility/cost function, where k_{dq} is the rank of d in the result list for q. f is defined to return a value of 1 if the document is successfully retrieved below rank c, and 0 otherwise. In summary, $r(d)$ counts for how many queries $q \in Q$ a document d is retrieved at a rank lower than a chosen cutoff c.

Using TREC collections and simulated queries, Azzopardi et al. demonstrated the effectiveness of retrievability as a measure for bias, and how retrievability can be used to compare the bias of different retrieval models [1]. We add to their findings by examining the effectiveness of the retrievability measure, and the query simulation procedure in a more realistic setting and we answer the following research questions:

• *RQ1: Is the access to the digitized newspaper collection influenced by a retrievability bias?*

We use the retrievability measure following a similar experimental setup as described in [1] to the digitized historic newspaper archive of the National Library of the Netherlands. This allows us to investigate the retrievability inequality of documents on a digitized – and therefore error-prone – corpus.

• *RQ2: Can we correlate features of a document (such as document length, time of publishing, and type of document) with its retrievability score?*

We investigate whether documents with specific features are particularly susceptible or resistant towards retrievability bias. This allows to better understand the origin of retrievability bias.

• *RQ3: To what extent are retrievability experiments using simulated queries representative of the search behavior of real users of a digital newspaper archive?*

The availability of user logs allows us to compare retrievability patterns of simulated queries to those generated with real user queries. We investigate how the results differ, for example, what types of documents the queries favor most. Finally, we compare the retrieved document sets with the documents viewed by users to explore how well the results match with users' interests.

Our study investigates the applicability of the retrievability concept to a digitized newspaper collection and the representativeness of simulated query sets of user queries.

2. RELATED WORK

The *Gini coefficient* and the *Lorenz curve* were introduced as means to assess and express potential bias in the accessibility of documents in a collection [1]. Both indicators were originally developed to measure and visualize a degree of inequality in societies [7], such as deprivation and satisfaction [14]. A "perfect tyranny", where one "tyrant" owns the entire fortune, is represented by a Gini coefficient of $G = 1$, whereas for the "perfect communist" scenario $G = 0$. Both have been used in several studies to facilitate the comparison of retrievability inequality of different IR models, subsets of the document collection, parameter sets and cutoff values [1, 2, 12, 11]. We follow these examples and use Lorenz curves and Gini coefficients to assess the retrievability inequality in a digitized newspaper archive, but we also show what other indicators could be used to better understand the *source* of the inequality.

Several additional studies investigated different aspects of retrievability. Most of these studies largely followed the approach introduced in [1], as well as its metrics. Subdomains of IR that are very sensitive to recall are legal and patent retrieval. An IR model that performs poorly on a specific patent collection can therefore have a devastating effect on the result of the search task. A study comparing the retrievability of documents in the MAREC[1] collection through different retrieval models [2] adapted the process used in [1] to generate queries to better simulate the search behavior of patent searchers. They included only bi-term queries as it allowed them to use Boolean operators. Our study shows that even more improvements to the query simulation process are necessary.

To facilitate comparisons across corpora, Bache and Azzopardi suggest that the document to query ratio (DQR) should be kept constant [2]. A high DQR, meaning that a relatively small number of queries is applied to a large data set, may lead to an unrealistically high *Gini* coefficient as a large fraction of documents is never retrieved. Low DQR values are very difficult for experiments with large corpora and real queries. None of the studies we found addresses this problem. The main reason for this being that most studies on retrievability make use of TREC document collections [6, 3, 4, 12, 13, 11], or a freely available corpus of patents from the US patent and trademark office [5]. As these data collections are not provided with query logs from real users, the queries for these studies were generated from the terms in the collection, which allows the researchers to create any number of queries to meet a predefined DQR. We show how a high DQR influences the results of a retrievability study with queries based on user logs and suggest compensation strategies.

3. APPROACH

To answer *RQ1*, we explore whether we can identify a retrievability bias with an approach similar to that reported in [1]. We assess the bias by calculating retrievability scores for every document in the collection for three different IR models, two different query sets (real and simulated), and several cutoff values c. For all of these conditions, we calculate the Gini coefficient. Additionally, we visualize the bias in the retrievability results using Lorenz curves.

To verify that the retrievability scores we generated are meaningful, we test in a known-item-search setup, whether documents with a lower $r(d)$ score are actually harder to find than documents with a higher $r(d)$ score. This is achieved by comparing the mean reciprocal ranks (MRR) of target documents of low scoring and high scoring documents for significant differences.

Understanding how specific document features contribute to a potential retrievability bias would allow a data custodian or a user to make a prediction of how likely they would be able to find documents with this feature in a specific retrieval task. We analyze whether features, such as time of publishing, estimated OCR quality or the newspaper title a document originates from, correlates with a higher or lower retrievability of a document (*RQ2*). Furthermore, we investigate the influence of different parameters (specifically stemming, use of Boolean operators and stopwords) on the retrievability of documents.

As queries play an essential role in any retrieval task, we compare how representative simulated queries are for real user queries. We analyze and compare the composition and length of simulated and real queries and how their result sets differ (*RQ3*). To find out which setup best caters to the users' interests, we compare how well the result sets we obtained in our previous experiments overlap with the documents that were actually viewed.

4. EXPERIMENTAL SETUP

We describe the collection of historic newspapers, the query sets and the parameters we used. To obtain comparable results, we followed the experimental setup of [1] as closely as possible, namely to assess the retrievability of documents through a cumulative scoring model. This means that a document score is given for each query for which a docu-

[1]www.ir-facility.org/prototypes/marec

ment ranks above a pre-specified cutoff rank (c). We quantified the extent to which the retrievability scores of different retrieval models vary using Lorenz curves and Gini coefficients. To verify the meaningfulness of the retrievability scores, we measure the effectiveness of queries designed to retrieve previously selected documents. An analysis of document features and their correlation with retrievability scores concludes our exploration of the bias in our document collection. The second part of our research investigates the representativeness of retrievability results by comparing the results with view data from the user logs.

