The Archive Query Log: Mining Millions of Search Result Pages of Hundreds of Search Engines from 25 Years of Web Archives

Jan Heinrich Reimer
Friedrich-Schiller-Universität Jena

Sebastian Schmidt
Leipzig University

Maik Fröbe
Friedrich-Schiller-Universität Jena

Lukas Gienapp
Leipzig University

Harrisen Scells
Leipzig University

Benno Stein
Bauhaus-Universität Weimar

Matthias Hagen
Friedrich-Schiller-Universität Jena

Martin Potthast
Leipzig University and ScaDS.AI

ABSTRACT

The Archive Query Log (AQL) is a previously unused, comprehensive query log collected at the Internet Archive over the last 25 years. Its first version includes 356 million queries, 166 million search result pages, and 1.7 billion search results across 550 search providers. Although many query logs have been studied in the literature, the search providers that own them generally do not publish their logs to protect user privacy and vital business data. Of the few query logs publicly available, none combines size, scope, and diversity. The AQL is the first to do so, enabling research on new retrieval models and (diachronic) search engine analyses. Provided in a privacy-preserving manner, it promotes open research as well as more transparency and accountability in the search industry.

CCS CONCEPTS

• **Information systems → Query log analysis**.

KEYWORDS

query log, search engine result page, information retrieval history

ACM Reference Format:
Jan Heinrich Reimer, Sebastian Schmidt, Maik Fröbe, Lukas Gienapp, Harrisen Scells, Benno Stein, Matthias Hagen, and Martin Potthast. 2023. The Archive Query Log: Mining Millions of Search Result Pages of Hundreds of Search Engines from 25 Years of Web Archives. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '23), July 23–27, 2023, Taipei, Taiwan.* ACM, New York, NY, USA, 12 pages. https://doi.org/XX.XXXX/XXXXXX.XXXXXX

1 INTRODUCTION

Search engine query logs are a rich resource for many information retrieval applications [4], such as user behavior and user experience analysis, or query suggestions and query reformulations. When a query log also includes users' clicks and dwell time on search results, this is a valuable source of implicit relevance feedback about their information needs. Modern search engines use this feedback to train retrieval models for re-ranking [94, 118]. However, query logs are also highly sensitive data that affect a number of stakeholders [11, 73]: First and foremost are user privacy concerns. Over time, if a user frequently uses a search engine, their query log can be enough to personally identify them and reveal a lot about their state of mind and health. To some extent, this also applies to persons or organizations mentioned or implied in queries or search results. Not least, relevance feedback from a query log is an important asset for search providers, commercial or otherwise.

Table 1: The Archive Query Log 2022 (AQL-22) at a glance.

Search provider		**URLs**	**Queries**	**Queries**	**SERPs**	**Results**
(known domains)		(total)	(total)	(unique)	(estimate)	(estimate)
G	Google	89.4 M	72.7 M	20.0 M	34.0 M	270.9 M
▶	YouTube	41.8 M	41.4 M	11.3 M	19.3 M	411.8 M
🐾	Baidu	78.5 M	69.6 M	2.9 M	32.5 M	130.7 M
🐧	QQ	0.5 M	0.5 M	0.1 M	0.2 M	2.6 M
f	Facebook	3.1 M	0.2 M	0.0 M	0.1 M	0.8 M
y!	Yahoo!	8.8 M	2.8 M	1.2 M	1.3 M	11.2 M
a	Amazon	66.8 M	0.8 M	0.3 M	0.4 M	9.5 M
W	Wikipedia	68.5 M	1.7 M	0.6 M	0.8 M	8.5 M
JD	JD.com	4.4 M	3.9 M	0.4 M	1.8 M	19.4 M
●	360	1.5 M	1.1 M	0.1 M	0.5 M	4.2 M
⋮	540 others	646.6 M	161.8 M	27.7 M	75.4 M	839.5 M
∑	550	1,010.0 M	356.5 M	64.5 M	166.4 M	1,709.0 M

All of the above are good reasons for not publishing query logs. Another, yet questionable, reason for major search providers is that governments and civil societies around the world, as well as affected users and third parties, expect more transparency and accountability from them due to their market dominance. Access to their query logs would enable independent investigations on a large scale into the accuracy and fairness of their search results [11], as well as help to promote competition [12] and law enforcement [35]. Not least, relevance feedback from a query log would be an important asset for public information retrieval research.

We have uncovered and acquired an extensive query log that has accumulated at the Internet Archive over the last 25 years. We call this new resource Archive Query Log (AQL). Table 1 gives an overview of the first version of 2022 and the top ten search providers as per fused snapshots of Alexa website traffic rankings. Shown are the respective numbers of archived URLs, the queries extracted from them, and archived search result pages (SERPs)

SIGIR '23, July 23–27, 2023, Taipei, Taiwan

ACM ISBN XXX-X-XXX-XXXX-X/XX/XX…$15.00
https://doi.org/XX.XXXX/XXXXXX.XXXXXX

Table 2: Overview of large-scale query logs used in previous work. Private logs are grouped by source. Each source is referenced by the paper using the largest sample. Number of usages is given under №. Timespan indicates crawled duration; year indicates date of last included query. Fields with '–' are either not available or not specified. Languages are estimated. The ★ marks logs still available for download; ⚙ marks industry; 🎓 academic; ⤨ mixed affiliations.

	Source (name)	Clickable link: 🔗	**Queries**	**Queries** (uniq.)	**Sessions**	**Users**	**Clicks**	**Results**	**Task**	**Lang.**	**Span**	**Year**	**Aff.**	**Ref.**	**№**
Private (sample)	AltaVista		575,244,993	153,645,050	–	285,000,000	–	–	Web	en	1m	1998	⚙	[103]	3
	Dreamer		2,184,256	228,566	–	–	–	–	Web	zh	3m	1998	🎓	[26]	1
	Excite		51,473	18,098	–	18,113	–	–	Web	en	–	1998	🎓	[63]	3
	Infoseek		19,933,187	–	–	–	–	–	Web	zh	1m	1998	⤨	[46]	1
	GAIS		475,564	114,182	–	–	–	–	Web	zh	2w	1999	🎓	[26]	1
	Lycos		500,000	243,595	–	–	500,000	361,906	Web	en	1d	2000	⤨	[17]	1
	Yahoo!		2,369,282	–	–	–	21,421	–	Web	en,zh	4m	2000	🎓	[59]	9
	OpenFind		2,493,211	–	–	–	–	–	Web	en	1y	2000	🎓	[26]	1
	Encarta		2,772,615	–	2,772,615	–	–	–	Web	en	1m	2002	⤨	[110]	1
	Utah State Gov.		792,103	575,389	458,962	161,042	323,285	–	Web	en	5m	2003	🎓	[24]	1
	Timway		1,255,633	–	–	–	–	–	Web	zh	3m	2004	🎓	[81]	1
	MetaSpy		580,000	–	–	–	–	–	Web	en	5d	2005	🎓	[6]	1
	arXiv		44,399	–	–	13,304	48,976	–	Edu	en	3m	2006	🎓	[100]	1
	TodoCL		192,924	–	348,035	–	360,641	360,641	Web	en	9m	2006	⤨	[48]	1
	kunstmuseum.nl		7,531	1,183	–	–	–	–	Lib.	nl	2y	2007	🎓	[9]	1
	Microsoft AdCenter		27,922,224	27,922,224	–	–	7,820,000	–	Ads	en	2m	2007	🎓	[74]	2
	Europeana		3,024,162	1,382,069	–	–	–	–	Lib.	en	6m	2011	⤨	[22]	1
	GNU IFT		2,099	–	–	–	–	4,754	Img.	en	1y	2011	🎓	[86]	1
	INDURE		14,503	2,923	85,215	6,434	–	–	Edu	en	3m	2011	🎓	[49]	1
	Bing Videos		1,218,936	445,859	174,955	174,955	–	–	Vid.	en	1w	2011	⤨	[72]	1
	Baidu		362,994,092	10,413,491	87,744,130	–	–	13,126,252	Web	zh	–	2012	⚙	[116]	4
	CADAL Library		45,892	–	81,759	–	–	164,822	Lib.	en	–	2012	🎓	[113]	1
	Taobao		1,410,960	–	–	4,285	–	–	Prod.	zh	1m	2013	🎓	[117]	2
	Startpagina		10,000,000	–	–	–	–	–	Web	nl	1m	2014	⤨	[55]	1
	parsijoo.ir		27,000,000	–	–	–	–	–	Web	fa	2y	2017	🎓	[85]	1
	CiteSeerX		78,124,884	14,759,852	–	–	–	–	Edu	en	4y	2021	🎓	[101]	1
⋮	15 query logs from undisclosed sources, all private, 5 🎓, 2 ⚙, 8 ⤨													[14, 15, 18, 20, 54, 58–60, 66, 67, 79, 80, 99, 107, 112]	–
Public (exhaustive)	PubMed		2,996,301	–	–	627,455	–	–	Med.	en	1d	2005	🎓	[57]	3
	AOL (AOL Query Log '06)	🔗	36,389,567	10,154,742	–	657,426	19,442,629	19,442,629	Web	en	3m	2006	⚙	[96]	14
	MSN (MS RFP'06)	🔗	14,921,285	–	14,921,285	–	–	–	Web	en	1m	2006	⚙		17
	Belga News Agency	🔗	1,402,990	–	–	–	5,697,287	498,039	Img.	en	1y	2008	🎓	[87]	2
	bildungsserver.de (DBMS)	🔗	98,512	31,347	65,513	–	68,604	–	Edu.	de	–	2009	🎓	[84]	4
	Gov2 Crawl (LETOR 4.0)	🔗	2,500	–	–	–	–	–	Web	en	–	2009	⚙	[98]	2
	Sogou	🔗	18,393,652	4,580,463	–	8,168,051	–	14,075,717	Web	zh	1m	2009	🎓	[115]	4
	Tumba!		458,623	–	–	–	–	–	Web	en	–	2009	🎓	[83]	1
	European Library (TEL)	🔗	162,642	–	75,100	–	–	–	Lib.	en	1y	2010	🎓	[83]	4
	Yandex	🔗	10,139,547	–	–	956,536	–	49,029,185	Web	ru	–	2011	⚙		2
	Bing (MS Image Ret. Ch.)	🔗	11,701,890	–	–	–	–	–	Img.	en	–	2013	⚙		1
	Bing (ORCAS)	🔗★	18,823,602	–	–	–	18,823,602	18,823,602	Web	en	–	2020	⤨	[38]	3
	AOL (AOLIA)	🔗★	11,337,160	–	–	–	–	1,525,524	Web	en	–	2022	🎓	[82]	1
	StackOverflow	🔗★	9,046,179	–	16,164,506	–	–	–	Q&A	en	–	2022	⚙		2
	Archive Query Log (AQL)	🔗★	356,450,494	64,544,345	–	–	–	1,709,027,339	Multi	Multi	25y	2022	🎓		

and results linked to them. The SERPs of many queries have been archived multiple times, enabling diachronic analysis. At the time of writing, we collect this data for a total of 550 search providers. The Archive Query Log 2022 includes 356 million queries (65 million unique), 166 million search result pages, and 1.7 billion search results—an unprecedented scale for a public query log. Based on a comprehensive review of public and private query logs used in the literature (Section 2), we detail our acquisition method (Section 3), initial analyses (Section 4), and discuss our plan to share the data with the information retrieval community in a privacy-preserving manner, as well as limitations and ethical considerations (Section 5).

2 BACKGROUND AND RELATED WORK

We take an in-depth look at the use of query logs and search result pages in information retrieval research as well as a brief one at search transparency and accountability and at the Internet Archive.

