Generative Information Retrieval Evaluation

Marwah Alaofi[†], Negar Arabzadeh[‡], Charles L. A. Clarke[‡], and Mark Sanderson[†]

[†] *RMIT University, Australia*
[‡] *University of Waterloo, Canada*

Abstract

In this chapter, we consider generative information retrieval evaluation from two distinct but interrelated perspectives. First, large language models (LLMs) themselves are rapidly becoming tools for evaluation, with current research indicating that LLMs may be superior to crowdsource workers and other paid assessors on basic relevance judgement tasks. We review past and ongoing related research, including speculation on the future of shared task initiatives, such as TREC, and a discussion on the continuing need for human assessments. Second, we consider the evaluation of emerging LLM-based generative information retrieval (GenIR) systems, including retrieval augmented generation (RAG) systems. We consider approaches that focus both on the end-to-end evaluation of GenIR systems and on the evaluation of a retrieval component as an element in a RAG system. Going forward, we expect the evaluation of GenIR systems to be at least partially based on LLM-based assessment, creating an apparent circularity, with a system seemingly evaluating its own output. We resolve this apparent circularity in two ways: 1) by viewing LLM-based assessment as a form of "slow search", where a slower IR system is used for evaluation and training of a faster production IR system; and 2) by recognizing a continuing need to ground evaluation in human assessment, even if the characteristics of that human assessment must change.

1 Introduction

Both the structure of Generative Information Retrieval (GenIR) systems and the capabilities of Large Language Models (LLMs) are evolving rapidly. It would appear from an evaluation perspective, GenIR presents both challenges and opportunities both concrete and speculative.

- **Challenges** stem from evaluating the prosodic form of GenIR output: a written synthesis of answers, and sometimes, hallucinated text replacing the classic search response, a ranking of documents.

- **Opportunities** arise from the prospect of automating components of the methodology to evaluate current document retrieval systems. With the apparent ability of generative methods to simulate human actions, we speculate on a range of potential rapid assessments of the worth of a technology prior to actual user trials.

As with any document written at the start of a revolution, it is too early to say what will come. The functionalities and limitations of GenIR are not yet well understood. In many cases we can only provide a sketch of ongoing research and emerging opportunities. In general, we err on the side of describing future potential rather than surveying the current state of the art, since the latter has changed significantly even between the time we first wrote these words and this, our final proofreading pass. We examine past work to try to contextualise the challenges, opportunities, and speculations in more detail.

We interpret *GenIR evaluation* in two ways: (1) the use of generative methods to aid evaluation practices in Information Retrieval (IR), such as generating document relevance labels, and (2) evaluating the output of a GenIR system, which is likely employing some form of Retrieval Augmented Retrieval (RAG) architecture. Figure 1 provides a brief description of this chapter's sections and outlines their subsections. We start the chapter by reflecting on past assumptions and challenges within evaluation practices and explore how LLMs can challenge these assumptions and contribute to the development of better practices (section 2). We then address the challenges associated with evaluating the output of GenIR systems (section 3). Across both sections, we speculate on possible challenges.

2 Generative Methods for IR Evaluation

The arrival of GenIR systems prompts the question of whether the traditional ranking of search results, known as the ten blue links[1], will be replaced. Ranking, however, remains a common means of search result presentation, which is likely to persist in some form either as an internal component of a RAG system

---

[1]This phrase seems to emerge around 2007 from the makers of Ask Jeeves – one of the earliest question answering systems – seeking to contrast their system's written output with what they saw as a traditional search response [38].
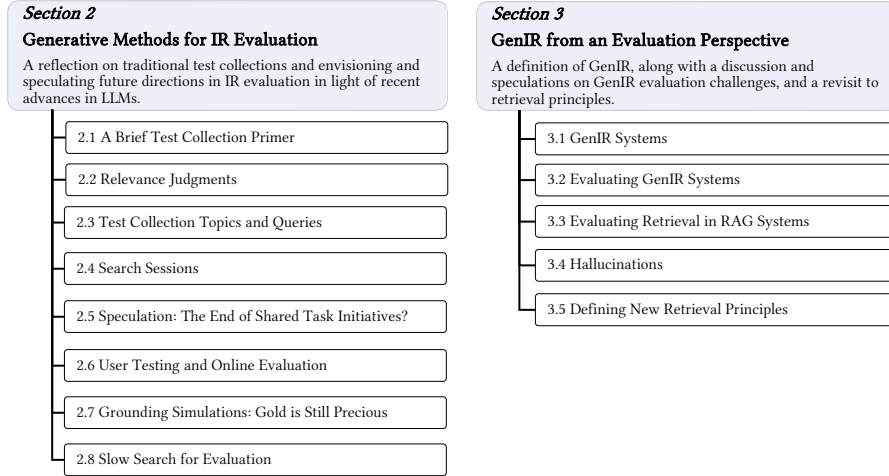
Figure 1: An overview of the main two sections and their subsections.

or in applications where the purpose of a search is to identify items as part of seeking information: e.g. mapping applications returning locations in rank order, music applications matching songs, job search engines sorting employment opportunities in order, etc. In this section, we detail the impact of LLMs on the common offline document ranking evaluation methodology, a test collection [65]. We first provide a brief primer to test collections before detailing the impact of LLMs on a number of test collection components and approaches to testing. We then outline the impact of LLMs on the capturing of relevance judgement, on the creation of topics and queries for test collections, and on search sessions, before speculating on the role of shared tasks initiatives in future. The impact of LLMs on wider user testing is described, before the section concludes with a discussion of the continued role of human labeling in IR evaluation.

2.1 A Brief Test Collection Primer

A typical test collection includes a set of search topics (expressed through queries), a corpus of documents, and relevance judgments that record the documents that are relevant to the topics, often referred to as *qrels*. To test an IR system, queries from the collection are run through the system and its ability to locate relevant documents is measured. Evaluation using a test collection is fully automated, allowing systems to be optimized at low cost to the experimenter. However, there is a substantial cost involved in creating test collections.

According to Voorhees, offline evaluation practices have mainly operated with the following simplifying assumptions:

- *"relevance can be approximated by topical similarity, which implies all relevant documents are equally desirable; relevance of one document is inde-*

pendent of the relevance of any other document; and the user information need is static.

- *a single set of judgments for a topic is representative of the user population.*

- *(essentially) all relevant documents for a topic are known"*, Voorhees [80, p. 47]. To this list, we might add that

- there is one representation of an information need.

Under these assumptions the costs associated with constructing test collections are manageable, but still substantial enough to make it nearly impossible for individual academics or research groups to generate such large comprehensive collections by themselves. Consequently, several initiatives were started to share the costs and labor involved in their creation, such as Text Retrieval Conference (TREC) [79], Cross Language Evaluation Forum (CLEF) [59], and NII Testbeds and Community for Information Access Research (NTCIR) [54]. The challenge of constructing test collections meant that researchers predominantly relied on those produced by these initiatives. The subsequent sections discuss the challenges in more detail and present opportunities for using LLMs to address these challenges.

Early test collections consisted of a few hundred to a few thousand documents.[2] Creating relevance judgments for all documents in such a collection was practically possible. For instance, when creating Cranfield II, Cleverdon employed a team of individuals to manually scan the entire collection to identify all relevant documents for each topic. When building the LISA test collection during the 1980s, one person was employed to search the physical issues of a journal to find relevant documents, which was supplemented with some online search.[3] The scale of test collections was limited by the cost of creating relevance judgments. There was a need among IR researchers to find a way '*to produce larger test collections while at the same time locate as many relevant documents as possible*' [65, p. 271].

Multiple strategies were explored, though not practically implemented, for creating larger collections, most notably Spärck Jones and Van Rijsbergen [69] introduced document pooling. This technique, which involves sampling documents for relevance assessment through multiple participating searches (now *runs*), subsequently became the conventional method for building IR test collections and is the standard within TREC. Although this approach has its limitations, primarily due to missing some relevant documents [88], which in turn raises concerns about the reusability of collections [81, 17], it has facilitated the expansion of test collections, giving us access to test collections with massive corpora, such as the ClueWeb series, with millions of documents.

---

[2]Readers can refer to the website hosted by the University of Glasgow, which archives some of the early test collections, to gain a sense of their modest scale: `https://ir.dcs.gla.ac.uk/resources/test_collections/`

[3]See the readme file for further information: `https://ir.dcs.gla.ac.uk/resources/test_collections/lisa/`

2.2 Relevance Judgments

Recent studies have demonstrated that LLMs can be used to produce relevance judgments (or labels, distinguishing them from those generated by humans). In May 2023, researchers at Microsoft Bing announced their use of GPT-4 in generating relevance labels, which was later shared in a paper [72]. The LLM generated labels were found to be as accurate as labels created by crowd source workers and were being used to train the production system of Bing. Around the same time, Faggioli et al. [35] reported promising results from using LLMs for generating relevance labels. Although these findings have not been extensively tested and may have limitations, they prompt a reevaluation of the need for document pooling, originally adopted to manage the costs associated with human labor. That is, it might be now feasible to create complete relevance judgments on a large scale, or at least, create deeper pools as the cost of generating relevance labels has substantially decreased.

