MacAvaney, S. , Macdonald, C. and Ounis, I. (2022) Reproducing Personalised Session Search over the AOL Query Log. In: 44th European Conference on Information Retrieval (ECIR 2022), Stavanger, Norway, 10-14 Apr 2022, pp. 627-640.
(doi: 10.1007/978-3-030-99736-6_42)

https://doi.org/10.1007/978-3-030-99736-6_42

https://eprints.gla.ac.uk/259616/

Date deposited: 21 Jan 2022

Reproducing Personalised Session Search over the AOL Query Log

Sean MacAvaney, Craig Macdonald, and Iadh Ounis

University of Glasgow, United Kingdom
{first.last}@glasgow.ac.uk

Abstract. Despite its troubled past, the AOL Query Log continues to be an important resource to the research community—particularly for tasks like search personalisation. When using the query log these ranking experiments, little attention is usually paid to the document corpus. Recent work typically uses a corpus containing versions of the documents collected long after the log was produced. Given that web documents are prone to change over time, we study the differences present between a version of the corpus containing documents as they appeared in 2017 (which has been used by several recent works) and a new version we construct that includes documents close to as they appeared at the time the query log was produced (2006). We demonstrate that this new version of the corpus has a far higher coverage of documents present in the original log (93%) than the 2017 version (55%). Among the overlapping documents, the content often differs substantially. Given these differences, we re-conduct session search experiments that originally used the 2017 corpus and find that when using our corpus for training or evaluation, system performance improves. We place the results in context by introducing recent adhoc ranking baselines. We also confirm the navigational nature of the queries in the AOL corpus by showing that including the URL substantially improves performance across a variety of models. Our version of the corpus can be easily reconstructed by other researchers and is included in the `ir-datasets` package.

1 Introduction

When released in 2006, the AOL Query Log [27] drew harsh criticism from the media over privacy concerns [4]. Since then, however, it has been an important resource to the research community (e.g., [25, 30]). Even to this day, the AOL Query Log continues to enable studies in analysis of data leaks [13], search autocompletion [14], weak supervision for adhoc search [7, 21], search result personalisation [9, 16], and session-based search [1, 2, 6, 28, 33].

A key limitation of the AOL Query Log is that it does not include document contents; it only provides a user identifier, query text, query date/time, and the URL and rank of clicked documents (if any). This means that for studies that use the logs as a training and benchmark data for tasks like search result personalisation and session search, a document corpus needs to be constructed.

Often the approach used for constructing the corpus is unspecified. To the best of our knowledge, when the approach is specified, it always involves scraping *current* versions of the documents. There are two main problems with this approach. First, given that the contents of web documents are highly prone to change over time, recent versions of the documents may not reflect the contents of the documents as they appeared to the users. Second, this approach impedes reproduciblity and replicability efforts in the area, since the contents of the documents cannot be released publicly due to the potential that they contain copyrighted material.

In this paper, we study the effect that the document corpus used for AOL Query Log experiments has on reproducibility. We start by building a new document corpus that attempts to better reflect the documents present in the AOL Query Log as they appeared when the log was collected. This is accomplished by using the Internet Archive,[1] and thus we refer to our corpus as `AOLIA`. We find that this approach is able to cover far more of the documents that appeared in the AOL Query log (93%) when compared to a commonly-used version of the corpus that was collected in 2017 (55%, shared on request by Ahmad et al. [2]). Based on the timestamps from the Internet Archive, we are confident that the documents in `AOLIA` also better reflect the content of the documents as they appeared at the time, with 86% of the documents coming from during or in the three months prior to the log. We find that the content of the overlapping documents changed substantially in the 11-year period, with 28% of documents having no token overlap in the title (which is often used for session-based search [2, 9, 28]).

We further conduct a reproducibility and replicability study[2] of personalised session search tasks based on the AOL Query Log. We are unable to *reproduce* results using the 2017 version of the corpus, but our *replication* results (using `AOLIA`) are more in line with the original findings. To put the results in context, we also include a neural adhoc ranking baseline, which ultimately outperforms the methods we investigate. We also study the effect of using the document's URL as additional text and find that it improves the performance of all methods we investigate (often by a large margin), further confirming the navigational nature of the queries in the AOL Query Log. In summary, our contributions are:

1. We provide an alternative document corpus (`AOLIA`) for the AOL Query Log based on versions of the documents as they were likely to have appeared at the time the query log was collected.
2. We release artifacts and software such that other researchers will be able to construct `AOLIA` themselves, promoting reproducibility.
3. We study the reproducibility and replicability of three session-based search approaches, and find that using `AOLIA` alone can improve the performance of session-based search systems due to higher-quality documents, and that the training and evaluation datasets constructed from `AOLIA` can be considerably larger due to the increased coverage of the dataset.

