Recent Advances in Generative Information Retrieval

Yubao Tang
CAS Key Lab of Network Data Science and Technology, ICT, CAS
University of Chinese Academy of Sciences
Beijing, China
tangyubao21b@ict.ac.cn

Ruqing Zhang
CAS Key Lab of Network Data Science and Technology, ICT, CAS
University of Chinese Academy of Sciences
Beijing, China
zhangruqing@ict.ac.cn

Zhaochun Ren
Leiden University
Leiden, The Netherlands
z.ren@liacs.leidenuniv.nl

Jiafeng Guo
CAS Key Lab of Network Data Science and Technology, ICT, CAS
University of Chinese Academy of Sciences
Beijing, China
guojiafeng@ict.ac.cn

Maarten de Rijke
University of Amsterdam
Amsterdam, The Netherlands
m.derijke@uva.nl

ABSTRACT

Generative retrieval (GR) has witnessed significant growth recently in the area of information retrieval. Compared to the traditional "index-retrieve-then-rank" pipeline, the GR paradigm aims to consolidate all information within a corpus into a single model. Typically, a sequence-to-sequence model is trained to directly map a query to its relevant document identifiers (i.e., docids). This tutorial offers an introduction to the core concepts of the GR paradigm and a comprehensive overview of recent advances in its foundations and applications. We start by providing preliminary information covering foundational aspects and problem formulations of GR. Then, our focus shifts towards recent progress in docid design, training approaches, inference strategies, and applications of GR. We end by outlining challenges and issuing a call for future GR research.

Throughout the tutorial we highlight the availability of relevant resources so as to enable a broad audience to contribute to this topic. This tutorial is intended to be beneficial to both researchers and industry practitioners interested in developing novel GR solutions or applying them in real-world scenarios.

CCS CONCEPTS

• **Information systems** → **Retrieval models and ranking**.

KEYWORDS

Generative retrieval

ACM Reference Format:
Yubao Tang, Ruqing Zhang, Zhaochun Ren, Jiafeng Guo, and Maarten de Rijke. 2024. Recent Advances in Generative Information Retrieval. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '24), July 14–18, 2024, Washington, DC, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3626772.3661379

SIGIR '24, July 14–18, 2024, Washington, DC, USA

ACM ISBN 979-8-4007-0431-4/24/07.
https://doi.org/10.1145/3626772.3661379

1 MOTIVATION

First, we describe the topic of the tutorial and highlight its importance and relevance to the Web Conference. Then, we summarize the presenters' qualifications to deliver a high-quality introduction.

Information retrieval (IR) is a core task in a wide range of real-world applications, such as web search and question answering. It aims to retrieve information from a large repository that is relevant to an information need. Most existing IR methods follow a common pipeline paradigm of "index-retrieve-then-rank," which includes (i) building an index for each document in the corpus [16]; (ii) retrieving an initial set of candidate documents for a query [12]; and (iii) determining the relevance degree of each candidate [16]. Despite its wide usage, this paradigm has limitations: (i) during training, heterogeneous modules with different optimization objectives may lead to sub-optimal performance, and capturing fine-grained relationships between queries and documents is challenging; and (ii) during inference, a large document index is needed to search over the corpus, which may come with substantial memory and computational requirements.

Recently, a fundamentally different paradigm, known as *generative retrieval* (GR) [19], has garnered attention to replace the long-standing "index-retrieve-then-rank" paradigm. The key idea of the GR paradigm is to parameterize the indexing, retrieval, and ranking components of traditional IR systems into a single consolidated model. Based on [20], GR includes closed-book and open-book GR. Closed-book GR refers to the scenario where the language model is the only source of knowledge leveraged during generation. Open-book GR allows the language model to draw on external memory prior to, during, and after generation. Our focus is on closed-book GR. A sequence-to-sequence (Seq2Seq) model is trained to directly map queries to their relevant document identifiers (docids). Such a single-step generative model dramatically simplifies the search process, can be optimized in an end-to-end manner, and can better leverage the capabilities of large language models.