The use of LLMs to reduce the cost of relevance judgments echoes a significant historical shift in the value of a material we now take for granted: aluminum. In middle school, American children learn that on the top of the Washington Monument is a relatively small pyramid of solid aluminum. At the time it was placed there, in 1884, aluminum was as rare, and as precious, as silver. The pyramid was the largest piece of solid aluminum in the world. Two years later, Paul Héroult and Charles Hall independently invented a process that would eventually make aluminum cheap enough that when buying an aluminum can of drink, most of the price pays for the contents, not the container. IR is having its Hall-Héroult moment: human judgment was once a rare and precious resource. Now, it appears we can simply ask the LLM anything we might ask a human searcher or assessor, but at a much lower cost. This opens many new opportunities for evaluation.

One opportunity is to tailor the definition of relevance to be more specific, including additional dimensions of information utility to different users. Voorhees et al. highlight the score saturation problem in the TREC Deep Learning Track (2021), where many systems are already capable of retrieving ten relevant documents for a wide range of queries from large corpora, calling for *"different metrics or a more focused definition of relevance"* [81]. Relevance can vary across users and contexts, and it is often assessed based on topicality without considering other dimensions, such as understandability. This underscores the need to consider other dimensions of relevance to create test collections that can distinguish among systems. For example, a document might be topically relevant to a query but could exhibit different levels of utility to users based on their domain expertise or operating contexts. It now seems feasible to explore the utility of LLMs to make relevance labels more specific, enabling a detailed and most importantly realistic system evaluation.

Another potential benefit of using LLMs to produce relevance labels is their consistency in the generated labels for documents. Unlike humans, LLMs do not get tired as they generate more labels, nor are they influenced by judgements previously made. There is evidence that there is a great level of inconsistency in

human relevance assessment [67, 66], whether due to forgetting earlier decisions, re-calibrating assessments based on the documents already seen, or simply making errors, leading to varying assessments even for almost identical documents [13, 67]. Using a recognised model of LLM with controlled parameters to ensure a deterministic behaviour would enable consistency and reproducibility of relevance labels.

While it might be conceivable that the need for collecting these relevance labels and distributing them in test collections could diminish, given that system effectiveness can now be evaluated dynamically and at substantially lower than those incurred using human labor (see [72, Figure 5] for cost-accuracy relative comparison), this approach would undermine the core principle of having test collections serving as static, shared, and reusable resources for system evaluation. See further discussion in section 2.5.

2.3   Test Collection Topics and Queries

In test collections, the convention is that each information need (search topic) is represented using a *single* query. The queries are generated by either (1) consulting a group of people to generate queries given information need statements, or by (2) obtaining a sample from a query log. Going beyond one query to represent a broader spectrum of users employing different query variants was expensive and thought to be unnecessary. However, research suggests that when seeking a common information need, users tend to use a large number of query formulations (often referred to as *query variants*). In studies of user populations, over fifty variants were found per information need [11, 47]. Previous research has demonstrated that factors – such as the used device [26, 42], domain expertise [52, 83], age [75, 14], and language proficiency [25] – influence query formulation. These consequentially impact the quality of search results and overall user satisfaction. Culpepper et al. [32] showed that the impact of query variants on system effectiveness is substantially greater than that due to topic or ranking models. Yet, the effect of query variation on IR system effectiveness is often overlooked. Evaluations typically rely on test collections with single queries, leaving the performance of systems for a broader range of users largely unexamined. Given the recent studies demonstrating an important role of query variants in system evaluation, how such variants might be generated in a cost effective manner is challenge that LLMs may be able to help with.

Unlike with relevance judgements where LLMs have been shown to be a valid substitute for human labels, the work on query variants in more in its infancy. Using artificially created and manually verified query variants, Penha et al. [56] showed a significant drop in the effectiveness of both neural and transformer-based retrieval models. Likewise, Alaofi et al. [3] undertook an empirical investigation into the effects of query variants on a commercial search engine and some inverted indexes. Their research revealed inconsistency in search results across different query variants and shed light on the impact of variants on document retrievability. Similarly, inconsistencies in search results were also demonstrated in the context of searches conducted by children [58].

6

Crowdsourcing and click graphs have been used to gather query variants. However, both methods have their limitations: crowdsourcing is expensive to scale, and click graphs are noisy and lack information about users. User simulation has been a prevalent instrument in IR, but its application for generating query variants has not been as extensively explored. For example, Penha et al. [56] proposed a taxonomy for query variants and use multiple techniques to artificially create query variants. More recently, research has shown that LLMs can, to a limited extent, reproduce human query variants, yielding a similar pool of documents of that obtained by using human generated query variants [4]. Engelmann et al. [34] also used LLMs to simulate query variants in an interactive manner, taking into account user sessions and the results seen as feedback for the query generation process. This approach yields more effective search sessions, but does not necessarily reflect how humans engage with search sessions. Another line of research explores using LLMs to generate queries but not for simulations but as a way of generating more query variants to train better rankers (e.g. [15]), generate query expansions (e.g., [48]), and improve document retrievablity (e.g., [57]). Giving the ability of LLMs to align its generation to certain properties, an important question arises regarding how effectively they can align with how humans engage with information seeking tasks, reflecting the diverse user properties identified in the literature as influencing query formulation.

2.4 Search Sessions

There has long been a recognition that there is more to evaluation than the initial query that establishes a search. Many attempts [8] to extend offline evaluation to include sessions have been tried [20], but as with most efforts to 'shift the dial' of offline evaluation, those efforts have not been successful in starting a new standard.

Many of the reasons underpinning the lack of movement in the design of offline evaluation has been a question of cost. The current approach to evaluation while expensive to set up, is cheap to use when built. Most approaches to extending the evaluation of search have been more expensive to create, or require higher on-going costs to use. The arrival of generative methods and the ability of generative to apparently simulate human behaviour to a convincing degree, suggests a shifting of the dial. This has already been demonstrated with relevance assessments, but it may also be possible to have viable simulations of interactive sessions with a search engine including effective simulations of document selection and query reformulation, as well as simulations that determine when a search would stop seeking more documents.

2.5 Speculation: The End of Shared Task Initiatives?

The arrival of generative systems has the potential to completely redefine how evaluation is conducted in the field of Information Access. Much of this chapter has focused on existing innovations and future speculations on what might be possible using generative methods. It is worth asking if generative systems may

also alter the way that researchers behave. For decades, the field of IR has been characterised by the creation and sharing of large resources that can be used for evaluation. Key amongst these resources is the test collection. Initially something that was just created by one group and shared with others, this evolved into large scale shared evaluation tasks starting with TREC in the early 1990s. The tasks were formed in order to share the work required to build large evaluation resources. However, if it is possible to construct evaluation resources individually through the use of generative methods, one might question if these large scale collaborative evaluation exercises will continue. The costs to researchers of building bespoke data sets with human generated labels has come down substantially thanks to the rise of crowd sourcing services. Consequently, participation rates at exercises around the world have dropped substantially in recent years. The rise of generative methods simulating the behaviour of users and data labelling may be the final nail in the coffin of these long standing mainstays of our research ecosystem.

2.6   User Testing and Online Evaluation

Traditional IR systems returned just a ranked list of documents, see for example, Harman's review of pre-web systems [41]. Over time, the sophistication of ranked output grew. The way ranked documents were displayed depended on the text of the query, thanks to snippets [73], a summary composed of query focused content extracted from the body of the document [27]. Commercial search engines further augmented the output with direct answers [84], quick links [21, 39], entity cards [16, 53], query suggestions [19] and other components [55]. The sophistication of the output prompted work on so-called whole-page relevance [10], but in the academic community, this approach was not widely adopted, most likely due to the costs of using it.

    In the speculations detailed so far in this chapter, the main focus has been on the way that offline evaluation is being redefined through the use of generative methods to label documents as relevant and to generate queries arising from an information need. However, there may be the potential for such replacements to expand into other aspect of evaluation. [40] detailed how LLMs could be used to simulate many qualitative human responses to the use of and the reactions to systems employed in usability experiments finding that LLMs can "*yield believable accounts of HCI experiences*". It may be possible to revisit whole-page relevance evaluation using generative methods.

2.7   Grounding Simulations: Gold is Still Precious

Evaluation outcomes of systems using test collections reflect 'anticipated' real-world performance. Although these test collections appear concrete, featuring human queries and relevance judgments, they are fundamentally abstract and considerably simplified *simulations* of real-world search scenarios. Use of so-called *offline evaluation* imagines a simplified process of a searcher browsing the sorted list, top to bottom, identifying relevant pages and at some point

stopping[50, 51, 5, 68, 82]. The extent to which evaluation outcomes reflect actual user satisfaction is crucial; yet, it has not received much attention within the community. It was not until 2006 that Turpin and Scholer [76] demonstrated that test collections may poorly reflect reality. This was further investigated by Al-Maskari and Sanderson [2].

The use of LLMs to simulate users in creating test collections raises questions about the validity of this simulation and necessitates further exploration of how well LLMs are aligned with real users. Before going further in simulation and drawing conclusions about how well systems perform, we need to first substantiate the validity of our user simulations. This requires datasets, tools, metrics, and procedures.