4.1 Data Sets

We used three different data sets. The National Library of the Netherlands[2] (KB) provided us with the data of their entire digitized newspaper archive along with server logs from which we could extract the queries users issued via the library's webinterface, Delpher[3]. Additionally, we generated a set of simulated queries from the body text of the documents.

4.1.1 Historic Newspaper Collection

The newspaper data set made available to us ranges from 1618 to 1995[4] and consists of more than 102 million OCRed newspaper items. This comprises articles, advertisements, official notifications, and the captions of illustrations (see Table 1 for details).

As the archive spans almost four centuries, the newspaper pages vary strongly in visual appearance which is known to influence the performance of OCR software [8, 9]. The very high vocabulary size (see Table 1) indicates that the corpus might contain a high number of OCR errors, which can impact retrieval tasks [10]. The OCR quality has not been evaluated, therefore the actual error rates for the documents in this collection are unknown. An estimation of the quality by the OCR engine, however, is included in the metadata in the form of page confidence values.

From the KB data, we extracted and tokenized the body text of the newspaper items, which excludes the headings and meta data. We removed all stopwords and terms with fewer than three characters and kept only numbers with four digits, as these are likely to represent years and can therefore be used as query terms by users. The large majority of items (98%) are written in Dutch. As a stemmer for Dutch text was not available in the Indri[5] search engine, we created a stemmed version during preprocessing. We used the default Snowball stemmer for Dutch[6].

4.1.2 Real Queries

Under conditions of strict confidentiality, the KB made user logs available to us that were collected between March 2015 and July 2015. In order to protect the privacy of the users, the logs had been anonymized by hashing the IP addresses, which enabled us to trace queries that originated from the same address without identifying the user. Delpher provides an advanced search interface, which allows users to apply boolean operators and facets based on metadata to their search queries. We processed the query logs the same way as the document collection by removing operators and stopwords, and stemming. For the latter, we again used the Snowball stemmer[9] (see Table 2 for details).

Newspaper Collection		**1618 - 1995**
Total Size[7]		102,718,528
Vocabulary Size[8]		353,086,358
Articles	67%	69,237,655
Advertisements	29%	29,591,599
Official Notifications	2%	1,918,375
Captions	2%	1,970,899
User Logs		**March - July 2015**
Log Size (No. HTTP Requests)		107,684,434
No. Queries		4,169,379
No. Unique Queries		1,051,676
No. Unique IPs		162,536
No. Document Views		3,328,090
No. Unique Documents Viewed		2,732,139

Table 1: Data sets used based on the historic newspaper collection from KB.

Query Set	**Composition**	**Size**	**DQR**
Sim. Queries	single term	2,000,000	
	bi-term	2,000,000	
	total	4,000,000	26
Real Queries	no op., no stopw., st.	957,239	107

Table 2: Sizes and document to query ratios (DQR) of the query sets.

4.1.3 Simulated Queries

To be able to compare our results with those reported in [1], we created a simulated query set. For this, we counted the unique terms and bigrams in the preprocessed documents and extracted the top 2 million terms as single term queries and the top 2 million bigrams as bi-term queries (see Table 2). The frequencies for the two query sets ranged from more than 180 million to 5 for the single term queries and from more than 10 million to 20 for the bi-term queries. We did not filter for OCR errors, therefore frequently occurring misspellings can still be found in the simulated queries.

4.1.4 Document Query Ratio

Azzopardi et al. use query sets of which the size are comparable to the size of the corpus [1]. In this setting all documents have a fair chance to be retrieved. As we used real user queries in a very large corpus, it was not possible for us to influence the DQR. Consequently, the DQR values in our experiments vary greatly for the different query sets (see Table 2), as opposed to the study reported in [1], where the DQRs were 0.57 (AQUAINT) and 0.43 (.GOV). This issue has not been addressed in previous studies investigating retrievability of large document collections.

[2] www.kb.nl
[3] www.delpher.nl
[4] A small number of documents from the 20th century is incorrectly dated to 2011 in the metadata.
[5] http://www.lemurproject.org/indri.php
[6] https://lucene.apache.org/core/4_0_0/analyzers-common/org/apache/lucene/analysis/nl/DutchAnalyzer.html
[7] Number of all articles, advertisements, official notifications and captions
[8] Stopwords removed, length of term at least 2 characters
[9] https://pypi.python.org/pypi/PyStemmer/1.3.0

4.2 Setup for Retrievability Analysis

We compute retrievability scores based on three of the retrieval models used in [1]: TFIDF, Language Model using Bayes Smoothing with $\mu = 1{,}000$ (LM1000), and BM25.

Azzopardi et al. chose to report their results for $c = 100$ [1], therefore we also included these values for comparison. Additionally, we report on a cutoff value of $c = 10$ as it best represents the behavior of our users. The default number of results per page the Delpher interface shows is 10 and an analysis of the user logs showed that only a small fraction of users go beyond this. For the results based on the real queries, we also report on $c = 1000$, as this result set was of comparable size to the $c = 100$ results for the simulated queries.

We did not apply the query weights o_q as the by far largest fraction of real queries were issued only once.

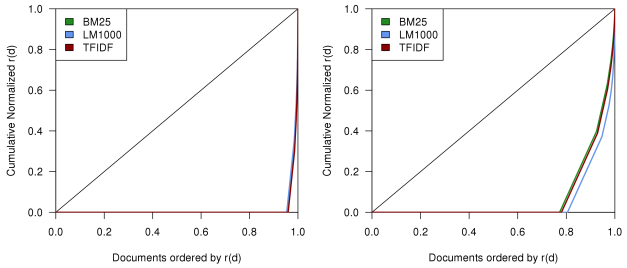
4.3 Setup for Retrievability Validation

We validated the effectiveness of the retrievability scores for the newspaper collection. We examined whether documents with a low retrievability score are harder to retrieve than documents with a high score when a query is *specifically* designed to return the targeted document. We performed one experiment per query set. For simulated queries we follow [1] and use BM25 at $c = 100$ (stemmed, stopwords and operators removed). For the smaller set of real queries, we chose the same parameters but with a cutoff at $c = 1,000$, as the result set is more similar in size to the chosen set for the real queries. We included the documents with $r(d) = 0$, as they represent the group of documents that is supposedly the least accessible one.

