Impact of Crowdsourcing OCR Improvements on Retrievability Bias

Myriam C. Traub
Centrum Wiskunde & Informatica
Amsterdam, The Netherlands
myriam.traub@cwi.nl

Thaer Samar
Centrum Wiskunde & Informatica
Amsterdam, The Netherlands
thaer.samar@cwi.nl

Jacco van Ossenbruggen
Centrum Wiskunde & Informatica
Amsterdam, The Netherlands
VU University Amsterdam
Amsterdam, The Netherlands
jacco.van.ossenbruggen@cwi.nl

Lynda Hardman
Centrum Wiskunde & Informatica
Amsterdam, The Netherlands
Utrecht University
Utrecht, The Netherlands
lynda.hardman@cwi.nl

ABSTRACT

Digitized document collections often suffer from OCR errors that may impact a document's readability and retrievability. We studied the effects of correcting OCR errors on the retrievability of documents in a historic newspaper corpus of a digital library. We computed retrievability scores for the uncorrected documents using queries from the library's search log, and found that the document OCR character error rate and retrievability score are strongly correlated. We computed retrievability scores for manually corrected versions of the same documents, and report on differences in their total sum, the overall retrievability bias, and the distribution of these changes over the documents, queries and query terms. For large collections, often only a fraction of the corpus is manually corrected. Using a mixed corpus, we assess how this mix affects the retrievability of the corrected and uncorrected documents. The correction of OCR errors increased the number of documents retrieved in all conditions. The increase contributed to a less biased retrieval, even when taking the potential lower ranking of uncorrected documents into account.

CCS CONCEPTS

• **Information systems → Query log analysis**; **Evaluation of retrieval results**;

KEYWORDS

Retrievability Bias, Digital Library, Data Quality, OCR

ACM Reference Format:
Myriam C. Traub, Thaer Samar, Jacco van Ossenbruggen, and Lynda Hardman. 2018. Impact of Crowdsourcing OCR Improvements on Retrievability Bias. In *JCDL '18: The 18th ACM/IEEE Joint Conference on Digital Libraries, June 3–7, 2018, Fort Worth, TX, USA,* Editor1, Editor2, and Editor3 (Eds.). ACM, New York, NY, USA, Article 4, 8 pages. https://doi.org/10.1145/3197026.3197046

1 INTRODUCTION

Digitized collections are the foundation for services and research tasks that would be much more difficult (if not impossible) to perform on collections of physical items. Examples of such tasks are full-text search and quantification of changes in textual features over long time periods. Most of these services, however, rely on the use of retrieval systems.

How well these systems perform has been investigated with regard to many different aspects, such as precision and recall, and based on many different types of corpora, such as community-created TREC collections, digital libraries or Web archives. The *retrievability measure* as introduced by Azzopardi et al. [1] extends these evaluation measures by means to detect and assess bias when retrieving documents.

In a previous study, we used retrievability to investigate whether a retrievability bias influences access to a digitized collection of historic newspapers and to measure the extent of this bias [14]. We found a relation between document features, such as document length, and retrievability. In this study, we focus on the effects of OCR quality on retrievability and how a (partial) manual correction of the OCR errors impacts the accessibility of document. We investigate the following research questions.

- *RQ1: What is the relation between a document's OCR character error rate and its retrievability score?* By relating the retrievability scores of documents with the character error rates of their content, we investigate how the quality of OCR processing impacts a document's retrievability.
- *RQ2: How does the correction of OCR errors impact the retrievability bias of the corrected documents (*direct impact*)?* Assuming that the complete set of documents has been corrected, we investigate if the correction makes retrieval more or less biased in terms of retrievability, and how differences in retrievability scores are distributed over documents, queries and query terms.
- *RQ3: How does the correction of a fraction of error-prone documents influence the retrievability of non-corrected ones (*indirect

JCDL '18, June 3–7, 2018, Fort Worth, TX, USA

ACM ISBN 978-1-4503-5178-2/18/06…$15.00
https://doi.org/10.1145/3197026.3197046

impact*)?* Typically, only small fractions of a collection are corrected. We investigate how this affects the other documents in the collection by comparing the retrievability scores in a mixed collection where 50% of the collection has been corrected with those of an uncorrected only collection.

2 RELATED WORK

2.1 OCR Quality and Retrieval

In 2015, we conducted a series of interviews with digital humanities scholars on their use of digital archives for their research. All agreed that the (OCR) quality of digitized documents makes digital libraries unsuited for "distant reading" and other computational approaches [15]. Several studies investigated the applicability of crowdsourcing tasks to transcribe documents [7, 8] or the use of a tool that combines the search in a digitized corpus with correction of OCR errors [10]. While the results from these studies can help improve data quality more efficiently, it remains unclear how this correction affects a scholar's research.

Mittendorf et al. investigated how robust IR systems are toward OCR errors in digitized documents [9]. They found that longer documents describing a single topic redundantly have a better chance of retrieval than documents that are either short or discuss different topics.

Taghva et al. investigated the performance of the vector space model on OCRed documents [13]. They found that for their full text collection neither average precision, nor recall of the documents is affected by OCR errors. 674 documents were used in a OCR processed version and a manually corrected ground truth version. The character error rate was estimated to be around $10-20\%$ and the average length of the documents is reported to be around 40 pages. This confirms the findings of Mittendorf et al. that the effect of OCR errors on long documents can be expected to be very low. Since our corpus is characterized by relatively short documents with a high estimated error rate, we expect a higher impact of OCR errors than in the studies of Mittendorf et al. and Taghva et al.