2.1 Query Logs

Table 2 compiles an overview of 14 public and 31 private query logs from a focused literature review. Using the DBLP[1] title search, we screened all publications that mention "query log", "click log", or "clickthrough" in their title—a high-precision heuristic to ensure logs play a role, at the expense of recall. From the 642 publications found, the 492 related to information retrieval (e.g., not databases) were downloaded. We then searched for occurrences of the pattern "`<number>` `<qualifier>` 'queries' " in them with regular expressions, assuming that virtually all researchers using query logs also specify how large they are.[2] This facilitated the manual extraction of passages and tables from 120 random randomly selected publications for the table. Some were manually added to cover public logs.

[1] https://dblp.org/

[2] Examples: "1 million queries", "386,879 queries", "386 879 queries", or "386k queries". A qualifier is a sequence of up to 20 characters excluding end of sentence punctuation.

Despite the fact that query logs are rarely published, researchers in academia have sought alternative means of access, usually by collaborating with search providers of many kinds. Weighted by the number of publications per log, research with very large query logs was conducted in industry and at major search providers (the AOL log being the exception). The AOL log [96] and its recent extension AOLIA [82] are the largest query logs ever made publicly available. The AltaVista log and the Baidu log are the two largest private logs. Our Archive Query Log is on par with the latter two. The ratio of unique queries to all queries averages 0.24. With 0.18 our log is slightly lower due to its multilingual nature. In addition to queries, organic search engine query logs may include information about users, clicks, sessions, and results, while our log only includes queries and results (SERPs and the result documents themselves).

Given the main tasks for which query logs are used, the AQL can be used to study many—though not all—of them at a scale not easily attainable for academic researchers: Query understanding involves analyzing user information behavior. Subtasks include determining the user's search intent [61] and examining user populations [63], particularly with respect to geographic [105] and temporal [64] dimensions. In addition, much research has focused exclusively on how people search for health information, from both consumer [95] and expert perspectives [102]. Query suggestion involves exploiting query logs to recommend alternative queries to the user. Subtasks include clustering [17], query similarity [25], and query expansion using relevance feedback [44, 45, 59]. The AQL can support both tasks in general, in particular as pre-training data. However, model transfer to a specific application domain will be required.

Session analysis examines how users reformulate their queries across one or more sessions [5], a key subtask being session detection [52, 56]. User modeling involves analyzing logs to build models of user interaction. Subtasks include examining the distributions of query lengths and query terms [65, 103, 106, 111], relevance feedback mechanisms [104], and what users consider relevant [62]. The AQL does not support these tasks; it lacks session or user data.

Learning to rank is about exploiting query logs to derive effective ranking models. Subtasks include developing click models [70, 71] and models that incorporate implicit feedback such as dwell time on pages [3, 114]. The AQL supports this task despite the lack of click data. Craswell et al.'s [38] rationale for the design of the MS MARCO benchmark corroborates this claim. Here, only passages (judged for relevance) from the top-ranked documents returned by Bing for a query are included as ground truth for training, which has proven to be sufficient to yield effective retrieval models. The same is true for the AQL, where the ranked results of third-party retrieval models encode the domain expertise and the implicit relevance feedback from query logs that the respective search providers incorporated into their development.

2.2 Search Engine Result Pages

Search engine result pages (SERPs) are how search engines present results to users in response to a query. SERPs for web search typically consist of a list of links to web pages ranked by their relevance to the user's query, along with additional information such as snippets, images, and other features designed to help users meet their information needs. SERPs have been studied in information retrieval research for many years to understand how users interact with them, how they can be improved, and how they can present information more effectively to better meet user needs. The AQL contains the SERP for the majority of its queries.

One area of SERP research has focused on understanding how users interact with search results. Researchers have used techniques such as eye-tracking [10, 68, 69] and brain monitoring [88] to study how users perceive SERPs. These studies have led to a deeper understanding of how to improve the presentation of results and the ranking algorithms used by search engines. Another area of SERP research is the study of their design and layout [90, 91]. These longitudinal studies show how SERPs evolve in response to new technologies. The AQL provides millions of archived SERPs which include all necessary assets for showing them in a browser, so that they can be used for user studies and offline experiments.

2.3 Transparency and Accountability

In November 2022, the Digital Market Act [1] and the Digital Services Act [2] came into force in the European Union. The former applies primarily to so-called "gatekeepers" in digital markets, such as Google for the search market, the latter to all digital services that act as so-called (information) intermediaries. Both laws contain provisions that, among other things, require search providers to increase data privacy, transparency, and accountability, with the goal of ensuring fair and open digital markets. In particular, legislators are allowed to exercise regulatory and market investigation powers, which may include looking into the algorithms used. The AQL complements these measures and also gives civilian initiatives the means to conduct independent investigations of search providers. Previous studies on search accountability raise the question of how to inform users about a search engine's retrieval algorithms to raise awareness of how they work [37, 76, 77] and to ensure unbiased results [51, 78]. While we cannot consider all previous work in this context, a recent overview was provided at the FACTS-IR workshop [92, 93] on fairness, accountability, confidentiality, transparency, and safety in information retrieval. In terms of both algorithm transparency and search engine accountability, archived search result pages are perhaps one of the best representations of a search engine's behavior, and archiving them on a large scale allows for corresponding post-hoc analyses.

2.4 Internet Archive

The Internet Archive is a nonprofit digital library that has grown to become the largest and most comprehensive digital library in the world since its inception in 1996. In addition to providing access to extensive archives of books, audio recordings, videos, images, and software, the Internet Archive's best-known service is probably the Wayback Machine, which provides a digital archive of the web.[3] At the time of writing, it contains 788 billion web pages. We believe that the AQL accumulated both due to accidental crawling by their crawlers and intentional archiving by their users, since any user can request archiving of any publicly accessible URL. AOLIA [82] extends the original AOL log by providing links to archived versions of its search results, originally specified as URLs only.

[3] https://web.archive.org/

3 MINING THE ARCHIVE QUERY LOG

Besides general-purpose search engines, many other websites such as social media platforms offer a search function (a query field) for users. The answer to a query is often encoded as URL linking to a SERP, which is displayed to the user. Like other URLs, these "SERP URLs" can be linked to by web pages and are thus included in automated web crawls. The Internet Archive, as the world's largest digital library of archived web pages, is likely to include many SERPs, a fact which can be exploited for large-scale query log mining. This section describes a multi-step process to mine a query log from the Internet Archive's Wayback Machine, which eventually becomes the Archive Query Log (see Figure 1).

First, a list of popular search providers including general-purpose search engines and all kinds of media platforms is compiled (see upper part of Figure 1; Section 3.1). Second, for each provider the top-level domains, subdomains, and URL patterns under which SERPs are likely to be found are semi-automatically generated and the URL captures found in the Internet Archive (using the CDX API[4] of the Wayback Machine) are aggregated (Section 3.2). Third, the queries are extracted from the URLs using provider-specific parsers (Section 3.3). Fourth, the HTML content of archived SERPs is downloaded and its search results snippets are extracted (Section 3.4). Both queries and snippets form the AQL 2022 Corpus (Section 3.5); the corpus will be made accessible for shared tasks and experiments via the TIRA platform [97] as discussed in Section 4.

3.1 Search Provider Collection

Our search provider collection shall contain both (1) websites that primarily act as search engines, and (2) highly relevant websites that have been identified by their Alexa Rank.[5]

Regarding (1), we exploit a dedicated list of search engines on Wikipedia which we extend manually.[6] Regarding (2), we take the 3,088 archived snapshots of the Alexa top-1M ranking between June 2010 and November 2022[7] and apply reciprocal rank fusion [36] considering the 1,000 highest ranked domains of each snapshot. The resulting list of 13,647 domains is narrowed down to 951 search providers by checking whether a search bar is present on the website's landing page of the respective provider. For this purpose we load the landing page directly or from an Internet Archive snapshot from 2022, render the page if JavaScript content is found, and check for HTML forms or `<div>` elements containing the pattern "search" in its attributes.

The merged list of 1,028 unique candidate search providers is used to identify relevant URL patterns as well as suitable approaches for query extraction (see Section 3.2). Further manual curation steps weed out providers because they have been identified as spam, do not encode the search query in their URL, or offer only an autocomplete search that links directly to a page. Also, search providers are merged because more than one of their second-level domains appears in the merged Alexa list (e.g., `so.com` is merged into `360.com`). After these analyses and curation steps, the list still includes 793 search providers.

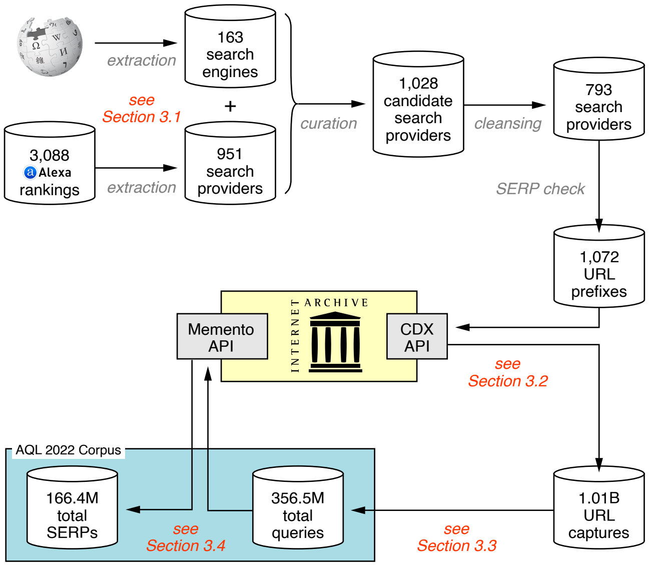


Figure 1: Overview of the AQL creation process.

3.2 Provider Domains and URLs

Since many search providers are available under multiple top-level domains and/or subdomains,[8] we expand the list of the 793 provider domains manually as well as with publicly available lists.[9]

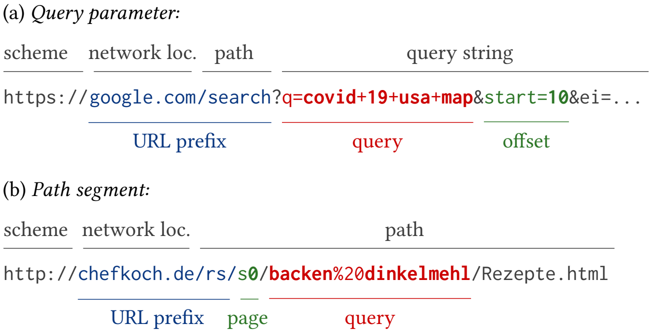
However, on high-traffic domains, only a small fraction of all archived URLs is likely to link SERPs and thus is relevant for our purposes. To identify common prefixes of URLs that contain queries, we submit multiple test queries using the search provider's query field and examine the URL the request is redirected to. For discontinued or otherwise inaccessible websites, we resort to the most recent functional snapshot of the search provider's homepage in the Internet Archive. The final list of 1,072 URL prefixes is used to filter the list of available captures with the help of the Internet Archive's CDX API. This API allows to request a list of URLs by the crawling date they were archived in the Wayback Machine for a certain domain or URL prefix. Via the CDX API, a list of all available captures for each of the 1,072 URL prefixes is retrieved and filtered for successful captures with HTML content (i.e., HTTP status code 200). This process further narrows the search provider list with archived SERPs to 550.

Altogether, 1.1B URL captures along with their crawling date are collected, an average of 1.8M URLs per provider. Most of the captures originate from search engines (226M URLs, 22 %) with Google being the largest contributor of archived URLs (89M URLs, 9 %).

