The Power of Anchor Text in the Neural Retrieval Era

Maik Fröbe,[1] Sebastian Günther,[1] Maximilian Probst,[1]
Martin Potthast,[2] Matthias Hagen[1]

[1] Martin-Luther-Universität Halle-Wittenberg
[2] Leipzig University

Abstract In the early days of web search, a study by Craswell et al. [11] showed that anchor texts are particularly helpful ranking features for navigational queries and a study by Eiron and McCurley [24] showed that anchor texts closely resemble the characteristics of queries and that retrieval against anchor texts yields more homogeneous results than against documents. In this reproducibility study, we analyze to what extent these observations still hold in the web search scenario of the current MS MARCO dataset, including the paradigm shift caused by pre-trained transformers. Our results show that anchor texts still are particularly helpful for navigational queries, but also that they only very roughly resemble the characteristics of queries and that they now yield less homogeneous results than the content of documents. As for retrieval effectiveness, we also evaluate anchor texts from different time frames and include modern baselines in a comparison on the TREC 2019 and 2020 Deep Learning tracks. Our code and the newly created Webis MS MARCO Anchor Texts 2022 datasets are freely available.[3]

Keywords: Anchor text · MS MARCO · ORCAS · TREC Deep Learning track

1 Introduction

Almost from the beginning, search engines have exploited the Web's link structure to improve their result rankings. But besides the actual links, also the anchor texts (i.e., the clickable texts of the links) were an important ranking feature, since they "often provide more accurate descriptions of web pages than the pages themselves" [2].

The seminal works of Craswell et al. [11] and Eiron and McCurley [24] from 2001 and 2003 examined two important aspects of anchor text. Craswell et al. showed that anchor text especially helps for navigational queries (i.e., queries to find a specific document [3]). This result explained why commercial search engines heavily used anchor text even though no positive effect was observed in TREC scenarios [27, 49]: more than 20% of the traffic of commercial search engines were navigational queries [3], but hardly any TREC topic was navigational. Eiron and McCurley showed that retrieval against anchor texts yields more homogeneous results than against documents and that anchor texts closely resemble the characteristics of queries. This result later inspired others to use anchor texts as a replacement for proprietary query logs [7, 20, 36, 38].

In the two decades since the studies of Craswell et al. and Eiron and McCurley were published, the Web and the search behavior of users have changed. We thus analyze to

[3] Code and data: https://github.com/webis-de/ECIR-22. Data is integrated in ir_datasets [39]. Data on Zenodo: https://zenodo.org/record/5883456

what extent the original findings can be reproduced on current web crawls and query logs. Additionally, given the recent success of pre-trained transformers [52], we also analyze whether anchor text is still a valuable ranking feature or whether it might be "obsolete" for retrieval pipelines using BERT [43], MonoT5 [44], or DeepCT [18].

As reproducibility scenario for our study, we employ the two available versions of the MS MARCO datasets (3.2 and 12 million documents, 367,013 queries with relevance judgments) [15], the ORCAS query log (18.8 million query–click entries related to MS MARCO documents) [8], and extract anchor texts from Common Crawl snapshots of the last six years to construct the Webis MS MARCO Anchor Texts 2022 dataset: it contains billions of anchor texts for about 1.7 million documents from MS MARCO Version 1 (about 53% of all documents), and for about 4.82 million documents from MS MARCO Version 2 (about 40% of all documents).

The results of our reproducibility study are dichotomous. While we can reproduce Craswell et al.'s observation that anchor text is particularly helpful for navigational queries (details in Section 5), we find substantial differences for the results of Eiron and McCurley. In the MS MARCO scenario, the anchor texts are pretty different to queries (e.g., number of distinct terms) and retrieval against them yields less (not more) homogeneous results than against the content of documents (details in Section 4). We attribute both changes to the fact that Eiron and McCurley conducted their study in the corporate IBM intranet with queries and anchor texts both formulated by employees of IBM, whereas, in our reproducibility scenario, we have "arbitrary" searchers and anchor text authors from the Web. In the reproducibility experiments for the study of Craswell et al., we also evaluate the effectiveness of anchor text from different time frames and include modern baselines in a comparison on the topics of the TREC 2019 and 2020 Deep Learning tracks. The results still confirm the observation that anchor text only slightly improves the effectiveness in TREC scenarios [11, 27, 49]. All our code and data is published under a permissible open-source license.

2 Related Work

Exploiting link structure has a long tradition in IR [16]. Already in 1993, Dunlop and van Rijsbergen [23] used text referring to non-textual objects like images to retrieve those non-textual objects for text queries. McBryan [41] refined this process by only including terms from the clickable texts of links: the *anchor texts*. Anchor texts were later reported to be heavily used by commercial search engines [2, 24] but had no positive effect in TREC scenarios [1, 26, 27, 49]. Craswell et al. [11] resolved this dichotomy by showing that anchor text is particularly useful for navigational queries (i.e., queries to find a specific document [3]) while hardly any TREC topics were navigational.

After Craswell et al.'s result, dedicated shared tasks like homepage finding or named page finding evolved [9, 12, 10] and more and more systems incorporated anchor text for navigational queries. For instance, Westerveld et al. [49] combined anchor text with a document's content, URL, and link count, and Ogilvie and Callan [45] showed that anchor text can also be combined with poor-performing features without harming the overall effectiveness for navigational queries. Since links may "rot" over time [34]—resulting in possibly outdated anchor texts—, several approaches used historical infor-

mation [17] or importance estimation [22, 42] to weight anchor text. Finally, the anchor text source and quantity were shown to be very important. Kamps et al. [29] found that anchor text from the Wikipedia is more effective than anchor text from the general Web while Koolen and Kamps [35] showed that more anchor text led to higher early precision on the TREC 2009 Web track [6], which includes 66 navigational subtopics.

Anchor text became an important retrieval feature also used in lieu of query logs [24, 20, 36, 7, 38]. But with the recent paradigm shift due to transformers [52], the IR community's main focus changed from feature engineering to neural re-ranking and dense retrieval models [30]. The MS MARCO datasets, utilized by the TREC Deep Learning tracks [14, 8], particularly enabled this shift, but since they lack anchor texts, our goal of reproducing the seminal anchor text studies by Craswell et al. [11] and Eiron and McCurley [24] requires the extra effort of collecting anchor texts for its documents.

3 The Webis MS MARCO Anchor Text 2022 Dataset

MS MARCO does not feature anchor texts, and its documents are only sparsely linked. To overcome this shortcoming for the reproduction of the results of Craswell et al. and Eiron and McCurley on MS MARCO, we compile the Webis MS MARCO Anchor Text 2022 dataset by extracting anchor texts from web pages linking to MS MARCO documents found in Common Crawl snapshots. A high recall has been achieved by processing one randomly selected snapshot from each year between 2016 and 2021 (between 1.7–3.4 billion documents each). Unlike Craswell et al. and Eiron and McCurley, we applied the three filtering steps developed by Chen et al. [5] to remove low-quality anchor texts. An anchor text has been omitted, if it consisted of (1) one or more of the manually selected "stop words" 'click', 'read', 'link', 'mail', 'here', and 'open'; (2) more than 10 words, since these are often due to parsing errors; or, if it (2) originated from an intra-site link (i.e., same source and target domain), since anchor texts of inter-site links are usually more descriptive [42]. These filtering steps removed about 50% of all anchor texts pointing to MS MARCO documents.