Ohta et al. studied whether the effect of OCR errors on document retrieval can be compensated by generating additional search terms based on a character confusion matrix [11]. They based their study on two collections of documents obtained from the *Elsevier Electronic Subscriptions* service and published between 1995 and 1996. The document collection in this case can therefore be expected to be very homogenous in terms of layout, fonts, document length, quality of the physical copy and as a consequence cause little variation in error rates and error types. In our case, documents vary strongly in all of these aspects and therefore errors are less systematic as in the documents of Ohta et al. A statistical approach would be difficult, as it could only be applied to subsets of very similar documents.

Chiron et al. investigated for the AmeliOCR project how OCR errors are distributed in a large and diverse digitized corpus [5]. They found that about 15% of the misspelled terms represent named entities and that even 80% of the top 500 queries contain at least a mention of one. In a manual inspection of the 100 most frequent terms in the query set we used for [14], we found that 56 were named entities. The frequency of named entities in the document collection, however, can be very low and they may not even be found in common dictionaries. This makes them particularly susceptible to OCR errors. This combination, i.e. terms that occur very frequently in queries, but very infrequently in documents, is the reason that OCR errors in these terms can have a disproportional effect on the retrieval results [5].

2.2 Retrievability Assessment

The foundation for the assessment of retrievability bias in document collections is the work by Azzopardi et al [1]. They introduced *retrievability* as an extension to traditional IR measures, one that does not require the availability of relevance judgments. It considers the number of results that a user is willing to examine (c). If the rank k_{dq} of a document d is retrieved within the cutoff value c, the utility/cost function f returns a score of 1, otherwise 0.

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

o_q allows different weighting of queries according to their importance. We use $o_q = 1$ for all queries. To measure a potential bias among *r(d)* scores, [1] suggested to use the Gini coefficient, which was introduced to measure inequalities in societies [6]. Wilkie et al. later compared it to other inequality measures and confirmed its aptitude for retrievability analyses [18].

Follow-up studies confirmed the applicability of the retrievability measure to assess bias in retrieval models [16] and its relatedness to retrieval effectiveness [3, 4]. Several studies found that Okapi BM25 induces the least bias and can therefore be considered to be the fairest retrieval model [14, 16, 17]. While [1] and most subsequent retrievability studies (e.g. [2, 4, 12, 18]) made use of *simulated user queries*, we follow the line of our previous study and use queries collected from real users of the digital library [14]. In [14], we investigated the applicability of the retrievability metric on a digitized newspaper collection and questioned the representativeness of simulated queries for the search behavior of real users. Our findings revealed significant differences in number of query terms used and the frequency of named entity queries.

The current study extends the findings of [14] in several aspects. Our first study was based on the complete archive which comprises more than 102 million documents. The relatively high *document - query ratio* (DQR) had a large impact on the inequality in the *r(d)* scores because a large fraction of the documents was never retrieved. By focusing on a small subset of the newspaper collection in this study, we prevented a high DQR rate, and analyze an inversed scenario where the number of queries exceeds by far the number of documents. Finally, the availability of a ground truth data set enables us to investigate retrieval results on a corrected document collection, a collection containing errors, and a mixed collection.

3 APPROACH

To investigate whether and how errors in OCRed documents influence their retrievability, we performed a series of experiments that make use of the concept of *retrievability* as introduced by Azzopardi et al. [1]. For this, we used different subsets of a digitized newspaper collection and search queries that were collected from users of the online access portal of the archive.

The National Library of the Netherlands (KB)[1] made a ground truth data set available that contains the manually corrected versions of 100 newspaper issues. By comparing these documents with their original versions, we were able to assess the number of incorrect characters and compute the character error rates (*CER*) for each document. This allowed us to investigate a relation between the documents' quality and their retrievability scores (*RQ1*).

The manual correction of OCR errors *directly* impacts the retrievability of these documents. We investigated this effect with two retrievability experiments based on a small document collection and two versions of query sets that were originally collected from users of the digital archive. By comparing the *r(d)* scores, we investigate which documents and queries gained or lost *r(d)* scores through the correction and how this influences the total number of retrieved documents (*wealth*) and retrieval bias (*inequality*) of the results (*RQ2*).

Since correction of OCR errors is often performed manually, it is a costly process. As a consequence only relatively small fractions of a collection are corrected. The same document may score lower in a corpus consisting of only highly findable documents than it would as part of a collection of documents that are difficult to find. Therefore, we explored how the correction of only a part of the collection *indirectly* impacts the retrievability of documents that remain uncorrected (*RQ3*).

4 EXPERIMENTAL SETUP

The setup we used for our experiments follows the setup used in [1, 14], modifications are explicitly described in this section.

4.1 Document Collections

We use different subsets of the historic newspaper archive, a manually corrected ground truth subset and queries collected from the online users of the archive.

OCR Ground Truth Corpus (822_{GTcor}) For a small subset of the newspaper collection of the National Library of the Netherlands, the OCR text has been manually corrected. This subset covers 100 newspaper issues published between 14-06-1618 and 26-10-1624 ($17^{th}century$ subset) and between 04-10-1940 and 29-09-1944 ($WWII$ subset). The $17^{th}century$ sub-collection constitutes the part of the archive with the oldest documents. It is prone to OCR errors as the decay of the physical material, the layout and the (gothic) fonts make character recognition very difficult. The $WWII$ collection includes illegal newspapers, printed secretly, often in non-professional settings. Some of these articles therefore have a lower OCR quality than pro-German papers of the same period with better print quality. Combined, they include a total of 822 newspaper items. Note that this corpus is very small compared to 100M item corpus used in our first study [14].