3.3 Query Extraction

To extract the query from a SERP URL, the URL is parsed into its components[10] and the query encoding is identified as one of three possible categories: (1) URL parameter, (2) path segment, or (3) fragment identifier. Examples of the first two patterns are illustrated in Figure 2.

[4] https://github.com/internetarchive/wayback/tree/master/wayback-cdx-server
[5] The Alexa Rank was a ranking system that reflected the global popularity of websites based on estimated visits; it was shut down end of 2022.
[6] See https://en.wikipedia.org/wiki/List_of_search_engines
[7] See https://web.archive.org/web/*/s3.amazonaws.com/alexa-static/top-1m.csv.zip
[8] Google lists 190 supported domains https://www.google.com/supported_domains
[9] E.g., https://github.com/JamieFarrelly/Popular-Site-Subdomains
[10] RFC 2396; https://datatracker.ietf.org/doc/html/rfc2396.html



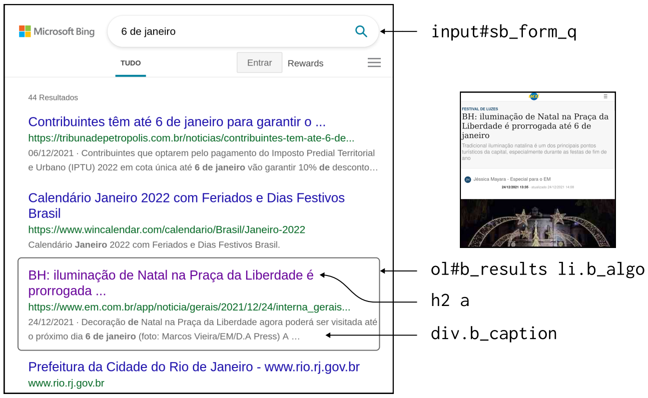
(a) *Query parameter:*

scheme
network loc.
path
query string
https://google.com/search?q=covid+19+usa+map&start=10&ei=...
URL prefix
query
offset

(b) *Path segment:*

scheme
network loc.
path
http://chefkoch.de/rs/s0/backen%20dinkelmehl/Rezepte.html
URL prefix
page
query

Figure 2: Illustration of the components of SERP URLs and the relevant parts for query extraction by (a) query parameter or (b) path segment.



Microsoft Bing 6 de janeiro

TUDO Entrar Rewards

44 Resultados

Contribuintes têm até 6 de janeiro para garantir o ...
https://tribunadepetropolis.com.br/noticias/contribuintes-tem-ate-6-de...
06/12/2021 · Contribuintes que optarem pelo pagamento do Imposto Predial Territorial e Urbano (IPTU) 2022 em cota única até **6 de janeiro** vão garantir 10% **de** desconto…

Calendário Janeiro 2022 com Feriados e Dias Festivos Brasil
https://www.wincalendar.com/calendario/Brasil/Janeiro-2022
Calendário **Janeiro** 2022 com Feriados e Dias Festivos Brasil.

BH: iluminação de Natal na Praça da Liberdade é prorrogada ...
https://www.em.com.br/app/noticia/gerais/2021/12/24/interna_gerais...
24/12/2021 · Decoração **de** Natal na Praça da Liberdade agora poderá ser visitada até o próximo dia **6 de janeiro** (foto: Marcos Vieira/EM/D.A Press) A …

Prefeitura da Cidade do Rio de Janeiro - www.rio.rj.gov.br
www.rio.rj.gov.br

FESTIVAL DE LUZES
BH: iluminação de Natal na Praça da Liberdade é prorrogada até 6 de janeiro
Tradicional iluminação natalina é um dos principais pontos turísticos da capital, especialmente durante as festas de fim de ano
Jéssica Mayara - Especial para o EM
24/12/2021 13:35 - atualizado 24/12/2021 14:08

Figure 3: Screenshot of an archived Bing SERP along with the CSS selectors for the query and result items. The nearest archived version of the referenced article is shown on the right.

For each of the three patterns, configurable query parsers are built using the `urllib` package:[11] (1) parsing a query parameter by its name (e.g., name `q` for the first URL in Figure 2), (2) parsing a segment from the URL's path component by its index (e.g., index 2 for the second URL in Figure 2), or (3) parsing a parameter from the fragment identifier by treating it like a query parameter. In addition to the query, URLs can include a page number or offset. These help in reconstructing longer rankings from separate SERPs for the same query that are captured at nearly the same time. Google's SERPs, for instance, are paginated with 10 results per page. Thus, the page number can be used to infer the continued ranks of documents on the next page.

To determine the query, page, and offset of the parsers for each search provider, we manually examined the captured URLs in a similar way as for URL prefixes (see Section 3.2) and derive suited URL parser types and parameters. Regular expressions are optionally used to limit parsing to specific URLs, and to further refine the parsers (e.g., removing prefixes such as `page-` in `/search/page-4`). The resulting set of parsers is applied to all available captures of all search providers, ordered by preference, so that the first parser that returns a non-empty query, page, or offset is used.

Altogether a total of 356.5M URLs containing queries are collected. Again, the majority of queries stems from search engines (162M queries, 46 %) such as Google (73M queries, 20 %) and Baidu (70M queries, 20 %). On average, 648,092 queries are extracted per search provider. This unfiltered set of queries contains large amounts of duplicates (288M, 81 %) for which we identify three reasons: (1) the same query is captured at different times, (2) the query is captured at approximately the same time but with different result page offsets, and (3) the same query is captured as issued from different users (e.g., if a user identifier is included in the URL). This is supported by the fact that search engines are the main contributors of duplicates (131M); however, government sites have the highest share (91 %).

We create a set of unique queries for each search provider by selecting a representative query URL (the capture with the shortest query string) from each group with the same parsed query.[12] If a group of duplicates has multiple captures with the same query parameter and URL length, the representative URL is chosen by lexicographic order. We refrain from using the capture's timestamp as a tie-breaker to not favor older or newer captures. The deduplication results in 64.5M unique queries across the final list of 550 search providers. There are 117,353 unique queries per each provider on average. Again, the search engines make up the majority of deduplicated queries (31M, 45 %).

3.4 SERP Acquisition and Parsing

Previous query logs rarely contain results for the logged queries (see Section 2). The AQL however, since it is obtained from web page captures in the Internet Archive, naturally contains the full ranking of results for the majority of queries. By downloading and parsing the archived SERPs, one has access to the full ranking of results for each search query (including the processed query itself, as it appears in the query field of the SERP). Parsing the search results including result titles, referenced URLs, snippets, and the query from a SERP facilitates not only the comparison of different search provider's ranking functions but also the evaluation of their query understanding and reformulation techniques.

We download the SERP HTML content for the unique search queries identified in the previous step and save it in WARC format:[13] For the 20 most popular search providers SERPs are downloaded for all unique URLs; for the remaining providers the download is limited to a maximum of 25,000 due to resource constraints. Connection timeouts and other errors during download are handled by repeating the download up to 10 times, after which we consider the archived SERP snapshot to be unavailable. Altogether, a total of 166.4M SERPs are collected most of which originate from search engines (ca. 40 %) and media sharing platforms (ca. 20 %). The downloads are ongoing, and we plan on scaling them up (see Section 5) to compile the full set of estimated SERPs available (see Section 4).

From the downloaded SERPs the search result ranking is extracted and processed using a configurable parser pipeline based on

[11] https://docs.python.org/3/library/urllib.html
[12] As split according to RFC 2396.
[13] ISO 28500:2017; https://iipc.github.io/warc-specifications/

FastWARC [19], Beautiful Soup,[14] and Approval Tests.[15] In detail, the CSS path or selector[16] to the result list items is specified, as well as the path from each individual result item to its title element, the referenced URL anchor, and the snippet text. Similarly, the processed query is parsed based on the CSS path to the query field. Figure 3 shows an archived Bing SERP and highlights how CSS paths are used to select relevant HTML tags. Each search provider can have multiple parser configurations, ordered by preference, that, for example, account for a changed HTML structure after redesigns of a search provider's SERPs.

We derive parser configurations (CSS paths) for the 50 most popular search providers by generating Approval Tests according to following workflow for a provider: (1) Randomly sample 10 SERPs from the downloaded SERPs. (2) For each SERP manually annotate the expected ranking and query. (3) Apply the existing parser configurations to the sampled SERPs. (4) Compare the parsed results to the annotations. (5) If the annotations do not match, inspect the HTML page in a browser, adapt or extend missing patterns, and add them to the provider's parser configurations. New configurations are added iteratively until all sampled SERPs are correctly parsed. Altogether, 70 parser configurations for SERPs and 57 parsers for processed queries of the 50 most popular providers are derived. With additional manual tests for the 10 most popular providers (see Table 1), the parsers pass a test suite of 444 Approval Tests. The code base is available open source.[17]

3.5 The Archive Query Log 2022 (AQL-22)

We merge the filtered URL captures, queries, and SERPs into a single corpus to be used in subsequent analyses (see Section 4). This corpus, the Archive Query Log 2022, consists of two artifacts: (1) a set of queries and (2) a set of ranked documents (search result snippets). Both artifacts are stored in a GZIP-compressed, newline-delimited JSON format.[18] To create the query set, each captured URL is assigned a unique identifier based on the full URL string and timestamp of the capture.[19] The captured URL is associated with its parsed query, the location of the stored copy of the SERP, and the processed query and search results parsed from that SERP. In addition, we include the URL to the SERP's archived snapshot on the Wayback Machine and tag the query language based on the parsed query text using `cld3`.[20]

The set of result documents is created by concatenating all ranked search results (i.e., rank, snippet text with title, and document URL to the referenced web page) from all parsed SERPs. Each document is assigned a unique identifier based on the document URL, the timestamp of its origin query, and the rank of the snippet on the SERP. We also associate each document with the attributes of the corresponding query in the query set. Two additional fields are the URL to the nearest available snapshot of the SERP on the Wayback Machine and the snippet language as tagged using `cld3` based on the snippet's title and text.

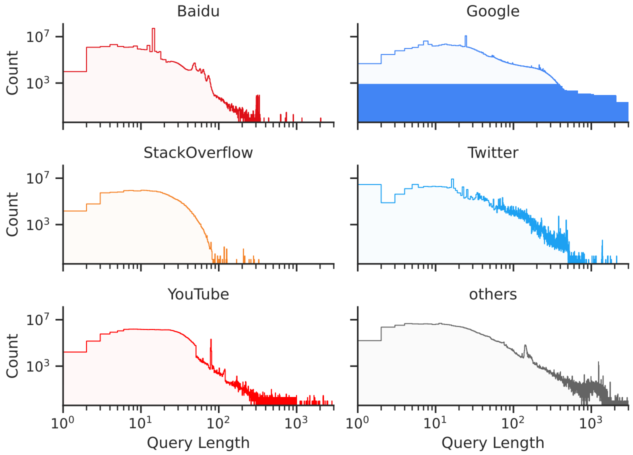


Figure 4: Distribution of query lengths for 5 search providers contributing the highest amount of queries. The remaining search providers are grouped as "others".

4 ANALYSIS

In order to provide a better understanding of the corpus, we conduct some analyses on the query and SERP characteristics and highlight potential use cases. Detailed analyses of the AQL will be the subject of future work. At the time of writing, we could download and parse all URLs and queries. However, due to computational constraints, only a subset of all available SERPs could be parsed yet (cf. Section 4.4) and are used for our analyses.