User relevance judgments and queries are abundantly available through numerous iterations of shared tasks. Consequently, the approximation of queries and relevance labels to human generated ones can be examined. However, if personalized relevance labels are to be simulated, for example, taking into account other dimensions of information utility, then we have almost no way to validate their performance since such ground truth data is not widely available. For instance, in a context where we would like to evaluate how well a system performs in response to an expert user as opposed to a non-expert, such data is not readily available. Similarly, when simulating query variants issued by multiple users, very few sources of data are available for validation, and demographic data is often missing. Real human data that fits the definition of gold [9], where both the query and relevance assessments are produced by a diverse set of humans operating in different contexts and demographic data is collected are highly needed in order to facilitate the research of simulation validation.

In terms of measuring the accuracy of simulations, that is how closely the LLM aligns with human searchers, one can consider if the simulated data exhibits similar properties to human-generated data or leads to comparable conclusions [12], as exact matches may not be feasible in tasks involving language, where queries can be formulated in various ways. Statistical properties of queries, such as length and complexity, can serve as indicators. Other metrics may assess the impact of simulated data compared to human-generated data. For instance, do the generated queries demonstrate similar effectiveness to human queries and/or produce similar pools of documents? Do the relevance labels result in the same system rankings as if those produced by humans are used?

2.8 Slow Search for Evaluation

In 2023, researchers proposed replacing human relevance assessments with LLM assessments [72, 35]. A common objection to these proposals recognizes their circularity. Using automated methods to assess other automated methods is not without its dangers. If, as is common in this chapter, one looks at historical precedence for current events. One could look at the way in which automated relevance assessments were attempted earlier in the history of IR. A classic example is pseudo relevance feedback [31]. This is a technique that assumes a

query from a user will be sufficiently accurate that one can make the assumption that top ranked documents returned by that initial query are themselves likely to be relevant. The text of those documents can then be used in an internal reformulation of the query to produce better results. While rarely seen in commercial systems, pseudo relevance feedback is a well known technique.

If LLM assessment is sufficient to replace a human assessment, then why not treat the LLM as a ranker, ranking items according to their LLM assessed relevance? If an LLM-based evaluation is generating the labels for evaluation, ranking by those labels always produces an ideal result.

One way to avoid this circularity is to consider the difference in time and resources needed by a production GenIR system vs. the time and resources required for LLM-based evaluation. For evaluation purposes, we can take all the time we need to find the best response, and then use that response to evaluate the efficiency vs. effectiveness trade-off between, for example, a production system that responds in 100ms and one that responds in 500ms. From the standpoint of an efficiency vs. effectiveness trade-off, for the purposes of evaluation we can essentially ignore efficiency.

The trade-off between retrieval efficiency and effectiveness has long been a subject of academic research [18, 28, 7, 87, 6] and a key consideration for commercial search engines, which aim for an average query latency in the hundreds of milliseconds [49, 7]. However, in the past we have had relatively few methods for tuning the trade-off between efficiency and effectiveness beyond a narrow range. Efficiency vs. effectiveness trade-offs might be measured in terms of tiny percentages of effectiveness improvements at the cost of milliseconds of query latency, but we could never improve effectiveness enough to justify a latency of seconds or longer

Teevan et al. [71] in advocating for "Slow Search" write, "*With even just a little extra time to invest, search engines can relax existing restrictions to improve search result quality. For example, complex query processing can be done to identify key concepts in the query, and multiple queries derived from the initial query can be issued to broaden the set of candidate documents to cover different aspects of the query.*"

Unfortunately, it was never fully demonstrated that investing more time would ever achieve these goals. We had no way to operationalize the proposal of Teevan et al. [71]. If a search engine is fast, the searcher can quickly see if the results are not relevant and immediately reformulate their query [62]. If a query is missing a key concept, the searcher can add it. Low latency is an important feature of search engines, since it facilitates rapid interaction. We can only justify higher latency if rapid interaction is not required.

We have now entered an era where deriving multiple queries and other complex query processing might genuinely improve the results in more than a trivial way. With more time, our GenIR system might prompt an LLM to make relevance judgments, determine what aspects of a document make it relevant, and automatically refine queries in light of these determinations. A Gen-IR system might compare one item against another, until it identifies the best overall result. In some cases, it might be worth the time of the searcher to wait for

this result, but if not, it can still be used to evaluate the faster result actually returned to the searcher.

In some sense, evaluation has always been slow search with a human in the loop. In a traditional TREC ad hoc task, we build a pool and humans assess items in the pool, creating an ideal response. Now we can use an LLM to replace these humans. However, unless we determine that taking all the time we need always produces the best possible response, we still need a way to evaluate the results of slow search. If the quality of LLM assessment can reach the level of traditional human assessment, do we consider this as our peak achievement? Or do we recognize that there is still room for improvement by involving humans to perhaps monitor LLMs or revisit our ideal definition of relevance.

3   GenIR from an Evaluation Perspective

In the previous section, we considered the use of generative methods to aid evaluation practices in current IR systems, and in particular for generating document relevance labels. In this section, we consider the evaluation of emerging IR systems that may not adhere to conventional assumptions about ranking and result presentation.

3.1   GenIR Systems

The current and potential capabilities of GenIR systems were engendered by the increasing capabilities of LLMs especially their ability to conduct zero-shot natural language tasks, including summarization, query understanding, and query expansion. Most GenIR systems replace the query and ranked list with a conversation and a written synthesis of information, similar to that shown in Figure 2. At the time of writing, these systems include Perplexity[4] and newer versions of Bing[5]. The TREC 2024 RAG Track, which supersedes the Deep Learning Track, also assumes this interface format[6]. The searcher poses a question in potentially a longer, more natural and conversational form. The system responds with a single coherent answer, which may be supported by links to sources. Gienapp et al. [37] view a GenIR system as a "synthetical" search engine that searches for sources, "compiles them, synthesizes missing information, presents it coherently, and grounds its claims in the retrieved sources." The system provides searchers with a single unified answer "that covers a complex topic with in-depth analysis from varied perspectives"(Figure 3). Such interactions and outputs will require us to seek a new evaluation model.

For evaluation purposes, we need not make any assumptions about the internal architecture of a GenIR system, which may simply be a single large neural model. In this case, our evaluation must focus on the end-to-end interaction. A query or question is entered by the searcher and the system responds with

---

[4]`perplexity.ai`
[5]`bing.com`
[6]`trec-rag.github.io`

Figure 2: A GenIR user interface from March 2024 (from `perplexity.ai`)

an answer, which may reflect a larger conversational context, including personalization. Under this view, our core search metric becomes: How good is this overall response? In traditional IR evaluation, the focus was often on the output of a ranker. While whole-page relevance was a factor in evaluation, especially in industry contexts [10], it was one of many factors. If we view a GenIR system as a black box, whole-page relevance becomes a central factor.

While we can view a GenIR system as a black box for evaluation purposes, a RAG architecture [36] often underlies a GenIR system. In Figure 4, we have simplified the architecture of a RAG system to its key components. At the front end, a searcher interacts with a *generative component*, which in turn interacts with both an *LLM* and a *retrieval component*. The retrieval component is used to search a corpus, which is assumed as a source of ground truth – although, like any IR system, the corpus itself may contain spam and documents of varying quality. Information provided by the RAG system to the searcher requires support from the corpus. The generative component interacts with the LLM for purposes of query understanding, query expansion, summarization, and similar tasks, while it interacts with the retrieval component through keyword or other queries to find sources for its response. The system may interact with both the LLM and the retrieval component multiple times before responding to a user's query, where the overall approach may be retrieve-then-generate, generate-then-
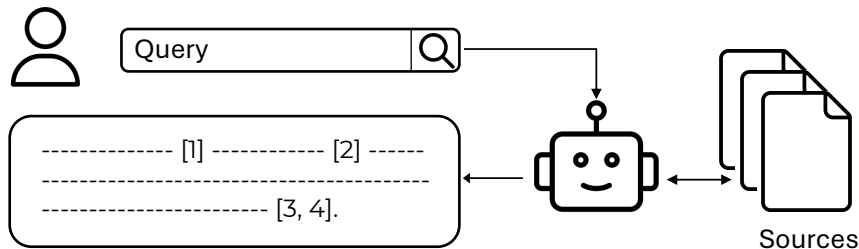
Figure 3: A GenIR system as a synthetical search engine

retrieve [1] , or a combination of multiple generation and retrieval steps.

A competing definition of "Generative Information Retrieval" systems describes a GenIR system as one that does not generate the answer to the searchers query, instead it replaces a traditional search engine by using a neural model to directly generate the identifiers of documents that answer the query [85, 23, 60, 70, 24]. While these systems are "generative" in the sense that a neural model is directly generating document identifiers, from an evaluation perspective they are no different from any other retrieval component that returns document identifiers, except in one respect. Since the document identifiers are generated, it is conceivable for such systems to "hallucinate" document identifiers that do not exist. Nonetheless, for the purpose of our discussion, we view them as a type of retrieval component.