Processing the total 17.12 billion Common Crawl documents (343 TiB compressed WARC files) on our 3000 CPU Hadoop cluster [48] yielded 8.16 billion anchor texts for MS MARCO documents. A first data analysis revealed that most links point to only a few very popular documents. To obtain a sensible dataset size both for our experiments and future users, we applied min-wise sampling of 1,000 anchor texts for documents that are targeted by more links than that. This stratified sampling still ensured the inclusion of all anchor texts for most of the documents (94% for MS MARCO version 1; 97% for version 2), downsampling only the most popular documents.

Table 1 shows an overview of all extracted anchor texts (column group 'Anchors') and the downsampled subsets for the two MS MARCO versions ('Sample@V1' and 'Sample@V2'). Overall, the combined samples cover 1.70 million documents of Version 1 (53% of all documents) and 4.82 million documents of Version 2 (40%). For each anchor text, our datasets also contain the source URL, the target URL, and the MS MARCO ID of the target document. Besides releasing the dataset to the community, we employ it to reproduce the main findings of Eiron and McCurley [24] (next section) and the retrieval effectiveness results of Craswell et al. [11] (Section 5).

Table 1. The Webis MS MARCO Anchor Text 2022 dataset at a glance. The samples for Versions 1 and 2 (Sample@V1 / V2) include at most 1,000 anchor texts per MS MARCO document.

Common Crawl snapshot			**Anchors**		**Sample@V1**		**Sample@V2**	
Snapshot	Docs	Size	V1	V2	Anchors	Docs cov.	Anchors	Docs cov.
2016-07	1.73 b	28.57 TiB	1.05 b	0.75 b	54.05 m	0.83 m	65.04 m	1.49 m
2017-04	3.14 b	53.95 TiB	0.95 b	0.91 b	61.19 m	1.18 m	94.35 m	2.34 m
2018-13	3.20 b	67.66 TiB	0.83 b	0.68 b	81.24 m	1.27 m	116.59 m	2.45 m
2019-47	2.55 b	53.95 TiB	0.55 b	0.41 b	65.60 m	1.16 m	90.18 m	2.83 m
2020-05	3.10 b	59.94 TiB	0.67 b	0.48 b	78.46 m	1.24 m	108.16 m	3.10 m
2021-04	3.40 b	78.98 TiB	0.52 b	0.36 b	60.62 m	1.14 m	84.93 m	3.18 m
$\sum$	17.12 b	343.05 TiB	4.57 b	3.59 b	207.28 m	1.70 m	341.17 m	4.82 m

4 Properties of Anchor Texts, Queries, and Documents

In 2003, Eiron and McCurley [24] studied properties of anchor texts, queries, and documents on the IBM intranet (2.95 million documents, 2.57 million anchor texts, and 1.27 million queries). They found that anchor texts closely resembled query length, that terms in document titles/bodies and in anchor texts often have different meanings, and that retrieval against anchor text yielded more homogeneous results than against document content. Eiron and McCurley also conducted a study on retrieval effectiveness but we do not reproduce their setup (without relevance judgments) but instead reproduce the retrieval experiments of Craswell et al. [11] with relevance judgments (cf. Section 5).

Analyzing to what extent the similarity of anchor texts and queries that Eiron and McCurley observed can be reproduced in a current retrieval scenario is particularly important, since the observation had inspired others to replace proprietary query logs by anchor texts [7, 20, 38]. We repeat the study of Eiron and McCurley on the MS MARCO Version 1 dataset and the ORCAS query log [8] linked to it. Interestingly, in our "modern" web search scenario with about 27 times more anchor texts (81.24 million in the 2018 subset matching the MS MARCO Version 1 crawling date) and 15 times more queries (18.82 million from ORCAS), we obtain some substantially different results.

Number of Distinct Terms. The plots in Figure 1 show the distributions of the number of distinct terms per anchor text, query, or document title as reported by Eiron and McCurley for their IBM dataset (left plot) and what we observe for MS MARCO (right). While Eiron and McCurley reported the distributions for anchor texts and queries as highly similar, we find them to be rather dissimilar on MS MARCO.

To assess the similarity of the distributions, we calculate the symmetric Jensen-Shannon distance [25] for all pairs (right plot of Figure 2; a distance of 0 indicates equal distributions). The anchor text distributions are very similar for the MS MARCO and the IBM data (distance of 0.10) as are the distributions of anchor texts and queries for the IBM data (0.14). However, on the MS MARCO data, anchor texts and queries are more dissimilar (0.28), probably mainly due to the more "web-like" query distribution: the IBM query distribution is pretty different to the ORCAS queries (distance of 0.34; most IBM queries have one term, most ORCAS queries have three terms, etc.).

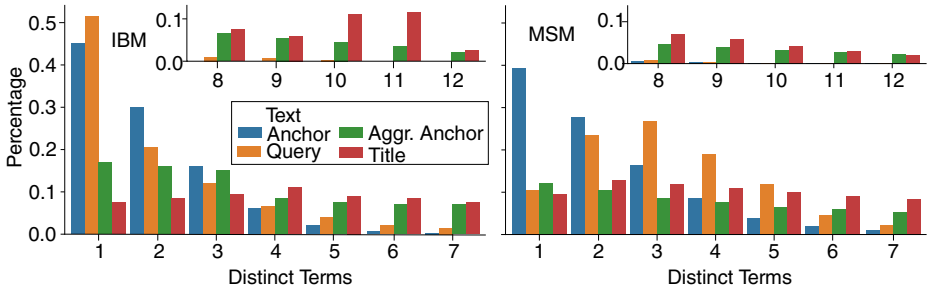
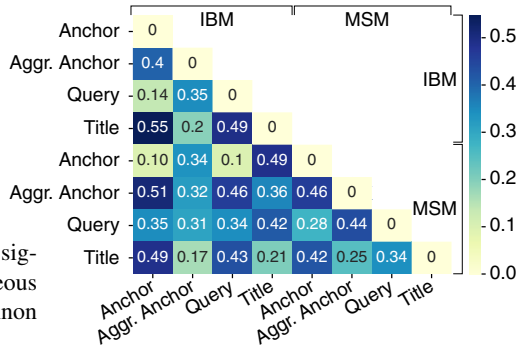


Figure 1. Distributions of the number of distinct terms in anchor texts, queries, document titles, and aggregated anchor texts (all anchors combined that point to a document) on the IBM data (left) and MS MARCO (MSM; right).

Queries w/ significant differences			
p-value	More	Less	Equally
0.05	6,770	1,121	71
0.01	6,764	1,113	85
0.001	6,748	1,098	116

Figure 2. Left: Number of queries with significantly more, less, or equally homogeneous content-based results. Right: Jensen-Shannon distance of all pairs (0 = identical).