---

[1] https://archive.org/   [2] ACM version 1.1 definitions of reproducibility and replicability: https://www.acm.org/publications/policies/artifact-review-and-badging-current.

The remainder of this paper is organised as follows. In Section 2 we provide additional background information about the problem. Then, Section 3 details our process for constructing `AOLIA`. Section 4 provides a comparison between `AOLIA` and a version produced in 2017 that is used by several recent works. Section 5 then focuses on reproducing prior works using `AOLIA`. Finally, Section 6 details the limitations of our approach, and Section 7 draws final conclusions.

2 Background

Past works that make use of the AOL Query Log use recent versions of the log's clicked documents. Because the content of web pages can change over time, using recent versions necessitates a filtering process, which removes query-document pairs that are no longer relevant. For instance, Ahmad et al. [2] reports *"...in our preliminary experiments, we observed that many recorded clicks do not have lexical overlap concerning the queries. One possible reason is that we crawled the recorded clicks from the AOL search log in 2017 and many of the clicked documents' content updated since 2006 when the AOL log was recorded."*

Several alternatives to The Internet Archive exist as sources of data for a reproducible AOL corpus. Although the Common Crawl[3] would provide a more comprehensive corpus (i.e., it includes a more natural selection of documents, rather than only documents clicked by the user), we show in Section 3 that the AOL corpus at the time likely did not contain a representative sample of documents from the web, but rather focused heavily on home pages. Moreover, since the oldest version of the Common Crawl is from 2008–09, the content of the documents may already have changed since the time of the log. Finally, the size of the relevant archives (hundreds of terabytes) could add substantial difficulty in downloading and working with the data. The ClueWeb 2009[4] and 2012[5] corpora would be another option, and are appealing given that many research groups already have a copy of them. However, like the Common Crawl, they reflect the contents of documents several years after the log was constructed. Furthermore, there is low coverage of the target URLs in the ClueWeb corpora.

Other efforts investigate the stability of using mutable web resources as document corpora in IR. McCreadie et al. [23] find that naturally-occurring deletions from the Twitter corpus used by the TREC Microblog tasks do not have a substantial effect on the results of experiments that use the corpus. However, the situation for general web pages is different because the content can change over time (tweets can only be deleted, not updated). Despite these findings, Sequiera and Lin [31] investigate the use of the Internet Archive as an alternative source of data for the TREC 2013–14 Microblog corpus. Our work not only differs in terms of the document corpus targeted, but also the download mechanism; the Twitter stream they use is conveniently bundled by month by the Internet Archive, whereas there is no such bundle available for the documents present in the AOL Query Log. Consequently, the steps involved to build our version of the AOL corpus are necessarily more complicated.

---

[3] https://commoncrawl.org/   [4] https://lemurproject.org/clueweb09/
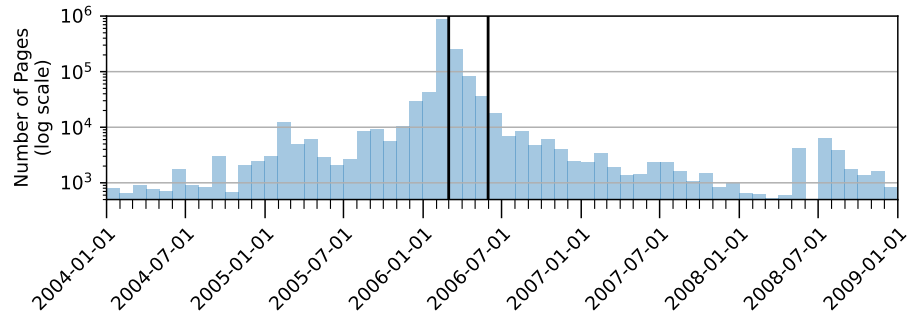[5] https://lemurproject.org/clueweb12/

Fig. 1. Distribution of the archive dates of the web pages in the scraped collection from 2004–2008. The vast majority of pages are recovered from the time period during or right before the AOL Query Log (marked by vertical lines).

3   Reconstructing the AOL Document Corpus

In this section, we reconstruct a document corpus that better reflects the documents as they appeared to the users at the time. Through this process, we create artifacts and software that others can use to construct this dataset themselves, further promoting reproducibility in this area.

We start by building a set of all unique URLs that appear in the AOL Query Log.[6] Importantly, we acknowledge that this only represents documents that users clicked; the full list of documents indexed by AOL at the time is not available. This process results in 1,632,620 unique URLs. Nearly half (48.7%) of the URLs were only clicked a single time. All but 15 URLs specify either `http` or `https` URI schemes (14 specify `ftp` and one specifies `about`). 98.4% of the URLs refer to the home page of a website (i.e., have no path), which suggests that the search engine primarily functioned as a navigational tool at the time.