OCRed Corpus containing Errors (822_{GTerr}) We used the uncorrected versions of the articles in 822_{GTcor} to build the 822_{GTerr} corpus.

Mixed Documents Corpus (1644_{mix} and 1644_{err}) We extended 822_{GTcor} and 822_{GTerr} with an equal number of articles that originate from the same newspaper titles as in the 822_{GTerr} collection. We selected the 503 earliest articles from the KB collection and a random sample of 319 articles from the WWII period (822_{mixin}). These documents added to 822_{GTcor} yields the 1644_{mix} corpus, and added to 822_{GTerr} yields the 1644_{err} corpus.

4.2 Query Set

The queries we used were collected from the users of the library's Web interface (Delpher.nl) between March and July 2015. The data set comprises a total number of $1,008,915$ queries from $162,536$ unique users with an average length of three terms. We removed stopwords[2] and terms shorter than three characters from the queries. The final, deduplicated, query set comprises $859,716$ *multi-term queries*. Additionally, we created a *single-term query set* by extracting all $259,091$ unique terms.

4.3 OCR Quality Assessment

We measured the OCR quality of 822_{GTerr} set using the OCRevalUAtion tool[3] developed by the IMPACT project. It allowed us to compute the character error rates (*CER*) for each article in 822_{GTerr}.

4.4 Setup for Retrievability Analysis

We investigated whether and how OCR quality impacts retrievability by comparing how retrievability scores ($r(d)$) differ between documents containing errors and their corrected versions. To compute the $r(d)$ score for each document, we issued all queries against the document collections using the Indri search engine[4] and BM25 as retrieval model. For each document we calculated how often it was retrieved in the top c results (for cut-off values of $c = 1, 10$ and 100) and how often it was retrieved at all ($c = \infty$).

The *wealth*, or the total sum of all *r(d)* scores, depends on the number of queries issued and the number of results taken into account (c). To assess differences between the results obtained from the different corpora we calculated the wealth for each corpus for all values of c. An increase or decline in retrieval bias is determined using the *Gini coefficient*, which is a measure developed to express inequalities in societies [6].

4.5 Impact Analysis

Assessment of Query Impact We investigated the *impact* each unique query *term* has on the total wealth of a document collection. For this, we issued all unique single query terms against the document collections and recorded the matching query - document pairs. We used these to assess, for every multi-term query - document pair, which of the terms in the query was responsible for retrieving the documents that appeared on the result list for said multi-term query. We then assigned each successful term a score of $\frac{1}{n_t}$ where n_t is the number of successful query terms t for a document - multi-term query pair. The sum of all of these scores for all occurrences of a query term is its *impact score* and the sum of all impact scores equals the total wealth of all $r(d)$ scores for a corpus.

Assessment of Direct Impact We investigated the differences in the retrievability of documents before (822_{GTerr}) and after

[1] www.kb.nl

[2] http://snowball.tartarus.org/algorithms/dutch/stop.txt

[3] http://www.digitisation.eu/training/succeed-training-materials/ocr-evaluation/ocrevaluation/the-ocrevaluation-tool/

[4] http://www.lemurproject.org/indri.php

(822_{GTcor}) error correction. For this, we evaluate the total number of documents retrieved (*wealth*), the equality of the *r(d)* scores' distribution, and we analyze qualitatively the documents and queries for which the differences between the experimental conditions are the largest. We measure the difference in inequality among the *r(d)* scores for the two versions of the document collection using the Gini coefficient. A high Gini coefficient indicates a large inequality in the distribution, a low Gini coefficient indicates a more equal and therefore less biased distribution. Then we investigate the difference in *r(d)* scores for each document in both versions. A gain in *r(d)* scores indicates that the document benefited from the correction of its content. A decrease in *r(d)* scores shows that its corrected version was retrieved by fewer queries than the original version. We manually assessed the documents with the largest differences and the queries that retrieved those documents to find out what caused the drop or increase in *r(d)* scores.

Assessment of Indirect Impact For this experiment we used the 1644_{mix} and the 1644_{err} data sets. Again, we evaluated differences in the overall wealth of distributed *r(d)* scores, the inequality between documents in terms of *r(d)* scores and the differences between documents in direct comparison. Differences in the results are caused by the interlace between the rankings of the corrected and unchanged documents. The analyses we perform for this section are similar to those of the *direct impact* experiments.

4.6 Limitations

Since relevance judgments are not available for this document collection, we were not able to explore how OCR errors correlate with precision and recall. In our mixed experiment, we only evaluated a correct/incorrect ratio of 50:50, other ratios are planned for future work.

5 RESULTS

5.1 OCR Quality versus Retrievability

First, we studied to what extent a document's OCR error rate and its *r(d)* score are related (*RQ1*).

QCR Quality We evaluated the OCR quality using the OCRevalUAtion tool[5]. The results showed that the mean character error rate (*CER*) of the collection is high: 29% (with a median CER of even 37%). We found a clear difference in the *CER* distributions of the two sub-collections (see Fig. 1). As expected, the more recent documents from *WWII* suffer from far fewer mis-recognized characters (median CER = 3.97%) than the documents from the 17^{th}*century* (median CER = 42.00%).