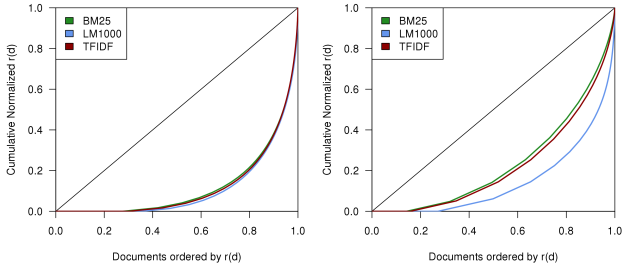
For both result sets we generate queries from the target documents which contain OCR misspellings. In the experiment described in Subsection 4.2 the impact of these misspellings was lowered as a side effect of selecting the most frequent terms in the large corpus. Here, we select terms from a single document, which required us to apply filters as very rare misspellings being part of queries led to very high mean reciprocal rank (MRR) values, but are very unlikely to be used as queries by users. First, we created a dictionary of terms that occurred in more than one document, but in fewer than 25% of all documents and for which the document frequency was *not* equal to collection frequency. This allowed us to exclude extremely rare misspellings that occur in only one document or only once in multiple documents, and very generic terms. The dictionary we created from these terms was used to determine a list of suitable documents. We removed all words from the documents that did not appear in the dictionary or appeared only once in the document. All documents with fewer than four unique words were discarded for the experiment. By applying these filters, we removed 38,026,541 documents from the collection, leaving 64,691,987.

We divided the remaining documents into four bins, the same number of bins as used in [1]. For the division into bins, however, we diverged from the description given in [1] (where documents were ordered by retrievability and then divided into quartiles) because due to a different distribution of $r(d)$ values, the lower scores would have dominated the lower quartiles. Instead of binning on $r(d)$, we used a strategy that is inspired by the distribution of wealth measurements in economics. In our case, wealth is represented as the number of data points per $r(d)$. It is calculated for each $r(d)$ score by multiplying the score with its number of documents. Then we successively merged the $r(d)$ bins, until their summed up wealth reached the threshold of 25% of the total wealth. This led to four bins that roughly correspond to quartiles.

From each bin, we picked a random sample of 1,000 documents. We randomly selected 2 to 3 of the most frequent terms of each document to use as a query, as the mean number of terms issued by users was 2.32. The 1,000 queries we created this way were issued against the collection using the same IR model as before, BM25. We determined the rank of the target documents in the result lists and calculated the MRR for each bin as a measure of its retrieval performance.



$c = 10$, real queries $c = 10$, simulated queries



$c = 1000$, real queries $c = 100$, simulated queries

Figure 1: Lorenz curves visualize the inequality of retrievability scores for the *real queries* (left) and the simulated queries (right) at different cutoff values c.

5. RETRIEVABILITY ASSESSMENT

The high DQR value for our setup suggested that the fraction of documents with $r(d) = 0$ will be relatively high, especially for low cutoff values. Therefore, a large inequality in the retrievability scores was to be expected (*RQ1*). We describe the measured retrievability bias in different result sets and explore how to deal with the non-retrieved documents.

5.1 Assessment of Retrievability Inequality

We first look at the retrievability bias for both query sets at $c = 10$, which is the most realistic representation for the bias users of the archive are confronted with. The *Lorenz* curves depict a high inequality in the retrievability scores (see Fig. 1), with almost identical curves for the TFDIF, BM25 and LM1000 models. This is also reflected in the high *Gini* coefficients ranging from 0.97 to 0.98 for the real and from 0.85 to 0.89 for the simulated queries (see Table 3). The largest part of both curves consists of a flat line, which represents documents that were not retrieved. The setup

	Model	C					
		10		100		1000	
		G	Z	G	Z	G	Z
Real	TFIDF	0.98	96%	0.91	78%	0.77	30%
	BM25	0.97	95%	0.89	75%	0.76	28%
	LM1000	0.97	95%	0.90	77%	0.78	35%
Sim.	TFIDF	0.86	78%	0.55	16%	-	-
	BM25	0.85	77%	0.52	14%	-	-
	LM1000	0.89	80%	0.71	27%	-	-

Table 3: Gini coefficients (G) and fractions of documents with $r(d) = 0$ (Z) for the complete data set.

	Model	C					
		10		100		1000	
		G	Z	G	Z	G	Z
Real	TFIDF	0.71	47%	0.74	36%	0.71	13%
	BM25	0.64	40%	0.69	29%	0.70	10%
	LM1000	0.63	39%	0.71	33%	0.73	20%
Sim.	TFIDF	0.52	26%	0.50	5%	-	-
	BM25	0.48	24%	0.46	3%	-	-
	LM1000	0.63	34%	0.67	18%	-	-

Table 4: Gini coefficients (G) and fractions of documents with $r(d) = 0$ (Z) for the $Union_c$ data set.

with the highest *Gini* coefficient (TFIDF at $c = 10$, real queries) also contains the highest fraction of non-retrieved documents (96%).

By contrast, the *Lorenz* curves for the higher cutoff values depicted in Fig. 1 indicate a more balanced distribution of $r(d)$ values. The curves for all models show a smaller deviation from the equality diagonal and both the *Gini* coefficient, as well as the fractions of documents with $r(d) = 0$, are lower. This suggests that the large number of documents with $r(d) = 0$ has a strong influence on both the shape of the *Lorenz* curve and the *Gini* coefficient. As never-retrieved documents are inevitable in a realistic scenario such as ours, it is important to find a way to address this problem.