We then query archive.org's WayBack Machine's availability API[7] to request a version of the page as it appeared as close as possible to 1 March 2006 (the beginning of the AOL Query Log). Remarkably, we find that 93% of URLs were archived. Fig. 1 shows the distribution of the dates of the archived pages. The vast majority of the found URLs (84%) are from the period during or in the three months prior to the log (January to May 2006). A further 8% are from before 2006, and a total of 96% of pages have an archived copy before 2007. Based on these dates, we feel that the corpus represents a reasonable approximation of the documents present in the query log at the time it was collected.

We then fetch the archived versions of the documents and parse the resulting HTML using a libxml2-based parser. The title and body text are extracted (discarding content that appears in non-content tags, such as `<script>`). A small number of documents (0.3%) encountered either parsing errors or persistent

---

[6] http://www.cim.mcgill.ca/~dudek/206/Logs/AOL-user-ct-collection/aol-data.tar.gz
[7] API Endpoint: https://archive.org/wayback/available

Table 1. Top languages present in the `AOLIA` corpus, compared to the prevalence of the language of queries in the log.

Language	Corpus	Queries	Language	Corpus	Queries
English	92.5 %	79.6 %	Japanese	0.4 %	0.1 %
French	1.6 %	3.2 %	Portuguese	0.3 %	0.8 %
Spanish	1.4 %	2.1 %	Dutch	0.3 %	0.9 %
German	1.1 %	2.6 %	Russian	0.2 %	0.4 %
Italian	0.5 %	1.8 %	All Others	1.5 %	8.5 %

download errors. We discard these documents. The median title length is 5 tokens (interquartile: (3, 9)), while the median body length is 198 tokens (interquartile: (47, 454)). As is the case for web content, some documents are substantially longer (up to 1.7M tokens). When compressed, the corpus is 3.4G in size. Using a FastText [12] language classifier[8] over the document title and body, we find that the vast majority of documents (92.5%) likely contain English text, as expected. Table 1 presents a further breakdown of the top languages in `AOLIA`. The breakdown is similar to that of the queries that appear in the log, when considering that language identification is more prone to errors for short texts like keyword queries.

Though we cannot distribute the contents of this corpus directly due to potentially copyrighted content, we take the following steps to facilitate reproducibility using this dataset:

1. We publicly release a mapping of the Internet Archive URLs so that others can fetch the same versions of the original documents.[9]
2. We provide software to download and extract the contents of these documents.[10]
3. We include a new `aol-ia` dataset in the `ir-datasets` [20] package, which provides easy access to this document corpus and the AOL log records. The package automatically downloads the log records from a public source, the Internet Archive mapping (from 1), and provides instructions to the user on how to run the extraction software (from 2). Once built, the dataset can easily be used by tools like PyTerrier [22] and OpenNIR [17].

4   Comparing `AOLIA` with `AOL17`

In this section, we compare `AOLIA` with a version of the corpus that used more recent versions of the documents present in the log. Specifically, we use the version first used by Ahmad et al. [1], which uses documents from the AOL Query Log as they appeared in 2017 (so we call this corpus `AOL17`). This corpus has been used by other works (e.g., [2, 9, 28]). We treat `AOL17` as a representative example

---

Table 2. Comparisons of URLs present in the `AOLIA` and `AOL17` datasets. The Total column indicates the percentage of all URLs present in the AOL corpus.

	Count	Total
$|\texttt{AOLIA}|$	1,525,524	93.4%
$|\texttt{AOL17}|$	897,984	55.0%
$|\texttt{AOL17} \setminus \texttt{AOLIA}|$	12,207	0.7 %
$|\texttt{AOLIA} \setminus \texttt{AOL17}|$	639,747	39.2 %
$|\texttt{AOLIA} \bigcup \texttt{AOL17}|$	1,537,731	94.2 %
$|\texttt{AOLIA} \bigcap \texttt{AOL17}|$	885,777	54.3 %

of a contemporary version of the AOL corpus, noting that more recent versions of the corpus are likely to diverge even further from the original documents.

Table 2 provides a comparison between the URLs present in the two datasets. In terms of absolute coverage, `AOLIA` provides a high (albeit still incomplete) coverage of 93%. Meanwhile, `AOL17` contains only 55% of the URLs found in the log. There are roughly 12k URLs found in `AOL17` but not in `AOLIA`. *Content pages* (i.e., non-homepages) are over-represented among these documents, constituting 13% (1,637) of pages (compared to 2% of the overall corpus). `AOLIA` compensates for this disparity simply by virtue of size, filling in 639,747 documents missing from `AOL17` (9,803 of which are content pages). Even though adding missing documents to `AOLIA` from `AOL17` would increase the total coverage from 93.4% to 94.2%, doing so would reduce reproducibility, since those documents may contain copyrighted material and therefore cannot be distributed publicly.