2 OBJECTIVES

A flourishing body of research into this new retrieval paradigm reflects the growing interest in the area. We have organized and presented a tutorial dedicated to GR at SIGIR-AP'23 on November 26, 2023, in Beijing, China, and at ECIR'24 on March 28, 2024, in Glasgow, Scotland. At SIGIR'24, we will offer a new edition that has been revised based on the feedback received and incorporates new relevant work. The scope is as follows.

1. Introduction. We start by reminding our audience of the required background and examining the motivation behind GR.

2. Preliminaries. With GR, the document retrieval task is formulated as a Seq2Seq problem, i.e., directly generating identifiers of relevant documents with respect to the given query. To achieve this functionality, GR encompasses two fundamental training tasks [26], based on an encoder-decoder architecture: (i) *indexing* – this task aims to establish associations between each document and its corresponding docid; the GR model takes each original document as input and generates its docid as output in a straightforward Seq2Seq fashion; and (ii) *retrieval* – this task focuses on mapping each query to its relevant docids; given a query, the GR model learns to generate its relevant docid string.

It is crucial to store document information as comprehensively as possible during the indexing process, thus ensuring that the subsequent retrieval process is not hindered by information loss [8]. Using these two operations, a GR model can be trained to index a corpus of documents and optionally fine-tune with an available set of labeled query-document pairs. Thereafter, during inference, the optimized generative retriever can be used to efficiently retrieve relevant documents within a single neural model. Building on these preliminaries, we will cover docid design, training approaches, inference strategies, and applications of GR in downstream scenarios.

3. Docid designs. With GR, employing identifiers, rather than generating original documents directly, could reduce irrelevant information in documents and make it easier for the model to memorize the corpus. Therefore, one of the key challenges in GR is how to assign a high-quality identifier to represent a document. An effective docid should be unique to enable effective distinction among different documents and concise for ease of generation. Therefore, we proceed to discuss the work related to docid designs.

Most existing GR approaches utilize pre-defined static docids, i.e., these docids are fixed and are not learnable during training the indexing and retrieval tasks. To be specific, these works usually leverage a single docid to represent the document, and several types of identifiers have been explored, including number-based and word-based docids. The number-based docids encompass atomic unique integers [18, 26, 34], structured integer strings [26], semantically structured strings [21, 26, 28], product quantization code [4, 33], while the word-based docids primarily involve document titles [6, 7, 10, 13, 27], n-grams [3, 5, 14], important word sets [32], pseudo-queries [24], and URLs [33]. Given that a document has the potential to answer multiple queries from different views, some research advocates the use of multiple types of identifiers to comprehensively represent a document [14, 15].

Although pre-defined static docids have demonstrated some effectiveness, they are not tailored to the retrieval objectives, limiting their capacity to adapt to semantic relationships within documents during the training process. Consequently, recent research [23, 29] has introduced document tokenization learning methods to acquire learnable docids for GR.

4. Training approaches. Here, we consider two main scenarios for training the GR model. The first, a more straightforward one, assumes a stationary learning scenario where the document collection is fixed and no longer updates. The second, a more practical scenario, is a dynamic corpora setting where information changes and new documents emerge incrementally over time.

The majority of GR research [3, 10, 26, 28, 35] primarily focuses on implementing GR in a stationary learning scenario. These works can be further categorized into supervised learning methods and pre-training methods, depending on the availability of labeled query-docid pairs. (i) For supervised learning methods, Tay et al. [26] introduced fundamental training strategies, jointly optimizing indexing and retrieval tasks using the standard Seq2Seq objective, i.e., maximum likelihood estimation with teacher forcing. Building upon this foundation, a series of improvements [22, 24, 28, 35] have been proposed, significantly enhancing performance. (ii) In IR research, limited labeled data is often a challenge. Some researchers explore the design of self-supervised pre-training objectives to generate a large number of pseudo pairs of queries and docids [7].

In many scenarios, document collections are dynamic, with new documents continuously being added to the corpus, old documents being removed, or updated. A significant challenge in GR is how to enable the model to remember information from new documents while minimizing the forgetting of information from previously learned documents. Mehta et al. [18] demonstrate that continually memorizing new documents leads to considerable forgetting of old documents. Several follow-up approaches [4, 31] have been proposed to address this issue.