Retrievability in 822_{GTerr} An analysis of the *r(d)* scores showed that we retrieved $4,521,030$ documents from 822_{GTerr} ($c = \infty$) in total. The scores ranged from *r(d)* = 0 (16 documents, of which two are part of the *WWII* sub-collection and 14 are part of the 17^{th}*century* sub-collection) to *r(d)* = $65,347$. Most documents are in the lowest bin ($r(d) < 674$), as shown in the margin histogram on the right of Fig. 1. The median scores are

- *r(d)* = 991 for 822_{GTerr},
- *r(d)* = 447 for 17^{th}*century*, and
- *r(d)* = $8,237$ for the *WWII* sub-collection.

Corpus	c = 1	c = 10	c = 100	c = ∞
822_{GTerr}	0.75	0.72	0.74	0.74
822_{GTcor}	0.68	0.59	0.61	0.61
1644_{err}	0.78	0.70	0.73	0.73
1644_{mix}	0.73	0.63	0.66	0.66

Table 1: Gini coefficients indicating to which extent the distribution of *r(d)* scores among documents for different c's is biased (higher values indicate more bias).

This confirms the hypothesis that the *WWII* documents are easier to retrieve due to their better OCR quality.

We found a strong correlation between OCR quality and retrievability of a document for results with $c = \infty$. Documents with a low *CER* generally obtained higher *r(d)* scores (see Figure 1). The correlations of -0.57 (Pearson) and -0.61 (Spearman) were both strong and significant with $p < 0.001$. While this correlation may suggest that low $r(d)$ scores are *caused* by high OCR error rates, other explanations could be that our modern query set just better matches the *WWII* sub-collection, or that 17^{th}*century* documents are harder to retrieve in general. To establish a causal relation, we study the direct impact of the crowd-sourced improvements on the *r(d)* scores in the next section.

5.2 Direct Impact Assessment

Next, we studied how the correction of OCR errors influences retrievability bias (*RQ2*). For this, we measure the direct impact of correcting OCR errors by comparing the $r(d)$ scores over 822_{GTcor} with the corresponding scores in 822_{GTerr}.

Wealth We found that more documents were retrieved from 822_{GTcor} than from 822_{GTerr} and that the relative difference increases for larger values of c (see Fig. 2). The total wealth at $c = 1$ indicates how many queries could be matched with at least one document. For $c = 1$, 8% more documents are retrieved from 822_{GTcor} than from 822_{GTerr}, which means that fewer queries retrieved no documents at all. For $c = \infty$ the total wealth increases by 34% (see Fig. 2). This suggests that for users willing to examine *all* search results (which is not uncommon in a research library) the impact of the error-correction is much larger. Correcting the OCR errors thus indeed leads to higher numbers of documents retrieved, even for small $c's$, and the effect increases when more results are taken into account.

Equality We computed and compared Gini coefficients for 822_{GTerr} and 822_{GTcor} to find out whether the increase in wealth contributed to a more equal or more biased distribution of *r(d)* scores (see Table 1). Gini coefficients for 822_{GTcor} are consistently lower than for 822_{GTerr} for all $c's$. The correction of the documents thus contributed to *less biased* retrieval for all $c's$. In contrast to other studies [1–3] and our earlier findings in [14], Gini coefficients do *not* show a clearly decreasing trend for larger cutoff values c. This suggests that in this experiment, ranking does not contribute much additional bias. This may be caused by the relatively small corpus size.

Increased retrieval per document We investigated how the changes of *r(d)* scores were distributed among documents, i.e. whether many documents gained a little or whether very few documents

[5] https://github.com/impactcentre/ocrevalUAtion

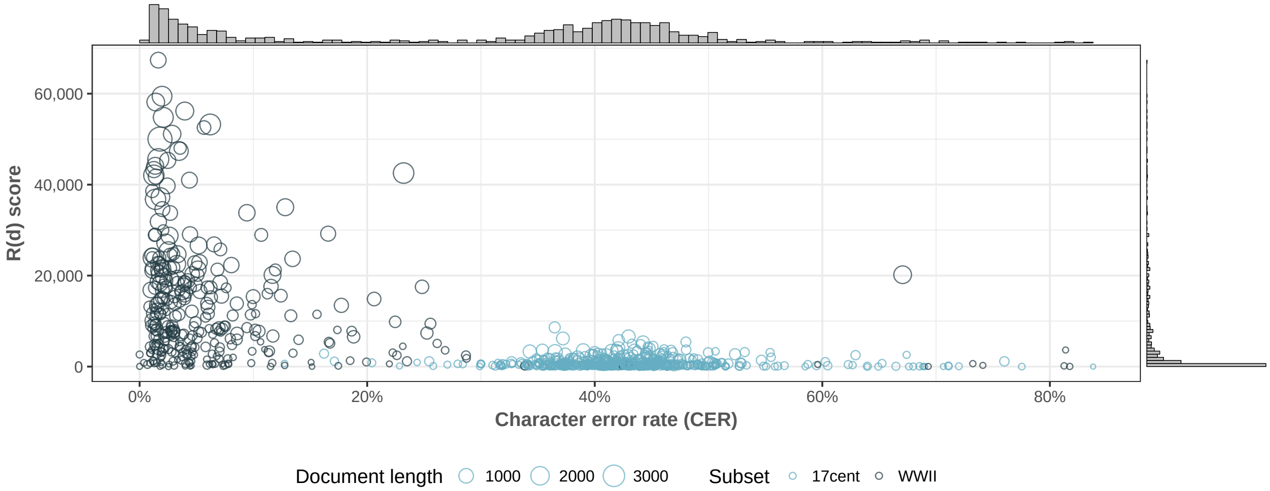


Figure 1: The 17^{th}*century* collection has a higher character error rate (*CER*) than the *WWII* collection. The *r(d)* scores and *CER* for $c = \infty$ are strongly correlated: the higher the error rate, the less retrievable is a document.