We now dig into the characteristics of the 885,777 documents that overlap between the corpora. Fig. 2 presents the Jaccard similarity between the set of title tokens[11] present in each version of the document. Only 17% of the titles have a perfect token overlap. Among these, 87% are exact case-insensitive sequence matches, with typical differences being the replacement or addition of punctua-

---

[11] Tokens considered are case-folded, alphanumeric strings separated by whitespace or punctuation.
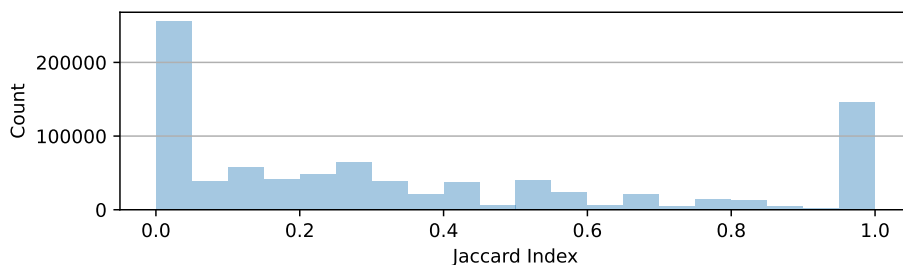

Fig. 2. Distribution of Jaccard similarities over title tokens from overlapping documents in `AOLIA` and `AOL17`.

Table 3. Examples of different titles found on pages between the `AOLIA` and `AOL17` corpora and their corresponding Jaccard index.

#	AOLIA	AOL17	Jac.
1	Welcome To Atlanta Music Group !	Welcome to Atlanta Music Group	1.0
2	Vinopolis Wine Shop - Portland , Oregon	Vinopolis Wine Shop | Portland , Oregon	1.0
3	Mechanics Savings Bank	Mechanics Savings Bank - Mechanics Savings Bank	1.0
4	Indalo Productions	InMotion Hosting	0.0
5	Kennebec Valley Organization – Home Page	*(empty)*	0.0
6	UK TV Guide	Homepage | UKTV	0.0
7	nutone // welcome	Nutone Records - Home | Facebook	0.2
8	Vedanta Press and Catalog	Books on Vedanta Philosophy	0.1
9	Venning Graphic Utilities for blending images	Venning Graphic Utilities	0.5
10	Steinway Musical Instruments , Inc .	Steinway Musical Instruments - Steinway & Sons	0.6

tion in the titles, but sometimes involves the repetition of words. Table 3 shows such examples in rows 1–3.

The majority of documents (53%) have at most a Jaccard index of 0.25, indicating low overlap, with 28% having no overlap at all. In the cases when no overlap is present, semantically dissimilar content is often present, such as a placeholder document or a replacement with information about the web hosting provider. In semantically dissimilar cases, the queries that resulted in clicks for these documents are usually no longer relevant in `AOL17`. For instance, the query "indalo" resulted in a click of the documented represented by #4 in Table 3, which is a reasonable document in `AOLIA` but not in `AOL17`. In some cases, the content is semantically similar, such as in example #6, and still likely a reasonable document for associated queries. Based on a manually-annotated random sample of 100 documents with Jaccard similarities of 0, 23 appeared to be semantically-related (i.e., would likely satisfy similar information needs), while the remainder were not.

5 Reproduction and Replication

The AOL Query Log has been used for training and evaluating numerous search tasks. In this section, we explore the effect of `AOLIA` on one such task: session-based personalisation. In this setting, a user's sequence of searches and clicks are broken into sessions (or tasks [11]). Within each session, the prior queries and clicks can act as additional context to help disambiguate the information needs of a query. For instance, if a user searches for "cars" followed by "jaguar", it is reasonable to tailor the results for the second query towards the luxury car brand rather than the animal. Although a few datasets are available for training and/or evaluating these systems (e.g., TREC Sessions [15]), the AOL Query Log remains a popular (and often exclusive) choice for conducting these experiments.

5.1 Methods

We focus on three neural session-based personalisation techniques: M-NSRF [1], M-MatchTensor [1], and CARS [2]. We select these methods because numerous

recent works in the area use them as baselines (e.g., [6, 28, 32, 33]). All three models function as multi-task models, jointly learning to predict both document relevance and to predict the next query in the sequence. The models differ in the neural network architecture used to accomplish this. M-MatchTensor adapts the multi-task approach to the MatchTensor model [10], where MatchTensor builds a query-document similarity matrix between LSTM-encoded query and documet text and aggregates the results using CNN filters and max-pooling. M-NSRF encodes the query and document in separate bi-directional LSTM networks and combines them using a feed-forward layer to produce ranking scores. CARS builds upon M-NSRF by modeling the session interactions hierarchically using attention-based LSTM networks. For these three approaches, we use the authors' released code[12] with default parameters. In line with the code, the number of training iterations is tuned on dev data.