5. Inference strategies. During inference, in cases where a single docid represents a document, the trained GR model autoregressively generates a ranked list of candidate docids in descending order of output likelihood conditioned on each query. To ensure the validity of the generated docids, three classical approaches are commonly used: constrained beam search [6, 7, 10, 23, 24], constrained greedy search [32] and FM-index [3, 5, 29]. In cases where multiple docids represent a single document, some research [14, 15] combines the aforementioned approaches and designs heuristic scoring functions to determine the ranking order of relevant docids.

6. Applications. We then will demonstrate how GR models are adapted to downstream applications. First, we will discuss methods for specific offline tasks. Then, we will explore methods tailored for industrial applications. These examples underscore the tremendous promise and value of the GR paradigm in IR.

7. Conclusions and future directions. We conclude our tutorial by discussing several important questions and future directions, including (i) What are the differences and connections between GR models and discriminative models in terms of fundamental indexing and retrieval mechanisms? (ii) How can we enhance the scalability of GR models to support complex, diverse, and dynamically changing retrieval tasks without compromising performance?

(iii) How can we achieve controllability over the black-box integrated generative retrieval process to enhance interpretability and trustworthiness? (iv) How can we integrate GR models for document retrieval with large language models for answer generation?

3 RELEVANCE TO THE COMMUNITY

3.1 Importance and timeliness

In 2021, Metzler et al. [19] envisioned a model-based IR approach that replaces the long-standing "index-retrieve-then-rank" paradigm with a single consolidated model. A plethora of publications have emerged in reputable conferences, e.g., SIGIR [5, 6], CIKM [4, 7, 29], KDD [24], NeurIPS [3, 23, 26, 28], ICLR [10], and ACL [8, 13, 15], in Gen-IR@SIGIR2023 [17, 21, 22, 35], in journals [34], and on arXiv [14, 32, 33]. The first workshop on GR at SIGIR'23 (Gen-IR@SIGIR2023) [1] welcomed lots of submissions and attendees, underscoring the research community's current keen interest.

3.2 Relevance to SIGIR

One of the core research area at SIGIR is search and ranking. GR is a novel IR paradigm that aligns well with the theme of SIGIR in particular. Our tutorial will describe recent advances in GR and shed light on future research directions. It would benefit the IR community and help to encourage further research into GR.

3.3 Previous editions

We have presented this tutorial on GR at SIGIR-AP'23 on November 26, 2023, in Beijing, China, and at ECIR'24 on March 28, 2024, in Glasgow, Scotland. This tutorial has also been accepted by TheWebConf'24 and will be presented in May 2024. At SIGIR'24 we offer a new edition that has been revised based on the feedback received and incorporates coverage of new relevant work in this rapidly evolving area (from SIGIR'24, KDD'24, ACL'24, and TheWebConf'24). In addition, the SIGIR'24 edition will have a focus on resources and sharing resources to make the research topic more widely accessible to a broad group of researchers. Finally, after presentations in Asia and Europe, we are keen to present this material to a North-American audience – especially with the Large Language Model Day that is planned as part of SIGIR'24.[1]

3.4 Commitment

We, the authors, have all dedicated our research to IR, with a substantial emphasis on GR recently. We have published research papers about GR at SIGIR [5, 6, 30], CIKM [4, 7], WSDM [11], NeurIPS [23] KDD [24], and TOIS [25]. And we have actively participated in organizing relevant workshops and tutorials in IR, e.g., the first workshop on neural IR at SIGIR 2016 and the first workshop on generative IR at SIGIR'23, thereby helping to bring the community together around this topic [1, 2, 9]. We believe that our collective and diverse experience makes us well-qualified to deliver a high-quality GR tutorial.

4 FORMAT AND DETAILED SCHEDULE

This is a 3-hour and lecture-style tutorial.