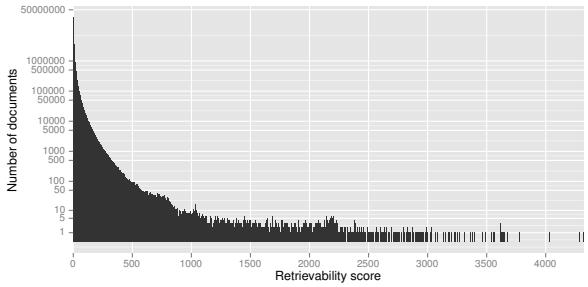
To further explore the influence of the $r(d) = 0$ values, we created a $Union_c$ result set, that contains only documents retrieved by at least one of the models. While this removed most of the documents with $r(d) = 0$, a surprisingly large number of zeros still remained in the subset. The number of zero-scoring documents for TFIDF at $c = 10$, for example, was only reduced from 96% to 47%. Even with never-retrieved documents removed, the inequality in the $Union_c$ data set remains quite high for $c = 10$ with *Gini* coefficients ranging from 0.48 (BM25) to 0.63 (LM1000) (see Table 4). The remaining zero-scoring documents are a first indication that, while their Lorenz curves and Gini coefficients are similar, the models actually retrieve very different sets of documents.

We finally removed *all* documents with $r(d) = 0$ to measure the inequality among the retrieved documents. This caused the *Gini* coefficients to drop to values between 0.40 and 0.46 (real queries at $c = 10$). This again shows the large influence of a high fraction of zeros on the overall *Gini* score.

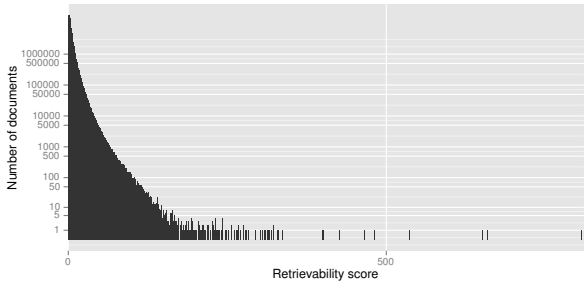
The similarity of the different models' *Lorenz* curves indicates a similar degree of bias in the $r(d)$ scores, but it does not allow insights into the type of bias, i.e. whether it originates from the high DQR, from the users' interest, or from a technological bias towards particular document features.

	Model	C		
		10	100	1000
		G	G	G
Real	TFIDF	0.46	0.59	0.67
	BM25	0.40	0.56	0.67
	LM1000	0.40	0.56	0.66
Sim.	TFIDF	0.35	0.47	-
	BM25	0.32	0.44	-
	LM1000	0.43	0.60	-

Table 5: Gini coefficients (G) for the *Non Zero* data set from which all documents with $r(d) = 0$ were removed.



Real queries, $c = 1000$



Simulated queries, $c = 100$

Figure 2: Log scale representation of the distribution of retrievability scores $r(d)$ for BM25 based on the complete KB dataset.

Fig. 2 shows the frequencies of $r(d)$ values (log scale), with a long tail distribution for both query sets. The maximum $r(d)$ value for the real queries is $r(d) = 4319$, while for the simulated queries this is much smaller (max $r(d) = 807$). This shows one possible cause for the bias towards higher fractions of documents with $r(d) = 0$ within the real queries: they tend to retrieve the same documents more often, leading to a smaller number of unique retrieved documents. This indicates that the query sets themselves may be biased, the real query set towards the users' interest and the simulated query set towards the language use in the document collection.

5.2 Validation of the Retrievability Scores

We validated our results using a known-item-search experiment (see Subsection 4.3) to confirm that documents with low $r(d)$ scores are indeed harder to find.

The results show that the MRR values indeed increase for the bins containing the documents with the higher $r(d)$ values (see Table 6). With one exception the differences in the ranks between the bins proved to be significant in a Kolmogorov-Smirnov test. This suggests that documents in

Query Set		Bin			
		1st	2nd	3rd	4th
Simulated	MRR	0.19	0.28	0.36	0.45
	D	0.20	0.12	0.08	-
Real	MRR	0.17	0.26	0.34	0.38
	D	0.20	0.11	0.05*	-

Table 6: MRR values are higher for items in the quartiles with higher $r(d)$ scores. An * indicates that the Kolmogorov-Smirnov test did not confirm a significant difference ($p > 0.05$) between the indicated bin and the fourth bin. D is the maximum vertical deviation as computed by the KS test.

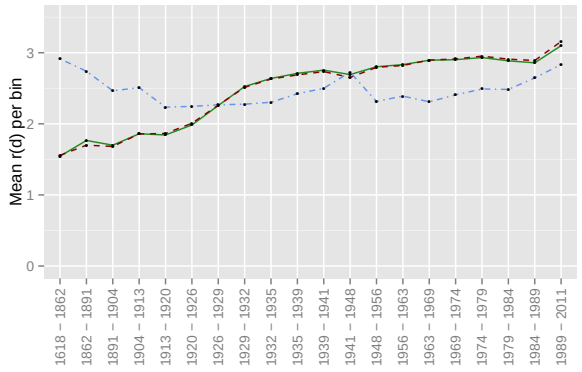


Figure 3: The mean $r(d)$ scores (20 equally sized bins, based on $Union_c$ data, real queries for $c = 100$) for BM25 (green) and TFIDF (red) are nearly identical and double in value over time. LM1000 (blue) does not show this upward trend.

the first bin are significantly more difficult to retrieve than documents in the fourth bin.

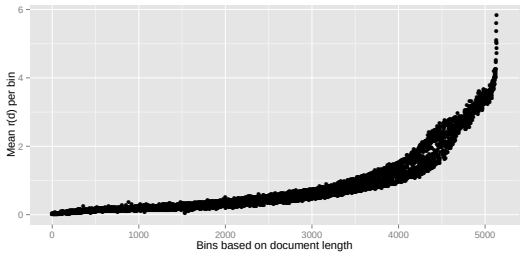
This pattern is similar to the findings in [1] and confirms that a document's retrievability score is a good indication of how hard it is to retrieve the document by a user.