4.1 Query Characteristics

A central feature of the AQL is its diversity. In addition to the variety of search providers, it also features a total of 104 different languages[21] with Cantonese and English being the most frequently used query languages (see Table 3). The query length in the AQL follows a skewed distribution, with most queries containing between 5 and 20 characters. Figure 4 provides a visualization for the 5 search providers with the most queries. We inspect samples of queries with 5, 10, 100, and 1000 characters. Very short queries are often Mandarin keywords (e.g, `长袖衬衫男`, "men's long sleeve shirt') or single English words (e.g., `video`). Queries with 10 characters are mostly keyword-style queries from Latin languages (e.g., `comic font`) or hashtags (e.g. `#чемпионат`, "championship"). Most longer queries extensively use search operators like `site:` and `order:`, include literature references, or include long multi-line text like stack traces from errors in programming.

Second, we evaluate whether obscene or unwanted terms comprise a large share of the AQL. We use lists of obscene words for 27 languages[22] and expand the list of English terms with new expressions found in the downloaded queries. We check each query from the two most dominant languages, Cantonese and English, for their lists of obscene terms. Overall, only 1.30 % of all queries contain obscene terms. The highest share of these obscene queries

[14] https://pypi.org/project/beautifulsoup4/
[15] https://pypi.org/project/approvaltests/
[16] See https://facelessuser.github.io/soupsieve/ for a list of supported selectors. The CSS path to an HTML element can be inferred using a web browser's developer tools.
[17] AQL code: https://github.com/webis-de/archive-query-log
[18] https://jsonlines.org/
[19] Name-based SHA-1 UUID according to RFC 4122: https://rfc-editor.org/rfc/rfc4122
[20] Google's Compact Language Detector; https://github.com/google/cld3
[21] Out of 107 detectable with `cld3`.
[22] Compiled by Shutterstock; https://github.com/LDNOOBW/List-of-Dirty-Naughty-Obscene-and-Otherwise-Bad-Words.

Table 3: The Archive Query Log 2022 (AQL-22) in detail. Categories manually annotated. Top-3 languages tagged by `cld3`. Ticks in the timelines indicate days of archival from Jan 1999 to Dec 2022. Number of SERPs and results estimated (c.f. Section 4.4).

Search provider	Category	URLs	Queries				SERPs	Results		
		Total	Total	Unique	Lang.	Timeline	Estimate	Estimate	Lang.	Timeline
Google	Search engine	89,364,948	72,673,044	19,953,592	en, th, zh		33,974,648	270,874,852	en, de, pt	
YouTube	Media sharing	41,846,525	41,365,166	11,250,179	ru, ko, ja		19,338,215	411,842,988	ru, en, ko	
Baidu	Search engine	78,506,825	69,619,339	2,900,878	zh, ga, ja		32,547,041	130,685,140	zh, en, mr	
QQ	Web portal	515,895	513,608	51,228	zh, ja, lb		240,112	2,552,346	–	
Facebook	Social media	3,131,212	159,087	35,492	ca, en, bs		74,373	790,571	–	
Yahoo!	Web portal	8,787,707	2,827,103	1,232,589	en, la, de		1,321,671	11,223,594	en, es, pt	
Amazon	E-commerce	66,795,164	776,127	315,068	en, ja, zh		362,839	9,458,241	en, ja, it	
Wikipedia	Wiki	68,547,509	1,707,058	621,971	sv, zh, en		798,049	8,483,111	–	
JD.com	E-commerce	4,370,884	3,902,604	370,473	zh, hr, ja		1,824,467	19,393,742	–	
360	Search engine	1,495,365	1,090,152	65,596	zh, ja, mg		509,646	4,234,775	zh, mr, en	
Weibo	Social media	6,245,012	5,324,385	1,886,458	zh, ja, en		2,489,150	26,459,198	–	
Reddit	Forum	94,162	89,492	36,852	en, la, de		41,837	444,719	–	
Vk.com	Social media	643,354	153,642	46,134	ru, sr, ky		71,828	763,518	–	
CSDN	Social media	21,863	946	736	zh, en, vi		405	4305	–	
Bing	Search engine	11,263,539	6,152,425	2,253,965	en, zh, pt		2,876,259	15,330,377	en, pt, fr	
Twitter	Social media	55,499,532	48,084,528	3,869,382	ja, en, gl		22,479,517	293,657,729	en, ja, es	
Twitch	Streaming	21,931	15,225	11,445	en, zh, de		6,294	66,904	–	
eBay	E-commerce	7,927,123	5,507,532	1,379,646	zh, en, la		2,574,771	30,515,373	en, es, de	
Naver	Search engine	1,063,991	756,153	400,490	ja, ko, vi		353,502	3,568,671	ko, en, hi	
AliExpress	E-commerce	4,620,331	1,861,642	55,849	en, lb, fy		870,318	6,944,542	en, fr, ru	
⋮ 530 others		559,225,614	93,871,236	17,806,322	en, zh, de		43,677,091	461,732,643	en, zh, de	
∑ 550		1,009,988,486	356,450,494	64,544,345	zh, en, ga		166,432,033	1,709,027,339	en, ru, ko	

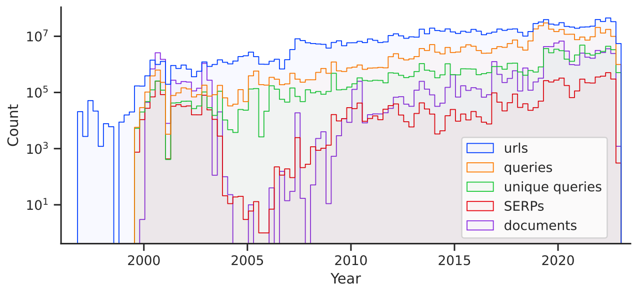


Figure 5: Time coverage of the total amount of different data types collected for the AQL-22, per quarter.

were observed on pornography (19.48 %), torrent (3.73 %), and forum (2.87 %) websites. For non-pornographic search providers, most stem from `heroturko.org` (16.13 %, e-commerce), `reddit.com` (5.05 %, forum), and `kat.cr` (4.08 %, forum).

Regarding time coverage, Figure 5 shows that the archival of SERPs dropped between 2004 and 2010 for unknown reasons, which might indicate that using more specialized SERP parsers are required, or that results were loaded using Javascript. The queries, however, extend over the whole timespan, with tens of thousands of queries recorded for the early 2000's as well. Table 3 contains an overview of the 20 most popular services' time coverage.

4.2 SERP Characteristics

In Table 4, we consider the most frequently referenced URLs from search results as an indicator of plausible rankings. Excluding frequent self-references (e.g., to internal redirect pages), by far the most frequently ranked domain is `wikipedia.org` contributing 2.9 % of all top-5 results and 2.2 % of the top-10. Other popular domains like `youtube.com` and `facebook.com` also frequently appear on high ranks. The most frequent languages are shown in Table 3. Interestingly, Cantonese is not among the top-3 languages of search results, even though it is the most frequently used query language, representing a bias that should be evaluated more thoroughly in future work.

Table 4: Most frequent document domains in the top-5 or the top-10 search results compared to references to the search provider's own domain (↻) or 791,646 other domains (···).

Top	W	▶	f	in	IMDb	◎	a	P	···	↻
5	2.9 %	0.8 %	0.6 %	0.4 %	0.3 %	0.3 %	0.3 %	0.2 %	24.3 %	69.6 %
10	2.2 %	0.7 %	0.5 %	0.3 %	0.3 %	0.2 %	0.3 %	0.3 %	24.6 %	70.4 %

4.3 Use Cases

The AQL opens up a variety of use cases for the IR community. We highlight two promising applications.

First, we evaluate the exact overlap of the queries in the AQL with the collections used in various TREC tracks from 2004 to 2022 [7, 8, 21, 27–34, 39–43, 108, 109]. As shown in Figure 6, the highest overlap exists with the Web tracks, specifically in 2010 (74 %), 2003, and 2009 (both 72 %). The lowest overlap was found with the Deep Learning tracks, ranging between 0–2 %. The high overlap on older Web tracks poses an interesting opportunity for enriching existing benchmarks. While query logs have been used previously in shared community tasks [83], shared tasks often specify only one query for each topic. We propose to sample semantically similar queries

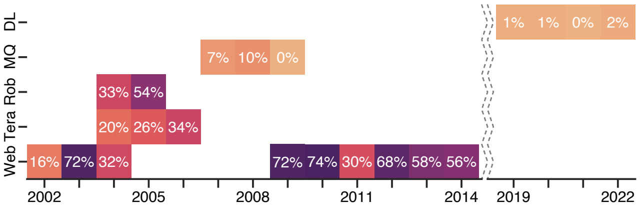


Figure 6: Query overlap with TREC Robust (Rob), Terabyte (Tera), Million Query (MQ), Web, and Deep Learning (DL).

from the AQL to generate topics with user query variations [16] automatically. On the other hand, the low overlap with the Deep Learning tracks highlights a sampling bias in creating the Deep Learning topics. The topics were sampled from the official eval set of MS MARCO, which includes only natural language questions from a Bing query log [43, 89]. The AQL, on the other hand, contains a much broader range of queries, including queries from other search providers and non-question-like queries. Therefore, we propose using the AQL to create new, "harder" Deep Learning topics that are more representative of other kinds of queries users submit.

Second, we demonstrate how global trends are reflected in the AQL on the example of the Covid-19 pandemic. In Figure 7, we count the occurrences of the terms `covid 19`, `sars cov 2`, and `corona virus` each month since the outbreak in 2019. A peak can be observed during the first global lockdowns in early 2020, but overall, interest in the pandemic has yet to stagnate. The example showcases how the AQL enables unique opportunities for diachronically analysing global trends.

4.4 Total Size Estimates

As Section 3 explains, we have only downloaded and parsed a subset of all available SERPs. Based on our results so far, we estimate 85 % of all SERP snapshots to be available for download. Assuming an estimated parsing success rate of 55 % and 10.6 results per SERP, we expect the total number of parsed SERPs in the AQL-22 to be 166.4M with 1.7B search results. As outlined in Section 5, we continue to download and parse SERPs and look forward to expanding the AQL with the IR community.

5 DISCUSSION

Access to query logs has long been an insurmountable barrier to answering critical questions about the search economy at large—if not to ask them in the first place. As a consequence, the media and general public were left with no choice but to trust search engines on questions such as "How accountable are organizations operating search engines in terms of measures of interest, like representation and fairness?", "How have these accountability measures changed for these organizations over time?", and "How honest has self-reported accountability of these organizations been?". As the most extensive public query log to date, the AQL enables detailed analyses of and thus facilitates the public discourse on the search industry. It also furthers the research on information retrieval, whose retrieval models are often presumed to be behind or at least detached from those of industry players (e.g., [13]). Using the AQL, researchers will also be able to answer questions such as "How far is academic information retrieval research behind industry?", "How much do query logs contribute compared to other ranking signals?", and not least "What are domain-specific differences?".

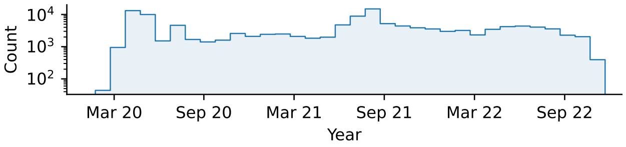


Figure 7: Timeline of Covid-19-related terms in the AQL-22.

However, with the scale of the AQL, several ethical and legal considerations also arise, particularly around personally identifiable information or illegal content. To address inherent risks, we release the data by imposing a barrier to access that minimizes potential harms while giving the information retrieval community as much freedom to conduct their research as possible. In addition, we also acknowledge the challenges of creating such extensive collections and discuss our plans for opening contributions to the AQL.