3.2 Evaluating GenIR Systems

As shown in Figure 4, a RAG system may be evaluated at three points: (1) at the front end, where we are evaluating the end-to-end performance of the system; (2) at the top of the retrieval component, where we are evaluating the retrieval component in the context of the overall GenIR system; and (3) at the point of interaction with the LLM. In the context of the overall system, the retrieval component (#2) is essentially a subordinate system returning a ranked list of items for the generative component. While a human searcher may eventually be given links to items in the corpus, but these will be selected by the generative component. Evaluating interactions with the LLM (#3) falls slightly outside our scope into the broader topic of NLP evaluation, including the evaluation of summarization and information extraction.

Evaluation of an end-to-end GenIR system (#1) introduces challenges beyond those of traditional IR evaluation. Gienapp et al. [37] argue that the key difference between a traditional search engine and a GenIR system is that the GenIR system is essentially searching an infinite corpus of all possible responses that could be synthesized by the system [33]. Traditional IR test collections, such as those created by the TREC, try to be reusable, with a nearly complete set of relevance judgments. With a finite corpus, this approach is conceptually possible, with an infinite corpus it is not.
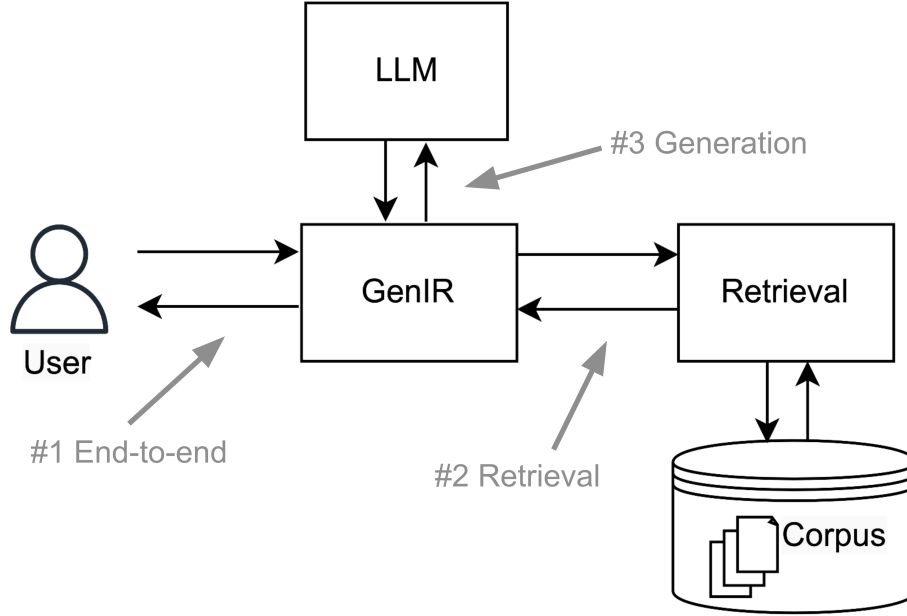
Figure 4: RAG Architecture overview

One approach to evaluating the retrieval component (#2) would be to evaluate it as a traditional search engine. Its role is to execute a query over a corpus of items and return a ranked list of those items. To evaluate the retrieval component of a RAG system we may be able to adapt existing offline evaluation methods. Even if the interface seen by the searcher is no longer "ten blue links", internally we can imagine a similar interface between the generative component and the retrieval component, although the browsing models assumed by offline evaluation metrics no longer apply. These browsing models often assume that the searcher has limited patience [51] or that the searcher will stop scanning the ranked list after a relevant item is found [22]. A generative component might be assumed to dig deeper into the ranked list and seek information from more multiple sources.

3.3 Evaluating Retrieval in RAG Systems

RAG systems include a retrieval component (Figure 4), which supports retrieval over a corpus that provides ground truth for our GenIR system. For evaluation purposes, we might treat the retrieval component as an old-fashioned search engine, even if it itself includes generative components. A query goes into the retrieval component and a ranked list comes out. However, since this response is entirely internal to the GenIR system, it need not only be a ranked list. It could be richer and more complex. The output of the retrieval component must

be tailored to the needs of the overall system, and not to the needs of a human searcher.

If we view the retrieval component as an old-fashioned search engine, returning a ranked list, we might employ traditional evaluation methods. If we think about the GenIR system as internally browsing down the output of the retrieval component, we could use NDCG@10 as our metric. However, the GenIR has more "patience" than a human searcher, so the Normalized Discounted Cumulative Gain (NDCG) discount function might not be the right one to use.

The purpose of the retrieval component is to return the items that the overall GenIR system needs to craft its response. Traditional ranking stacks often use a BM25 based first stage that returns a large collection of items, maybe 1000, for re-ranking by a second-stage ranker [87]. The output of this second stage is then filtered, re-ranked, and processed by more stages until a final stage produces a ranked list that can be shown to the searcher. A typical metric for the first stage is recall@1000. Perhaps recall might be a better metric for evaluating the retrieval component, since the overall GenIR system essentially act as the upper stages.

3.4   Hallucinations

Even when supported by a retrieval component, GenIR systems might generate factually inaccurate or misleading responses. In traditional IR evaluation we assume that the corpus is curated and can be trusted. If we can not trust it, then we filter it for spam and other misinformation. While in traditional web search some pages are higher quality than others, the output of the search engine is a list of pages, which the searcher can ultimately inspect for themselves. They are not depending on the search engine to summarize the information for them.

Since GenIR systems can hallucinate[74], it is not sufficient to filter the corpus for spam and misinformation. We must also evaluate the accuracy of the end-to-end response. The final generated response can be false, or contain falsehoods, even if the retrieved material is true. Fact checking must become a standard component of GenIR evaluation.

The situation has already happened[7] "in the wild". A chatbot on the Air Canada website incorrectly advised a customer, Jake Moffatt, that he could receive a reduced bereavement rate by submitting a claim within 90 days of ticket issue. The response from the chatbot included a link to a static page on the company's website that provided the correct information, indicating that the claim had to be submitted in advance of ticket issue. Air Canada refused the Moffatt's claim. Moffatt took the matter to the Civil Resolution Tribunal of the province of British Columbia who allowed the claim, writing:

> Air Canada argues it cannot be held liable for information provided by one of its agents, servants, or representatives – including a chatbot. It does not explain why it believes that is the case. In effect, Air Canada suggests the chatbot is a separate legal entity that is

---

[7]https://www.canlii.org/en/bc/bccrt/doc/2024/2024bccrt149/2024bccrt149.html

responsible for its own actions. This is a remarkable submission. While a chatbot has an interactive component, it is still just a part of Air Canada's website. It should be obvious to Air Canada that it is responsible for all the information on its website. It makes no difference whether the information comes from a static page or a chatbot.

I find Air Canada did not take reasonable care to ensure its chatbot was accurate. While Air Canada argues Mr. Moffatt could find the correct information on another part of its website, it does not explain why the webpage titled "Bereavement travel" was inherently more trustworthy than its chatbot. It also does not explain why customers should have to double-check information found in one part of its website on another part of its website.

While technical details of the chatbot are not available, we can view it as a RAG system since it returned both a generated answer and a link intended to support the answer. While this is a minor matter from a legal standpoint, it demonstrates that a RAG system can generate materially false information, even when supported by retrieved information that is correct. Extracted webpage summaries have long been a feature of web search results [27]. While extracted summaries may not always provide the information the searcher requires, they generally provide an accurate quote from the page or its metadata.

The accuracy of a traditional search engine depends on the accuracy of the information in its corpus. The search engine may not be able to find relevant information, but when it does, it does not alter or interfere with it. If the corpus contains misinformation, we attempt to filter it. For evaluation purposes, we measure the quality of the filter. Since a GenIR system can hallucinate misinformation, we must now evaluate accuracy of its output, along with relevance and other traditional considerations.

3.5 Defining New Retrieval Principles

Such is the ubiquity of documents in retrieval system design and evaluation, many of the fields key principles are grounded in documents. We briefly detail three of the best known retrieval: Robertson's Probability Ranking Principle (PRP) [63], Jardine and van Rijsbergen's Cluster Hypothesis [44], and Craswell et al.'s Cascade Model [30].

Robertson's PRP is widely viewed as a fundamental goal of ranking in IR. It is most commonly expressed as: "*If an IR system's response to each query is a ranking of the documents in the collection in order of decreasing probability of relevance, then the overall effectiveness of the system to its users will be maximized.*" The notion of an ideal ranking, which is built into traditional evaluation metrics such as NDCG [45] depend on the PRP, that the best result is to order items according to their probability of relevance.

The Cluster Hypothesis was defined twice, first as: '*It is intuitively plausible that the associations between documents convey information about the relevance*

of documents to requests.". Later, van Rijsbergen [77, Chapter 3] simplified the hypothesis as "*closely associated documents tend to be relevant to the same requests*". The hypothesis inspired many later approaches to the clustering of documents [78, 43] as well as result diversification [46].

Seeking a simplified model of user behavior, Craswell et al. examined large user interaction logs in an attempt to capture a broad form of behaviour of user interaction. They produced the cascade model: "*where users view results from top to bottom and leave as soon as they see a worthwhile document*". This model has underpinned a great many modern evaluation measures and also inspired many subsequent studies developing extensions to this model.