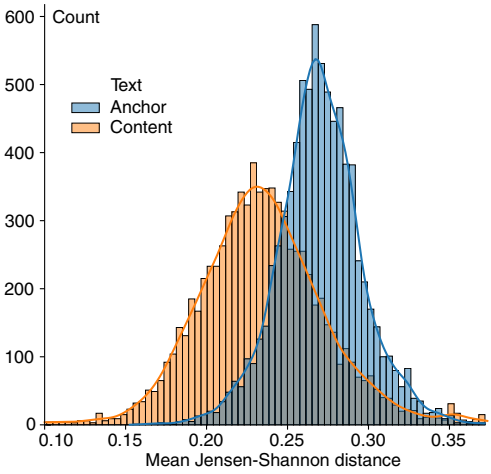
Frequent Terms. Eiron and McCurley also compared the 16 most frequent terms in document titles, queries, and anchor texts and found that these are rather different. Some terms like 'of' are frequent in all types but most terms frequent in one type are rare in the other types. Eiron and McCurley then argued that the different frequencies indicate that anchor texts should be kept separate and not mixed with document content such that methods depending on term frequencies could better exploit the different contexts of a term's frequencies. We can confirm the observed substantial differences also for the MS MARCO scenario. For example, the frequent terms 'you', 'it', and 'are' for titles, 'meaning', 'online', and 'free' for ORCAS queries, as well as 'home', 'university', or 'website' for anchor texts very rarely occur in the other types.

Search Result Homogeneity. Eiron and McCurley reported that most of the queries in their log were navigational (e.g., `benefits` or `travel` to find respective IBM guidelines) and that matching queries in the document content tended to retrieve results for every possible meaning of the query terms while matching only in the anchor texts retrieved more homogeneous results—but in an experiment with only 14 queries.

On 10,000 randomly sampled ORCAS queries, we follow the setup of Eiron and McCurley: we rank the MS MARCO documents by either matching their anchor texts or their content, we remove queries with less than 800 results (7,962 queries remain), and we measure the results' homogeneity using the method of Kilgarriff and Rose [33] to compute the mean Jensen-Shannon distances; distributions shown in Table 2a.

Table 2. Homogeneity of anchor text and content-based search results: (a) mean Jensen-Shannon (JS) distance, (b) result excerpts for query with largest distance

(a) Distributions of mean JS distance.



(b) Ranking excerpts for query `kelly clarkson fan club`.

Rank	Page description
Index: Anchor	**Mean JS distance:** 0.32
1	Homepage of Kelly Clarkson
2	Wikipedia article on Kelly Clarkson
3	Latest News on Kelly Clarkson
⋮	
98	Statistics on movie star Grace Kelly
99	Vacancy by the Kelly Services company
100	Login Page to Facebook
Index: Content	**Mean JS distance:** 0.13
1	News on Kelly Clarkson's career
2	Wikipedia article on a Kelly Clarkson single
3	IMDb biography of Kelly Clarkson
⋮	
98	News article on a Kelly Clarkson album
99	Review of Kelly Clarkson at American Idol
100	Article on weight gain of Kelly Clarkson

In contrast to Eiron and McCurley, we observe that retrieval against document content yields more homogeneous results than against anchor text (cf. Figure 2 (left table); content yields more homogeneous results for more than 6,700 queries). For example, the top-100 content-based results for the query `kelly clarkson fan club` all refer to Kelly Clarkson while the anchor text-based results are more "diverse" (cf. the excerpts in Table 2b). An explanation for the difference to the observation of Eiron and McCurley probably is twofold: (1) our large-scale dataset has rather diverse authors and queries from different searchers while in the IBM data anchor text writers and searchers probably were IBM employees with experience in intranet search, and, probably more importantly, (2) Eiron and McCurley have experimented with 14 queries only.

5 Anchor Text and Retrieval Effectiveness

To reproduce the result of Craswell et al. [11] (that anchor text helps for navigational queries), we compare the effectiveness of traditional and modern content-based retrieval for navigational queries to the effectiveness of focused retrieval in the MS MARCO anchor text datasets. We also further extend the experiment to the queries with judgments from the TREC Deep Learning tracks [14, 13, 15]—all of them informational queries.

5.1 Navigational Queries for MS MARCO

Craswell et al. [11] experimented with three sets of navigational queries to demonstrate the effectiveness of anchor text for web search. For a web crawl with 18.5 million pages, they created 100 navigational queries for random entry pages and 100 navigational queries for random popular entry pages (selected from a manually maintained Yahoo!

list of popular entry pages). Additionally, for a crawl of 0.4 million documents from the domain of the Australian National University, they created 100 navigational queries pointing to academic persons or institutions—we omit those academic queries from our reproduction to focus on general web search.

Following Craswell et al. [11], we created 100 navigational queries for random entry pages and 100 for popular entry pages in the MS MARCO document sets as follows. We extracted all MS MARCO Version 1 documents that potentially are entry pages by applying the respective rules of Westerveld et al. [49] (URL-path must be empty or must be index.html). From the resulting 92,562 candidates, we selected 100 pages at random and 100 documents at random with domains listed in the Alexa top-1000 ranking of 2018 (probable crawl date of the MS MARCO Version 1 document set). To actually create the 200 navigational queries, we manually inspected each of the 200 target pages and formulated a query that searchers would probably use to search for that page. We then also checked whether the page is still present in MS MARCO Version 2 and whether the same navigational query still applies. For 194 query–document pairs, the transfer was easily possible while for the 6 remaining ones we manually had to correct changed URLs (e.g., calendar.live.com $\rightarrow$ outlook.live.com).

5.2 Retrieval Models and Training

For navigational queries, Craswell et al. [11] compared the effectiveness of BM25-based retrieval using document content to BM25-based retrieval using anchor texts. In our reproducibility study, we substantially extend this setup by employing 18 different retrieval systems. We use different anchor text sets to evaluate the effectiveness of anchor text over time and include novel retrieval models that did not exist during the evaluation of Craswell et al. back in 2001.

Seven of the systems in our study retrieve results only against anchor texts using BM25 as the retrieval model; six systems for six different Common Crawl versions of our anchor text dataset and a seventh system that uses all the combined anchor texts. From the other eleven systems that we use for comparison, six solely use the documents' content (one is BM25-based), while the remaining five systems use combinations of document content, anchor text, and ORCAS query–click information [8]. Nine of the eleven comparison systems employ approaches that did not exist during the evaluation of Craswell et al.: DeepCT [18, 19], MonoBERT [43], MonoT5 [44], and LambdaMART [4] (cf. left column of Table 4 for a list of all the 18 systems). For DeepCT, we use different training setups (with or without access to query log information and anchor texts), and for LambdaMART, we use different sets of features (with or without access to query log information and anchor texts) such that we can assess the importance of anchor texts in such models as an additional case study.