5.1 Accessing the AQL

Our goal with releasing the AQL is to do so responsibly and in a privacy-preserving manner. We work towards that goal using the TIRA platform [50, 97] for any analysis on the AQL one wishes to conduct. TIRA has been used since 2012 [53] to facilitate shared tasks with software submissions while ensuring that the submitted software can process the data without giving the participants themselves access. TIRA achieves this through sandboxing, i.e., disconnecting the software from the internet while it is running and thus ensuring that it can not leak data. We added the AQL to TIRA to allow researchers to submit their analysis software as Docker images. The platform is open to the public, and we provide examples and documentation on how to perform analyses on the AQL.[23] Specific shared tasks will be developed as well.

TIRA allows running arbitrary software packaged in Docker images on the AQL dataset in a privacy-preserving way, as the analysis results are blinded until reviewed. Specifically, we review the output and the software installed in the Docker image to ensure no sensitive data is leaked. TIRA runs the software in a Kubernetes cluster (1,620 CPU cores, 25.4 TB RAM, 24 GeForce GTX 1080 GPUs) with a timeout of 24 hours, so that almost any evaluation is supported.[24] In summary, TIRA provides the ideal means to work with the AQL, ensuring sensitive query log data remains secure and is responsibly used for academic research.

5.2 Limitations and Scalability

While creating the AQL, we encountered several technical limitations that guide future optimizations and improvements. First, the various parsers for creating the AQL-22 corpus were written semi-automatically. This approach was error-prone and inefficient, requiring much manual work. When building future versions of the AQL, we plan to train token classification models like BERT [47] to automatically generate query parsers based on a training set derived from our existing parsers. Manually finding the correct CSS

[23] https://tira.io/task/archive-query-log

[24] We can extend the timeouts and available resources individually if the need arises.

paths for snippets on a SERP is a similarly tedious process that can benefit from wrapper generation [75], which has successfully been applied to web page parsing [23].

Second, dynamic content cannot be interpreted by our existing parsers. The SERPs of DuckDuckGo, for instance, are loaded dynamically using JavaScript and thus cannot be parsed from just the archived HTML snapshot, yet the search results are still archived as a different record. To overcome this limitation, we plan to use a headless browser to render the SERPs and then parse them. A helpful library for this type of content extraction is Scriptor.[25]

However, the last and most pressing limitation is that all URL captures and SERP contents must first be downloaded from the Internet Archive, which is restricted by both rate limits and network bandwidth. As estimated in Section 4.4, currently 96 % of the SERPs still need to be fully downloaded from the Internet Archive and thus could not yet be parsed. Hence, more computational resources are required to use this extensive collection. In this regard, we will reach out to the Internet Archive whose privileged access to their infrastructure will allow for much faster compilation of the data.

5.3 Contributing to the AQL

There is an inherent boundary between search providers and researchers when using query logs. The AQL lowers this boundary by exploiting the web archival process of the Internet Archive. As we have described above, physical limitations such as network speed restrict the rate at which we can further grow the AQL. One way to overcome such limitations is to distribute computations across an open community, an approach that has been successfully employed in mathematics.[26] We will therefore open source our code to allow the community to contribute query and SERP parsers.

6 CONCLUSION

The Archive Query Log provides an unparalleled academic resource for information retrieval researchers. It consists of over 356 million queries, over 166 million SERPs, and over 1.7 billion results extracted from the SERPs, all coming from 550 search providers spanning 25 years. The AQL is the largest and most diverse query log ever publicly available. From an academic perspective, the AQL will enable researchers to tackle challenges in information retrieval that were not possible until now, ranging from the development of new retrieval models, the development of query suggestion or query prediction models, to large-scale diachronic analyses of search engines; to name the most salient research avenues. Furthermore, our release plan for accessing the AQL ensures that we minimize the harm to society and will allow researchers to safely research the transparency and accountability of commercial search engines while protecting user privacy.

In this paper, we have documented the initial version of the AQL (i.e., AQL-22). We have plans to release future versions of the AQL that will further expand the collection. First, we will continue to add to the long tail of search providers and continue our efforts to download and extract more data from the Internet Archive. Continuing to grow the types of data provided, the next version of the AQL will also include the content of web pages for each result in a SERP. Not least, we will investigate the training of large re-ranking models based on this data.

Altogether, the AQL is an exceedingly valuable resource for researchers and will enable advances in information retrieval research that were previously insurmountable due to the relatively low scale of query logs. Because of its scope, size, and diversity we consider the AQL a significant contribution to the community, and these dimensions will continue to grow as we build upon and expand future versions of the AQL.

[25] https://github.com/webis-de/scriptor

[26] E.g., at the Great Internet Mersenne Prime Search: https://mersenne.org/

REFERENCES

[1] 2022. Regulation (EU) 2022/1925 of the European Parliament and of the Council of 14 September 2022 on contestable and fair markets in the digital sector and amending Directives (EU) 2019/1937 and (EU) 2020/1828 (Digital Markets Act). *OJ* L 265 (2022), 1–66.

[2] 2022. Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services and amending Directive 2000/31/EC (Digital Services Act). *OJ* L 277 (2022), 1–102.

[3] Eugene Agichtein, Eric Brill, and Susan T. Dumais. 2006. Improving web search ranking by incorporating user behavior information. (2006), 19–26.

[4] Maristella Agosti, Franco Crivellari, and Giorgio Maria Di Nunzio. 2012. Web log analysis: a review of a decade of studies about information acquisition, inspection and interpretation of user interaction. *Data Min. Knowl. Discov.* 24, 3 (2012), 663–696.

[5] Maristella Agosti and Giorgio Maria Di Nunzio. 2007. Web Log Mining: A study of user sessions. In *Proceedings of the 10th DELOS Thematic Workshop on Personalized Access, Profile Management, and Context Awareness in Digital Libraries (PersDL)*.

[6] Farooq Ahmad and Grzegorz Kondrak. 2005. Learning a Spelling Error Model from Search Query Logs. In *HLT/EMNLP 2005, Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference, 6-8 October 2005, Vancouver, British Columbia, Canada*. The Association for Computational Linguistics, 955–962. https://aclanthology.org/H05-1120

[7] James Allan, Javed A. Aslam, Ben Carterette, Virgil Pavlu, and Evangelos Kanoulas. 2008. Million Query Track 2008 Overview. In *TREC*.

[8] James Allan, Ben Carterette, Javed A. Aslam, Virgil Pavlu, Blagovest Dachev, and Evangelos Kanoulas. 2007. Million Query Track 2007 Overview. In *TREC*.

[9] Avi Arampatzis, Jaap Kamps, Marijn Koolen, and Nir Nussbaum. 2007. Deriving a Domain Specific Test Collection from a Query Log. In *Proceedings of the Workshop on Language Technology for Cultural Heritage Data, LaTeCHACL 2007, Prague, Czech Republic, June 28, 2007*, Caroline Sporleder, Antal van den Bosch, and Claire Grover (Eds.). Association for Computational Linguistics, 73–80. https://aclanthology.org/W07-0910

[10] Ioannis Arapakis, Joemon M. Jose, and Philip D. Gray. 2008. Affective feedback: an investigation into the role of emotions in the information seeking process. In *SIGIR*. ACM, 395–402.

[11] Cédric Argenton and Jens Prüfer. 2007. The Structure of Search Engine Law. *Iowa Law Review* 93, 1 (11 2007), 1–63. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=979568

[12] Cédric Argenton and Jens Prüfer. 2012. Search Engine Optimization with Network Externalities. *Journal of Competition Law & Economics* 8, 1 (03 2012), 73–105. https://doi.org/10.1093/joclec/nhr018

[13] Leif Azzopardi, Yashar Moshfeghi, Martin Halvey, Rami Suleiman Alkhawaldeh, Krisztian Balog, Emanuele Di Buccio, Diego Ceccarelli, Juan M. Fernández-Luna, Charlie Hull, Jake Mannix, and Sauparna Palchowdhury. 2016. Lucene4IR: Developing Information Retrieval Evaluation Resources using Lucene. *SIGIR Forum* 50, 2 (2016), 58–75.

[14] Ricardo Baeza-Yates. 2015. Incremental Sampling of Query Logs. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, Santiago, Chile, August 9-13, 2015*, Ricardo Baeza-Yates, Mounia Lalmas, Alistair Moffat, and Berthier A. Ribeiro-Neto (Eds.). ACM, 1093–1096. https://doi.org/10.1145/2766462.2776780

[15] Ricardo A. Baeza-Yates and Felipe Saint-Jean. 2003. A Three Level Search Engine Index Based in Query Log Distribution. In *String Processing and Information Retrieval, 10th International Symposium, SPIRE 2003, Manaus, Brazil, October 8-10, 2003, Proceedings (Lecture Notes in Computer Science, Vol. 2857)*, Mario A. Nascimento, Edleno Silva de Moura, and Arlindo L. Oliveira (Eds.). Springer, 56–65. https://doi.org/10.1007/978-3-540-39984-1_5

[16] Peter Bailey, Alistair Moffat, Falk Scholer, and Paul Thomas. 2016. UQV100: A Test Collection with Query Variability. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval, SIGIR 2016, Pisa, Italy, July 17-21, 2016*, Raffaele Perego, Fabrizio Sebastiani, Javed A. Aslam, Ian Ruthven, and Justin Zobel (Eds.). ACM, 725–728.

https://doi.org/10.1145/2911451.2914671

[17] Doug Beeferman and Adam L. Berger. 2000. Agglomerative clustering of a search engine query log. In *KDD*. ACM, 407–416.

[18] Steven M. Beitzel, Eric C. Jensen, Abdur Chowdhury, David A. Grossman, and Ophir Frieder. 2004. Hourly analysis of a very large topically categorized web query log. In *SIGIR 2004: Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Sheffield, UK, July 25-29, 2004*, Mark Sanderson, Kalervo Järvelin, James Allan, and Peter Bruza (Eds.). ACM, 321–328. https://doi.org/10.1145/1008992.1009048

[19] Janek Bevendorff, Martin Potthast, and Benno Stein. 2021. FastWARC: Optimizing Large-Scale Web Archive Analytics. In *3rd International Symposium on Open Search Technology (OSSYM 2021)*, Andreas Wagner, Christian Guetl, Michael Granitzer, and Stefan Voigt (Eds.). International Open Search Symposium. https://doi.org/10.5281/zenodo.6840911

[20] Bin Cao, Dou Shen, Kuansan Wang, and Qiang Yang. 2010. Clickthrough Log Analysis by Collaborative Ranking. In *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2010, Atlanta, Georgia, USA, July 11-15, 2010*, Maria Fox and David Poole (Eds.). AAAI Press. https://aaai.org/ocs/index.php/AAAI/AAAI10/paper/view/1906

[21] Ben Carterette, Virgil Pavlu, Hui Fang, and Evangelos Kanoulas. 2009. Million Query Track 2009 Overview. In *TREC*.

[22] Diego Ceccarelli, Sergiu Gordea, Claudio Lucchese, Franco Maria Nardini, and Gabriele Tolomei. 2011. Improving Europeana Search Experience Using Query Logs. In *Research and Advanced Technology for Digital Libraries - International Conference on Theory and Practice of Digital Libraries, TPDL 2011, Berlin, Germany, September 26-28, 2011. Proceedings (Lecture Notes in Computer Science, Vol. 6966)*, Stefan Gradmann, Francesca Borri, Carlo Meghini, and Heiko Schuldt (Eds.). Springer, 384–395. https://doi.org/10.1007/978-3-642-24469-8_39

[23] Chia-Hui Chang, Mohammed Kayed, Moheb R. Girgis, and Khaled F. Shaalan. 2006. A Survey of Web Information Extraction Systems. *IEEE Trans. Knowl. Data Eng.* 18, 10 (2006), 1411–1428. https://doi.org/10.1109/TKDE.2006.152

[24] Michael Chau, Xiao Fang, and Olivia R. Liu Sheng. 2005. Analysis of the query logs of a Web site search engine. *J. Assoc. Inf. Sci. Technol.* 56, 13 (2005), 1363–1376. https://doi.org/10.1002/asi.20210

[25] Steve Chien and Nicole Immorlica. 2005. Semantic similarity between search engine queries using temporal correlation. In *WWW*. ACM, 2–11.