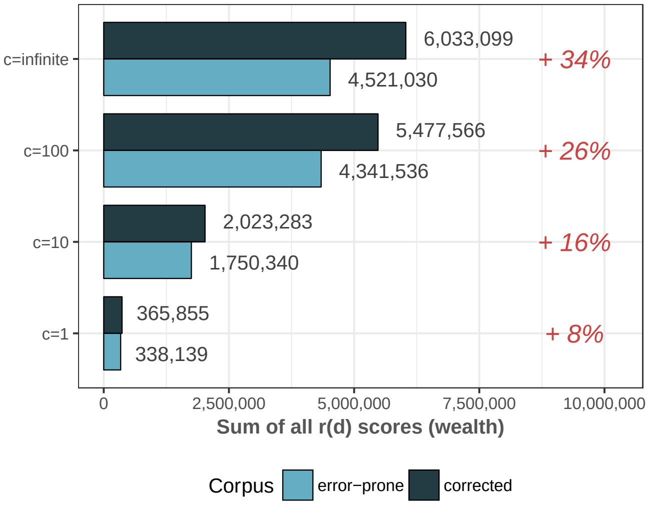


Figure 2: Difference in distributed wealth between the uncorrected and corrected corpus.

c	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
1	-5,039	-56	8	34	99	7,160
10	-7,124	-153	177	332	647	8,408
100	-7,040	24	652	1,382	1,941	25,647
∞	-7,019	275	912	1,840	2,292	27,926

Table 2: Summary statistics of differences in *r(d)* scores between the two corpora.

gained a lot. For Fig. 3 we ordered documents according to their difference in *r(d)* scores between 822_{GTerr} and 822_{GTcor}. We see a few documents on the left of the 0-axis, these documents had a

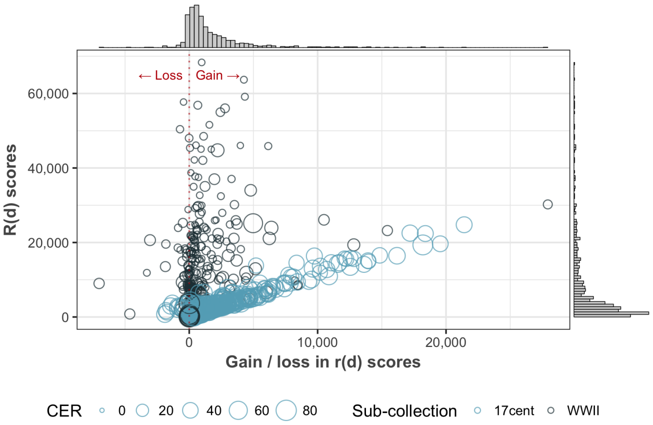


Figure 3: Documents ordered by their gain/loss in *r(d)* scores ($c = \infty$). The position on the y-axis represents their *r(d)* scores for 822_{GTcor}.

higher *r(d)* score in the uncorrected corpus. Closer inspection indicated that these were false positive matches *caused by* OCR errors. Their decreasing scores can therefore be interpreted as a potential improvement in precision. For most documents, OCR correction increased their $r(d)$ score, and they are therefore found on the right of the 0-axis. This can be interpreted as a potential improvement in recall. We see clearly different patterns for the two corpora, with many 17^{th}*century* documents improving more but scoring overall lower than the *WWII* documents. Several documents scored very low in 822_{GTerr}, but gained a lot from the correction. This is one explanation for why Gini coefficients for 822_{GTcor} show less bias than for 822_{GTerr}. Most documents, however, have a modest *r(d)* score and gained a modest amount, as shown in the margin

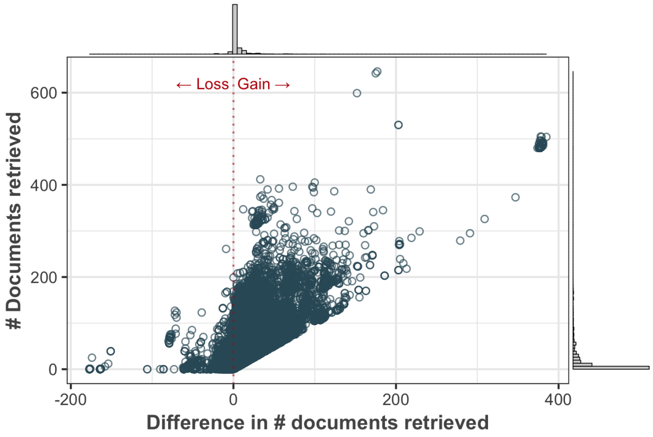
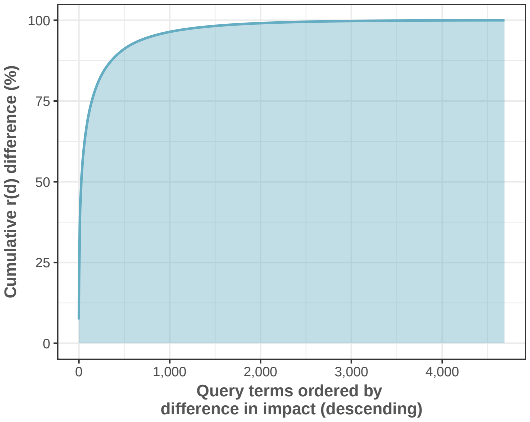