5.3 Document Features' Influence on Retrievability

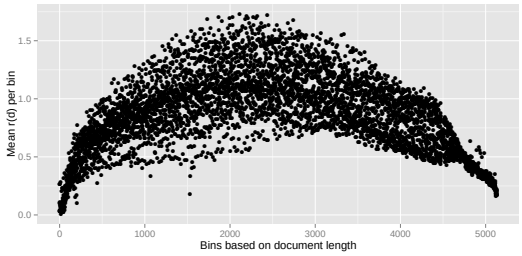
To better understand the inequality in our document collection, we explored whether we can identify subsets within the archive that are particularly susceptible or resistant towards retrievability bias (*RQ2*).

• The *time of publishing* of the newspapers in our collection spans a period of nearly 400 years. Newspapers that belong to the early issues are very different from today's newspapers in terms of content as well as visual appearance. This affects the performance of OCR software, which results in high OCR error rates in older newspapers. We are therefore interested if this is reflected in the $r(d)$ values. For the analysis, we ordered the newspaper items in the $Union_c$ set by publishing date, divided them into 20 equally sized bins (1,7M items per bin) and calculated the mean retrievability score for each bin. Note that due to the much lower number of documents in the early periods of the archive, the 20th century occupies by far the most bins. The results for BM25 and TFIDF show a very small upward trend for later documents (see Fig. 3). This trend is, however, not visible for LM1000 and could also not be confirmed in an analysis of the raw data.

• The *document length* in our collection varies from 33



LM1000



BM25

Figure 4: Document length vs. r(d) for c=100, bins of 20,000 documents

to 381,563 words with a mean length of 362 words. As [1] found that longer documents in their collections were more retrievable than short ones, we were interested in finding out whether the same holds for our collection. We sorted all items in the collection according to their length and divided them into bins of 20,000 documents, leading to 5,135 bins in total. For each bin, we calculated the *mean* $r(d)$. While the pattern we obtained for LM1000 shows an upwards trend for longer documents and thereby confirms this assumption (see Fig. 4), the results for BM25 and TFIDF[10] indicate that documents of medium length are most retrievable, whereas documents at both extremes are less retrievable. We can see a bias in both patterns, while LM1000 clearly favors longer documents, BM25 and TFIDF overcompensate for long documents, while they seem to fail to compensate for short ones.

• The library's OCR engine assigns *confidence scores* to each page (PC), word (WC) and character (CC) in the corpus. This is intended to give an indication of the quality of the OCR processing. From our contacts with the KB we learned that, during the post-processing, the scores were adapted based on the occurrence of a term in a Dutch word list. A formal evaluation of error rates in the KB data has not yet been performed, therefore we do not know to what extent these PC values are realistic. We divided the collection into bins of 20,000 documents based on their PC value and plotted the mean $r(d)$ score for each bin. The resulting plot shows an upward trend for increasing confidence values (see Fig. 5). Documents with an $r(d)$ score very close to 1.0, however, seem to be less retrievable. A closer look revealed that these documents often contain only very short texts, which makes them harder to find.

• *Newspaper titles* do not only vary with respect to their political orientation, but also concerning the content they provide to their readers. The mean number of articles per newspaper title in the archive is 82,638, with a median of

[10]The pattern for TFIDF looks very similar to BM25, therefore we did not include the plot.

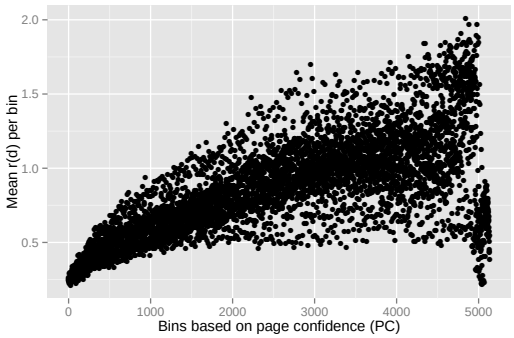


Figure 5: Mean $r(d)$ scores versus page confidence (PC) scores for bins of 20,000 documents

Top 10 Newspaper Titles	Mean r(d)
Rotterdamsch nieuwsblad*	0.05
Algemeen Handelsblad	0.06
De Telegraaf	0.06
Het Vaderland: staat- en letterkundig nieuwsblad	0.07
Leeuwarder courant*	0.07
De Tijd: godsdienstig-staatkundig dagblad	0.08
Het vrije volk: democratisch-socialistisch dagblad	0.10
Limburgsch dagblad*	0.12
Nieuwsblad van het Noorden*	0.14
Leeuwarder courant: hoofdblad van Friesland*	0.15

Table 7: Mean $r(d)$ values for the most prevalent newspaper titles for BM25 at $c = 10$, real queries. An * indicates a regional newpaper title.

127 and a range from one to 16,348,557 documents. We list differences in retrievability scores of the 10 most prevalent newspaper titles in our collection (see Table 7). While the differences seem small, three regional titles have a higher mean $r(d)$ than the seven national titles. Again, this may be caused by a bias in user preferences.

• We computed the mean $r(d)$ scores of the four *types of documents* in the archive for the two query sets. The means resulting from simulated queries show relatively small differences (see Table 8), whereas the mean scores obtained through real queries show a much higher score for official notifications. This again shows the large difference in the document sets retrieved by the two query sets.

From these results we can conclude that the large fraction of never retrieved documents is inevitable in realistic setups and needs to be addressed when assessing retrievability bias. We found evidence for a relation between low OCR confidence values, and short document length and a lower retrievability of documents. When comparing the degree of bias among the three IR models, we found LM1000

	Real	Simulated
Article	0.90	3.89
Advertisement	0.51	3.32
Official notification	4.80	3.22
Caption	0.84	3.06

Table 8: Mean $r(d)$ for different types of articles (BM25, c=100).