[26] Shui-Lung Chuang and Lee-Feng Chien. 2003. Enriching Web taxonomies through subject categorization of query terms from search engine logs. *Decis. Support Syst.* 35, 1 (2003), 113–127. https://doi.org/10.1016/S0167-9236(02)00099-4

[27] Charles L. A. Clarke, Nick Craswell, and Ian Soboroff. 2004. Overview of the TREC 2004 Terabyte Track. In *Proceedings of the Thirteenth Text REtrieval Conference, TREC 2004, Gaithersburg, Maryland, USA, November 16-19, 2004 (NIST Special Publication, Vol. 500-261)*, Ellen M. Voorhees and Lori P. Buckland (Eds.). National Institute of Standards and Technology (NIST).

[28] Charles L. A. Clarke, Nick Craswell, and Ian Soboroff. 2009. Overview of the TREC 2009 Web Track. In *Proceedings of The Eighteenth Text REtrieval Conference, TREC 2009, Gaithersburg, Maryland, USA, November 17-20, 2009 (NIST Special Publication, Vol. 500-278)*, Ellen M. Voorhees and Lori P. Buckland (Eds.). National Institute of Standards and Technology (NIST).

[29] Charles L. A. Clarke, Nick Craswell, Ian Soboroff, and Gordon V. Cormack. 2010. Overview of the TREC 2010 Web Track. In *Proceedings of The Nineteenth Text REtrieval Conference, TREC 2010, Gaithersburg, Maryland, USA, November 16-19, 2010 (NIST Special Publication, Vol. 500-294)*, Ellen M. Voorhees and Lori P. Buckland (Eds.). National Institute of Standards and Technology (NIST).

[30] Charles L. A. Clarke, Nick Craswell, Ian Soboroff, and Ellen M. Voorhees. 2011. Overview of the TREC 2011 Web Track. In *Proceedings of The Twentieth Text REtrieval Conference, TREC 2011, Gaithersburg, Maryland, USA, November 15-18, 2011 (NIST Special Publication, Vol. 500-296)*, Ellen M. Voorhees and Lori P. Buckland (Eds.). National Institute of Standards and Technology (NIST).

[31] Charles L. A. Clarke, Nick Craswell, and Ellen M. Voorhees. 2012. Overview of the TREC 2012 Web Track. In *Proceedings of The Twenty-First Text REtrieval Conference, TREC 2012, Gaithersburg, Maryland, USA, November 6-9, 2012 (NIST Special Publication, Vol. 500-298)*, Ellen M. Voorhees and Lori P. Buckland (Eds.). National Institute of Standards and Technology (NIST).

[32] Charles L. A. Clarke, Falk Scholer, and Ian Soboroff. 2005. The TREC 2005 Terabyte Track. In *Proceedings of the Fourteenth Text REtrieval Conference, TREC 2005, Gaithersburg, Maryland, USA, November 15-18, 2005 (NIST Special Publication, Vol. 500-266)*, Ellen M. Voorhees and Lori P. Buckland (Eds.). National Institute of Standards and Technology (NIST).

[33] Kevyn Collins-Thompson, Paul N. Bennett, Fernando Diaz, Charlie Clarke, and Ellen M. Voorhees. 2013. TREC 2013 Web Track Overview. In *Proceedings of The Twenty-Second Text REtrieval Conference, TREC 2013, Gaithersburg, Maryland, USA, November 19-22, 2013 (NIST Special Publication, Vol. 500-302)*, Ellen M. Voorhees (Ed.). National Institute of Standards and Technology (NIST).

[34] Kevyn Collins-Thompson, Craig Macdonald, Paul N. Bennett, Fernando Diaz, and Ellen M. Voorhees. 2014. TREC 2014 Web Track Overview. In *Proceedings of The Twenty-Third Text REtrieval Conference, TREC 2014, Gaithersburg, Maryland, USA, November 19-21, 2014 (NIST Special Publication, Vol. 500-308)*, Ellen M. Voorhees and Angela Ellis (Eds.). National Institute of Standards and Technology (NIST).

[35] Alissa Cooper. 2008. A survey of query log privacy-enhancing techniques from a policy perspective. *ACM Trans. Web* 2, 4 (2008), 19:1–19:27. https://doi.org/10.1145/1409220.1409222

[36] Gordon V. Cormack, Charles L. A. Clarke, and Stefan Büttcher. 2009. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In *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*, James Allan, Javed A. Aslam, Mark Sanderson, ChengXiang Zhai, and Justin Zobel (Eds.). ACM, 758–759. https://doi.org/10.1145/1571941.1572114

[37] Vittoria Cozza, Van Tien Hoang, Marinella Petrocchi, and Rocco De Nicola. 2019. Transparency in Keyword Faceted Search: An Investigation on Google Shopping. In *Digital Libraries: Supporting Open Science - 15th Italian Research Conference on Digital Libraries, IRCDL 2019, Pisa, Italy, January 31 - February 1, 2019, Proceedings (Communications in Computer and Information Science, Vol. 988)*, Paolo Manghi, Leonardo Candela, and Gianmaria Silvello (Eds.). Springer, 29–43. https://doi.org/10.1007/978-3-030-11226-4_3

[38] Nick Craswell, Daniel Campos, Bhaskar Mitra, Emine Yilmaz, and Bodo Billerbeck. 2020. ORCAS: 18 Million Clicked Query-Document Pairs for Analyzing Search. *CoRR* abs/2006.05324 (2020). https://doi.org/10.48550/arXiv.2006.05324 arXiv:2006.05324

[39] Nick Craswell and David Hawking. 2002. Overview of the TREC-2002 Web Track. In *Proceedings of The Eleventh Text REtrieval Conference, TREC 2002, Gaithersburg, Maryland, USA, November 19-22, 2002 (NIST Special Publication, Vol. 500-251)*, Ellen M. Voorhees and Lori P. Buckland (Eds.). National Institute of Standards and Technology (NIST).

[40] Nick Craswell and David Hawking. 2004. Overview of the TREC 2004 Web Track. In *Proceedings of the Thirteenth Text REtrieval Conference, TREC 2004, Gaithersburg, Maryland, USA, November 16-19, 2004 (NIST Special Publication, Vol. 500-261)*, Ellen M. Voorhees and Lori P. Buckland (Eds.). National Institute of Standards and Technology (NIST).

[41] Nick Craswell, David Hawking, Ross Wilkinson, and Mingfang Wu. 2003. Overview of the TREC 2003 Web Track. In *Proceedings of The Twelfth Text REtrieval Conference, TREC 2003, Gaithersburg, Maryland, USA, November 18-21, 2003 (NIST Special Publication, Vol. 500-255)*, Ellen M. Voorhees and Lori P. Buckland (Eds.). National Institute of Standards and Technology (NIST), 78–92.

[42] Nick Craswell, Bhaskar Mitra, Emine Yilmaz, and Daniel Campos. 2020. Overview of the TREC 2020 Deep Learning Track. In *Proceedings of the 29th Text REtrieval Conference, TREC 2020, Virtual Event, Gaithersburg, MD, USA, November 16-20, 2020 (NIST Special Publication, Vol. 1266)*, Ellen M. Voorhees and Angela Ellis (Eds.). National Institute of Standards and Technology (NIST).

[43] Nick Craswell, Bhaskar Mitra, Emine Yilmaz, Daniel Campos, and Ellen M. Voorhees. 2019. Overview of the TREC 2019 Deep Learning Track. In *28th International Text Retrieval Conference, TREC 2019, Gaithersburg, Maryland, USA (NIST Special Publication)*, Ellen M. Voorhees and Angela Ellis (Eds.). National Institute of Standards and Technology (NIST).

[44] Hang Cui, Ji-Rong Wen, Jian-Yun Nie, and Wei-Ying Ma. 2002. Probabilistic query expansion using query logs. In *WWW*. ACM, 325–332.

[45] Hang Cui, Ji-Rong Wen, Jian-Yun Nie, and Wei-Ying Ma. 2003. Query Expansion by Mining User Logs. *IEEE Trans. Knowl. Data Eng.* 15, 4 (2003), 829–839.

[46] Erika F. de Lima and Jan O. Pedersen. 1999. Phrase Recognition and Expansion for Short, Precision-Biased Queries Based on a Query Log. In *SIGIR '99: Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, August 15-19, 1999, Berkeley, CA, USA*, Fredric C. Gey, Marti A. Hearst, and Richard M. Tong (Eds.). ACM, 145–152. https://doi.org/10.1145/312624.312669

[47] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *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 (Long and Short Papers)*, Jill Burstein, Christy Doran, and Thamar Solorio (Eds.). Association for Computational Linguistics, 4171–4186. https://doi.org/10.18653/v1/n19-1423

[48] Georges Dupret, Benjamin Piwowarski, Carlos A. Hurtado, and Marcelo Mendoza. 2006. A Statistical Model of Query Log Generation. In *String Processing and Information Retrieval, 13th International Conference, SPIRE 2006, Glasgow, UK, October 11-13, 2006, Proceedings (Lecture Notes in Computer Science, Vol. 4209)*, Fabio Crestani, Paolo Ferragina, and Mark Sanderson (Eds.). Springer, 217–228. https://doi.org/10.1007/11880561_18

[49] Yi Fang, Naveen Somasundaram, Luo Si, Jeongwoo Ko, and Aditya P. Mathur. 2011. Analysis of an expert search query log. In *Proceeding of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2011, Beijing, China, July 25-29, 2011*, Wei-Ying Ma,

Jian-Yun Nie, Ricardo Baeza-Yates, Tat-Seng Chua, and W. Bruce Croft (Eds.). ACM, 1189–1190. https://doi.org/10.1145/2009916.2010113

[50] Maik Fröbe, Matti Wiegmann, Nikolay Kolyada, Bastian Grahm, Theresa Elstner, Frank Loebe, Matthias Hagen, Benno Stein, and Martin Potthast. 2023. Continuous Integration for Reproducible Shared Tasks with TIRA.io. In *Advances in Information Retrieval. 45th European Conference on IR Research (ECIR 2023) (Lecture Notes in Computer Science)*. Springer, Berlin Heidelberg New York.

[51] Fernando Galindo and Javier García Marco. 2017. Freedom and the Internet: empowering citizens and addressing the transparency gap in search engines. *Eur. J. Law Technol.* 8, 2 (2017). https://ejlt.org/index.php/ejlt/article/view/476

[52] Daniel Gayo-Avello. 2009. A survey on session detection methods in query logs and a proposal for future evaluation. *Inf. Sci.* 179, 12 (2009), 1822–1843.