1. **Introduction** (15 minutes)
   - An overview of the tutorial
   - Why generative retrieval?
2. **Preliminaries** (15 minutes)
   - Retrieval task formulation: generative models vs. discriminative models
   - Basic concepts in generative retrieval
   - Resources
3. **Generative retrieval: Docid design** (30 minutes)
   - Pre-defined static docids
     - Single docids: number-based and word-based docids
     - Multiple docids
   - Learnable docids: jointly with retrieval tasks
   - Resources
4. **Generative retrieval: Training approaches** (40 minutes)
   - Static corpora: supervised learning with labeled data, and pre-training with unlabeled data
   - Dynamic corpora: continual learning
   - Resources
5. **Generative retrieval: Inference strategies** (25 minutes)
   - For a single docid: constrained beam search, constrained greedy search and FM-index
   - For multiple docids: aggregation scoring functions
   - Resources
6. **Generative retrieval: Applications** (35 minutes)
   - Offline application: e.g., entity retrieval, fact checking, recommender systems, multi-hop retrieval and code generation
   - Industry applications
7. **Conclusions and future directions** (20 minutes)
   - Challenges
   - Resources

5 TUTORIAL MATERIALS

We plan to share the following materials on this website:[2] (i) Slides: All slides are made publicly available. (ii) Annotated bibliography: An annotated compilation of references that lists all works discussed in the tutorial and provides a good basis for further study. (iii) Code: An annotated list of pointers to open source code bases and datasets. (iv) Lists of publicly available resources to support research into generative information retrieval. We agree to allow the publication of slides and videos in the ACM anthology.

ACKNOWLEDGMENTS

This work was funded by the National Key Research and Development Program of China under Grants No. 2023YFA1011602, the Strategic Priority Research Program of the CAS under Grants No. XDB0680102, the project under Grants No. JCKY2022130C039, and the Lenovo-CAS Joint Lab Youth Scientist Project. This work was also (partially) funded by the Hybrid Intelligence Center, a 10-year program funded by the Dutch Ministry of Education, Culture and Science through the Netherlands Organisation for Scientific Research, https://hybrid-intelligence-centre.nl, project LESSEN with project number NWA.1389.20.183 of the research program NWA ORC 2020/21, which is (partly) financed by the Dutch Research Council (NWO), project ROBUST with project number KICH3.LTP.-20.006, which is (partly) financed by the Dutch Research Council (NWO), DPG Media, RTL, and the Dutch Ministry of Economic

[1] https://sigir-2024.github.io/call_for_participation_llm_day.html

[2] https://ecir2024-generativeir.github.io/

Affairs and Climate Policy (EZK) under the program LTP KIC 2020-2023, and the FINDHR (Fairness and Intersectional Non-Discrimination in Human Recommendation) project that received funding from the European Union's Horizon Europe research and innovation program under grant agreement No 101070212. All content represents the opinion of the authors, which is not necessarily shared or endorsed by their respective employers and/or sponsors.