All three ideas assume the fundamental unit in retrieval is the document. In the case of GenIR, the entirety of the system's end-to-end response should be relevant and nothing should be redundant, the boundaries between documents hold far less importance. Everything in the response should be there for a reason, and in many cases the response should include more than just the bare answer. The response might link to background articles that support the response. It might provide opposing perspectives. It might suggest cheaper or higher quality alternatives to a product. It might synthesise similar responses from multiple sources into a single sentence. It might ask for clarification or disambiguation.

We might ask what replaces these principles in a GenIR system. One idea is provided by the work of Rajput et al. [61]. They propose *nuggets* as a basis for evaluation, where we might think of nuggets as an atomic unit of relevance, e.g., some fact, relationship, or concept that a perfectly relevant document would contain [64, 29]. Rajput et al. [61] propose to build a reusable test collection in a two-phase process. In the first phase, human assessors would identify and extract nuggets from relevant documents. In the second phase, these nuggets would be automatically matched against unjudged documents to measure relevance, providing a reusable test collection that does not depend on a fixed corpus with relevance labels for individual items.

While they provide experimental support demonstrating both the feasibility and benefits of this approach, it was not widely adopted for either academic or industry assessment. Possible reasons include the need for reliable and trained assessors to identify nuggets, as well as the need to automatically match the nuggets against documents. In 2012, they could only suggest a surface-level, lexical approach to matching, and of course humans are expensive. Crowdsourcing might reduce the cost, but might increase noise and decrease reliability.

In 2024, an LLM might be expected to reliably and cheaply extract nuggets and match them against documents. All it takes is a few calls to an API, costing fractions of a cent per call. It's now almost trivial to realize the vision of Rajput et al. [61], and this proposal is just one of many such proposals in the literature. All the proposals for IR evaluation in terms of diversity, novelty, fairness, completeness, conciseness, effort, or whatever are now both cheap and straightforward to implement.

We can already see nugget-based evaluation emerging as a basis for GenIR

evaluation. For example, the new TREC 2024 RAG track [8] takes a nugget-based approach. To formulate a general principle we turn to Zhai et al. [86]. They propose *subtopic evaluation*, which is closely related to nugget-based evaluation. Evaluation with subtopics is "based on *dependent relevance*, instead of *independent relevance*, as has been assumed in most traditional retrieval methods. The subtopic retrieval problem has to do with finding documents that cover as many different subtopics as possible". To extend this idea to GenIR, we might articulate a principle that the system's response should cover as many nuggets or subtopics as possible.

4 Conclusions

Evaluation lies at the core of so much of IR research. If there is any aspect that separates this field from others, it is focus on high quality evaluation of systems. In this chapter, we examined the impact of LLM on the evaluation of IR both from the perspective of exploiting the models to speed up traditional evaluation methodologies and to consider the more challenging prospect of evaluating a fully generated response following a conversational interaction. There are some clear early wins such as the revelation that LLMs can be used to generate relevance labels, however, as with any technology when it is first introduced, the boundaries of what the technology can achieve and – more importantly what it can't – are still being drawn. We have attempted to describe what currently sits within those boundaries, what is yet to be known, and what might change in our field.

References

[1] Abbasiantaeb, Z. and Aliannejadi, M. [2024], 'Generate then retrieve: Conversational response retrieval using llms as answer and query generators'.

[2] Al-Maskari, A. and Sanderson, M. [2010], 'A review of factors influencing user satisfaction in information retrieval', *Journal of the American Society for Information Science (JASIST)* **61**(5), 859–868.
URL: *https://doi.org/10.1002/asi.21300*

[3] Alaofi, M., Gallagher, L., McKay, D., Saling, L. L., Sanderson, M., Scholer, F., Spina, D. and White, R. W. [2022], Where do queries come from?, *in* E. Amigó, P. Castells, J. Gonzalo, B. Carterette, J. S. Culpepper and G. Kazai, eds, 'The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022', ACM, pp. 2850–2862.
URL: *https://doi.org/10.1145/3477495.3531711*

---

[8]`trec-rag.github.io`

[4] Alaofi, M., Gallagher, L., Sanderson, M., Scholer, F. and Thomas, P. [2023], Can generative llms create query variants for test collections? an exploratory study, *in* H. Chen, W. E. Duh, H. Huang, M. P. Kato, J. Mothe and B. Poblete, eds, 'Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2023, Taipei, Taiwan, July 23-27, 2023', ACM, pp. 1869–1873.
URL: *https://doi.org/10.1145/3539618.3591960*

[5] Arabzadeh, N., Kmet, O., Carterette, B., Clarke, C. L. A., Hauff, C. and Chandar, P. [2023], A is for adele: An offline evaluation metric for instant search, *in* M. Yoshioka, J. Kiseleva and M. Aliannejadi, eds, 'Proceedings of the 2023 ACM SIGIR International Conference on Theory of Information Retrieval, ICTIR 2023, Taipei, Taiwan, 23 July 2023', ACM, pp. 3–12.
URL: *https://doi.org/10.1145/3578337.3605115*

[6] Arabzadeh, N., Yan, X. and Clarke, C. L. A. [2021], Predicting efficiency/effectiveness trade-offs for dense vs. sparse retrieval strategy selection, *in* G. Demartini, G. Zuccon, J. S. Culpepper, Z. Huang and H. Tong, eds, 'CIKM '21: The 30th ACM International Conference on Information and Knowledge Management, Virtual Event, Queensland, Australia, November 1 - 5, 2021', ACM, pp. 2862–2866.
URL: *https://doi.org/10.1145/3459637.3482159*

[7] Arapakis, I., Bai, X. and Cambazoglu, B. B. [2014], Impact of response latency on user behavior in web search, *in* S. Geva, A. Trotman, P. Bruza, C. L. A. Clarke and K. Järvelin, eds, 'The 37th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '14, Gold Coast , QLD, Australia - July 06 - 11, 2014', ACM, pp. 103–112.
URL: *https://doi.org/10.1145/2600428.2609627*

[8] Azzopardi, L., Järvelin, K., Kamps, J. and Smucker, M. D. [2010], 'Report on the SIGIR 2010 workshop on the simulation of interaction', *SIGIR Forum* **44**(2), 35–47.
URL: *https://doi.org/10.1145/1924475.1924484*

[9] Bailey, P., Craswell, N., Soboroff, I., Thomas, P., de Vries, A. P. and Yilmaz, E. [2008], Relevance assessment: are judges exchangeable and does it matter, *in* S. Myaeng, D. W. Oard, F. Sebastiani, T. Chua and M. Leong, eds, 'Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2008, Singapore, July 20-24, 2008', ACM, pp. 667–674.
URL: *https://doi.org/10.1145/1390334.1390447*

[10] Bailey, P., Craswell, N., White, R. W., Chen, L., Satyanarayana, A. and Tahaghoghi, S. M. M. [2010], Evaluating whole-page relevance, *in* F. Crestani, S. Marchand-Maillet, H. Chen, E. N. Efthimiadis and J. Savoy, eds, 'Proceeding of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2010, Geneva,

Switzerland, July 19-23, 2010', ACM, pp. 767–768.
URL: *https://doi.org/10.1145/1835449.1835606*

[11] Bailey, P., Moffat, A., Scholer, F. and Thomas, P. [2017], Retrieval consistency in the presence of query variations, *in* N. Kando, T. Sakai, H. Joho, H. Li, A. P. de Vries and R. W. White, eds, 'Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, Shinjuku, Tokyo, Japan, August 7-11, 2017', ACM, pp. 395–404.
URL: *https://doi.org/10.1145/3077136.3080839*

[12] Balog, K. and Zhai, C. [2023], 'User simulation for evaluating information access systems'.
URL: *https://doi.org/10.48550/arXiv.2306.08550*

[13] Bernstein, Y. and Zobel, J. [2005], Redundant documents and search effectiveness, *in* O. Herzog, H. Schek, N. Fuhr, A. Chowdhury and W. Teiken, eds, 'Proceedings of the 2005 ACM CIKM International Conference on Information and Knowledge Management, Bremen, Germany, October 31 - November 5, 2005', ACM, pp. 736–743.
URL: *https://doi.org/10.1145/1099554.1099733*

[14] Bilal, D. and Gwizdka, J. [2018], 'Children's query types and reformulations in google search', *Information Processing & Management (IP&M)* **54**(6), 1022–1041.
URL: *https://www.sciencedirect.com/science/article/pii/S0306457317308889*

[15] Bonifacio, L. H., Abonizio, H. Q., Fadaee, M. and Nogueira, R. F. [2022], Inpars: Unsupervised dataset generation for information retrieval, *in* E. Amigó, P. Castells, J. Gonzalo, B. Carterette, J. S. Culpepper and G. Kazai, eds, 'SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022', ACM, pp. 2387–2392.
URL: *https://doi.org/10.1145/3477495.3531863*

[16] Bota, H. S., Zhou, K. and Jose, J. M. [2016], Playing your cards right: The effect of entity cards on search behaviour and workload, *in* D. Kelly, R. Capra, N. J. Belkin, J. Teevan and P. Vakkari, eds, 'Proceedings of the 2016 ACM Conference on Human Information Interaction and Retrieval, CHIIR 2016, Carrboro, North Carolina, USA, March 13-17, 2016', ACM, pp. 131–140.
URL: *https://doi.org/10.1145/2854946.2854967*