In addition to the above task-specific methods, we include three additional adhoc ranking baselines to put the results in context. First, we use the Terrier [26] BM25 implementation to re-rank the candidate documents with default BM25 parameters. This corresponds to the (unspecified) BM25 baseline conducted in [2]. Furthermore, in light of recent findings in *adhoc* retrieval, we also include two neural re-ranking baselines based on the T5 model [24, 29]. Given that transferring relevance signals from one dataset/task to another using contextualised language models has generally been shown to be an effective technique (e.g., [18]), we include a T5 "transfer" baseline. This version is tuned on the MS MARCO dataset [3]. We also use a "tuned" baseline, which continues model tuning on the AOL session data from the MS MARCO checkpoint (batch size 8, learning rate 5×10^{-5}). In line with Nogueira et al. [24], we simply train for a fixed number of batches without tuning this or other settings (here, 10,000 batches). For all three adhoc ranking baselines, we use the PyTerrier [22] implementation.

5.2   Experimental Settings

We test the above six systems on three settings. (1) Using both sessions and documents from the `AOL17` dataset. Here, we use the sessions constructed and provided by Ahmad et al. [2]. This reflects the original experimental setup and is therefore a study of reproducibility. (2) Using the sessions from `AOL17` (provided by [2]), but replacing the document titles with those from `AOLIA`. In the case where there is not a corresponding document in `AOLIA`, we leave the title blank (blanks present in 1.8% of documents across training, dev, and test sets). Since the same sessions are used as in (1), these results are directly comparable and isolate the impact of the document text itself. (3) Using sessions from `AOLIA`. In this setting, we re-create sessions as described in [2] (using same session elimination strategy and date ranges in each split), but using the full `AOLIA` corpus. Due to the higher coverage of documents, this results in considerably more data and longer sessions across all three splits. Table 4 presents the characteristics of the

---

[12] https://github.com/wasiahmad/context_attentive_ir

Table 4. Session search dataset characteristics provided by `AOL17` compared to `AOLIA`.

	Train			Dev			Test		
	AOL17	AOLIA	%	AOL17	AOLIA	%	AOL17	AOLIA	%
# Sessions	219,748	311,877	+42 %	34,090	49,522	+45 %	29,369	50,944	+73 %
# Queries	566,967	1,099,568	+94 %	88,021	170,095	+93 %	76,159	167,497	+120 %
Avg. Queries per Session	2.58	3.53	+37 %	2.58	3.43	+33 %	2.59	3.29	+27 %

sessions built by each dataset. Both (2) and (3) are replicability studies because the experimental setting differs from the original paper (a different corpus is used than the original papers).

As is commonplace for the task, we use the document title for the document content. Given that many of the queries are navigational in nature, we also test a variant of each of the 3 above settings that also appends the tokenised URL to the title, which can allow models to distinguish between pages that have the same title content and to match queries that ask for a specific URL.

In all three settings, each of the supervised methods are each trained, tuned, and tested using the data from the corresponding setting, while the unsupervised (BM25) and transfer (T5) baselines are simply run on the test set without tuning. We evaluate the results using MAP, MRR[13], which are measures commonly-used for evaluation of this task. We calculate the measures using the `trec_eval` implementation provided by `ir-measures` [19]. Note that this evaluation tool differs from the original work, which used their own implementation of the measures. Qu et al. [28] notes that this can result in differences in measures due to tie-breaking behaviour. We conduct significance tests between all pairs of systems within each setting (paired t-test, $p < 0.05$, with Bonferroni correction). We do not use a tool like `repro_eval` [5] to compare our results with those from the original papers because the rankings provided from the original papers are not available.

5.3   Results

Table 5 presents the results for the three settings when using the document title as its contents. In Setting (1), where we use the same data and code as [2], we are unable to reproduce the performance reported by Ahmad et al. [2]. However, we note that the performance for CARS is not far from the results reported by Qu et al. [28], who report that the discrepancies with the original work are due to using the `trec_eval` measure implementation. Setting (2), where we use `AOLIA` document titles, yields performances closer to those reported by Ahmad et al. [2], with the CARS model outperforming M-MatchTensor and M-NSRF. Finally, when using both sessions and documents from `AOLIA` (3), M-NSRF outperforms the other two session search methods. In short, we validate the findings of Ahmad et al. [2] that CARS significantly outperforms M-MatchTensor and M-NSRF, but only when using `AOL17` sessions and `AOLIA` documents; in the other two

---

[13] Though this measure has been criticised [8], we report it to compare with past work.