Figure 4: Queries ordered by their gain/loss in number of retrieved documents. The position on the y-axis represents the number of documents retrieved from 822_{GTcor}.

histograms. The distributions of the differences in *r(d)* scores in Table 2, show that for all cutoff values, the median of the differences is positive, and increases from 8 ($c = 1$) to 912 ($c = \infty$). The maximum loss and the maximum gain in *r(d)* scores increase for larger cutoff values c, the latter to a much larger extent. Note that for $c = 1$ and $c = 10$ the entire first quartile is filled with documents that scored *worse* in the corrected version. This shows that the competition in the top results makes the gain of some documents the loss of others.

Increased retrieval per query In a final step, we investigated how the changes of *r(d)* scores were distributed among the queries, i.e. if many queries contributed a little or if only a few queries that contributed a lot to the change in wealth. The large majority of queries does not match with any of the documents in our collection. Only 384, 486 out of 859, 716 queries retrieved at least one document from either of the document collections. This is due to misspellings from users, invalid words, numbers, words in foreign languages or simply queries that are unrelated to our (small) corpus. In Figure 4 we ordered these queries by how many more (or less) documents they retrieved in 822_{GTcor}. Note that despite the small corpus size, we still see outliers with very large gains (to over 400 documents more retrieved for some queries). Also note that some queries have a negative gain, which means that for these queries, the OCR errors caused more false positive matches than false negatives.

Finally, we were interested in finding out which query *terms* are responsible for most of the increase in wealth. Figure 5 shows that most of the increase can be attributed to *very few query terms*. The top ten queries[6] (see table adjacent to Fig 5) contribute 35% of the increase. This disproportionately large impact originates from a combination of the terms' high frequency in the users' queries and the large extent to which they are susceptible to errors in OCR processing.



Query	Frequency in			Cum.
Term	Queries	822_{GTerr}	822_{GTcor}	Impact
nieuwe	1,903	99	166	7.36%
amsterdam	7,885	41	57	14.65%
ende	185	103	480	18.69%
heer	826	20	89	21.99%
overleden	3,698	5	18	24.78%
groot	1,573	125	153	27.33%
willem	5,375	5	13	29.81%
twee	319	64	175	31.83%
drie	401	34	120	33.81%
oude	991	50	78	35.41%

Figure 5: The accumulated impact scores of single-term queries show that very few query term contribute a large fraction of the overall wealth. The top ten query terms account for more than a third of the increase (see Table).

5.3 Results of Indirect Impact Assessment

Finally, we investigated the influence of OCR error correction on the retrieval of documents that remain uncorrected (*RQ3*). We investigate for the typical case of a partial error-correction how the improved retrievability of the corrected documents impacts the *r(d)* scores of the documents that have not (yet) been corrected.

Wealth When looking at the *r(d)* scores of the mixed collection, we see that the correction of half the documents still leads to an increase in wealth for the complete corpus for all values of c (see Fig. 6). We first focused on the 822_{GTcor} documents within the mixed corpus. These are retrieved for the same queries as in the previous section. The mixed-in documents only cause differences in ranking. For $c = \infty$, we thus see identical *r(d)* scores and total wealth as in Section 5.2. For the lower c values, we see lower wealth due to competition in the ranking with the unaltered documents, but also large gains caused by the manual OCR correction.

In the remainder of this section, we focus solely on the documents that remain uncorrected, 822_{mixin}. In terms of distributed *r(d)* scores we found a decrease in wealth for the mixed-in documents for values of c from $c = 1$ to $c = 100$. This is because the corrected versions push many mixed-in documents to higher ranks that exceed the number of documents we take into account (c). This difference in wealth is largest for $c = 1$ (-13%), followed by $c = 10$

[6] Translations: new, Amsterdam, end, Mister, died/dead, grand/large, Willem (name), two, three, old