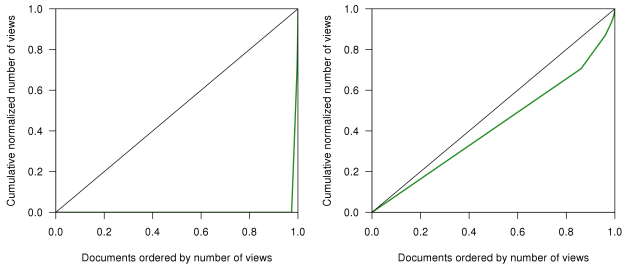


Figure 6: The *Lorenz* curve of viewed documents shows that only a small fraction of the collection was accessed (left) (Gini = 0.98). When non-viewed documents are removed, the inequality largely disappears, because most documents that are viewed, are viewed only once (right) (Gini = 0.16).

to show a greater bias for simulated queries. A comparison of the distributions of retrievability scores indicated a higher variety in $r(d)$ scores for real queries, and a bias towards official notifications for real queries which is not present in the simulated queries.

6. REPRESENTATIVENESS OF THE RETRIEVABILITY EXPERIMENT

We explore to what extent the different types of bias we see in the retrievability experiments are representative for bias in the documents actually viewed on the library's website (*RQ3*). For this purpose we compare the reported results with click data from the user logs, and revisit the use of simulated queries versus real queries.

6.1 Retrieved versus Viewed

The *Lorenz* curve in Fig. 6 (left) shows the inequality in the corpus with respect to the number of views. With only 2.7M out of 102M documents that are viewed, the fraction of documents that is never viewed by users is even larger than the fraction of never retrieved documents in our $c = 10$ experiments. This confirms that a large fraction of not-accessed documents is not only an artifact in our retrievability experiments caused by a relatively small query set: it also reflects the fact that in most large digital libraries, the number of views in any reasonable observation period will be small in comparison to the number of documents in the collection. Since the retrievability and the viewing scores are dominated by the large number of never accessed documents, neither the *Lorenz* curves nor the *Gini* coefficients are very informative measures of bias.

Distribution of r(d) scores and view frequencies.

For documents that are never accessed, it is hard to classify whether this is indeed the result of the small number of user views, the result of bias in user interest, or the result of technical bias in the retrieval system. Focussing only on the accessed documents would ignore the latter type of bias. However, even if we discard the non-accessed documents, the *Lorenz* curve of only the 2.7M viewed documents (see Fig. 6 (right)) is not much more informative. Here we see the opposite: extremely low inequality, which results from the fact that the large majority (86%) of the viewed documents is only viewed once.

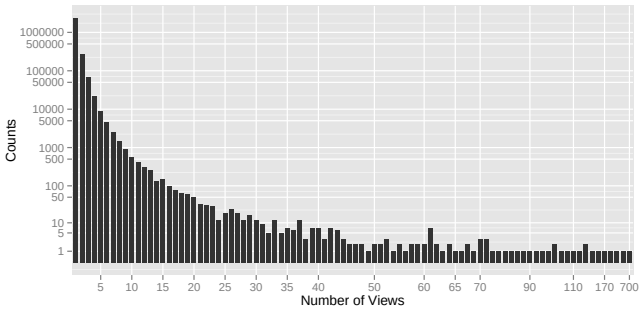


Figure 7: Log scale representation of the frequencies of document views based on the query logs.

A log scale bar chart of the (non-zero) viewing frequencies (as in Fig. 7) provides more insight than the Lorenz curves. While the viewed documents dataset is smaller, the shape of the view frequency distribution is very similar to that of the retrievability score of the real queries in Fig. 2, and even more similar than the scores of the simulated queries. Again, this suggests that simulated queries do not necessarily represent real user behavior.

Viewed but not retrieved.

To explore if the unique documents retrieved in our experiment using real queries are representative for the 2.7M unique documents actually viewed by the users, we investigated the overlap between the two. Given that most users only look at the first page with 10 results, we looked at the overlap for BM25 at $c = 10$, where we have 4.7M unique documents that are retrieved at least once. Less than 0.6M of these were also viewed, leaving 2.1M documents that were viewed but not retrieved in our top 10.

To find out what the reasons for the small overlap were, we performed a preliminary manual assessment of the top viewed documents that had *not* been retrieved by BM25 at $c = 10$. The most viewed document in this subcollection is a very short article describing an incident, in which a cow accidentally "caught" a rabbit ("Men kan niet weten hoe een koe een haas vangt."[11]). From the user logs, we learned that this was caused by deep linking: the article was accessed in response to a hyperlink in a newsletter, not in response to a direct search action. The second most viewed article[12] was retrieved in response to a direct search action, but by making use of the search interface's time facet which allows users to narrow down the search results to specific time periods.

Other often viewed documents were retrieved in our experiment, but with a ranking slightly above the $c = 10$ cutoff. That this is not just anecdotical but a larger issue is confirmed by the much larger overlap for the higher cutoff values. $c = 100$ retrieved 1.5M viewed documents, and $c = 1000$ retrieved more than 2.4M of the 2.7M viewed documents.

These results can be interpreted in two ways. First, small differences in the ranking scheme can have quite dramatic effects due to the all-or-nothing scoring function. This suggests that a smoother cost function based on the ranking might be worthwhile. Another potential interpretation is that the experimental setup needs to reflect the real search engine better, and also take the faceted search parameters, pagination, search operators and other more complex search settings into account.

[11] http://resolver.kb.nl/resolve?urn=ddd:110540686:mpeg21:a0015

[12] http://resolver.kb.nl/resolve?urn=ddd:000011882:mpeg21:a0004

Model	C	Real	Simulated
BM25	10	56.19 %	91.19 %
	100	7.94 %	73.51 %
TFIDF	10	53.48 %	91.44 %
	100	8.19 %	75.53 %
LM1000	10	54.74 %	89.24 %
	100	8.75 %	70.62 %

Table 9: The percentages of results from query logs and simulated queries that are *not* found by the other query set show that for small values of c the results vary strongly.