We use the Anserini toolkit [51] in our experiments and follow Craswell et al. [11] by not tuning the parameters of BM25—keeping them at Anserini's defaults of $k = 0.9$ and $b = 0.4$. In general, we preprocess queries and the indexed texts via Porter stemming and stopword removal using Lucene's default stopwords for English but for re-ranking documents using MonoT5 and MonoBERT, we follow Nogueira et al. [44] and omit stemming and stopword removal. For all rankers, we break score ties within a ranking via alphanumeric ordering by document ID as implemented in Anserini (given

random document IDs, this leads to a random distribution with respect to other document properties such as text length [37]).

BM25 on Anchor Text. Following Craswell et al. [11], we concatenate all anchor texts pointing to the same target page and index these aggregated anchor text "documents" in dedicated Anserini BM25 indexes for all 14 anchor text samples (6 individual Common Crawl versions and their combination for MS MARCO Version 1 and Version 2; see Table 1). At query time, the actual documents are returned in the order of their retrieved aggregated anchor text "documents". With this setup, we mimic the corresponding baseline of Craswell et al. with the novel aspect that we can compare the retrieval effectiveness for the individual anchor text subsets and their combination.

BM25 on Content. Mimicking the baseline of Craswell et al. [11], we concatenate the title and body of the documents and create a respective Anserini BM25 index.

DeepCT on Content. DeepCT [18, 19] estimates the importance of terms in their context, removing unimportant terms while including multiple copies of important terms. With its focus on precision, DeepCT could be particularly suited for navigational queries. We train three DeepCT models: on the training data of MS MARCO Version 1, on the ORCAS data, and on our combined anchor texts. Interestingly, Dai and Callan [18] designed DeepCT to use anchor text as training data but had not tried it for MS MARCO since no anchor text dataset existed—a gap that we now close with the release of our anchor text data and our respective results for DeepCT.

Following Dai and Callan [18], we compute the importance of a term t in a document d as the fraction of queries with clicks on d that contain t as a query term or the fraction of anchor texts pointing to d that contain t. The three different DeepCT-based systems in our comparison are trained on the queries in the official MS MARCO Version 1 training data, on the queries in the ORCAS data, and on our new anchor text data. To avoid any train/test leakage, we remove the 270,511 MS MARCO documents from the training for which any query or anchor text in the training data contains a term from any of the 200 navigational queries used in our evaluation. The DeepCT systems thus are trained on 249,046 documents for the official MS MARCO training data, on 876,950 documents for the ORCAS data, and on 1,432,621 documents for the combined anchor texts. Following a suggestion of Dai and Callan [19], each document is split into fixed-length passages of 250 terms since working with fixed-length passages is more effective than variable-length original passages [31] (passage splitting done with the TREC CAsT tools[4]). Table 3a shows the characteristics of the training datasets including the number of passages that do not contain any important term.

Table 3b shows the correlations (Kendalls τ and Pearsons ρ) of the term importance scores derived from the three training datasets and also the Jaccard similarity of the term sets with non-zero importance scores. Interestingly, anchor texts and the ORCAS queries lead to more similar scores than the two query sets. Still, the differences for any pair are large enough so that we decided to train and compare three individual models. For the training, we use the implementation of Dai and Callan [18] and follow their suggestions: each DeepCT model is trained with a maximum input length of 512 tokens for 100,000 steps with a batch size of 16 and a learning rate of 2e-5. For

[4]https://github.com/grill-lab/trec-cast-tools

Table 3. (a) Characteristics of the train/test leakage filtered term importance datasets for DeepCT: MS MARCO training data (MARCO), ORCAS data (ORCAS), and the combined anchor texts from the Common Crawls (Anchor). (b) Pairwise comparison of the importance scores' correlations (Kendall's τ, Pearson's ρ) and the Jaccard similarity (J) of terms with non-zero weights.

(a) Term importance training datasets.

Dataset	**Docs**	**Passages**	w/o imp. term
MARCO	0.25 m	2.08 m	0.29 m
ORCAS	0.88 m	8.17 m	0.92 m
Anchor	1.43 m	11.64 m	2.02 m

(b) Comparison of importance scores.

Compared datasets	τ	ρ	J
Anchor vs. ORCAS	**0.39**	**0.61**	**0.53**
ORCAS vs. MARCO	0.35	0.46	0.51
Anchor vs. MARCO	0.26	0.41	0.45

inference, we process all passages with PyTerrier [40] and index the documents (processed passages concatenated again) in an Anserini BM25 index.

MonoBERT and MonoT5 on Content. Since Transformer-based re-rankers recently caused a paradigm shift in information retrieval [52], we include two such systems in our experiments: MonoBERT [43], the first re-ranker based on BERT [21], and MonoT5 [44] that outperforms MonoBERT on MS MARCO and Robust04 [52] by classifying the relevance of a document to a given query using the sequence-to-sequence Transformer T5 [46]. For both, MonoBERT and MonoT5, we use the implementations of PyGaggle[5] and let the default trained castorini/monobert-large-msmarco model and the castorini/monot5-base-msmarco model re-rank the top-100 BM25 results via the maximum score of a passage as the document score.

BM25 on ORCAS. For each document d, we concatenate all queries that have clicks on d in the ORCAS data and index these aggregated query "documents" with Anserini's BM25 implementation. At query time, the actual documents are returned in the order of their retrieved aggregated query "documents". Note that in the TREC 2021 Deep Learning track that uses MS MARCO Version 2 the ORCAS query log should not be used since it might cause train/test leakage.[6] However, since we do not evaluate the effectiveness of retrieval models on the topics of the TREC 2021 Deep Learning track, this potential train/test leakage can not occur in our situation and we can use the ORCAS query log also for MS MARCO Version 2 in our navigational query scenario without the risk of train/test leakage.

LambdaMART. To study the effectiveness of anchor text in combination with other features and to analyze whether the observation still holds that anchor text adds only small or no effectiveness in TREC scenarios [24], we train four LambdaMART [4] models—the state-of-the-art for feature-based learning to rank [4, 28, 50]—on the training and validation labels of MS MARCO Version 1. Again, since we removed the MS MARCO documents from the training for which any query or anchor text contains a term from any of the 200 navigational queries used in our evaluation, there is no risk of train/test leakage. In our setup, we distinguish four feature sources: anchor texts, ORCAS queries, document titles, and document bodies. For each of the four sources, we calculate the following eight feature types using Anserini: TF, TF $\cdot$ IDF, BM25, F2exp, QL, QLJM, PL2, and SPL. Four LambdaMART models are trained

[5]https://github.com/castorini/pygaggle

[6]https://microsoft.github.io/msmarco/TREC-Deep-Learning.html

Table 4. Effectiveness of the 18 retrieval systems in our comparison as mean reciprocal rank (MRR), recall at 3 (R@3), and recall at 10 (R@10) on 100 navigational queries for random entry pages and 100 navigational queries for popular entry pages in MS MARCO version 1 (V1) and version 2 (V2). Bold: highest scores per group.

