Generative Information Retrieval

Marc Najork
Google DeepMind
Mountain View, California, USA
najork@google.com

Abstract

Historically, information retrieval systems have all followed the same paradigm: information seekers frame their needs in the form of a short query, the system selects a small set of relevant results from a corpus of available documents, rank-orders the results by decreasing relevance, possibly excerpts a responsive passage for each result, and returns a list of references and excerpts to the user. Retrieval systems typically did not attempt fusing information from multiple documents into an answer and displaying that answer directly. This was largely due to available technology: at the core of each retrieval system is an index that maps lexical tokens or semantic embeddings to document identifiers. Indices are designed for retrieving responsive documents; they do not support integrating these documents into a holistic answer.

More recently, the coming-of-age of deep neural networks has dramatically improved the capabilities of large language models (LLMs). Trained on a large corpus of documents, these models not only memorize the vocabulary, morphology and syntax of human languages, but have shown to be able to memorize facts and relations [3]. Generative language models, when provided with a prompt, will extend the prompt with likely completions – an ability that can be used to extract answers to questions from the model. Four years ago, Metzler et al. argued that this ability of LLMs will allow us to rethink the search paradigm: to answer information needs directly rather that directing users to responsive primary sources [1]. Their vision was not without controversy; the following year Shaw and Bender argued that such a system is neither feasible nor desirable [4]. Nonetheless, the past four years have seen the emergence of such systems, with offerings from established search engines and many new entrants to the industry.

This keynote (an updated version of [2]) will summarize the history of these generative information retrieval systems, and focus on the many open challenges: ensuring that answers are grounded, attributing answer passages to a primary source, providing nuanced answers to non-factoid-seeking questions, avoiding bias, and going beyond simple regurgitation of memorized facts. It will also touch on the changing nature of the content ecosystem. LLMs are starting to be used to generate web content. Should search engines treat such derived content equal to human-authored content? Is it possible to distinguish generated from original content? And how should we view hybrid authorship where humans contribute ideas and LLMs shape these ideas into prose?

KDD '25, Toronto, ON, Canada

ACM ISBN 979-8-4007-1454-2/2025/08
https://doi.org/10.1145/3711896.3736806

CCS Concepts

• **Information systems** → **Question answering**; • **Computing methodologies** → **Natural language generation**.

Keywords

Generative Information Retrieval, Large Language Models, Question Answering, Tool-Augmented Generation

ACM Reference Format:
Marc Najork. 2025. Generative Information Retrieval. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '25), August 3–7, 2025, Toronto, ON, Canada.* ACM, New York, NY, USA, 1 page. https://doi.org/10.1145/3711896.3736806

1 Biography

Marc Najork is a Distinguished Research Scientist in Google DeepMind, working on new techniques to make it easier for people to obtain relevant and useful information when and where they need it. Marc is interested in using generative language models to answer questions directly, rather than referring users to relevant sources. Direct answers represent a major paradigm shift in Information Retrieval, affecting the user experience, the fundamental architecture of the retrieval system, and the economic foundation of commercial web search and the entire web content ecosystem. Prior to joining Google, Marc was a Principal Researcher at Microsoft Research, and a Research Scientist at Digital Equipment Corporation. He is an ACM Fellow, IEEE Fellow, AAAS Fellow and a SIGIR Academy member.

References

[1] Donald Metzler, Yi Tay, Dara Bahri, and Marc Najork. 2021. Rethinking search: making domain experts out of dilettantes. *ACM SIGIR Forum* 55, 1 (2021), 1–27.
[2] Marc Najork. 2023. Generative Information Retrieval. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval.* 1.
[3] Adam Roberts, Colin Raffel, and Noam Shazeer. 2020. How Much Knowledge Can You Pack Into the Parameters of a Language Model?. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing.* 5418–5426.
[4] Chirag Shah and Emily M. Bender. 2022. Situating Search. In *Proceedings of the 2022 Conference on Human Information Interaction and Retrieval.* 221–232.