6.2 Real versus Simulated Queries

Since real query logs for large document collections are hard to obtain, most retrievability experiments reported in the literature use simulated queries, typically based on sampling the most popular n-grams. However, our results seem to suggest such queries might not be representative of real user queries.

Qualitative comparison of often retrieved documents.

To get a better intuition of the type of documents retrieved, we manually explored the top 10 articles for both query sets (for BM25 at $c = 10$). The top results for the real queries completely consisted of articles that contained lists of names [13]. This is because the logs from the KB contain a large number of queries with names and locations.

We compared this finding to the top results set retrieved by the simulated queries. Here, the top scoring documents either contain a very repetitive text pattern (e.g. repetitive poems[14]), or the documents themselves are near duplicates of other documents (e.g. chain letters, advertisements with identical text, or other documents that were published multiple times[15]). This finding might indicate another drawback of the way the simulated queries are traditionally sampled: frequently occurring terms are more likely to be included in the query set.

Overlap in retrieved documents.

The variety of $r(d)$ values is much larger on the real queries, indicating that the two query sets might retrieve very different documents (see Fig. 2). We explored the overlap of documents that were retrieved by the real queries and the (larger) set of simulated queries. For all three models, at $c = 10$, more than half of the documents retrieved by the real queries are *not* found in the results from the simulated queries (see Table 9). This again suggests that we should improve the construction of our simulated query set to better represent real queries. Note that the fraction of documents that are retrieved by both approaches is considerably higher for $c = 100$, where less than 9% of the documents in the

[13] see for example http://resolver.kb.nl/resolve?urn=ddd:010179873:mpeg21:a0001

[14] http://resolver.kb.nl/resolve?urn=ddd:010210514:mpeg21:a0150

[15] see http://resolver.kb.nl/resolve?urn=ddd:010691557:mpeg21:a0069

result set of the real queries are not found in the results of the simulated queries.

Differences between query sets.

In addition to the difference between the documents retrieved by both types of queries we also looked at the characteristics of the query sets themselves. The two query sets differ not only in size (as indicated in Table 2). The mean length of the real queries is 2.32 and all queries use a total of 253,637 unique terms. As we followed [1] and only used single and bi-term queries for the simulated query set, its mean query length is much smaller (1.5). The number of unique terms (2,028,617) is, however, much higher. This suggests that even by sampling only the most popular (bi)terms, we would over estimate the vocabulary used by users to formulate their queries.

We manually assessed the number of terms that refer to named entities in the 100 most frequent terms in both query sets. For the simulated queries, we found only 5 mentions of persons or locations, as opposed to 56 named entities in the real queries, confirming again the large differences in this aspect between the two sets.

Table 8 shows a higher retrievability of *official notifications* for the real queries. We compare this finding with the fractions of viewed documents for each type. While these fractions are very low for articles (only 2.61% viewed), advertisements (2.07%) and captions (4.01%), a much higher fraction of the official notifications was viewed (40.10%). This again shows that retrievability measured by real queries are more representative than synthesized queries.

6.3 Representativeness of Parameters used

Apart from queries and document features, retrievability can also be influenced by the parameters used in the retrieval setup, namely the inclusion or exclusion of stopwords and operators, and stemming. While we followed the parameter settings used by [1] so far (PS1), we compare the results obtained with the real queries using two alternative parameter settings (PS2 and PS3):

PS1: operators removed, stopwords removed, stemmed (used by [1])

PS2: operators removed, stopwords kept, unstemmed

PS3: operators, stopwords removed, stemmed[16]

Parameter sets *PS2* and *PS3* resulted in nearly identical *Gini* coefficients to those we reported in Table 3 for *PS1*. This suggests that the removal of stopwords, or the use of stemming and operators, has no influence on the extent of inequality in the document retrieval. The question remains, however, whether and how the underlying retrieved document sets differ and how this relates to the documents the users found sufficiently relevant to view.

Differences in retrieved document sets.

We compared the retrieved document sets from *PS1* and *PS2* for their overlap and found that while the majority of documents retrieved in one setting is also retrieved in the other, still a large fraction is only found in one setting (see Table 10). Note that even though this difference is not reflected in the *Gini* coefficient, *Lorenz* curves or r(d) distribution plots, it is a form of retrieval bias that may have a huge impact on the user's task.

[16]As restrictions of the Indri toolkit (http://www.lemurproject.org/) did not allow us to run this set of parameters for BM25 and TFIDF, these results are available only for LM1000.

	PS1	Shared	PS2	C
BM25	1,939,710	2,758,599	1,971,087	10
TFIDF	1,667,374	2,485,412	1,689,125	
LM1000	2,141,563	2,620,988	1,317,420	
BM25	7,436,058	17,923,267	7,232,087	100
TFIDF	6,672,656	16,385,354	6,381,519	
LM1000	7,384,854	16,711,774	4,804,696	

Table 10: Numbers of documents retrieved only by one parameter set (*PS*) and number of documents retrieved by both sets.

	PS1	PS2	PS3	C
BM25	504,022	598,969	-	10
TFIDF	435,413	527,461	-	
LM1000	742,548	706,425	781,908	
BM25	1,422,231	1,511,973	-	100
TFIDF	1,323,284	1,423,589	-	
LM1000	1,788,719	1,741,290	1,840,285	

Table 11: Viewed documents that were retrieved by each model for the different parameter sets (*PS*) for a total of 2,732,139 viewed documents.

Again, as c increases, the fraction of shared documents between the parameter sets increases as well. To judge which of the document sets is the more favorable for our use case, we compare the overlaps of the result sets with documents that were viewed by users (e.g using views as a proxy for relevance judgements).

The combinations of IR model and parameter set vary strongly with respect to their ability to retrieve the viewed documents (see Table 11). BM25 and TFIDF achieved better results with *PS2* than with *PS1*, but both are outperformed by LM1000 in all settings. The best result is achieved by using LM1000 with *PS3* with 29% of the viewed documents retrieved, so that in this case, the retrieval model with the most bias also performs better. This is in contrast to results reported by [1], where better performing models typically also show less bias.