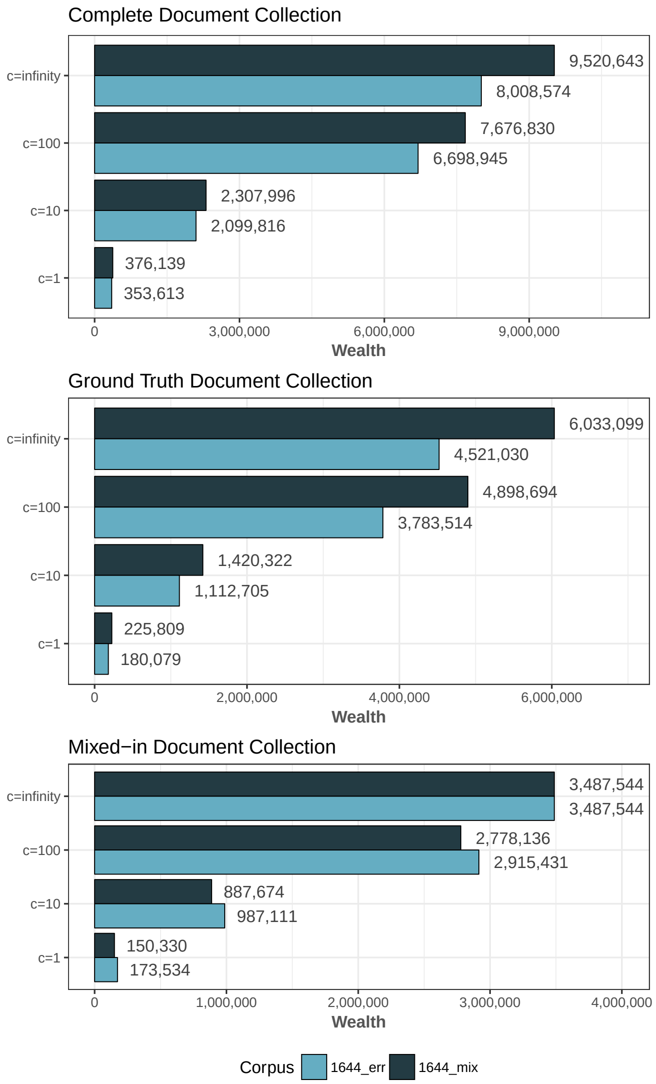


Figure 6: Wealth in *r(d)* scores for the complete collection (top), the 822_{GT} documents (middle) and the mixed in documents, 822_{mixin} (bottom).

(−10%) and $c = 100$ (−5%). For larger values the ranking does not take effect and the wealth remains the same.

Equality When we compare the Gini coefficients we obtained for different values of c, we see that they are lower for the corpus that was partially corrected, 1644_{mix}. Again, the correction of a part of the collection has reduced retrievability bias (see Table 1).

Retrieval per document The *r(d)* scores of most mixed-in documents changed very little after the correction of the other documents. Most documents' *r(d)* scores drop slightly (see Fig. 7), which could be expected as they now compete with corrected documents for low ranks. In total, 522 documents have lost in r(d) scores, of which 266 are from the $WWII$ sub-collection and 256 from $17^{th}century$. We also see that 171 documents gain in *r(d)* scores, of which 8 are from $WWII$ and 163 from $17^{th}century$ (see Fig. 7). These documents profit from false positive matches that disappeared through the correction.

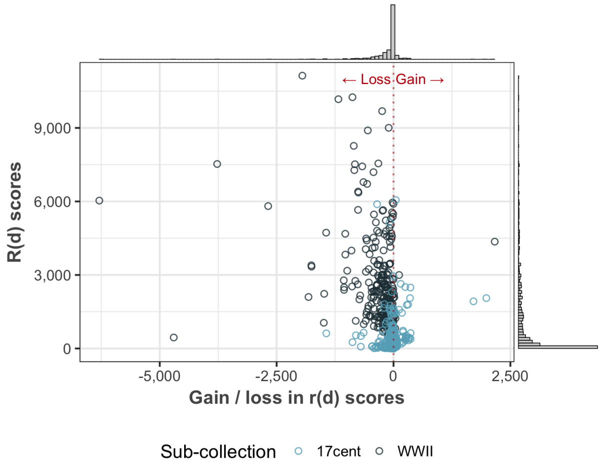


Figure 7: Documents ordered according to their difference in *r(d)* scores (non-GT documents at $c = 10$). Position on the y-axis indicated the *r(d)* score in the mixed condition. Documents in the left part of the graph lost *r(d)* scores.

Overall, we found that even in this mixed condition, the overall positive effect in improved retrievability for the corrected documents by far outweighs the slightly reduced retrievability of the unchanged documents. The net effect of the correction is still an overall *reduction of retrievability bias*.

6 CONCLUSIONS

Many text documents in digital libraries are affected by errors caused by OCR engines. It is therefore vital to understand how these errors and their (partial) correction impact retrieval tasks of digital library users. We investigated the relation between OCR quality of digitized newspaper articles and their retrievability and found a strong correlation: high error rates correlate with low retrievability scores. We compared the overall retrievability of a manually corrected ground truth document collection with the results obtained from the same documents but in their original, uncorrected version. Our analyses showed that error correction leads to both higher and more equally distributed retrievability scores.

The higher scores are mainly caused by a disproportionately small set of query terms, that are both very frequent in the query set and highly susceptible to OCR errors. This shows that for retrievability studies with real user queries, understanding the impact of a (biased) query set on the retrievability bias is important, while this is typically not considered in the literature, where synthetic query sets are more prevalent.

Our findings could be used for improving and evaluating automatic OCR-error correction techniques, or to improve query expansion techniques designed to deal with OCR-errors in uncorrected texts.

Furthermore, we looked at interference effects that the correction of a subset may have on documents that are excluded from the correction. We found that the reduced scores for the excluded documents do not outweigh the improved scores of the corrected version. The overall outcome is still a less biased retrieval result. Because we lack relevance judgments for this corpus, we cannot measure the improvement of the correction in terms of precision and recall. We can, however, conclude that the error correction has led to more documents being retrieved overall while reducing the retrievability bias in all experimental setups.

ACKNOWLEDGMENTS

We would like to thank the National Library of the Netherlands for their support and Alejandro Bellogín for his comments on an earlier version of the paper. This research is partly funded by the Dutch COMMIT/ program, the WebART project and the VRE4EIC project, a project that has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 676247. Part of the analysis work was carried out on the Dutch national e-infrastructure with the support of the SURF Foundation.

REFERENCES

[1] Leif Azzopardi and Vishwa Vinay. 2008. Retrievability: An Evaluation Measure for Higher Order Information Access Tasks. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management (CIKM '08)*. ACM, New York, NY, USA, 561–570. https://doi.org/10.1145/1458082.1458157

[2] Shariq Bashir. 2014. Estimating retrievability ranks of documents using document features. *Neurocomputing* 123, 0 (2014), 216 – 232. https://doi.org/10.1016/j.neucom.2013.07.011 Contains Special issue articles: Advances in Pattern Recognition Applications and Methods.

