The Effect of Content-Equivalent Near-Duplicates on the Evaluation of Search Engines

Maik Fröbe[1](✉), Jan Philipp Bittner[1], Martin Potthast[2], and Matthias Hagen[1]

[1] Martin-Luther-Universität Halle-Wittenberg, Halle, Germany
maik.froebe@informatik.unihalle.de
[2] Leipzig University, Leipzig, Germany

Abstract. Current best practices for the evaluation of search engines do not take into account duplicate documents. Dependent on their prevalence, not discounting duplicates during evaluation artificially inflates performance scores, and, it penalizes those whose search systems diligently filter them. Although these negative effects have already been demonstrated a long time ago by Bernstein and Zobel [4], we find that this has failed to move the community. In this paper, we reproduce the aforementioned study and extend it to incorporate all TREC Terabyte, Web, and Core tracks. The worst-case penalty of having filtered duplicates in any of these tracks were losses between 8 and 53 ranks.

1 Introduction

Web crawls contain pages that are duplicates or near-duplicate of other pages [15]. Although there can be legitimate reasons for a web publisher to host pages that duplicate other publishers' pages, the user of a web search engine gains nothing from viewing basically the same search result twice or more while browsing search results. Therefore, web search engines typically identify duplicates, either at crawl time, at indexing time, or at retrieval time, in order to remove all but one of them from their search results, showing only the "best" version of a piece of content according to some selection criteria.

The fact that duplicate results are not "useful" to the users of a web search engine has not been overlooked: Treating results as irrelevant when they are found to be *content-equivalent* to a document the user has already seen, Bernstein and Zobel [4] applied this so-called "novelty principle" during their analysis of content-equivalent documents within the GOV collections. With respect to the TREC 2004 Terabyte Track, their research shows (1) that 16.6% of all relevant documents in submitted runs are content-equivalent, and (2) that the application of the novelty principle causes MAP scores to decrease by 20% on average. This situation is an obstacle to progress, since doing the right thing and filtering duplicates is penalized. Fifteen years have passed since the report by Bernstein and Zobel, and we are curious as to whether anything has changed: Are there still duplicate documents in commonly used benchmarks, and if so, how do they affect the evaluation of retrieval systems?

J. M. Jose et al. (Eds.): ECIR 2020, LNCS 12036, pp. 12–19, 2020.
https://doi.org/10.1007/978-3-030-45442-5_2

Corpus	**Size**		**Equivalent**	
	Gzip	Docs	Docs	Classes
GOV1	4.6 GB	1.2 m	6.54%	24,394
GOV2	81 GB	25.1 m	23.39%	794,889
ClueWeb09	4.0 TB	1.04 b	7.74%	49.2 m
ClueWeb12	4.6 TB	0.73 b	14.71%	39.1 m
NYT-AC	8.3 GB	1.8 m	2.11%	28,493
WaPo	1.6 GB	0.59 m	12.52%	28,471

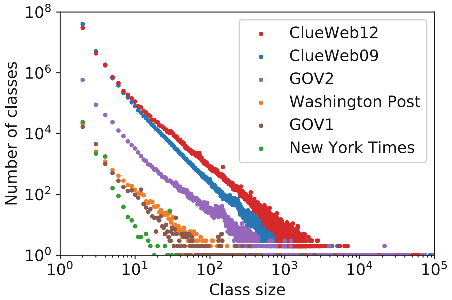

Fig. 1. Overview of the studied corpora (left), and the size of equivalence classes (right).

The two contributions of this paper are (1) a reproduction, and (2) an extension and generalization of the work of Bernstein and Zobel on the effects of content-equivalent documents on search engine evaluation. We independently confirm their findings on the TREC 2004 Terabyte Track, using our own reimplementation of their approach.[1] Thus validated, we go on to apply it to a selection of ad hoc retrieval tracks succeeding the one originally studied by Bernstein and Zobel: the Terabyte Track (2004–2006) [6,7,12], the Web Track (2009–2014) [8–11,13,14], and the recent Common Core Track (2017–2018) [1,2]. Applying the novelty principle causes changes in the evaluation scores of all shared tasks under consideration. These changes do not uniformly spread across participants. A participant who applies the novelty principle independently of others can loose up to 53 positions.