	Retrieval system	Random@V1			Popular@V1			Random@V2			Popular@V2		
		MRR	R@3	R@10	MRR	R@3	R@10	MRR	R@3	R@10	MRR	R@3	R@10
Anchor	BM25@2016-07	0.61	0.63	0.68	**0.62**	**0.72**	0.83	0.56	0.61	0.64	**0.57**	**0.64**	**0.80**
	BM25@2017-04	0.63	0.70	0.73	0.59	0.67	0.84	0.59	0.68	0.70	0.48	0.56	0.73
	BM25@2018-13	0.70	0.76	0.82	0.54	0.65	0.81	0.62	0.68	0.77	0.47	0.54	0.77
	BM25@2019-47	0.63	0.74	0.78	0.58	0.69	0.84	0.59	0.62	0.76	0.49	0.57	0.78
	BM25@2020-05	0.63	0.72	0.79	0.55	0.66	**0.86**	0.56	0.64	0.71	0.45	0.53	0.74
	BM25@2021-04	0.63	0.73	0.77	0.54	0.66	0.80	0.50	0.54	0.64	0.46	0.55	0.73
	BM25@Anchor	**0.74**	**0.83**	**0.89**	0.55	0.66	0.84	**0.67**	**0.73**	**0.85**	0.39	0.48	0.70
Content	BM25@Content	0.21	0.24	0.36	0.02	0.02	0.03	0.21	0.22	0.42	0.02	0.01	0.04
	DeepCT@Anchor	**0.43**	**0.46**	**0.58**	**0.03**	**0.03**	0.08	**0.43**	**0.49**	**0.66**	0.04	0.03	**0.13**
	DeepCT@ORCAS	0.38	0.42	0.57	0.02	0.00	**0.09**	0.36	0.40	0.60	**0.05**	**0.04**	0.10
	DeepCT@Train	0.27	0.28	0.44	0.02	0.01	0.05	0.32	0.34	0.49	0.03	0.02	0.08
	MonoT5	0.39	0.43	0.53	0.02	0.01	0.05	0.38	0.43	0.57	0.04	**0.04**	0.08
	MonoBERT	0.35	0.37	0.51	0.02	0.01	0.05	0.36	0.41	0.56	0.01	0.01	0.02
Other	BM25@ORCAS	**0.60**	**0.64**	**0.70**	**0.28**	**0.32**	**0.43**	**0.56**	0.59	0.66	**0.28**	**0.33**	**0.44**
	λ-MART@BTOA	0.48	0.55	0.63	0.08	0.07	0.18	0.52	0.57	**0.77**	0.12	0.12	0.21
	λ-MART@BTO	0.41	0.49	0.57	0.07	0.06	0.17	0.49	0.55	0.65	0.08	0.10	0.14
	λ-MART@BTA	0.43	0.51	0.61	0.06	0.06	0.19	0.55	**0.62**	0.75	0.14	0.15	0.24
	λ-MART@BT	0.27	0.31	0.46	0.04	0.03	0.09	0.40	0.44	0.60	0.05	0.05	0.08

with LightGBM [32] on different feature subsets: (1) using all 32 feature types (λ-MART@BTOA), (2) using body, title, and ORCAS (λ-MART@BTO), (3) using body, title, and anchor text (λ-MART@BTA), and (4) using body and title (λ-MART@BT).

5.3 Evaluation

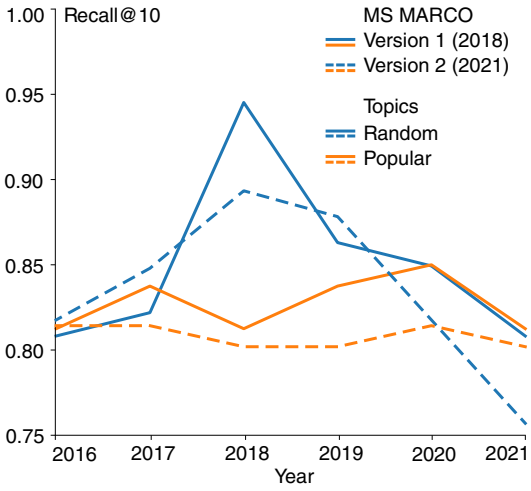
We experimentally compare the effectiveness of the 18 retrieval models on Version 1 and Version 2 of MS MARCO. In a first experiment, we use the above described 200 navigational queries created for MS MARCO to try to reproduce the result of Craswell et al. [11] that anchor text helps to improve MRR for navigational queries. We extend the original study by adding novel aspects like modern neural baselines and by evaluating the effectiveness of anchor text over time. In a second experiment, we then also evaluate the 18 retrieval models on the 88 informational topics from the TREC 2019 and 2020 Deep Learning tracks. Any reported significance test uses $p \leq 0.05$ and includes a Bonferroni correction in case of multiple comparisons.

Retrieval Effectiveness for Navigational Queries. Table 4 shows the retrieval effectiveness for the 200 navigational topics on MS MARCO Version 1 and Version 2.

For queries pointing to random entry pages (columns 'Random@V1' and 'Random@V2'), BM25 retrieval against the combined anchor texts (BM25@Anchor) achieves the best effectiveness scores. While the scores for BM25 on single anchor text snapshots are a little lower (the combination on average has 450 anchor texts per random entry page, each individual snapshot less than 250), the MRR differences from any anchor text-based BM25 retrieval to the best content-based retrieval, DeepCT with impor-

Table 5. (a) Overview of the effectiveness of anchor text on our navigational topics over the crawling period between 2016 and 2021. (b) Overview of the retrieval effectiveness on the TREC Deep Learning topics from 2019 and 2020 where we report nDCG@10 and nDCG@20.

(a) Effectiveness of anchor text over time.



(b) Effectiveness on TREC-DL topics.

	Retrieval systems	DL 2019		DL 2020	
	nDCG@k with $k=$	10	20	10	20
Anchor	BM25@2016-07	0.37	0.32	0.28	0.24
	BM25@2017-04	0.39	0.35	0.29	0.27
	BM25@2018-13	0.35	0.33	0.27	0.25
	BM25@2019-47	0.34	0.31	0.27	0.25
	BM25@2020-05	0.38	0.35	0.30	0.28
	BM25@2021-04	0.35	0.32	0.30	0.26
	BM25@Anchor	**0.41**	**0.38**	**0.34**	**0.32**
Content	BM25@Content	0.51	0.50	0.53	0.53
	DeepCT@Anchor	0.53	0.54	0.55	0.53
	DeepCT@ORCAS	0.52	0.53	0.54	0.54
	DeepCT@Train	0.54	0.53	0.51	0.52
	MonoT5	**0.68**	**0.64**	0.62	**0.63**
	MonoBERT	0.67	0.63	**0.63**	0.62
Other	BM25@ORCAS	0.45	0.41	0.36	0.33
	λ-MART@BTOA	**0.59**	**0.55**	**0.57**	0.57
	λ-MART@BTO	0.57	**0.55**	**0.57**	**0.58**
	λ-MART@BTA	0.57	0.54	**0.57**	0.57
	λ-MART@BT	0.57	**0.55**	0.56	0.56

tance scores trained on the anchor texts (DeepCT@Anchor), are significant. Within the content-based approaches, the recent improvements of neural approaches are also visible for our navigational queries: the score differences of DeepCT trained on anchor texts or ORCAS, of MonoT5, and of MonoBERT to the BM25 content-based retrieval all are statistically significant—as are the differences of the three better LambdaMART models to content-based BM25. Interestingly, also BM25 retrieval on ORCAS queries improves upon all content-only models (all MRR differences are significant), even reaching the effectiveness of some anchor text models. Still, BM25 against the combined anchor texts or the ones from 2018 significantly improves upon BM25 against ORCAS.

