Generative Information Retrieval: RAG and GAR

Jimmy X. Huang

Information Retrieval and Knowledge Management Research Lab
York University, Toronto, Canada

Abstract. Large language models (LLMs) are revolutionizing information retrieval (IR). Retrieval-Augmented Generation (RAG) and Generation-Augmented Retrieval (GAR) leverage LLMs to enhance IR. RAG improves response accuracy by integrating verified external information, addressing the "hallucinated" content. GAR refines search queries and enhances document representations, leading to more relevant results. The interaction between RAG and GAR creates a powerful synergy, improving both retrieval and generation processes. It is expected that rough set theory can be utilized to optimize the process in RAG and GAR via identifying the key features.

Keywords: Generative Information Retrieval · Generation-augmented Retrieval · Retrieval-augmented Generation.

1 Introduction

The emergence of large language models (LLMs) has led to significant advancements in information retrieval (IR) technologies, resulting in the development of methods such as Retrieval-Augmented Generation (RAG) and Generation-Augmented Retrieval (GAR) [4]. These methods utilize the advanced generative abilities and deep semantic understanding of LLMs to increase the precision and effectiveness of information systems.

Retrieval-Augmented Generation (RAG) aims to improve the reliability of responses generated by LLMs [5]. By dynamically retrieving and integrating external information during the inference process, RAG attempts to anchor the model's responses in verified contents. This approach addresses the issue of "hallucinated" information—coherent yet factually incorrect content produced by LLMs. The success of RAG depends on the model's ability to effectively use the retrieved information, which relies on the quality and completeness of the external sources.

Conversely, Generation-Augmented Retrieval (GAR) seeks to improve search results by utilizing the generative capabilities of LLMs. GAR employs these models to extend and refine search queries or to enhance document representations [2][6], thus better aligning user queries with the document corpus. This method not only increases the relevance of search results but also broadens the range of contents accessible in response to complex queries.

Rough sets have been successfully applied for Web mining (e.g. Web usage mining and web page classification) [1][3]. We expect that rough set theory can

also be utilized to optimize the process in RAG and GAR, ensuring that the retrieved contents are highly relevant and specific to the user's needs. This can be achieved by identifying the key features of the information that are most relevant to the generation task.

2 RAG and GAR

The interaction between RAG and GAR creates a powerful synergy where each component can inform and improve the other, forming a feedback loop that potentially enhances both retrieval and generation processes.

From RAG to GAR: The factual content retrieved and validated by RAG provides a reliable foundation for GAR tasks. By ensuring the accuracy of the information used to enhance queries or documents, RAG supports GAR in producing more focused and contextually relevant search enhancements. This leads to more effective and precise information retrieval, minimizing the occurrence of irrelevant results.

From GAR to RAG: In turn, the improved querying capabilities developed in GAR can substantially benefit RAG. Enhanced queries facilitate more effective and accurate information retrieval during the RAG process, ensuring that the external content integrated during generation is highly relevant and specific to the user's needs. This not only enhances the quality of the generated responses but also reduces the system's load by preventing the retrieval of unnecessary information.

The cyclical synergy between RAG and GAR presents a robust framework for leveraging the strengths of LLMs in both generating and retrieving information. Continuously refining these interactions is crucial for developing next-generation IR systems that are both intelligent and intuitive, capable of meeting the increasingly complex demands of users in the digital age. Further exploration and optimization of the dynamics between RAG and GAR are essential for achieving greater levels of accuracy and reliability in both generated content and search results.

References

1. An, A., Huang, Y., Huang, X. and Cercone, N.: "Feature selection with rough sets for web page classification". Transaction on Rough Sets. 2:1-13. 2004.
2. Huang, J. X., Miao, J. and He, B.: "High performance query expansion using adaptive co-training". IPM. 49(2):441-453, 2013.
3. Huang, X., Cercone, N. and An, A.: "Comparison of interestingness functions for learning web usage patterns". In Proc. of 2002 ACM CIKM, 617-620, 2002.
4. Huang, Y. and Huang, J. X.: "Exploring ChatGPT for next-generation information retrieval: Opportunities and challenges". Web Intelligence Journal. 1-16, 2024.
5. Huang, Y. and Huang, J. X.: "A Survey on Retrieval-Augmented Text Generation for Large Language Models". arXiv preprint arXiv:2404.10981. 2024.
6. Ye, Z., Huang, J. X. and Lin, H.: "Finding a good query-related topic for boosting pseudo-relevance feedback". JASIST. 62(4):748-760, 2011.