2 Corpus Analysis: Retrieval-Equivalent Documents

Following Bernstein and Zobel, we first analyzed the corpora employed in the shared tasks under consideration by grouping their documents into retrieval equivalence classes. Retrieval equivalence is determined using a fingerprint function based on selected elements of a standard indexing pipeline: the document string is lowercased, all HTML tags, punctuation, and stop words are replaced by a blank, all remaining words are stemmed, and all white space sequences are collapsed. The resulting string is fed to a cryptographic hash function, and the hash value is used as the document's fingerprint. We reimplement these steps using widespread open-source tools (e.g., Anserini [17]). Figure 1 shows the results of our analysis.

Terabyte Track. The Terabyte Track employed the GOV2 corpus, a crawl of web sites hosted under the .gov domain that took place in 2004 [7]. Bernstein and Zobel also analyzed its predecessor, the GOV1 corpus, which is why we do so as well. Although efforts were made to remove duplicates during crawling, this did not include retrieval-equivalent documents. As shown in the table in Fig. 1,

[1] Source code and resources: https://github.com/webis-de/trec-near-duplicates.

23.39% of the GOV2 documents are equivalent to at least one other document for a total of 794,889 equivalence classes. Our results for both GOV1 and GOV2 deviate by only about 1% from those reported by Bernstein and Zobel (22,870 and 865,362 classes, respectively), which, given the corpus sizes, is sufficiently close to say that their experiment can be successfully reproduced. This is further corroborated by the fact that the plot of the distribution of class sizes for GOV2, by visual inspection, has the same characteristics as that of Bernstein and Zobel. The observed differences are due to the lacking descriptions in the original paper of the normalization steps of the fingerprint function. Asking the authors for details was deemed unnecessary given the closeness of fit. Although the organizers of this track took note of the work of Bernstein and Zobel, they did not report on any activities to reduce the impact of near-duplicates [6,12].

Web Track. The Web Track employed the unrestricted web crawls ClueWeb09 and, as of 2013, ClueWeb12. Although the track organizers reported the removal of some duplicate documents up front, our analysis shows that both corpora still contain a large proportion of retrieval-equivalent documents (7.74% and 14.71%, respectively). The corpora are about 40 times larger than the GOV2 corpus, and so are the numbers of equivalence classes.

Core Track. The Core Track employed the New York Times Annotated Corpus (NYT-AC) in 2017, and the Washington Post (WaPo) corpus in 2018. One of this track's goals was to revisit the methodology of constructing evaluation corpora, implementing new ideas to avoid shortcomings of previous ones. The generation of relevance judgments for both corpora has been carefully carried out and documented. Unfortunately, we identify a proportion of retrieval-equivalent documents in the WaPo corpus similar to that of the ClueWeb12.

Altogether, the GOV2 corpus has the largest proportion of retrieval-equivalent documents, followed by ClueWeb12 and WaPo, ClueWeb09 and GOV1, and last the NYT-AC corpus with the least duplicates. Since each track has at least one corpus with significant amounts of duplication, this merits further investigation.

3 Assessment of Content-Equivalent Documents

Bernstein and Zobel consider two documents to be content-equivalent if they convey the same information. To quantify content equivalence, they employ a similarity measure that first fingerprints each document as a set of word 8-grams, and then divides the number of overlapping 8-grams between both fingerprints by the mean of their sizes (previously introduced as S_3 score in [3], and similar to the set-based resemblance measure of Broder [5]). An S_3 score of 0 indicates no syntactic overlap, and 1 retrieval-equivalence. From a user study, Bernstein and Zobel obtain an S_3 threshold of 0.58 above which content-equivalent pairs of GOV2 documents are identified with a precision of 0.95.

We repeated the user study for the GOV2 documents, as well as for the generic ClueWeb web pages and the news articles employed by the other tracks.

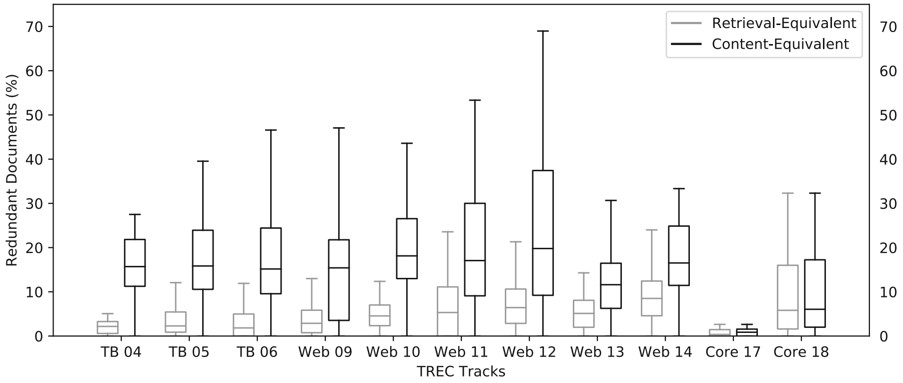

Fig. 2. Each box plots indicates the range of equivalent documents among the relevant documents across the topics of its respective track.