[3] Shariq Bashir and Andreas Rauber. 2014. Automatic ranking of retrieval models using retrievability measure. *Knowledge and Information Systems* 41, 1 (2014), 189–221. https://doi.org/10.1007/s10115-014-0759-6

[4] Shariq Bashir and Andreas Rauber. 2017. Retrieval Models Versus Retrievability. In *Current Challenges in Patent Information Retrieval*, Mihai Lupu, Katja Mayer, Noriko Kando, and Anthony J. Trippe (Eds.). Springer Berlin Heidelberg, 185–212. https://doi.org/10.1007/978-3-662-53817-3_7

[5] G. Chiron, A. Doucet, M. Coustaty, M. Visani, and J. P. Moreux. 2017. Impact of OCR Errors on the Use of Digital Libraries: Towards a Better Access to Information. In *JCDL 2017*. 1–4. https://doi.org/10.1109/JCDL.2017.7991582

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

[7] Rose Holley. 2009. How Good Can It Get? Analysing and Improving OCR Accuracy in Large Scale Historic Newspaper Digitisation Programs. *D-Lib Magazine* 15, 3/4 (2009). https://doi.org/10.1045/march2009-holley

[8] Kimmo Kettunen, Timo Honkela, Krister Lindén, Pekka Kauppinen, Tuula Pääkkönen, Jukka Kervinen, et al. 2014. Analyzing and Improving the Quality of a Historical News Collection using Language Technology and Statistical Machine Learning Methods. In *80th IFLA General Conference and Assembly*.

[9] Elke Mittendorf and Peter Schäuble. 2000. Information Retrieval Can Cope with Many Errors. *Inf. Retr.* 3, 3 (Oct. 2000), 189–216. https://doi.org/10.1023/A:1026564708926

[10] Günter Mühlberger, Johannes Zelger, and David Sagmeister. 2014. User-driven Correction of OCR Errors: Combining Crowdsourcing and Information Retrieval Technology. In *Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage (DATeCH '14)*. ACM, New York, NY, USA, 53–56. https://doi.org/10.1145/2595188.2595212

[11] M. Ohta, A. Takasu, and J. Adachi. 1997. Retrieval methods for English-text with missrecognized OCR characters. In *Proceedings of the Fourth International Conference on Document Analysis and Recognition*, Vol. 2. 950–956 vol.2. https://doi.org/10.1109/ICDAR.1997.620651

[12] Thaer Samar, Myriam C. Traub, Jacco van Ossenbruggen, Lynda Hardman, and Arjen P. de Vries. 2017. Quantifying retrieval bias in Web archive search. *International Journal on Digital Libraries* (18 Apr 2017). https://doi.org/10.1007/s00799-017-0215-9

[13] Kazem Taghva, Julie Borsack, and Allen Condit. 1996. Effects of OCR errors on ranking and feedback using the vector space model. *Information Processing & Management* 32, 3 (1996), 317 – 327. https://doi.org/10.1016/0306-4573(95)00058-5

[14] Myriam C. Traub, Thaer Samar, Jacco van Ossenbruggen, Jiyin He, Arjen de Vries, and Lynda Hardman. 2016. Querylog-based Assessment of Retrievability Bias in a Large Newspaper Corpus. In *Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries (JCDL '16)*. ACM, New York, NY, USA, 7–16. https://doi.org/10.1145/2910896.2910907

[15] Myriam C. Traub, Jacco van Ossenbruggen, and Lynda Hardman. 2015. *Impact Analysis of OCR Quality on Research Tasks in Digital Archives.* Springer International Publishing, Cham, 252–263. https://doi.org/10.1007/978-3-319-24592-8_19

[16] Colin Wilkie and Leif Azzopardi. 2014. Best and Fairest: An Empirical Analysis of Retrieval System Bias. In *Advances in Information Retrieval: 36th European Conference on IR Research, ECIR 2014, Amsterdam, The Netherlands, April 13-16, 2014. Proceedings*, Maarten de Rijke, Tom Kenter, Arjen P. de Vries, ChengXiang Zhai, Franciska de Jong, Kira Radinsky, and Katja Hofmann (Eds.). Springer International Publishing, Cham, 13–25. https://doi.org/10.1007/978-3-319-06028-6_2

[17] Colin Wilkie and Leif Azzopardi. 2014. Efficiently Estimating Retrievability Bias. In *Advances in Information Retrieval: 36th European Conference on IR Research, ECIR 2014, Amsterdam, The Netherlands, April 13-16, 2014. Proceedings*, Maarten de Rijke, Tom Kenter, Arjen P. de Vries, ChengXiang Zhai, Franciska de Jong, Kira Radinsky, and Katja Hofmann (Eds.). Springer International Publishing, Cham, 720–726. https://doi.org/10.1007/978-3-319-06028-6_82

[18] Colin Wilkie and Leif Azzopardi. 2015. Retrievability and Retrieval Bias: A Comparison of Inequality Measures. In *Advances in Information Retrieval*, Allan Hanbury, Gabriella Kazai, Andreas Rauber, and Norbert Fuhr (Eds.). Lecture Notes in Computer Science, Vol. 9022. Springer International Publishing, 209–214. https://doi.org/10.1007/978-3-319-16354-3_22