7. CONCLUSIONS AND OUTLOOK

Measuring the variation in the retrievability of documents in a collection complements standard IR evaluations that focus on efficiency and effectivity. No previous study has investigated how well retrievability studies represent the search behavior of real users and how they could be applied to a large collection of digitized documents that contain an unknown number of misspellings due to OCR processing. Our focus was on the exploration of the applicability of retrievability studies to a large digitized document collection and an evaluation of the representativeness of simulated queries for real users' search behavior.

While *Gini* coefficients and *Lorenz* curves allowed us to detect and quantify a retrievability bias in the document collection for three standard IR models, they were not sufficiently expressive to help us understand the source of it. We looked at the differences among the documents retrieved,

and showed that large differences are common even for models with similar *Gini* coefficients and *Lorenz* curves.

In addition, we explored several influencing factors: the document to query ratio, document features, characteristics of query sets and the use of different parameter sets.

When comparing the characteristics of simulated queries to those of real users' queries we found substantial differences with respect to composition of the query sets, number of (unique) terms used, and use of named entities. Real users' queries contained a much higher fraction of named entities than we found in the simulated query set.

Finally, we compared how effectively combinations of specific parameter settings could retrieve the documents users viewed. Based on the results from this study, the setup that best covers the users' information needs is the combination of real queries with operators on LM1000. Note that according to the inequality assessment, the least biased model is BM25. This shows, that switch to a model with a lower retrievability bias might hurt the system's performance in terms of retrieving the most relevant documents.

Simulated queries that are representative for the search behavior of real users are a key ingredient for a realistic assessment of retrievability bias. Future work should therefore focus on how the generation of simulated queries can be adapted in a way that they better represent the type of queries real users issue on a specific collection.

Acknowledgments

We would like to thank Marc Bron for pointing us to the retrievability literature, Emma Beauxis-Aussalet for her feedback on the statistical analyses, and the National Library of the Netherlands for their support. This research is funded by the Dutch COMMIT/ program and the WebART project. Part of the analysis work was carried out on the Dutch national e-infrastructure with the support of the SURF Foundation.

8. REFERENCES

[1] L. Azzopardi and V. Vinay. Retrievability: An evaluation measure for higher order information access tasks. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, CIKM '08, pages 561–570, New York, NY, USA, 2008. ACM.

[2] R. Bache and L. Azzopardi. Improving access to large patent corpora. In A. Hameurlain, J. Küng, R. Wagner, T. Bach Pedersen, and A. Tjoa, editors, *Transactions on Large-Scale Data- and Knowledge-Centered Systems II*, volume 6380 of *Lecture Notes in Computer Science*, pages 103–121. Springer Berlin Heidelberg, 2010.

[3] S. Bashir. Estimating retrievability ranks of documents using document features. *Neurocomputing*, 123(0):216 – 232, 2014. Contains Special issue articles: Advances in Pattern Recognition Applications and Methods.

[4] S. Bashir and A. Rauber. Improving retrievability and recall by automatic corpus partitioning. In A. Hameurlain, J. Küng, R. Wagner, T. Bach Pedersen, and A. Tjoa, editors, *Transactions on Large-Scale Data- and Knowledge-Centered Systems II*, volume 6380 of *Lecture Notes in Computer Science*, pages 122–140. Springer Berlin Heidelberg, 2010.

[5] S. Bashir and A. Rauber. Improving retrievability of patents in prior-art search. In C. Gurrin, Y. He, G. Kazai, U. Kruschwitz, S. Little, T. Roelleke, S. Rüger, and K. van Rijsbergen, editors, *Advances in Information Retrieval*, volume 5993 of *Lecture Notes in Computer Science*, pages 457–470. Springer Berlin Heidelberg, 2010.

[6] S. Bashir and A. Rauber. Automatic ranking of retrieval models using retrievability measure. *Knowledge and Information Systems*, 41(1):189–221, 2014.

[7] G. Garvy. Inequality of income: Causes and measurement. In *Studies in Income and Wealth, Volume 15*, pages 25–48. NBER, 1952.

[8] R. Holley. How good can it get? Analysing and improving OCR accuracy in large scale historic newspaper digitisation programs. *D-Lib Magazine*, 15(3/4), 2009.

[9] E. Klijn. The current state-of-art in newspaper digitization a market perspective. *D-Lib Magazine*, 14, January 2008.

[10] M. C. Traub, J. van Ossenbruggen, and L. Hardman. Impact analysis of ocr quality on research tasks in digital archives. In S. Kapidakis, C. Mazurek, and M. Werla, editors, *Research and Advanced Technology for Digital Libraries*, volume 9316 of *Lecture Notes in Computer Science*, pages 252–263. Springer International Publishing, 2015.

[11] C. Wilkie and L. Azzopardi. Best and fairest: An empirical analysis of retrieval system bias. In M. de Rijke, T. Kenter, A. de Vries, C. Zhai, F. de Jong, K. Radinsky, and K. Hofmann, editors, *Advances in Information Retrieval*, volume 8416 of *Lecture Notes in Computer Science*, pages 13–25. Springer International Publishing, 2014.

[12] C. Wilkie and L. Azzopardi. Efficiently estimating retrievability bias. In M. de Rijke, T. Kenter, A. de Vries, C. Zhai, F. de Jong, K. Radinsky, and K. Hofmann, editors, *Advances in Information Retrieval*, volume 8416 of *Lecture Notes in Computer Science*, pages 720–726. Springer International Publishing, 2014.

[13] C. Wilkie and L. Azzopardi. Retrievability and retrieval bias: A comparison of inequality measures. In A. Hanbury, G. Kazai, A. Rauber, and N. Fuhr, editors, *Advances in Information Retrieval*, volume 9022 of *Lecture Notes in Computer Science*, pages 209–214. Springer International Publishing, 2015.

[14] S. Yitzhaki. Relative deprivation and the Gini coefficient. *The Quarterly Journal of Economics*, 93(2):pp. 321–324, 1979.