We sampled 100 document pairs per track at random and judged their content equivalence, ensuring that the sample uniformly covers the range of S_3 scores between 0.4 and 1. For the threshold reported by Bernstein and Zobel, we achieved only a precision of 87% on our sample under our interpretation of content equivalence. We hence chose the threshold 0.68 for GOV2 documents, 0.84 for generic ClueWeb pages, and 0.68 for news articles to obtain a precision of 0.95 for all three text genres. Given the threshold difference we obtained for the GOV2 documents, we compared all 50 topics of the Terabyte Track 2004 in detail with the results reported by Bernstein and Zobel and found only two with discrepancies; a reasonable result given the difficulty of repeating a user study.

Following Bernstein and Zobel, we implemented the SPEX algorithm [3] to identify all pairs of documents for which relevance judgments have been collected during one of the eleven editions of the three tracks, which exceed the aforementioned S_3 thresholds, respectively. Figure 2 shows box plots for each of the tracks, contrasting retrieval- and content-equivalent documents, when regarding only documents judged as relevant. Each box plot indicates the range of numbers of equivalent documents across the topics of its corresponding track. Except for the Core Track 2017, all tracks have topics with a high number of equivalent documents. Particularly striking is Topic 194 in the Web Track 2012 for the query `designer dog breeds`: among 47 relevant documents, there are 40 content-equivalent ones derived from the same Wikipedia article.

4 Impact of the Novelty Principle on Retrieval Evaluation

The novelty principle states that a document, though relevant in isolation, is irrelevant if it is content-equivalent to a document the user has already seen in the result list. We quantify the effect of the novelty principle on the shared tasks under consideration. Like Bernstein and Zobel, we removed poorly performing runs—keeping the best 75%—to discount the effect of those runs.

Table 1. The impact of the novelty principle on the ranking of retrieval-systems under the scenarios that: (1) content-equivalent documents are marked as irrelevant, (2) content-equivalent documents are removed by the search engine. We report the average nDCG (avg_{nDCG}), the median (med_I) and maximum (max_I) ranking changes of the ideal participation model, changes in the average nDCG (Δ_{nDCG}), as well as Kendall's τ, and Kendall's τ of the top-5 systems (τ@5).

Track		Runs		Equiv. irrelevant					Equiv. removed		
		#	avg_{nDCG}	Δ_{nDCG}	τ	τ@5	med_I	max_I	Δ_{nDCG}	τ	τ@5
Terabyte	2004	70	0.425	–5.1%	0.96	0.80	–9.0	–19	+0.3%	0.98	1.00
	2005	58	0.586	–3.8%	0.95	0.20	–15.5	–27	+0.8%	0.98	0.80
	2006	80	0.654	–4.4%	0.94	1.00	–29.5	–53	–1.0%	0.94	0.86
Web	2009	71	0.323	–8.9%	0.89	0.80	-8.5	–24	–6.8%	0.91	0.80
	2010	56	0.302	–14.1%	0.49	0.42	–19.5	–39	–9.9%	0.57	0.33
	2011	37	0.341	–9.0%	0.85	0.40	–8.0	–13	–3.4%	0.92	0.80
	2012	28	0.295	–17.3%	0.72	0.61	–9.0	–16	–12.4%	0.81	0.73
	2013	34	0.324	–4.6%	0.86	0.80	–4.0	-8	–1.8%	0.90	0.80
	2014	30	0.380	–7.9%	0.87	1.00	–4.0	–11	–4.5%	0.94	1.00
Core	2017	75	0.560	–0.3%	0.99	1.00	–1.0	–9	+0.1%	1.00	1.00
	2018	72	0.541	–4.3%	0.92	1.00	–11.0	–26	–0.9%	0.93	0.73

We experiment with two strategies to model the novelty principle: local judgment manipulation as per Bernstein and Zobel, and a global judgment manipulation of our own design. In local judgment manipulation, judgments are manipulated for each run independently, so that a document that is content-equivalent to another document is judged irrelevant if the latter appears above in a search results list. Bernstein and Zobel employ only local judgment manipulation, which does not consistently implement the novelty principle.