[53] Tim Gollub, Benno Stein, Steven Burrows, and Dennis Hoppe. 2012. TIRA: Configuring, Executing, and Disseminating Information Retrieval Experiments. In *9th International Workshop on Text-based Information Retrieval (TIR 2012) at DEXA*, A Min Tjoa, Stephen Liddle, Klaus-Dieter Schewe, and Xiaofang Zhou (Eds.). IEEE, Los Alamitos, California, 151–155. https://doi.org/10.1109/DEXA.2012.55

[54] Siyu Gu, Jun Yan, Lei Ji, Shuicheng Yan, Junshi Huang, Ning Liu, Ying Chen, and Zheng Chen. 2011. Cross Domain Random Walk for Query Intent Pattern Mining from Search Engine Log. In *11th IEEE International Conference on Data Mining, ICDM 2011, Vancouver, BC, Canada, December 11-14, 2011*, Diane J. Cook, Jian Pei, Wei Wang, Osmar R. Zaïane, and Xindong Wu (Eds.). IEEE Computer Society, 221–230. https://doi.org/10.1109/ICDM.2011.44

[55] Dirk Guijt and Claudia Hauff. 2015. Using Query-Log Based Collective Intelligence to Generate Query Suggestions for Tagged Content Search. In *Engineering the Web in the Big Data Era - 15th International Conference, ICWE 2015, Rotterdam, The Netherlands, June 23-26, 2015, Proceedings (Lecture Notes in Computer Science, Vol. 9114)*, Philipp Cimiano, Flavius Frasincar, Geert-Jan Houben, and Daniel Schwabe (Eds.). Springer, 165–181. https://doi.org/10.1007/978-3-319-19890-3_12

[56] Matthias Hagen, Jakob Gomoll, Anna Beyer, and Benno Stein. 2013. From search session detection to search mission detection. In *OAIR*. ACM, 85–92.

[57] Jorge R. Herskovic, Len Y. Tanaka, William R. Hersh, and Elmer V. Bernstam. 2007. Research paper: A Day in the Life of PubMed: Analysis of a Typical Day's Query Log. *J. Am. Medical Informatics Assoc.* 14, 2 (2007), 212–220. https://doi.org/10.1197/jamia.M2191

[58] Yunhua Hu, Ya-nan Qian, Hang Li, Daxin Jiang, Jian Pei, and Qinghua Zheng. 2012. Mining query subtopics from search log data. In *The 35th International ACM SIGIR conference on research and development in Information Retrieval, SIGIR '12, Portland, OR, USA, August 12-16, 2012*, William R. Hersh, Jamie Callan, Yoelle Maarek, and Mark Sanderson (Eds.). ACM, 305–314. https://doi.org/10.1145/2348283.2348327

[59] Chien-Kang Huang, Lee-Feng Chien, and Yen-Jen Oyang. 2003. Relevant term suggestion in interactive web search based on contextual information in query session logs. *J. Assoc. Inf. Sci. Technol.* 54, 7 (2003), 638–649. https://doi.org/10.1002/asi.10256

[60] Po-Sen Huang, Xiaodong He, Jianfeng Gao, Li Deng, Alex Acero, and Larry P. Heck. 2013. Learning deep structured semantic models for web search using clickthrough data. In *22nd ACM International Conference on Information and Knowledge Management, CIKM'13, San Francisco, CA, USA, October 27 - November 1, 2013*, Qi He, Arun Iyengar, Wolfgang Nejdl, Jian Pei, and Rajeev Rastogi (Eds.). ACM, 2333–2338. https://doi.org/10.1145/2505515.2505665

[61] Bernard J Jansen and Amanda Spink. 2000. Methodological approach in discovering user search patterns through Web log analysis. *Bulletin of the American Society for Information Science and Technology* 27, 1 (2000), 15–17.

[62] Bernard J Jansen and Amanda Spink. 2003. An Analysis of Web Documents Retrieved and Viewed.. In *International Conference on Internet Computing*, Vol. 4. 65–69.

[63] Bernard J. Jansen, Amanda Spink, Judy Bateman, and Tefko Saracevic. 1998. Real Life Information Retrieval: A Study of User Queries on the Web. *SIGIR Forum* 32, 1 (1998), 5–17.

[64] Bernard J Jansen, Amanda Spink, and Jan Pedersen. 2005. A temporal comparison of AltaVista Web searching. *Journal of the American Society for Information Science and Technology* 56, 6 (2005), 559–570.

[65] Bernard J Jansen, Amanda Spink, and Tefko Saracevic. 2000. Real life, real users, and real needs: a study and analysis of user queries on the web. *Information processing & management* 36, 2 (2000), 207–227.

[66] Di Jiang, Kenneth Wai-Ting Leung, Wilfred Ng, and Hao Li. 2013. Beyond Click Graph: Topic Modeling for Search Engine Query Log Analysis. In *Database Systems for Advanced Applications, 18th International Conference, DASFAA 2013, Wuhan, China, April 22-25, 2013. Proceedings, Part I (Lecture Notes in Computer Science, Vol. 7825)*, Weiyi Meng, Ling Feng, Stéphane Bressan, Werner Winiwarter, and Wei Song (Eds.). Springer, 209–223. https://doi.org/10.1007/978-3-642-37487-6_18

[67] Di Jiang, Yongxin Tong, and Yuanfeng Song. 2016. Cross-Lingual Topic Discovery From Multilingual Search Engine Query Log. *ACM Trans. Inf. Syst.* 35, 2 (2016), 9:1–9:28. https://doi.org/10.1145/2956235

[68] Jimmy, Guido Zuccon, Bevan Koopman, and Gianluca Demartini. 2019. Health Cards for Consumer Health Search. In *SIGIR*. ACM, 35–44.

[69] Jimmy, Guido Zuccon, Bevan Koopman, and Gianluca Demartini. 2019. Health Cards to Assist Decision Making in Consumer Health Search. In *AMIA*. AMIA.

[70] Thorsten Joachims. 2002. Optimizing search engines using clickthrough data. In *KDD*. ACM, 133–142.

[71] Thorsten Joachims, Laura A. Granka, Bing Pan, Helene Hembrooke, and Geri Gay. 2005. Accurately interpreting clickthrough data as implicit feedback. In *SIGIR*. ACM, 154–161.

[72] Christoph Kofler, Linjun Yang, Martha A. Larson, Tao Mei, Alan Hanjalic, and Shipeng Li. 2012. When video search goes wrong: predicting query failure using search engine logs and visual search results. In *Proceedings of the 20th ACM Multimedia Conference, MM '12, Nara, Japan, October 29 - November 02, 2012*, Noboru Babaguchi, Kiyoharu Aizawa, John R. Smith, Shin'ichi Satoh, Thomas Plagemann, Xian-Sheng Hua, and Rong Yan (Eds.). ACM, 319–328. https://doi.org/10.1145/2393347.2393395

[73] Torsten Körber. 2015. Common Errors Regarding Search Eengine Regulation—And How to Avoid Them. *European Competition Law Review* 36 (2015). Issue 6.

[74] Chung-Lun Kuo and Hsin-Hsi Chen. 2016. Subtask Mining from Search Query Logs for How-Knowledge Acceleration. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC 2016, Portorož, Slovenia, May 23-28, 2016*, Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asunción Moreno, Jan Odijk, and Stelios Piperidis (Eds.). European Language Resources Association (ELRA). https://aclanthology.org/L16-1198

[75] Nicholas Kushmerick, Daniel S. Weld, and Robert B. Doorenbos. 1997. Wrapper Induction for Information Extraction. In *Proceedings of the Fifteenth International Joint Conference on Artificial Intelligence, IJCAI 97, Nagoya, Japan, August 23-29, 1997, 2 Volumes*. Morgan Kaufmann, 729–737.

[76] Emily B. Laidlaw. 2009. Private Power, Public Interest: An Examination of Search Engine Accountability. *Int. J. Law Inf. Technol.* 17, 1 (2009), 113–145. https://doi.org/10.1093/ijlit/ean018

[77] Lars Langer and Erik Frøkjær. 2008. Improving web search transparency by using a Venn diagram interface. In *Proceedings of the 5th Nordic Conference on Human-Computer Interaction 2008, Lund, Sweden, October 20-22, 2008 (ACM International Conference Proceeding Series, Vol. 358)*, Agneta Gulz, Charlotte Magnusson, Lone Malmborg, Håkan Eftring, Bodil Jönsson, and Konrad Tollmar (Eds.). ACM, 249–256. https://doi.org/10.1145/1463160.1463187

[78] Ruohan Li, Jianxiang Li, Bhaskar Mitra, Fernando Diaz, and Asia J. Biega. 2022. Exposing Query Identification for Search Transparency. In *WWW '22: The ACM Web Conference 2022, Virtual Event, Lyon, France, April 25 - 29, 2022*, Frédérique Laforest, Raphaël Troncy, Elena Simperl, Deepak Agarwal, Aristides Gionis, Ivan Herman, and Lionel Médini (Eds.). ACM, 3662–3672. https://doi.org/10.1145/3485447.3512262

[79] Zhen Liao, Daxin Jiang, Enhong Chen, Jian Pei, Huanhuan Cao, and Hang Li. 2011. Mining Concept Sequences from Large-Scale Search Logs for Context-Aware Query Suggestion. *ACM Trans. Intell. Syst. Technol.* 3, 1 (2011), 17:1–17:40. https://doi.org/10.1145/2036264.2036281

[80] Charles X. Ling, Jianfeng Gao, Huajie Zhang, Weining Qian, and HongJiang Zhang. 2001. Mining Generalized Query Patterns from Web Logs. In *34th Annual Hawaii International Conference on System Sciences (HICSS-34), January 3-6, 2001, Maui, Hawaii, USA*. IEEE Computer Society. https://doi.org/10.1109/HICSS.2001.926534

[81] Yan Lu, Michael Chau, and Xiao Fang. 2006. Mining the Query Logs of a Chinese Web Search Engine for Character Usage Analysis. In *Pacific Asia Conference on Information Systems, PACIS 2006, Kuala Lumpur, Malaysia, July 6-9, 2006*. AISeL, 17. https://aisel.aisnet.org/pacis2006/17

[82] Sean MacAvaney, Craig Macdonald, and Iadh Ounis. 2022. Reproducing Personalised Session Search Over the AOL Query Log. In *Advances in Information Retrieval - 44th European Conference on IR Research, ECIR 2022, Stavanger, Norway, April 10-14, 2022, Proceedings, Part I (Lecture Notes in Computer Science, Vol. 13185)*, Matthias Hagen, Suzan Verberne, Craig Macdonald, Christin Seifert, Krisztian Balog, Kjetil Nørvåg, and Vinay Setty (Eds.). Springer, 627–640. https://doi.org/10.1007/978-3-030-99736-6_42

[83] Thomas Mandl, Maristella Agosti, Giorgio Maria Di Nunzio, Alexander S. Yeh, Inderjeet Mani, Christine Doran, and Julia Maria Schulz. 2009. LogCLEF 2009: The CLEF 2009 Multilingual Logfile Analysis Track Overview. In *Multilingual Information Access Evaluation I. Text Retrieval Experiments, 10th Workshop of the Cross-Language Evaluation Forum, CLEF 2009, Corfu, Greece, September 30 - October 2, 2009, Revised Selected Papers (Lecture Notes in Computer Science, Vol. 6241)*, Carol Peters, Giorgio Maria Di Nunzio, Mikko Kurimo, Thomas Mandl, Djamel Mostefa, Anselmo Peñas, and Giovanna Roda (Eds.). Springer, 508–517. https://doi.org/10.1007/978-3-642-15754-7_62

[84] Thomas Mandl, Giorgio Maria Di Nunzio, and Julia Maria Schulz. 2010. LogCLEF 2010: the CLEF 2010 Multilingual Logfile Analysis Track Overview. In *CLEF 2010 LABs and Workshops, Notebook Papers, 22-23 September 2010, Padua,*

Italy (CEUR Workshop Proceedings, Vol. 1176), Martin Braschler, Donna Harman, and Emanuele Pianta (Eds.). CEUR-WS.org. https://ceur-ws.org/Vol-1176/CLEF2010wn-LogCLEF-MandlEt2010.pdf

[85] Behrooz Mansouri, Mohammad Sadegh Zahedi, Ricardo Campos, and Mojgan Farhoodi. 2018. Online Job Search: Study of Users' Search Behavior using Search Engine Query Logs. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR 2018, Ann Arbor, MI, USA, July 08-12, 2018*, Kevyn Collins-Thompson, Qiaozhu Mei, Brian D. Davison, Yiqun Liu, and Emine Yilmaz (Eds.). ACM, 1185–1188. https://doi.org/10.1145/3209978.3210125

[86] Donn Morrison, Stéphane Marchand-Maillet, and Eric Bruno. 2011. Query log simulation for long-term learning in image retrieval. In *9th International Workshop on Content-Based Multimedia Indexing, CBMI 2011, Madrid, Spain, June 13-15, 2011*, José M. Martínez (Ed.). IEEE, 55–60. https://doi.org/10.1109/CBMI.2011.5972520

[87] Donn Morrison, Theodora Tsikrika, Vera Hollink, Arjen P. de Vries, Eric Bruno, and Stéphane Marchand-Maillet. 2013. Topic modelling of clickthrough data in image search. *Multim. Tools Appl.* 66, 3 (2013), 493–515. https://doi.org/10.1007/s11042-012-1038-8

[88] Yashar Moshfeghi. 2021. NeuraSearch: Neuroscience and Information Retrieval. In *DESIRES (CEUR Workshop Proceedings, Vol. 2950)*. CEUR-WS.org, 193–194.