REFERENCES

[1] Gabriel Bénédict, Ruqing Zhang, and Donald Metzler. 2023. Gen-IR@ SIGIR 2023: The First Workshop on Generative Information Retrieval. In *SIGIR*. 3460–3463.
[2] Gabriel Bénédict, Ruqing Zhang, Donald Metzler, Andrew Yates, Romain Deffayet, Philipp Hager, and Sami Jullien. 2024. Report on the 1st Workshop on Generative Information Retrieval (Gen-IR 2023) at SIGIR 2023. *SIGIR Forum* 57, 2, Article 13 (jan 2024), 23 pages.
[3] Michele Bevilacqua, Giuseppe Ottaviano, Patrick Lewis, Wen-tau Yih, Sebastian Riedel, and Fabio Petroni. 2022. Autoregressive Search Engines: Generating Substrings as Document Identifiers. In *NeurIPS*. 31668–31683.
[4] Jiangui Chen, Ruqing Zhang, Jiafeng Guo, Maarten de Rijke, Wei Chen, Yixing Fan, and Xueqi Cheng. 2023. Continual Learning for Generative Retrieval over Dynamic Corpora. In *CIKM*. 306–315.
[5] Jiangui Chen, Ruqing Zhang, Jiafeng Guo, Maarten de Rijke, Yiqun Liu, Yixing Fan, and Xueqi Cheng. 2023. A Unified Generative Retriever for Knowledge-Intensive Language Tasks via Prompt Learning. In *SIGIR*. 1448–1457.
[6] Jiangui Chen, Ruqing Zhang, Jiafeng Guo, Yixing Fan, and Xueqi Cheng. 2022. GERE: Generative evidence retrieval for fact verification. In *SIGIR*. 2184–2189.
[7] Jiangui Chen, Ruqing Zhang, Jiafeng Guo, Yiqun Liu, Yixing Fan, and Xueqi Cheng. 2022. CorpusBrain: Pre-train a Generative Retrieval Model for Knowledge-Intensive Language Tasks. In *CIKM*. 191–200.
[8] Xiaoyang Chen, Yanjiang Liu, Ben He, Le Sun, and Yingfei Sun. 2023. Understanding Differential Search Index for Text Retrieval. In *ACL Findings*. 10701–10717.
[9] Nick Craswell, W. Bruce Croft, Jiafeng Guo, Bhaskar Mitra, and Maarten de Rijke. 2016. Neu-IR: The SIGIR 2016 Workshop on Neural Information Retrieval. In *SIGIR 2016: 39th international ACM SIGIR conference on Research and development in information retrieval*. ACM, 1245–1246.
[10] Nicola De Cao, Gautier Izacard, Sebastian Riedel, and Fabio Petroni. 2021. Autoregressive Entity Retrieval. In *Int. Conf. on Learning Representations*.
[11] Romain Deffayet, Thibaut Thonet, Jean-Michel Renders, and Maarten de Rijke. 2023. Generative Slate Recommendation with Reinforcement Learning. In *WSDM 2023: The Sixteenth International Conference on Web Search and Data Mining*. ACM, 580–588.
[12] Tom Kenter and Maarten de Rijke. 2015. Short Text Similarity with Word Embeddings. In *CIKM*. 1411–1420.
[13] Hyunji Lee, Jaeyoung Kim, Hoyeon Chang, Hanseok Oh, Sohee Yang, Vladimir Karpukhin, Yi Lu, and Minjoon Seo. 2023. Nonparametric Decoding for Generative Retrieval. In *Findings of the ACL 2023*. 12642–12661.
[14] Yongqi Li, Nan Yang, Liang Wang, Furu Wei, and Wenjie Li. 2023. Learning to Rank in Generative Retrieval. *arXiv preprint arXiv:2306.15222* (2023).
[15] Yongqi Li, Nan Yang, Liang Wang, Furu Wei, and Wenjie Li. 2023. Multiview Identifiers Enhanced Generative Retrieval. In *ACL*. 6636–6648.
[16] Shichen Liu, Fei Xiao, Wenwu Ou, and Luo Si. 2017. Cascade Ranking for Operational E-commerce Search. In *SIGKDD*. 1557–1565.
[17] Yu-An Liu, Ruqing Zhang, Jiafeng Guo, Wei Chen, and Xueqi Cheng. 2023. On the Robustness of Generative Retrieval Models: An Out-of-Distribution Perspective. In *Gen-IR@SIGIR*.
[18] Sanket Vaibhav Mehta, Jai Gupta, Yi Tay, Mostafa Dehghani, Vinh Q. Tran, Jinfeng Rao, Marc Najork, Emma Strubell, and Donald Metzler. 2023. DSI++: Updating Transformer Memory with New Documents. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. 8198–8213.