Consider a ranking that does not contain any document from a given class of relevant, content-equivalent documents. In local judgment manipulation, all documents of that class are considered relevant for that ranking, whereas this is not the case for an alternative ranking that contains all documents of that class, where only one of them would be relevant. To resolve this contradiction, we propose a global judgment manipulation in which all documents of the same equivalence class—except for a representative document—are marked as irrelevant for a query. If a ranking contains documents of a relevant equivalence class, we apply the local judgment manipulation to choose the representative document. Otherwise, we chose a representative at random. In all our experiments, local manipulation amplified score changes compared to global manipulation.

Bernstein and Zobel find that many content-equivalent documents have inconsistent relevance judgments, i.e., one being judged relevant, but not an equivalent one. We confirm this observation for *all* considered tracks. The minimum of 33 inconsistent classes was found in the Core Track 2018, and the maximum of 604 in the Terabyte Track 2004. The inconsistencies are fixed by assigning the entire equivalence class the most frequently occurring judgment.

In their analysis, Bernstein and Zobel examine the impact of the novelty principle on Mean Average Precision (MAP) scores. Meanwhile, the use of MAP has been discouraged [16]. We hence also analyze the novelty principle's impact on nDCG scores, observing much greater changes in MAP scores than for nDCG. Due to space limitations, we only discuss the novelty principle's impact on nDCG scores under global manipulation; a reproduction of the MAP-based analysis is included in our accompanying repository (see link above).

Table 1 shows the impact of the novelty principle on all considered tracks. Regarding the Terabyte Track 2004 (column group "Equiv. irrelevant"), we observe a reduction Δ_{nDCG} of the avg_{nDCG} by 5.1% from 0.425 to 0.403. This reduction may be ignored if it were rank-preserving with respect to the ranking of the participating systems. The original ranking of systems correlates to the cleansed one with 0.96 Kendall's τ. This seems acceptable, but we measure only 0.8 τ@5, regarding only the top five ranks; the best-performing systems are affected more strongly. We inspected how many ranks an "ideal system" I would drop, if it were the only one to remove content-equivalent documents from its rankings: In the median, it would drop 9 ranks (med_I) and in the worst case 19 (max_I). By contrast, if all systems had removed content-equivalent documents from their rankings/runs (column group "Equiv. removed"), the novelty principle's impact on the Terabyte Track 2004 would have been negligible. No ranking changes among the best-performing five systems occur, and the rank correlation of all systems increases to 0.98. Interestingly, even the average nDCG would have increased by 0.3%, caused by the fact that each run is only expected to return one document of a class of relevant content-equivalent documents under global judgment manipulation.

Similarly, Table 1 shows the novelty principle's impact on all other tracks under consideration. A complete discussion as exemplified for the Terabyte Track 2004 is beyond the space limitations; just a few more highlights: It turns out that the novelty principle has a strong impact on the Web Tracks of 2010 and 2012 in terms of Δ_{nDCG}. But we observe the maximum drop of ranks (53, max_I) in the Terabyte Track 2006. Unlike for the Terabyte Track 2004, for the Core Track 2018, if all participants were to remove duplicates ("Equiv. removed"), we still observe a large difference in Kendall's τ@5 of 0.73.

5 Conclusion

We successfully reproduced the work of Bernstein and Zobel [4], confirming their findings on the impact of near-duplicate and content-equivalent documents on the evaluation of the TREC Terabyte Track 2004. In addition, we extended their analysis to all Terabyte Tracks, the Web Tracks, and the two recent Core Tracks, and we improved upon their original implementation of the novelty principle. With the exception of the Core Track 2017, all of the tracks under consideration are (strongly) affected by the presence of content-equivalent duplicates.

Our findings are alarming. Not only are the evaluations carried out thus far invalidated to some extent, they also subdue newcomers: In practice, filtering duplicates from search results is done as a matter of course and without a second thought, and diligent participants may thus never learn that their retrieval systems would have actually outperformed the state of the art. One cannot expect anyone to realize that abstaining from filtering duplicates may result in better performance at TREC—a conclusion one can only draw from an in-depth run analysis. This is a call to action to all track organizers to henceforth take duplicates into account.