Table 5. Comparison of personalised session search baselines using various versions of the datasets. Note that the performances of systems using `AOL17` and `AOLIA` sessions cannot be directly compared. The top results we measure for each setting are bold (i.e., not including results reported by others). Non-significant differences between pairs of runs within a setting are indicated with superscript letters (paired t-test, $p < 0.05$, Bonferroni correction).

	Model	MAP	MRR	P@1
	(1) Sessions: AOL17, Documents: AOL17			
a	BM25 (unsupervised)	0.2457	0.2554	0.1454
	- from [2]	0.230	0.206	0.206
b	T5 (transfer)	0.3553	0.3649	0.2242
c	T5 (tuned)	**0.4538**	**0.4640**	**0.3001**
d	CARS	^{e}0.4280	^{e}0.4390	^{e}0.2787
	- from [2]	0.531	0.542	0.391
	- from [28]	-	0.4538	0.2940
e	M-MatchTensor	^{d}0.4335	^{df}0.4444	^{df}0.2830
	- from [2]	0.505	0.518	0.368
f	M-NSRF	0.4410	^{e}0.4521	^{e}0.2904
	- from [2]	0.491	0.502	0.391
	(2) Sessions: AOL17, Documents: AOLIA			
a	BM25 (unsupervised)	0.2942	0.3044	0.1914
b	T5 (transfer)	0.4228	0.4337	0.3021
c	T5 (tuned)	**0.5115**	**0.5223**	**0.3745**
d	CARS	0.4998	0.5107	0.3630
e	M-MatchTensor	^{f}0.4848	^{f}0.4961	^{f}0.3493
f	M-NSRF	^{e}0.4911	^{e}0.5023	^{e}0.3495
	(3) Sessions: AOLIA, Documents: AOLIA			
a	BM25 (unsupervised)	0.2413	0.2413	0.1462
b	T5 (transfer)	0.3620	0.3620	0.2260
c	T5 (tuned)	**0.4171**	**0.4171**	**0.2650**
d	CARS	0.3784	0.3784	0.2294
e	M-MatchTensor	0.3572	0.3572	0.2133
f	M-NSRF	0.4009	0.4009	0.2534

settings, we draw the conclusion that the (simpler) M-NSRF model significantly outperforms the other two approaches.

Across all three settings, however, the adhoc (i.e., session unaware) tuned T5 model outperforms all other methods. We acknowledge that comparing T5 with CARS, M-MatchTensor, and M-NSRF is not a completely fair comparison; the T5 model benefits from a much larger model and extensive pre-training, while CARS, M-MatchTensor, and M-NSRF benefit from access to past queries and clicks within the session. Techniques for adapting contextualised language models like T5 for session search have been explored in [28], though in pilot studies we had difficulty training effective models using the released code. We also note that although T5 benefits from tuning on the target domain, it can still perform reasonably well—especially in Setting (3), where the `AOLIA` documents and sessions are used.

Table 6. Results when including the tokenised URL in addition to the title. The Δ column indicates the improvements compared to the results without the URL (Table 5). Non-significant differences between pairs of runs within a setting are indicated with superscript letters (paired t-test, $p < 0.05$, Bonferroni correction).

Model	MAP	Δ	MRR	Δ	P@1	Δ
(1) Sessions: AOL17, Documents: AOL17						
a BM25 (unsupervised)	0.3204	+0.0747	0.3314	+0.0760	0.1991	+0.0537
b T5 (transfer)	0.5023	+0.1470	0.5135	+0.1486	0.3572	+0.1330
c T5 (tuned)	**0.7074**	**+0.2536**	**0.7190**	**+0.2550**	**0.6201**	**+0.3200**
d CARS	^{f}0.6530	+0.2250	^{f}0.6643	+0.2253	^{f}0.5493	+0.2706
e M-MatchTensor	0.6756	+0.2421	0.6871	+0.2427	0.5784	+0.2954
f M-NSRF	^{d}0.6634	+0.2224	^{d}0.6745	+0.2224	^{d}0.5602	+0.2698
(2) Sessions: AOL17, Documents: AOLIA						
a BM25 (unsupervised)	0.3484	+0.0542	0.3591	+0.0547	0.2360	+0.0446
b T5 (transfer)	0.5400	+0.1172	0.5514	+0.1177	0.3959	+0.0938
c T5 (tuned)	**0.7071**	**+0.1956**	**0.7183**	**+0.1960**	**0.6153**	**+0.2408**
d CARS	0.6665	+0.1667	0.6774	+0.1667	^{e}0.5660	+0.2030
e M-MatchTensor	^{f}0.6538	+0.1690	^{f}0.6654	+0.1693	^{df}0.5569	+0.2076
f M-NSRF	^{e}0.6520	+0.1609	^{e}0.6632	+0.1609	^{e}0.5501	+0.2006
(3) Sessions: AOLIA, Documents: AOLIA						
a BM25 (unsupervised)	0.2997	+0.0584	0.2997	+0.0584	0.1790	+0.0328
b T5 (transfer)	0.4260	+0.0640	0.4260	+0.0640	0.2693	+0.0433
c T5 (tuned)	**0.5679**	+0.1508	**0.5679**	+0.1508	**0.4418**	+0.1768
d CARS	0.5360	+0.1576	0.5360	+0.1576	0.4082	+0.1788
e M-MatchTensor	0.5458	**+0.1886**	0.5458	**+0.1886**	^{f}0.4297	**+0.2164**
f M-NSRF	0.5575	+0.1566	0.5575	+0.1566	^{e}0.4336	+0.1802