For queries pointing to popular entry pages (columns 'Popular@V1' and 'Popular@V2'), all anchor text-based BM25 models are statistically significantly more effective than any other model. Also BM25 on ORCAS queries is significantly better than all non-anchor-based models, again highlighting some similarity of anchor texts to queries.

Altogether, our results confirm the result of Craswell et al. [11] that retrieval against anchor texts is better than retrieval against document content for navigational queries—in our experiments now even including modern neural content-based approaches. However, in almost all of our experimental cases, retrieval for queries pointing to popular entry pages is less effective than for random entry pages. This contradicts an observation of Craswell et al. [11] who reported lower MRR scores for queries pointing to random entry pages than for queries pointing to popular entry pages. For content-based retrieval, the problem is that many other pages "talk" about popular entry pages and mention the respective query terms more often than the actual popular page does.

Retrieval Effectiveness of Anchor Text over Time. To further inspect the impact of crawling time on anchor text effectiveness, we look more deeply into navigational queries that yield at least 100 results against any anchor text snapshot. From the

200 queries, this filtering removes 47 for MS MARCO Version 1 (27 random, 20 popular) and 53 for Version 2 (34 random, 19 popular). Table 5a shows the Recall@10 over time for the remaining queries. For popular pages, there are only slight changes since they always have many anchors pointing to them. As for the random pages, the anchor text crawling time has a larger impact. In particular, the effectiveness peaks at 2018, reflecting the creation date of MS MARCO Version 1. We also observe this peak for Version 2 (crawled in 2021) since we use the same queries that we originally created by sampling pages from Version 1. Not surprisingly, anchor text indexes should thus be refreshed from time to time to match the temporal changes of navigational queries.

Retrieval Effectiveness for Informational Queries. In a final experiment, we evaluate the effectiveness of the 18 retrieval systems on the TREC Deep Learning tracks of 2019 [14] and 2020 [13] on MS MARCO Version 1 (judgments for Version 2 were not yet available)—the respective 88 topics all are informational. Since not all of the 18 systems did contribute to the judgment pools, we removed all unjudged documents from the rankings to mitigate bias as suggested by Sakai [47]. Table 5b shows the resulting nDCG@10 and nDCG@20 scores. Unsurprisingly, the modern Transformer-based MonoT5 and MonoBERT models achieve the overall best scores. For these informational queries, all models solely based on anchor texts or queries are less effective than BM25 on the content of the documents. Still, more anchor text is more effective (BM25@Anchor). Still, the LambdaMART results show that combining content-based retrieval with anchor texts and queries can very slightly improve the effectiveness. Overall, our experiments confirm the earlier observation [24] that anchor text alone is not effective in TREC-style scenarios with a focus on informational queries.

6 Conclusion

In the scenario of the MS MARCO dataset, we have successfully reproduced the result of Craswell et al. [11] that anchor text is very effective for navigational queries. Trying to also reproduce the other seminal anchor text study of Eiron and McCurley [24] we obtained rather different results. We found that the term distributions of anchor texts and queries today are rather dissimilar and that retrieval against anchor text now yields less homogeneous results than retrieval against the document content.

Besides the above positive and negative reproducibility results, another important result of our study is that Transformer-based approaches, be it in re-ranking scenarios or in the DeepCT context of estimating term importance, are less effective for navigational queries than a "basic" anchor text-oriented BM25 retrieval. Identifying navigational queries and switching to anchor text-based retrieval for them instead of neural models might thus improve the retrieval effectiveness of a general retrieval system. However, in the popular TREC Deep Learning tracks, the impact will be rather limited since the Deep Learning tracks do not involve navigational queries. Our code and the newly created Webis MS MARCO Anchor Texts 2022 datasets are freely available.[7]

[7] Code and data: https://github.com/webis-de/ECIR-22. Data is integrated in ir_datasets [39]. Data on Zenodo: https://zenodo.org/record/5883456

Bibliography

[1] Bailey, P., Craswell, N., Hawking, D.: Engineering a Multi-Purpose Test Collection for Web Retrieval Experiments. Inf. Process. Manag. **39**(6), 853–871 (2003)

[2] Brin, S., Page, L.: The Anatomy of a Large-Scale Hypertextual Web Search Engine. Comput. Networks **30**(1-7), 107–117 (1998)

[3] Broder, A.Z.: A Taxonomy of Web Search. SIGIR Forum **36**(2), 3–10 (2002)

[4] Burges, C.J.: From RankNet to LambdaRank to LambdaMART: An Overview. Learning **11**(23-581), 81 (2010)

[5] Chen, W.F., Syed, S., Stein, B., Hagen, M., Potthast, M.: Abstractive Snippet Generation. In: Huang, Y., King, I., Liu, T., van Steen, M. (eds.) Proceedings of the World Wide Web Conference, WWW 2020, San Francisco, CA, USA, April 20-24, 2020, pp. 1309–1319, ACM (Apr 2020), ISBN 978-1-4503-7023-3

[6] Clarke, C.L.A., Craswell, N., Soboroff, I.: Overview of the TREC 2009 Web Track. In: Voorhees, E.M., Buckland, L.P. (eds.) Proceedings of the 18th Text REtrieval Conference, TREC 2009, Gaithersburg, MD, USA, November 17-20, 2009, NIST Special Publication, vol. 500-278, National Institute of Standards and Technology (NIST) (2009)

[7] Craswell, N., Billerbeck, B., Fetterly, D., Najork, M.: Robust Query Rewriting Using Anchor Data. In: Leonardi, S., Panconesi, A., Ferragina, P., Gionis, A. (eds.) Proceedings of the 6th ACM International Conference on Web Search and Data Mining, WSDM 2013, Rome, Italy, February 4-8, 2013, pp. 335–344, ACM (2013)

[8] Craswell, N., Campos, D., Mitra, B., Yilmaz, E., Billerbeck, B.: ORCAS: 20 Million Clicked Query-Document Pairs for Analyzing Search. In: d'Aquin, M., Dietze, S., Hauff, C., Curry, E., Cudré-Mauroux, P. (eds.) Proceedings of the 29th ACM International Conference on Information and Knowledge Management, CIKM 2020, Virtual Event, Ireland, October 19-23, 2020, pp. 2983–2989, ACM (2020)

[9] Craswell, N., Hawking, D.: Overview of the TREC-2002 Web Track. In: Voorhees, E.M., Buckland, L.P. (eds.) Proceedings of the 11th Text REtrieval Conference, TREC 2002, Gaithersburg, MD, USA, November 19-22, 2002, NIST Special Publication, vol. 500-251, National Institute of Standards and Technology (NIST) (2002)