References

1. Allan, J., Harman, D., Kanoulas, E., Li, D., Gysel, C.V., Voorhees, E.M.: TREC 2017 common core track overview. In: Proceedings of The Twenty-Sixth Text REtrieval Conference, TREC 2017, Gaithersburg, Maryland, USA, 15–17 November 2017 (2017)
2. Allan, J., Harman, D., Kanoulas, E., Voorhees, E.M.: TREC 2018 common core track overview. In: Notebooks of The Twenty-Seventh Text REtrieval Conference (TREC 2018), Gaithersburg, Maryland, USA, 14–16 November 2018 (2018)
3. Bernstein, Y., Zobel, J.: A scalable system for identifying co-derivative documents. In: Apostolico, A., Melucci, M. (eds.) SPIRE 2004. LNCS, vol. 3246, pp. 55–67. Springer, Heidelberg (2004). https://doi.org/10.1007/978-3-540-30213-1_6
4. Bernstein, Y., Zobel, J.: Redundant documents and search effectiveness. In: Proceedings of the 2005 ACM CIKM International Conference on Information and Knowledge Management, Bremen, Germany, 31 October – 5 November 2005, pp. 736–743 (2005)
5. Broder, A.Z.: On the resemblance and containment of documents. In: Proceedings Compression and Complexity of SEQUENCES 1997, Positano, Amalfitan Coast, Salerno, Italy, 11–13 June 1997, pp. 21–29 (1997)
6. Büttcher, S., Clarke, C.L.A., Soboroff, I.: The TREC 2006 terabyte track. In: Proceedings of the Fifteenth Text REtrieval Conference (TREC 2006), Gaithersburg, Maryland, USA, 14–17 November 2006 (2006)
7. Clarke, C.L.A., Craswell, N., Soboroff, I.: Overview of the TREC 2004 terabyte track. In: Proceedings of the Thirteenth Text REtrieval Conference (TREC 2004), Gaithersburg, Maryland, USA, 16–19 November 2004 (2004)
8. Clarke, C.L.A., Craswell, N., Soboroff, I.: Overview of the TREC 2009 web track. In: Proceedings of The Eighteenth Text REtrieval Conference (TREC 2009), Gaithersburg, Maryland, USA, 17–20 November 2009 (2009)
9. Clarke, C.L.A., Craswell, N., Soboroff, I., Cormack, G.V.: Overview of the TREC 2010 web track. In: Proceedings of The Nineteenth Text REtrieval Conference (TREC 2010), Gaithersburg, Maryland, USA, 16–19 November 2010 (2010)
10. Clarke, C.L.A., Craswell, N., Soboroff, I., Voorhees, E.M.: Overview of the TREC 2011 web track. In: Proceedings of The Twentieth Text REtrieval Conference (TREC 2011), Gaithersburg, Maryland, USA, 15–18 November 2011 (2011)

11. Clarke, C.L.A., Craswell, N., Voorhees, E.M.: Overview of the TREC 2012 web track. In: Proceedings of The Twenty-First Text REtrieval Conference (TREC 2012), Gaithersburg, Maryland, USA, 6–9 November 2012 (2012)
12. Clarke, C.L.A., Scholer, F., Soboroff, I.: The TREC 2005 terabyte track. In: Proceedings of the Fourteenth Text REtrieval Conference (TREC 2005), Gaithersburg, Maryland, USA, 15–18 November 2005 (2005)
13. Collins-Thompson, K., Bennett, P.N., Diaz, F., Clarke, C., Voorhees, E.M.: TREC 2013 web track overview. In: Proceedings of The Twenty-Second Text REtrieval Conference (TREC 2013), Gaithersburg, Maryland, USA, 19–22 November 2013 (2013)
14. Collins-Thompson, K., Macdonald, C., Bennett, P.N., Diaz, F., Voorhees, E.M.: TREC 2014 web track overview. In: Proceedings of The Twenty-Third Text REtrieval Conference (TREC 2014), Gaithersburg, Maryland, USA, 19–21 November 2014 (2014)
15. Fetterly, D., Manasse, M.S., Najork, M.: On the evolution of clusters of near-duplicate web pages. In: 1st Latin American Web Congress (LA-WEB2003), Empowering Our Web, Sanitago, Chile, 10–12 November 2003, pp. 37–45 (2003)
16. Fuhr, N.: Some common mistakes in IR evaluation, and how they can be avoided. SIGIR Forum **51**(3), 32–41 (2017)
17. Yang, P., Fang, H., Lin, J.: Anserini: enabling the use of Lucene for information retrieval research. In: Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, Shinjuku, Tokyo, Japan, 7–11 August 2017, pp. 1253–1256 (2017)