Table 6 presents the results in each setting when appending the URL to the document text. We observe that in every case, this additional feature improves ranking effectiveness, sometimes by a considerable margin (up to +0.25 in MAP, +0.26 in MRR, and +0.32 in P@1). These findings underscore the importance of including this signal when queries are often navigational in nature.

Overall, we find that these experiments provide further evidence that AOLIA is well-constructed and useful. Between Settings (1) and (2), we see a consistent boost in ranking effectiveness across several models. Since the only thing we change between these settings is the document text, it suggests that the texts in AOLIA are more in line with the preferences of the users. Our experiments using T5 and URL features suggest that more care should be taken in future session-based search studies to construct evaluation data that focus on information needs that are less navigational in nature, as these can be addressed simply using established adhoc approaches and navigational signals.

6 Limitations

One limitation of using The Internet Archive are actions that the organisation takes in response to copyright claims made against archived content, which can

effectively remove documents from the archive.[14] Over an approximately one-month window, 51 documents originally present in a prior version of `AOLIA` were no longer available on the Internet Archive, presumably due to this policy. Though this has the potential for knock-on effects downstream, prior work [23] indicates that it will likely have little effect. Specifically, McCreadie et al. [23] find that the effect of a far greater proportion of documents being deleted from the TREC Microblog 2011 corpus had a minimal effect on system evaluation, hence we expect the same to be true for `AOLIA`. Furthermore, when compared with the vast proportion of documents missing in more recent versions of the AOL corpus (e.g., `AOL17` is missing 746,843 documents), the potential for several hundred removed documents per year seems preferable. Nevertheless, future work studying these effects may be warranted.

Although `AOLIA` improves the coverage and contents of documents that appear in the log, it does not attempt to fill in other documents that may have appeared in the corpus. In this way, it does not reflect a realistic sample of documents that likely existed in the entire AOL corpus at the time; it is more akin to the MS MARCO v1 passage corpus [3] (which only includes passages that were presented to annotators) than to corpora like ClueWeb (which includes documents scraped using typical web crawling techniques). As noted in Section 2, available web crawls would provide a less accurate picture of the documents as they appeared to the AOL users. Therefore, despite this limitation, we believe `AOLIA` is still valuable in many practical experimental settings.

7   Conclusions

In this work, we studied problems surrounding reproducibility of the AOL Query Log's corpus. We started by carefully constructing a new version of the corpus that better reflects the documents as they appeared to the users at the time of the log. We found that our approach increases the coverage of documents in the log considerably, when compared to a version that scraped documents eleven years after the log. We further found that the contents of documents are prone to considerable change over time, with the majority of document titles having very low token overlap between versions. When reproducing prior results for session search, we find that our new corpus improves the effectiveness across a variety of models (likely attributable to more realistic documents), and brings benchmarks based off the AOL Query Log more in line with adhoc ranking methods. We made access to our new version of the AOL corpus easily available to assist in future reproducibility efforts.

---

[14] See their official copyright policy here: https://archive.org/about/terms.php