[10] Craswell, N., Hawking, D.: Overview of the TREC 2004 Web Track. In: Voorhees, E.M., Buckland, L.P. (eds.) Proceedings of the 13th Text REtrieval Conference, TREC 2004, Gaithersburg, MD, USA, November 16-19, 2004, NIST Special Publication, vol. 500-261, National Institute of Standards and Technology (NIST) (2004)

[11] Craswell, N., Hawking, D., Robertson, S.E.: Effective Site Finding Using Link Anchor Information. In: Croft, W.B., Harper, D.J., Kraft, D.H., Zobel, J. (eds.) Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2001, New Orleans, LA, USA, September 9-13, 2001, pp. 250–257, ACM (2001)

[12] Craswell, N., Hawking, D., Wilkinson, R., Wu, M.: Overview of the TREC 2003 Web Track. In: Voorhees, E.M., Buckland, L.P. (eds.) Proceedings of the 12th Text REtrieval Conference, TREC 2003, Gaithersburg, MD, USA, November 18-21, 2003, NIST Special Publication, vol. 500-255, pp. 78–92, National Institute of Standards and Technology (NIST) (2003)

[13] Craswell, N., Mitra, B., Yilmaz, E., Campos, D.: Overview of the TREC 2020 Deep Learning Track. In: Voorhees, E.M., Ellis, A. (eds.) Proceedings of the 29th Text REtrieval Conference, TREC 2020, Virtual Event, Gaithersburg, MD, USA, November 16-20, 2020, NIST Special Publication, vol. 1266, National Institute of Standards and Technology (NIST) (2020)

[14] Craswell, N., Mitra, B., Yilmaz, E., Campos, D., Voorhees, E.M.: Overview of the TREC 2019 Deep Learning Track. In: Voorhees, E., Ellis, A. (eds.) 28th International Text Retrieval Conference, TREC 2019, Gaithersburg, Maryland, USA, NIST Special Publication, National Institute of Standards and Technology (NIST) (Nov 2019)
[15] Craswell, N., Mitra, B., Yilmaz, E., Campos, D., Voorhees, E.M., Soboroff, I.: TREC Deep Learning Track: Reusable Test Collections in the Large Data Regime. In: Diaz, F., Shah, C., Suel, T., Castells, P., Jones, R., Sakai, T. (eds.) Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2021, Virtual Event, Canada, July 11-15, 2021, pp. 2369–2375, ACM (2021)
[16] Croft, W.B., Metzler, D., Strohman, T.: Search Engines - Information Retrieval in Practice. Pearson Education (2009), ISBN 978-0-13-136489-9
[17] Dai, N., Davison, B.D.: Mining Anchor Text Trends for Retrieval. In: Gurrin, C., He, Y., Kazai, G., Kruschwitz, U., Little, S., Roelleke, T., Rüger, S.M., van Rijsbergen, K. (eds.) Proceedings of the 32nd European Conference on IR Research, ECIR 2010, Milton Keynes, UK, March 28-31, 2010, Lecture Notes in Computer Science, vol. 5993, pp. 127–139, Springer (2010)
[18] Dai, Z., Callan, J.: Context-Aware Sentence/Passage Term Importance Estimation For First Stage Retrieval. CoRR **abs/1910.10687** (2019)
[19] Dai, Z., Callan, J.: Context-Aware Document Term Weighting for Ad-Hoc Search. In: Huang, Y., King, I., Liu, T., van Steen, M. (eds.) Proceedings of the World Wide Web Conference, WWW 2020, Taipei, Taiwan, April 20-24, 2020, pp. 1897–1907, ACM / IW3C2 (2020)
[20] Dang, V., Croft, W.B.: Query Reformulation Using Anchor Text. In: Davison, B.D., Suel, T., Craswell, N., Liu, B. (eds.) Proceedings of the 3rd ACM International Conference on Web Search and Web Data Mining, WSDM 2010, New York, NY, USA, February 4-6, 2010, pp. 41–50, ACM (2010)
[21] Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: Burstein, J., Doran, C., Solorio, T. (eds.) Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1, pp. 4171–4186, Association for Computational Linguistics (2019)
[22] Dou, Z., Song, R., Nie, J., Wen, J.: Using Anchor Texts with their Hyperlink Structure for Web Search. In: Allan, J., Aslam, J.A., Sanderson, M., Zhai, C., Zobel, J. (eds.) Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2009, Boston, MA, USA, July 19-23, 2009, pp. 227–234, ACM (2009)
[23] Dunlop, M.D., van Rijsbergen, C.J.: Hypermedia and Free Text Retrieval. Inf. Process. Manag. **29**(3), 287–298 (1993)
[24] Eiron, N., McCurley, K.S.: Analysis of Anchor Text for Web Search. In: Clarke, C.L.A., Cormack, G.V., Callan, J., Hawking, D., Smeaton, A.F. (eds.) Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2003, Toronto, Canada, July 28-August 1, 2003, pp. 459–460, ACM (2003)
[25] Fuglede, B., Topsøe, F.: Jensen-Shannon Divergence and Hilbert Space Embedding. In: Proceedings of the 2004 IEEE International Symposium on Information Theory, ISIT 2004, Chicago Downtown Marriott, Chicago, IL, USA, June 27-July 2, 2004, p. 31, IEEE (2004)
[26] Hawking, D.: Overview of the TREC-9 Web Track. In: Voorhees, E.M., Harman, D.K. (eds.) Proceedings of the 9th Text REtrieval Conference, TREC 2000, Gaithersburg, MD,