[17] Buckley, C., Dimmick, D., Soboroff, I. and Voorhees, E. [2006], Bias and the limits of pooling, *in* 'Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval', SIGIR '06, Association for Computing Machinery, New York, NY, USA, p. 619–620.
URL: *https://doi.org/10.1145/1148170.1148284*

[18] Büttcher, S. and Clarke, C. L. A. [2005], Efficiency vs. effectiveness in terabyte-scale information retrieval, *in* E. M. Voorhees and L. P. Buckland, eds, 'Proceedings of the Fourteenth Text REtrieval Conference, TREC 2005, Gaithersburg, Maryland, USA, November 15-18, 2005', Vol. 500-266 of *NIST Special Publication*, National Institute of Standards and Technology (NIST).
URL: *http://trec.nist.gov/pubs/trec14/papers/uwaterloo-tera.pdf*

[19] Cao, H., Jiang, D., Pei, J., He, Q., Liao, Z., Chen, E. and Li, H. [2008], Context-aware query suggestion by mining click- through and session data, *in* 'Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining', KDD '08, Association for Computing Machinery, New York, NY, USA, p. 875–883.
URL: *https://doi.org/10.1145/1401890.1401995*

[20] Carterette, B., Clough, P. D., Hall, M. M., Kanoulas, E. and Sanderson, M. [2016], Evaluating retrieval over sessions: The TREC session track 2011-2014, *in* R. Perego, F. Sebastiani, J. A. Aslam, I. Ruthven and J. Zobel, eds, 'Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval, SIGIR 2016, Pisa, Italy, July 17-21, 2016', ACM, pp. 685–688.
URL: *https://doi.org/10.1145/2911451.2914675*

[21] Chakrabarti, D., Kumar, R. and Punera, K. [2009], Quicklink selection for navigational query results, *in* J. Quemada, G. León, Y. S. Maarek and W. Nejdl, eds, 'Proceedings of the 18th International Conference on World Wide Web, WWW 2009, Madrid, Spain, April 20-24, 2009', ACM, pp. 391–400.
URL: *https://doi.org/10.1145/1526709.1526762*

[22] Chapelle, O., Metlzer, D., Zhang, Y. and Grinspan, P. [2009], Expected reciprocal rank for graded relevance, *in* D. W. Cheung, I. Song, W. W. Chu, X. Hu and J. Lin, eds, 'Proceedings of the 18th ACM Conference on Information and Knowledge Management, CIKM 2009, Hong Kong, China, November 2-6, 2009', ACM, pp. 621–630.
URL: *https://doi.org/10.1145/1645953.1646033*

[23] Chen, J., Zhang, R., Guo, J., de Rijke, M., Chen, W., Fan, Y. and Cheng, X. [2023], Continual learning for generative retrieval over dynamic corpora, *in* I. Frommholz, F. Hopfgartner, M. Lee, M. Oakes, M. Lalmas, M. Zhang and R. L. T. Santos, eds, 'Proceedings of the 32nd ACM International Conference on Information and Knowledge Management, CIKM 2023, Birmingham, United Kingdom, October 21-25, 2023', ACM, pp. 306–315.
URL: *https://doi.org/10.1145/3583780.3614821*

[24] Chen, J., Zhang, R., Guo, J., de Rijke, M., Liu, Y., Fan, Y. and Cheng, X. [2023], A unified generative retriever for knowledge- intensive language tasks via prompt learning, *in* H. Chen, W. E. Duh, H. Huang, M. P. Kato,

J. Mothe and B. Poblete, eds, 'Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2023, Taipei, Taiwan, July 23-27, 2023', ACM, pp. 1448–1457.
URL: *https://doi.org/10.1145/3539618.3591631*

[25] Chu, P., Komlodi, A. and Rózsa, G. [2015], Online search in english as a non-native language, *in* 'Information Science with Impact: Research in and for the Community - Proceedings of the 78th ASIS&T Annual Meeting, ASIST 2015, St. Louis, Missouri, Missouri, USA, October 6-10, 2015', Vol. 52, Wiley, pp. 1–9.
URL: *https://doi.org/10.1002/pra2.2015.145052010040*

[26] Church, K., Smyth, B., Cotter, P. and Bradley, K. [2007], 'Mobile information access: A study of emerging search behavior on the mobile internet', *ACM Transactions on the Web (TWEB)* **1**(1), 4.
URL: *https://doi.org/10.1145/1232722.1232726*

[27] Clarke, C. L. A., Agichtein, E., Dumais, S. T. and White, R. W. [2007], The influence of caption features on clickthrough patterns in web search, *in* W. Kraaij, A. P. de Vries, C. L. A. Clarke, N. Fuhr and N. Kando, eds, 'Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Amsterdam, The Netherlands, July 23-27, 2007', ACM, pp. 135–142.
URL: *https://doi.org/10.1145/1277741.1277767*

[28] Clarke, C. L. A., Culpepper, J. S. and Moffat, A. [2016], 'Assessing efficiency–effectiveness tradeoffs in multi-stage retrieval systems without using relevance judgments', *Information Retrieval* **19**(4), 351–377.
URL: *https://doi.org/10.1007/s10791-016-9279-1*

[29] Clarke, C. L., Kolla, M., Cormack, G. V., Vechtomova, O., Ashkan, A., Büttcher, S. and MacKinnon, I. [2008], Novelty and diversity in information retrieval evaluation, *in* 'Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval', p. 659–666.

[30] Craswell, N., Zoeter, O., Taylor, M. J. and Ramsey, B. [2008], An experimental comparison of click position-bias models, *in* M. Najork, A. Z. Broder and S. Chakrabarti, eds, 'Proceedings of the International Conference on Web Search and Web Data Mining, WSDM 2008, Palo Alto, California, USA, February 11-12, 2008', ACM, pp. 87–94.
URL: *https://doi.org/10.1145/1341531.1341545*

[31] Croft, W. B. and Harper, D. J. [1979], 'Using probabilistic models of document retrieval without relevance information', *Journal of Documentation* **35**(4), 285–295.
URL: *https://doi.org/10.1108/eb026683*

[32] Culpepper, J. S., Faggioli, G., Ferro, N. and Kurland, O. [2022], 'Topic difficulty: Collection and query formulation effects', *ACM Transactions on Information Systems (TOIS)* **40**(1), 19:1–19:36.
URL: *https://doi.org/10.1145/3470563*

[33] Deckers, N., Fröbe, M., Kiesel, J., Pandolfo, G., Schröder, C., Stein, B. and Potthast, M. [2023], The infinite index: Information retrieval on generative text-to-image models, *in* J. Gwizdka and S. Y. Rieh, eds, 'Proceedings of the 2023 Conference on Human Information Interaction and Retrieval, CHIIR 2023, Austin, TX, USA, March 19-23, 2023', ACM, pp. 172–186.
URL: *https://doi.org/10.1145/3576840.3578327*

[34] Engelmann, B., Breuer, T., Friese, J. I., Schaer, P. and Fuhr, N. [2024], Context-driven interactive query simulations based on generative large language models, *in* N. Goharian, N. Tonellotto, Y. He, A. Lipani, G. McDonald, C. Macdonald and I. Ounis, eds, 'Advances in Information Retrieval - 46th European Conference on Information Retrieval, ECIR 2024, Glasgow, UK, March 24-28, 2024, Proceedings, Part II', Vol. 14609 of *Lecture Notes in Computer Science*, Springer, pp. 173–188.
URL: *https://doi.org/10.1007/978-3-031-56060-6_12*

[35] Faggioli, G., Dietz, L., Clarke, C. L. A., Demartini, G., Hagen, M., Hauff, C., Kando, N., Kanoulas, E., Potthast, M., Stein, B. and Wachsmuth, H. [2023], Perspectives on large language models for relevance judgment, *in* M. Yoshioka, J. Kiseleva and M. Aliannejadi, eds, 'Proceedings of the 2023 ACM SIGIR International Conference on Theory of Information Retrieval, ICTIR 2023, Taipei, Taiwan, 23 July 2023', ACM, pp. 39–50.
URL: *https://doi.org/10.1145/3578337.3605136*

[36] Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Guo, Q., Wang, M. and Wang, H. [2023], 'Retrieval-augmented generation for large language models: A survey', *CoRR* **abs/2312.10997**.
URL: *https://doi.org/10.48550/arXiv.2312.10997*

[37] Gienapp, L., Scells, H., Deckers, N., Bevendorff, J., Wang, S., Kiesel, J., Syed, S., Fröbe, M., Zuccon, G., Stein, B., Hagen, M. and Potthast, M. [2023], 'Evaluating generative ad hoc information retrieval', *CoRR* **abs/2311.04694**.
URL: *https://doi.org/10.48550/arXiv.2311.04694*

[38] Glover, E. [2007], The real world web search problem: bridging the gap between academic and commercial understanding of issues and methods, *in* R. Steinberger, F. Fogelman-Soulié, D. Perrotta and J. Piskorski, eds, 'Mining Massive Data Sets for Security', IOS Press, Oxford, pp. 115–129.