References

1. Ahmad, W.U., Chang, K.W., Wang, H.: Multi-task learning for document ranking and query suggestion. In: ICLR (2018)
2. Ahmad, W.U., Chang, K.W., Wang, H.: Context attentive document ranking and query suggestion. In: SIGIR (2019)
3. Bajaj, P., Campos, D., Craswell, N., Deng, L., Gao, J., Liu, X., Majumder, R., McNamara, A., Mitra, B., Nguyen, T., Rosenberg, M., Song, X., Stoica, A., Tiwary, S., Wang, T.: MS MARCO: A human generated machine reading comprehension dataset. In: CoCo@NIPS (2016)
4. Barbaro, M., Zeller, T.: A face is exposed for AOL searcher no. 4417749. The New York Times (Aug 2006), URL https://www.nytimes.com/2006/08/09/technology/09aol.html
5. Breuer, T., Ferro, N., Maistro, M., Schaer, P.: repro_eval: A python interface to reproducibility measures of system-oriented IR experiments. In: ECIR (2021)
6. Cheng, Q., Ren, Z., Lin, Y., Ren, P., Chen, Z., Liu, X.Y., de Rijke, M.: Long short-term session search: Joint personalized reranking and next query prediction. In: The Web Conference (2021)
7. Dehghani, M., Zamani, H., Severyn, A., Kamps, J., Croft, W.B.: Neural ranking models with weak supervision. In: SIGIR (2017)
8. Fuhr, N.: Some common mistakes in IR evaluation, and how they can be avoided. SIGIR Forum (2018)
9. Huang, M., Peng, W., Wang, D.: TPRM: A topic-based personalized ranking model for web search. ArXiv **abs/2108.06014** (2021)
10. Jaech, A., Kamisetty, H., Ringger, E.K., Clarke, C.: Match-tensor: a deep relevance model for search. ArXiv **abs/1701.07795** (2017)
11. Jones, R., Klinkner, K.L.: Beyond the session timeout: automatic hierarchical segmentation of search topics in query logs. In: CIKM (2008)
12. Joulin, A., Grave, E., Bojanowski, P., Mikolov, T.: Bag of tricks for efficient text classification. ArXiv **abs/1607.01759** (2016)
13. Kamara, S., Kati, A., Moataz, T., Schneider, T., Treiber, A., Yonli, M.: Cryptanalysis of encrypted search with LEAKER - a framework for leakage attack evaluation on real-world data. IACR Cryptol. ePrint Arch. (2021)
14. Kang, Y.M., Liu, W., Zhou, Y.: QueryBlazer: Efficient query autocompletion framework. In: WSDM (2021)
15. Kanoulas, E., Hall, M.M., Clough, P.D., Carterette, B., Sanderson, M.: Overview of the TREC 2011 session track. In: TREC (2011)
16. Ma, Z., Dou, Z., Bian, G., Wen, J.R.: PSTIE: Time information enhanced personalized search. In: CIKM (2020)
17. MacAvaney, S.: OpenNIR: A complete neural ad-hoc ranking pipeline. In: WSDM (2020)
18. MacAvaney, S., Cohan, A., Goharian, N.: SLEDGE-Z: A zero-shot baseline for covid-19 literature search. In: EMNLP (2020)
19. MacAvaney, S., Macdonald, C., Ounis, I.: Streamlining evaluation with ir-measures. In: ECIR (2022)

20. MacAvaney, S., Yates, A., Feldman, S., Downey, D., Cohan, A., Goharian, N.: Simplified data wrangling with ir_datasets. In: SIGIR (2021)
21. MacAvaney, S., Yates, A., Hui, K., Frieder, O.: Content-based weak supervision for ad-hoc re-ranking. In: SIGIR (2019)
22. Macdonald, C., Tonellotto, N., MacAvaney, S., Ounis, I.: Pyterrier: Declarative experimentation in python from bm25 to dense retrieval. In: CIKM (2021)
23. McCreadie, R., Soboroff, I., Lin, J.J., MacDonald, C., Ounis, I., McCullough, D.: On building a reusable twitter corpus. In: SIGIR (2012)
24. Nogueira, R., Jiang, Z., Pradeep, R., Lin, J.: Document ranking with a pretrained sequence-to-sequence model. In: Findings of EMNLP (2020)
25. Nunes, S., Ribeiro, C., David, G.: Use of temporal expressions in web search. In: ECIR (2008)
26. Ounis, I., Amati, G., Plachouras, V., He, B., Macdonald, C., Johnson, D.: Terrier information retrieval platform. In: ECIR (2005)
27. Pass, G., Chowdhury, A., Torgeson, C.: A picture of search. In: InfoScale (2006)
28. Qu, C., Xiong, C., Zhang, Y., Rosset, C., Croft, W.B., Bennett, P.N.: Contextual re-ranking with behavior aware transformers. In: SIGIR (2020)
29. Raffel, C., Shazeer, N.M., 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. ArXiv **abs/1910.10683** (2020)
30. Rafiei, D., Bharat, K., Shukla, A.: Diversifying web search results. In: WWW (2010)
31. Sequiera, R.D., Lin, J.: Finally, a downloadable test collection of tweets. In: SIGIR (2017)
32. Zhou, Y., Dou, Z., Wen, J.R.: Encoding history with context-aware representation learning for personalized search. In: SIGIR (2020)
33. Zhu, Y., Nie, J., Dou, Z., Ma, Z., Zhang, X., Du, P., Zuo, X., Jiang, H.: Contrastive learning of user behavior sequence for context-aware document ranking. ArXiv **abs/2108.10510** (2021)