USA, November 13-16, 2000, NIST Special Publication, vol. 500-249, National Institute of Standards and Technology (NIST) (2000)
[27] Hawking, D., Voorhees, E.M., Craswell, N., Bailey, P.: Overview of the TREC-8 Web Track. In: Voorhees, E.M., Harman, D.K. (eds.) Proceedings of the 8th Text REtrieval Conference, TREC 1999, Gaithersburg, MD, USA, November 17-19, 1999, NIST Special Publication, vol. 500-246, National Institute of Standards and Technology (NIST) (1999)
[28] Hu, Z., Wang, Y., Peng, Q., Li, H.: Unbiased LambdaMART: An Unbiased Pairwise Learning-to-Rank Algorithm. In: Liu, L., White, R.W., Mantrach, A., Silvestri, F., McAuley, J.J., Baeza-Yates, R., Zia, L. (eds.) Proceedings of the World Wide Web Conference, WWW 2019, San Francisco, CA, USA, May 13-17, 2019, pp. 2830–2836, ACM (2019)
[29] Kamps, J., Kaptein, R., Koolen, M.: Using Anchor Text, Spam Filtering and Wikipedia for Web Search and Entity Ranking. In: Voorhees, E.M., Buckland, L.P. (eds.) Proceedings of the 19th Text REtrieval Conference, TREC 2010, Gaithersburg, MD, USA, November 16-19, 2010, NIST Special Publication, vol. 500-294, National Institute of Standards and Technology (NIST) (2010)
[30] Karpukhin, V., Oguz, B., Min, S., Lewis, P.S.H., Wu, L., Edunov, S., Chen, D., Yih, W.: Dense Passage Retrieval for Open-Domain Question Answering. In: Webber, B., Cohn, T., He, Y., Liu, Y. (eds.) Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Virtual Event, November 16-20, 2020, pp. 6769–6781, Association for Computational Linguistics (2020)
[31] Kaszkiel, M., Zobel, J.: Passage Retrieval Revisited. In: Belkin, N.J., Narasimhalu, A.D., Willett, P., Hersh, W.R., Can, F., Voorhees, E.M. (eds.) Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 1997, Philadelphia, PA, USA, July 27-31, 1997, pp. 178–185, ACM (1997)
[32] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T.: LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In: Guyon, I., von Luxburg, U., Bengio, S., Wallach, H.M., Fergus, R., Vishwanathan, S.V.N., Garnett, R. (eds.) Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA, pp. 3146–3154 (2017)
[33] Kilgarriff, A., Rose, T.: Measures for Corpus Similarity and Homogeneity. In: Ide, N., Voutilainen, A. (eds.) Proceedings of the 3rd Conference on Empirical Methods for Natural Language Processing, Palacio de Exposiciones y Congresos, Granada, Spain, June 2, 1998, pp. 46–52, ACL (1998)
[34] Kobayashi, M., Takeda, K.: Information Retrieval on the Web. ACM Comput. Surv. **32**(2), 144–173 (2000)
[35] Koolen, M., Kamps, J.: The Importance of Anchor Text for Ad Hoc Search Revisited. In: Crestani, F., Marchand-Maillet, S., Chen, H., Efthimiadis, E.N., Savoy, J. (eds.) Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2010, Geneva, Switzerland, July 19-23, 2010, pp. 122–129, ACM (2010)
[36] Kraft, R., Zien, J.Y.: Mining Anchor Text for Query Refinement. In: Feldman, S.I., Uretsky, M., Najork, M., Wills, C.E. (eds.) Proceedings of the 13th International World Wide Web Conference, WWW 2004, New York, USA, May 17-20, 2004, pp. 666–674, ACM (2004)
[37] Lin, J., Yang, P.: The Impact of Score Ties on Repeatability in Document Ranking. In: Piwowarski, B., Chevalier, M., Gaussier, É., Maarek, Y., Nie, J., Scholer, F. (eds.) Proceedings of the 42nd International ACM SIGIR Conference on Research and

Development in Information Retrieval, SIGIR 2019, Paris, France, July 21-25, 2019, pp. 1125–1128, ACM (2019)
[38] Ma, Z., Dou, Z., Xu, W., Zhang, X., Jiang, H., Cao, Z., Wen, J.: Pre-training for Ad-hoc Retrieval: Hyperlink is Also You Need. In: 30th ACM International Conference on Information and Knowledge Management (CIKM 2021), ACM (Nov 2021)
[39] MacAvaney, S., Yates, A., Feldman, S., Downey, D., Cohan, A., Goharian, N.: Simplified Data Wrangling with ir_datasets. In: Diaz, F., Shah, C., Suel, T., Castells, P., Jones, R., Sakai, T. (eds.) SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021, pp. 2429–2436, ACM (2021)
[40] Macdonald, C., Tonellotto, N.: Declarative Experimentation in Information Retrieval using PyTerrier. In: Balog, K., Setty, V., Lioma, C., Liu, Y., Zhang, M., Berberich, K. (eds.) ICTIR '20: The 2020 ACM SIGIR International Conference on the Theory of Information Retrieval, Virtual Event, Norway, September 14-17, 2020, pp. 161–168, ACM (2020)
[41] McBryan, O.A.: GENVL and WWWW: Tools for Taming the Web. In: Proceedings of the 1st International World Wide Web Conference, WWW 1994, Geneva, Switzerland, May 25-27, 1994, vol. 341 (1994)
[42] Metzler, D., Novak, J., Cui, H., Reddy, S.: Building Enriched Document Representations Using Aggregated Anchor Text. In: Allan, J., Aslam, J.A., Sanderson, M., Zhai, C., Zobel, J. (eds.) Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2009, Boston, MA, USA, July 19-23, 2009, pp. 219–226, ACM (2009)
[43] Nogueira, R., Cho, K.: Passage Re-ranking with BERT. CoRR **abs/1901.04085** (2019)
[44] Nogueira, R., Jiang, Z., Pradeep, R., Lin, J.: Document Ranking with a Pretrained Sequence-to-Sequence Model. In: Cohn, T., He, Y., Liu, Y. (eds.) Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Virtual Event, November 16-20, 2020, Findings of ACL, vol. EMNLP 2020, pp. 708–718, Association for Computational Linguistics (2020)
[45] Ogilvie, P., Callan, J.P.: Combining Document Representations for Known-Item Search. In: Clarke, C.L.A., Cormack, G.V., Callan, J., Hawking, D., Smeaton, A.F. (eds.) Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2003, July 28-August 1, 2003, Toronto, ON, Canada, pp. 143–150, ACM (2003)
[46] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. J. Mach. Learn. Res. **21**, 140:1–140:67 (2020)
[47] Sakai, T.: Alternatives to Bpref. In: Kraaij, W., de Vries, A.P., Clarke, C.L.A., Fuhr, N., Kando, N. (eds.) Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2007, Amsterdam, The Netherlands, July 23-27, 2007, pp. 71–78, ACM (2007)
[48] Völske, M., Bevendorff, J., Kiesel, J., Stein, B., Fröbe, M., Hagen, M., Potthast, M.: Web Archive Analytics. In: Reussner, R.H., Koziolek, A., Heinrich, R. (eds.) 50. Jahrestagung der Gesellschaft für Informatik, INFORMATIK 2020 - Back to the Future, Karlsruhe, Germany, 28. September - 2. Oktober 2020, LNI, vol. P-307, pp. 61–72, GI (2020), https://doi.org/10.18420/inf2020_05
[49] Westerveld, T., Kraaij, W., Hiemstra, D.: Retrieving Web Pages Using Content, Links, URLs and Anchors. In: Voorhees, E.M., Harman, D.K. (eds.) Proceedings of the 10th Text REtrieval Conference, TREC 2001, Gaithersburg, MD, USA, November 13-16, 2001, NIST Special Publication, vol. 500-250, National Institute of Standards and Technology (NIST) (2001)

[50] Wu, Q., Burges, C.J.C., Svore, K.M., Gao, J.: Adapting Boosting for Information Retrieval Measures. Inf. Retr. **13**(3), 254–270 (2010)
[51] Yang, P., Fang, H., Lin, J.: Anserini: Enabling the Use of Lucene for Information Retrieval Research. In: Kando, N., Sakai, T., Joho, H., Li, H., de Vries, A.P., White, R.W. (eds.) Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2017, Shinjuku, Tokyo, Japan, August 7-11, 2017, pp. 1253–1256, ACM (2017)
[52] Yates, A., Nogueira, R., Lin, J.: Pretrained Transformers for Text Ranking: BERT and Beyond. In: Diaz, F., Shah, C., Suel, T., Castells, P., Jones, R., Sakai, T. (eds.) Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2021, Virtual Event, Canada, July 11-15, 2021, pp. 2666–2668, ACM (2021)