[39] Haas, K., Mika, P., Tarjan, P. and Blanco, R. [2011], Enhanced results for web search, *in* W. Ma, J. Nie, R. Baeza-Yates, T. Chua and W. B. Croft, eds, 'Proceeding of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2011, Beijing,

China, July 25-29, 2011', ACM, pp. 725–734.
URL: *https://doi.org/10.1145/2009916.2010014*

[40] Hämäläinen, P., Tavast, M. and Kunnari, A. [2023], Evaluating large language models in generating synthetic HCI research data: a case study, *in* A. Schmidt, K. Väänänen, T. Goyal, P. O. Kristensson, A. Peters, S. Mueller, J. R. Williamson and M. L. Wilson, eds, 'Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, CHI 2023, Hamburg, Germany, April 23-28, 2023', ACM, pp. 433:1–433:19.
URL: *https://doi.org/10.1145/3544548.3580688*

[41] Harman, D. [1992], 'User-friendly systems instead of user-friendly front-ends', *Journal of the American Society for Information Science (JASIST)* **43**(2), 164–174.
URL: *https://doi.org/10.1002/(SICI)1097-4571(199203)43:2<164::AID-ASI9>3.0.CO;2-W*

[42] Harvey, M. and Pointon, M. [2017], Searching on the go: The effects of fragmented attention on mobile web search tasks, *in* N. Kando, T. Sakai, H. Joho, H. Li, A. P. de Vries and R. W. White, eds, 'Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, Shinjuku, Tokyo, Japan, August 7-11, 2017', ACM, pp. 155–164.
URL: *https://doi.org/10.1145/3077136.3080770*

[43] Hearst, M. A. and Pedersen, J. O. [1996], Reexamining the cluster hypothesis: Scatter/gather on retrieval results, *in* H. Frei, D. Harman, P. Schäuble and R. Wilkinson, eds, 'Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR'96, August 18-22, 1996, Zurich, Switzerland (Special Issue of the SIGIR Forum)', ACM, pp. 76–84.
URL: *https://doi.org/10.1145/243199.243216*

[44] Jardine, N. and van Rijsbergen, C. J. [1971], 'The use of hierarchic clustering in information retrieval', *Inf. Storage Retr.* **7**(5), 217–240.
URL: *https://doi.org/10.1016/0020-0271(71)90051-9*

[45] Järvelin, K. and Kekäläinen, J. [2002], 'Cumulated gain-based evaluation of IR techniques', *ACM Transactions on Information Systems (TOIS)* **20**(4), 422–446.
URL: *http://doi.acm.org/10.1145/582415.582418*

[46] Kurland, O. [2014], The cluster hypothesis in information retrieval, *in* M. de Rijke, T. Kenter, A. P. de Vries, C. Zhai, F. de Jong, K. Radinsky and K. Hofmann, eds, 'Advances in Information Retrieval - 36th European Conference on IR Research, ECIR 2014, Amsterdam, The Netherlands, April 13-16, 2014. Proceedings', Vol. 8416 of *Lecture Notes in Computer Science*, Springer, pp. 823–826.
URL: *https://doi.org/10.1007/978-3-319-06028-6_105*

[47] Mackenzie, J. M., Benham, R., Petri, M., Trippas, J. R., Culpepper, J. S. and Moffat, A. [2020], Cc-news-en: A large english news corpus, *in* M. d'Aquin, S. Dietze, C. Hauff, E. Curry and P. Cudré-Mauroux, eds, 'The 29th ACM International Conference on Information and Knowledge Management, Virtual Event, Ireland, October 19-23, 2020', ACM, pp. 3077–3084.
URL: *https://doi.org/10.1145/3340531.3412762*

[48] Mackie, I., Chatterjee, S. and Dalton, J. [2023], Generative relevance feedback with large language models, *in* H. Chen, W. E. Duh, H. Huang, M. P. Kato, J. Mothe and B. Poblete, eds, 'Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2023, Taipei, Taiwan, July 23-27, 2023', ACM, pp. 2026–2031.
URL: *https://doi.org/10.1145/3539618.3591992*

[49] Maxwell, D. and Azzopardi, L. [2014], Stuck in traffic: how temporal delays affect search behaviour, *in* D. Elsweiler, B. Ludwig, L. Azzopardi and M. L. Wilson, eds, 'Fifth Information Interaction in Context Symposium, IIiX '14, Regensburg, Germany, August 26-29, 2014', ACM, pp. 155–164.
URL: *https://doi.org/10.1145/2637002.2637021*

[50] Moffat, A., Mackenzie, J., Thomas, P. and Azzopardi, L. [2022], A flexible framework for offline effectiveness metrics, *in* E. Amigó, P. Castells, J. Gonzalo, B. Carterette, J. S. Culpepper and G. Kazai, eds, 'The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022', ACM, pp. 578–587.
URL: *https://doi.org/10.1145/3477495.3531924*

[51] Moffat, A. and Zobel, J. [2008], 'Rank-biased precision for measurement of retrieval effectiveness', *ACM Transactions on Information Systems (TOIS)* **27**(1), 2:1–2:27.
URL: *https://doi.org/10.1145/1416950.1416952*

[52] Monchaux, S., Amadieu, F., Chevalier, A. and Mariné, C. [2015], 'Query strategies during information searching: Effects of prior domain knowledge and complexity of the information problems to be solved', *Information Processing & Management (IP&M)* **51**(5), 557–569.
URL: *https://doi.org/10.1016/j.ipm.2015.05.004*

[53] Navalpakkam, V., Jentzsch, L., Sayres, R., Ravi, S., Ahmed, A. and Smola, A. J. [2013], Measurement and modeling of eye-mouse behavior in the presence of nonlinear page layouts, *in* D. Schwabe, V. A. F. Almeida, H. Glaser, R. Baeza-Yates and S. B. Moon, eds, '22nd International World Wide Web Conference, WWW '13, Rio de Janeiro, Brazil, May 13-17, 2013', International World Wide Web Conferences Steering Committee / ACM, pp. 953–964.
URL: *https://doi.org/10.1145/2488388.2488471*

[54] Oard, D. W., Sakai, T. and Kando, N. [2019], Celebrating 20 years of NTCIR: the book, *in* N. Ferro, I. Soboroff and M. Zhang, eds, 'Proceedings of the 9th International Workshop on Evaluating Information Access co-located with the 14th NTCIR Conference on the Evaluation of Information Access Technologies (NTCIR 2019), Tokyo, Japan, June 10, 2019'.
URL: *https://research.nii.ac.jp/ntcir/workshop/OnlineProceedings14/pdf/evia/01-EVIA2019-BOOK-OardD.pdf*

[55] Oliveira, B. and Lopes, C. T. [2023], The evolution of web search user interfaces - an archaeological analysis of google search engine result pages, *in* J. Gwizdka and S. Y. Rieh, eds, 'Proceedings of the 2023 Conference on Human Information Interaction and Retrieval, CHIIR 2023, Austin, TX, USA, March 19-23, 2023', ACM, pp. 55–68.
URL: *https://doi.org/10.1145/3576840.3578320*

[56] Penha, G., Câmara, A. and Hauff, C. [2022], Evaluating the robustness of retrieval pipelines with query variation generators, *in* M. Hagen, S. Verberne, C. Macdonald, C. Seifert, K. Balog, K. Nørvåg and V. Setty, eds, 'Advances in Information Retrieval - 44th European Conference on IR Research, ECIR 2022, Stavanger, Norway, April 10-14, 2022, Proceedings, Part I', Vol. 13185 of *Lecture Notes in Computer Science*, Springer, pp. 397–412.
URL: *https://doi.org/10.1007/978-3-030-99736-6_27*

[57] Penha, G., Palumbo, E., Aziz, M., Wang, A. and Bouchard, H. [2023], Improving content retrievability in search with controllable query generation, *in* Y. Ding, J. Tang, J. F. Sequeda, L. Aroyo, C. Castillo and G. Houben, eds, 'Proceedings of the ACM Web Conference 2023, WWW 2023, Austin, TX, USA, 30 April 2023 - 4 May 2023', ACM, pp. 3182–3192.
URL: *https://doi.org/10.1145/3543507.3583261*

[58] Pera, M. S., Murgia, E., Landoni, M., Huibers, T. and Aliannejadi, M. [2023], Where a little change makes a big difference: A preliminary exploration of children's queries, *in* J. Kamps, L. Goeuriot, F. Crestani, M. Maistro, H. Joho, B. Davis, C. Gurrin, U. Kruschwitz and A. Caputo, eds, 'Advances in Information Retrieval', Springer Nature Switzerland, Cham, pp. 522–533.

[59] Peters, C. [2019], 'Information retrieval evaluation in a changing world lessons learned from 20 years of CLEF'.