[19] Donald Metzler, Yi Tay, Dara Bahri, and Marc Najork. 2021. Rethinking Search: Making Domain Experts Out of Dilettantes. *SIGIR Forum* 55, 1 (2021), 1–27.
[20] Marc Najork. 2023. Generative Information Retrieval. In *SIGIR*. 1–1.
[21] Thong Nguyen and Andrew Yates. 2023. Generative Retrieval as Dense Retrieval. In *Gen-IR@SIGIR*.
[22] Ronak Pradeep, Kai Hui, Jai Gupta, Adam D. Lelkes, Honglei Zhuang, Jimmy Lin, Donald Metzler, and Vinh Q. Tran. 2023. How Does Generative Retrieval Scale to Millions of Passages?. In *Gen-IR@SIGIR*.
[23] Weiwei Sun, Lingyong Yan, Zheng Chen, Shuaiqiang Wang, Haichao Zhu, Pengjie Ren, Zhumin Chen, Dawei Yin, Maarten de Rijke, and Zhaochun Ren. 2023. Learning to Tokenize for Generative Retrieval. In *NeurIPS*.
[24] Yubao Tang, Ruqing Zhang, Jiafeng Guo, Jiangui Chen, Zuowei Zhu, Shuaiqiang Wang, Dawei Yin, and Xueqi Cheng. 2023. Semantic-Enhanced Differentiable Search Index Inspired by Learning Strategies. In *SIGKDD*. 4904–4913.
[25] Yubao Tang, Ruqing Zhang, Jiafeng Guo, Maarten de Rijke, Wei Chen, and Xueqi Cheng. 2024. Listwise Generative Retrieval Models via a Sequential Learning Process. *ACM Transactions on Information Systems* (2024). To appear.
[26] Yi Tay, Vinh Q. Tran, Mostafa Dehghani, Jianmo Ni, Dara Bahri, Harsh Mehta, Zhen Qin, Kai Hui, Zhe Zhao, Jai Gupta, Tal Schuster, William W. Cohen, and Donald Metzler. 2022. Transformer Memory as a Differentiable Search Index. In *NeurIPS*, Vol. 35. 21831–21843.
[27] James Thorne. 2022. Data-efficient Autoregressive Document Retrieval for Fact Verification. In *SENLP*.
[28] Yujing Wang, Yingyan Hou, Haonan Wang, Ziming Miao, Shibin Wu, Hao Sun, Qi Chen, Yuqing Xia, Chengmin Chi, Guoshuai Zhao, Zheng Liu, Xing Xie, Hao Sun, Weiwei Deng, Qi Zhang, and Mao Yang. 2022. A Neural Corpus Indexer for Document Retrieval. In *NuerIPS*, Vol. 35. 25600–25614.
[29] Zihan Wang, Yujia Zhou, Yiteng Tu, and Zhicheng Dou. 2023. NOVO: Learnable and Interpretable Document Identifiers for Model-Based IR. In *CIKM*. 2656–2665.
[30] Shiguang Wu, Wenda Wei, Mengqi Zhang, Zhumin Chen, Jun Ma, Zhaochun Ren, Maarten de Rijke, and Pengjie Ren. 2024. Generative Retrieval as Multi-Vector Dense Retrieval. In *SIGIR 2024: 47th international ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM.
[31] Soyoung Yoon, Chaeeun Kim, Hyunji Lee, Joel Jang, and Minjoon Seo. 2023. Continually Updating Generative Retrieval on Dynamic Corpora. *arXiv preprint arXiv:2305.18952* (2023).
[32] Peitian Zhang, Zheng Liu, Yujia Zhou, Zhicheng Dou, and Zhao Cao. 2023. Term-Sets Can Be Strong Document Identifiers For Auto-Regressive Search Engines. *arXiv preprint arXiv:2305.13859* (2023).
[33] Yujia Zhou, Jing Yao, Zhicheng Dou, Ledell Wu, Peitian Zhang, and Ji-Rong Wen. 2022. Ultron: An Ultimate Retriever on Corpus with a Model-based Indexer. *arXiv preprint arXiv:2208.09257* (2022).
[34] Yu-Jia Zhou, Jing Yao, Zhi-Cheng Dou, Ledell Wu, and Ji-Rong Wen. 2023. DynamicRetriever: A Pre-trained Model-based IR System Without an Explicit Index. *Machine Intelligence Research* 20, 2 (2023), 276–288.
[35] Shengyao Zhuang, Houxing Ren, Linjun Shou, Jian Pei, Ming Gong, Guido Zuccon, and Daxin Jiang. 2023. Bridging the Gap between Indexing and Retrieval for Differentiable Search Index with Query Generation. In *Gen-IR@SIGIR*.