[89] Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. MS MARCO: A Human Generated MAchine Reading COmprehension Dataset. In *Proceedings of the Workshop on Cognitive Computation: Integrating neural and symbolic approaches 2016 co-located with the 30th Annual Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain, December 9, 2016 (CEUR Workshop Proceedings, Vol. 1773)*, Tarek Richard Besold, Antoine Bordes, Artur S. d'Avila Garcez, and Greg Wayne (Eds.). CEUR-WS.org. https://ceur-ws.org/Vol-1773/CoCoNIPS_2016_paper9.pdf

[90] Bruno Oliveira and Carla Teixeira Lopes. 2023. The Evolution of Web Search User Interfaces - An Archaeological Analysis of Google Search Engine Result Pages. *CoRR* abs/2301.08613 (2023).

[91] Bruno Oliveira and Carla Teixeira Lopes. 2023. From 10 Blue Links Pages to Feature-Full Search Engine Results Pages - Analysis of the Temporal Evolution of SERP Features. *CoRR* abs/2301.08042 (2023).

[92] Alexandra Olteanu, Jean Garcia-Gathright, Maarten de Rijke, and Michael D. Ekstrand. 2019. Workshop on Fairness, Accountability, Confidentiality, Transparency, and Safety in Information Retrieval (FACTS-IR). In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2019, Paris, France, July 21-25, 2019*, Benjamin Piwowarski, Max Chevalier, Éric Gaussier, Yoelle Maarek, Jian-Yun Nie, and Falk Scholer (Eds.). ACM, 1423–1425. https://doi.org/10.1145/3331184.3331644

[93] Alexandra Olteanu, Jean Garcia-Gathright, Maarten de Rijke, and Michael D. Ekstrand. 2019. Workshop on Fairness, Accountability, Confidentiality, Transparency, and Safety in Information Retrieval (FACTS-IR). In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2019, Paris, France, July 21-25, 2019*, Benjamin Piwowarski, Max Chevalier, Éric Gaussier, Yoelle Maarek, Jian-Yun Nie, and Falk Scholer (Eds.). ACM, 1423–1425. https://doi.org/10.1145/3331184.3331644

[94] Harrie Oosterhuis and Maarten de Rijke. 2018. Differentiable Unbiased Online Learning to Rank. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management, CIKM 2018, Torino, Italy, October 22-26, 2018*. ACM, 1293–1302. https://doi.org/10.1145/3269206.3271686

[95] João R. M. Palotti, Allan Hanbury, Henning Müller, and Charles E. Kahn Jr. 2016. How users search and what they search for in the medical domain - Understanding laypeople and experts through query logs. *Inf. Retr. J.* 19, 1-2 (2016), 189–224.

[96] Greg Pass, Abdur Chowdhury, and Cayley Torgeson. 2006. A picture of search. In *Proceedings of the 1st International Conference on Scalable Information Systems, Infoscale 2006, Hong Kong, May 30-June 1, 2006 (ACM International Conference Proceeding Series, Vol. 152)*, Xiaohua Jia (Ed.). ACM, 1. https://doi.org/10.1145/1146847.1146848

[97] Martin Potthast, Tim Gollub, Matti Wiegmann, and Benno Stein. 2019. TIRA Integrated Research Architecture. In *Information Retrieval Evaluation in a Changing World*, Nicola Ferro and Carol Peters (Eds.). Springer, Berlin Heidelberg New York. https://doi.org/10.1007/978-3-030-22948-1_5

[98] Tao Qin and Tie-Yan Liu. 2013. Introducing LETOR 4.0 Datasets. *CoRR* abs/1306.2597 (2013). https://doi.org/10.48550/arXiv.1306.2597 arXiv:1306.2597

[99] Filip Radlinski, Paul N. Bennett, and Emine Yilmaz. 2011. Detecting duplicate web documents using clickthrough data. In *Proceedings of the Forth International Conference on Web Search and Web Data Mining, WSDM 2011, Hong Kong, China, February 9-12, 2011*, Irwin King, Wolfgang Nejdl, and Hang Li (Eds.). ACM, 147–156. https://doi.org/10.1145/1935826.1935859

[100] Filip Radlinski and Thorsten Joachims. 2006. Minimally Invasive Randomization for Collecting Unbiased Preferences from Clickthrough Logs. *CoRR* abs/cs/0605037 (2006). https://doi.org/10.48550/arXiv.cs/0605037 arXiv:cs/0605037

[101] Shaurya Rohatgi, C. Lee Giles, and Jian Wu. 2021. What Were People Searching For? A Query Log Analysis of An Academic Search Engine. In *ACM/IEEE Joint Conference on Digital Libraries, JCDL 2021, Champaign, IL, USA, September 27-30, 2021*, J. Stephen Downie, Dana McKay, Hussein Suleman, David M. Nichols, and Faryaneh Poursardar (Eds.). IEEE, 342–343. https://doi.org/10.1109/JCDL52503.2021.00062

[102] Harrisen Scells, Connor Forbes, Justin Clark, Bevan Koopman, and Guido Zuccon. 2022. The Impact of Query Refinement on Systematic Review Literature Search: A Query Log Analysis. In *ICTIR '22: The 2022 ACM SIGIR International Conference on the Theory of Information Retrieval, Madrid, Spain, July 11 - 12, 2022*. ACM, 34–42. https://doi.org/10.1145/3539813.3545143

[103] Craig Silverstein, Monika Rauch Henzinger, Hannes Marais, and Michael Moricz. 1999. Analysis of a Very Large Web Search Engine Query Log. *SIGIR Forum* 33, 1 (1999), 6–12. https://doi.org/10.1145/331403.331405

[104] Amanda Spink, Bernard J Jansen, and H Cenk Ozmultu. 2000. Use of query reformulation and relevance feedback by Excite users. *Internet research* 10, 4 (2000), 317–328.

[105] Amanda Spink, Seda Ozmutlu, Huseyin C Ozmutlu, and Bernard J Jansen. 2002. US versus European Web searching trends. In *ACM Sigir Forum*, Vol. 36. ACM New York, NY, USA, 32–38.

[106] Amanda Spink, Dietmar Wolfram, Major BJ Jansen, and Tefko Saracevic. 2001. Searching the web: The public and their queries. *Journal of the American society for information science and technology* 52, 3 (2001), 226–234.

[107] Xu Sun, Jianfeng Gao, Daniel Micol, and Chris Quirk. 2010. Learning Phrase-Based Spelling Error Models from Clickthrough Data. In *ACL 2010, Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, July 11-16, 2010, Uppsala, Sweden*, Jan Hajic, Sandra Carberry, and Stephen Clark (Eds.). The Association for Computer Linguistics, 266–274. https://aclanthology.org/P10-1028

[108] Ellen Voorhees. 2004. Overview of the TREC 2004 Robust Retrieval Track. In *TREC*.

[109] Ellen M. Voorhees. 2005. Overview of the TREC 2005 Robust Retrieval Track. In *TREC*.

[110] Ji-Rong Wen, Jian-Yun Nie, and HongJiang Zhang. 2002. Query clustering using user logs. *ACM Trans. Inf. Syst.* 20, 1 (2002), 59–81. https://doi.org/10.1145/503104.503108

[111] Dietmar Wolfram, Amanda Spink, Bernard J Jansen, Tefko Saracevic, et al. 2001. Vox populi: The public searching of the web. *JASIST* 52, 12 (2001), 1073–1074.

[112] Ting Yao, Min Zhang, Yiqun Liu, Shaoping Ma, Yongfeng Zhang, and Liyun Ru. 2010. Investigating Characteristics of Non-click Behavior Using Query Logs. In *Information Retrieval Technology - 6th Asia Information Retrieval Societies Conference, AIRS 2010, Taipei, Taiwan, December 1-3, 2010. Proceedings (Lecture Notes in Computer Science, Vol. 6458)*, Pu-Jen Cheng, Min-Yen Kan, Wai Lam, and Preslav Nakov (Eds.). Springer, 85–96. https://doi.org/10.1007/978-3-642-17187-1_8

[113] Deng Yi, Yin Zhang, Haihan Yu, Yanfei Yin, Jing Pan, and Baogang Wei. 2012. Improving multi-faceted book search by incorporating sparse latent semantic analysis of click-through logs. In *Proceedings of the 12th ACM/IEEE-CS Joint Conference on Digital Libraries, JCDL '12, Washington, DC, USA, June 10-14, 2012*, Karim B. Boughida, Barrie Howard, Michael L. Nelson, Herbert Van de Sompel, and Ingeborg Sølvberg (Eds.). ACM, 249–258. https://doi.org/10.1145/2232817.2232864

[114] Dell Zhang and Yisheng Dong. 2002. A novel Web usage mining approach for search engines. *Comput. Networks* 39, 3 (2002), 303–310.

[115] Zhitao Zhang, Muyun Yang, Sheng Li, Haoliang Qi, and Chao Song. 2009. Sogou Query Log Analysis: A Case Study for Collaborative Recommendation or Personalized IR. In *2009 International Conference on Asian Language Processing, IALP 2009, Singapore, December 7-9, 2009*, Min Zhang, Haizhou Li, Kim-Teng Lua, and Minghui Dong (Eds.). IEEE Computer Society, 304–307. https://doi.org/10.1109/IALP.2009.72

[116] Shiqi Zhao, Haifeng Wang, and Ting Liu. 2012. User Behaviors Lend a Helping Hand: Learning Paraphrase Query Patterns from Search Log Sessions. In *COLING 2012, 24th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, 8-15 December 2012, Mumbai, India*, Martin Kay and Christian Boitet (Eds.). Indian Institute of Technology Bombay, 3137–3152. https://aclanthology.org/C12-1192

[117] Xiang Zhou, Pengyi Zhang, and Jun Wang. 2017. Identification and Analysis of Multi-tasking Product Information Search Sessions with Query Logs. *J. Data Inf. Sci.* 1, 3 (2017), 79–94. https://doi.org/10.20309/jdis.201621

[118] Shengyao Zhuang and Guido Zuccon. 2020. Counterfactual Online Learning to Rank. In *Advances in Information Retrieval - 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal, April 14-17, 2020, Proceedings, Part I (Lecture Notes in Computer Science, Vol. 12035)*. Springer, 415–430. https://doi.org/10.1007/978-3-030-45439-5_28