[60] Pradeep, R., Hui, K., Gupta, J., Lelkes, Á. D., Zhuang, H., Lin, J., Metzler, D. and Tran, V. Q. [2023], How does generative retrieval scale to millions of passages?, *in* H. Bouamor, J. Pino and K. Bali, eds, 'Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023', Association for Computational Linguistics, pp. 1305–1321.
URL: *https://aclanthology.org/2023.emnlp-main.83*

[61] Rajput, S., Pavlu, V., Golbus, P. B. and Aslam, J. A. [2011], A nugget-based test collection construction paradigm, *in* C. Macdonald, I. Ounis and I. Ruthven, eds, 'Proceedings of the 20th ACM Conference on Information and Knowledge Management, CIKM 2011, Glasgow, United Kingdom, October 24-28, 2011', ACM, pp. 1945–1948.
URL: *https://doi.org/10.1145/2063576.2063861*

[62] Rieh, S. Y. and Xie, H. I. [2006], 'Analysis of multiple query reformulations on the eeb: The interactive information retrieval context', *Information Processing & Management (IP&M)* **42**(3), 751–768.

[63] Robertson, S. [1977], 'The probability ranking principle in IR', *Journal of Documentation* **33**, 294–304.

[64] Sakai, T., Kato, M. P. and Song, Y.-I. [2011], Click the search button and be happy: Evaluating direct and immediate information access, *in* 'Proceedings of the 20th ACM International Conference on Information and Knowledge Management', p. 621–630.

[65] Sanderson, M. [2010], 'Test collection based evaluation of information retrieval systems', *Foundations and Trends® in Information Retrieval* **4**(4), 247–375.
URL: *https://doi.org/10.1561/1500000009*

[66] Sanderson, M., Scholer, F. and Turpin, A. [2010], Relatively relevant: Assessor shift in document judgements, *in* 'Australasian Document Computing Symposium'.
URL: *https://api.semanticscholar.org/CorpusID:14426189*

[67] Scholer, F., Turpin, A. and Sanderson, M. [2011], Quantifying test collection quality based on the consistency of relevance judgements, *in* W. Ma, J. Nie, R. Baeza-Yates, T. Chua and W. B. Croft, eds, 'Proceeding of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2011, Beijing, China, July 25-29, 2011', ACM, pp. 1063–1072.
URL: *https://doi.org/10.1145/2009916.2010057*

[68] Smucker, M. D. and Clarke, C. L. A. [2012], Time-based calibration of effectiveness measures, *in* W. R. Hersh, J. Callan, Y. Maarek and M. Sanderson, eds, 'The 35th International ACM SIGIR conference on research and development in Information Retrieval, SIGIR '12, Portland, OR, USA, August 12-16, 2012', ACM, pp. 95–104.
URL: *https://doi.org/10.1145/2348283.2348300*

[69] Spärck Jones, K. and Van Rijsbergen, C. [1975], Report on the need for and provision of an 'ideal' information retrieval test collection, Technical Report 5266, British Library Research and Development Report.
URL: *https://cir.nii.ac.jp/crid/1570572699089480448*

[70] Sun, W., Yan, L., Chen, Z., Wang, S., Zhu, H., Ren, P., Chen, Z., Yin, D., de Rijke, M. and Ren, Z. [2023], Learning to tokenize for generative retrieval, *in* A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt and S. Levine, eds, 'Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023'.
URL: *http://papers.nips.cc/paper_files/paper/2023/hash/91228b942a4528cdae031c1b68b127e8-Abstract-Conference.html*

[71] Teevan, J., Collins-Thompson, K., White, R. W. and Dumais, S. [2014], 'Slow search', *Communications of the ACM* **57**(8), 36–38.
URL: *https://doi.org/10.1145/2633041*

[72] Thomas, P., Spielman, S., Craswell, N. and Mitra, B. [2023], 'Large language models can accurately predict searcher preferences', *CoRR* **abs/2309.10621**.
URL: *https://doi.org/10.48550/arXiv.2309.10621*

[73] Tombros, A. and Sanderson, M. [1998], Advantages of query biased summaries in information retrieval, *in* W. B. Croft, A. Moffat, C. J. van Rijsbergen, R. Wilkinson and J. Zobel, eds, 'Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, August 24-28 1998, Melbourne, Australia', ACM, pp. 2–10.
URL: *https://doi.org/10.1145/290941.290947*

[74] Tonmoy, S. M. T. I., Zaman, S. M. M., Jain, V., Rani, A., Rawte, V., Chadha, A. and Das, A. [2024], 'A comprehensive survey of hallucination mitigation techniques in large language models', *CoRR* **abs/2401.01313**.
URL: *https://doi.org/10.48550/arXiv.2401.01313*

[75] Torres, S. D., Hiemstra, D. and Serdyukov, P. [2010], Query log analysis in the context of information retrieval for children, *in* F. Crestani, S. Marchand-Maillet, H. Chen, E. N. Efthimiadis and J. Savoy, eds, 'Proceeding of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2010, Geneva, Switzerland, July 19-23, 2010', ACM, pp. 847–848.
URL: *https://doi.org/10.1145/1835449.1835646*

[76] Turpin, A. and Scholer, F. [2006], User performance versus precision measures for simple search tasks, *in* E. N. Efthimiadis, S. T. Dumais, D. Hawking and K. Järvelin, eds, 'Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Seattle, Washington, USA, August 6-11, 2006', ACM, pp. 11–18.
URL: *https://doi.org/10.1145/1148170.1148176*

[77] van Rijsbergen, C. [1979], *Information retrieval*, Butterworth-Heinemann.

[78] Vdorhees, E. M. [2017], 'The cluster hypothesis revisited', *SIGIR Forum* **51**(2), 35–43.
URL: *https://doi.org/10.1145/3130348.3130353*

[79] Voorhees, E., Harman, D., of Standards, N. I. and (US), T. [2005], *TREC: Experiment and evaluation in information retrieval*, Vol. 63, MIT press Cambridge^ eMA MA.

[80] Voorhees, E. M. [2019], The evolution of cranfield, *in* N. Ferro and C. Peters, eds, 'Information Retrieval Evaluation in a Changing World - Lessons Learned from 20 Years of CLEF', Vol. 41 of *The Information Retrieval Series*, Springer, pp. 45–69.
URL: *https://doi.org/10.1007/978-3-030-22948-1_2*

[81] Voorhees, E. M., Craswell, N. and Lin, J. [2022], Too many relevants: Whither cranfield test collections?, *in* E. Amigó, P. Castells, J. Gonzalo, B. Carterette, J. S. Culpepper and G. Kazai, eds, 'The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022', ACM, pp. 2970–2980.
URL: *https://doi.org/10.1145/3477495.3531728*

[82] Wang, L., Lin, J. and Metzler, D. [2011], A cascade ranking model for efficient ranked retrieval, *in* W. Ma, J. Nie, R. Baeza-Yates, T. Chua and W. B. Croft, eds, 'Proceeding of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2011, Beijing, China, July 25-29, 2011', ACM, pp. 105–114.
URL: *https://doi.org/10.1145/2009916.2009934*

[83] White, R. W., Dumais, S. T. and Teevan, J. [2009], Characterizing the influence of domain expertise on web search behavior, *in* R. Baeza-Yates, P. Boldi, B. A. Ribeiro-Neto and B. B. Cambazoglu, eds, 'Proceedings of the Second International Conference on Web Search and Web Data Mining, WSDM 2009, Barcelona, Spain, February 9-11, 2009', ACM, pp. 132–141.
URL: *https://doi.org/10.1145/1498759.1498819*

[84] Wu, Z., Sanderson, M., Cambazoglu, B. B., Croft, W. B. and Scholer, F. [2020], Providing direct answers in search results: A study of user behavior, *in* M. d'Aquin, S. Dietze, C. Hauff, E. Curry and P. Cudré-Mauroux, eds, 'The 29th ACM International Conference on Information and Knowledge Management, Virtual Event, Ireland, October 19-23, 2020', ACM, pp. 1635–1644.
URL: *https://doi.org/10.1145/3340531.3412017*

[85] Yang, T., Song, M., Zhang, Z., Huang, H., Deng, W., Sun, F. and Zhang, Q. [2023], Auto search indexer for end-to-end document retrieval, *in* H. Bouamor, J. Pino and K. Bali, eds, 'Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023', Association for Computational Linguistics, pp. 6955–6970.
URL: *https://aclanthology.org/2023.findings-emnlp.464*

[86] Zhai, C., Cohen, W. W. and Lafferty, J. D. [2015], 'Beyond independent relevance: Methods and evaluation metrics for subtopic retrieval', *SIGIR Forum* **49**(1), 2–9.
URL: *https://doi.org/10.1145/2795403.2795405*

[87] Zhang, Y., Hu, C., Liu, Y., Fang, H. and Lin, J. [2021], Learning to rank in the age of Muppets: Effectiveness–efficiency tradeoffs in multi-stage ranking, *in* N. S. Moosavi, I. Gurevych, A. Fan, T. Wolf, Y. Hou, A. Marasović and S. Ravi, eds, 'Proceedings of the Second Workshop on Simple and Efficient Natural Language Processing', Association for Computational Linguistics, Virtual, pp. 64–73.
URL: *https://aclanthology.org/2021.sustainlp-1.8*

[88] Zobel, J. [1998], How reliable are the results of large-scale information retrieval experiments?, *in* W. B. Croft, A. Moffat, C. J. van Rijsbergen, R. Wilkinson and J. Zobel, eds, 'Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, August 24-28 1998, Melbourne, Australia', ACM, pp. 307–314.
URL: *https://doi.org/10.1145/290941.291014*