PhD position in Engineering and Computer Science, Criteo AI Lab, Paris, Fr

Deep Generative Models for Next-Generation Search and Recommendation

---

Contact : Alain Rakotomamonjy, a.rakotomamonjy@criteo.com, Alberto Lumbreras, a.lumbreras@criteo.com, Patrick Gallinari, patrick.gallinari@sorbonne-universite.fr

Location: Criteo AI Lab, Paris

Candidate profile: Master degree in computer science or applied mathematics, Engineering school. Background and experience in machine learning. Good technical skills in programming.

How to apply: please send a cv, motivation letter, grades obtained in master, recommendation letters when possible to the contacts

Start date: November/December 2024 for three years

Keywords: deep learning, language models, generative models, information retrieval, recommendation

---

Company context:

Criteo AI Lab, operates within the spectrum of two main areas: ML Engineering and ML Research. The position is opened for working in the research department. Research here is mainly dedicated to machine learning topics. The analysis of semantic data has become a major R&D topic for a variety of Criteo business. The PhD targets recent advances in the domain of generative deep learning.

Scientific context

Search and recommendation are at the core of Criteo's business. Generative Information Retrieval (GenIR) and Generative Recommendation (GenREC) are emerging paradigms based on foundation models, promising to transform how we search for and access information. GenIR integrates all components of traditional IR systems into a single generative model, directly generating relevant responses—such as document identities—from user queries. This approach eliminates the conventional distinction between the source of knowledge and the search engine. GenREC operates in a similar manner. Given this framework, search and recommendation—both critical problems for Criteo—can be structured in similar ways. Search becomes the process of matching user queries with catalog products, while recommendation involves matching a user's history and profile with relevant products in the catalog. Looking ahead, GenIR and GenREC will serve as a bridge to Large Language Models-based text generation applications, akin to the role foundation models play today. LLMs are emerging as a new way for users to access information (e.g., on the web or in retail catalogs) and may soon replace traditional search engines and recommendation interfaces. Therefore, developing generative search and recommendation capabilities within LLM-driven interfaces (such as chatbots) is a crucial challenge for Criteo.

Research directions

The goal of this PhD project is to explore this research direction. The first step will be to develop a unified generative engine for both search and recommendation, allowing for seamless alternation

between the two modes during interactive sessions using a single engine. This is also a step toward realizing foundation models that offer a variety of functions to enhance user interactions. The second step will involve adapting this model to the large-scale, dynamic corpora characteristic of recommendation systems in the adtech industry, which presents additional research challenges. A brief description of the two directions is provided below.

Task 1: Unifying Generative IR and Recommendation

This task aims to develop a unified engine for search and recommendation, allowing for alternating between the two modes in interactive sessions. The goal is to enhance performance in both domains through a multi-task framework, enriching training data for both. While search and recommendation share similarities, they also have key differences, such as query intent. Search is driven by user queries, while recommendation relies on past user behavior. We aim to address these differences by defining a joint architecture and multi-task training strategy that captures the semantic distinctions between search (similarity-based) and recommendation (collaborative).

Task 2: Enhancing ID Associations for Large and Dynamic Collections

In this task, the goal is to improve document and item ID representations in large-scale, dynamic collections for a joint search/recommendation system. We will explore methods such as hierarchical structures and prior knowledge (e.g., product taxonomies) to optimize ID design. By leveraging additional information like brands or categorizations, we aim to improve the retrieval and recommendation process, particularly for large and evolving datasets.

References

Hou, Y., Li, J., He, Z., Yan, A., Chen, X., & McAuley, J. (2024). Bridging Language and Items for Retrieval and Recommendation. 1. http://arxiv.org/abs/2403.03952

Hua, W., Xu, S., Ge, Y., & Zhang, Y. (2023). How to Index Item IDs for Recommendation Foundation Models. SIGIR-AP 2023, 195–204. https://doi.org/10.1145/3624918.3625339

Li, J., Wang, M., Li, J., Fu, J., Shen, X., Shang, J., & McAuley, J. (2023). Text Is All You Need: Learning Language Representations for Sequential Recommendation. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1258–1267. https://doi.org/10.1145/3580305.3599519

Li Y. , Yang N., Wang L., Wei F., & Li W., (2023). Generative retrieval for conversational question answering. Information Processing & Management 60, 5 (2023), 103475

Li, X., Jin, J., Zhou, Y., Zhang, Y., Zhang, P., Zhu, Y., & Dou, Z. (2024). From Matching to Generation: A Survey on Generative Information Retrieval. 2, 1–45. http://arxiv.org/abs/2404.14851

Li, Y., Lin, X., Wang, W., Feng, F., Pang, L., Li, W., Nie, L., He, X., & Chua, T.-S. (2024). A Survey of Generative Search and Recommendation in the Era of Large Language Models. http://arxiv.org/abs/2404.16924

Rajput, S., Mehta, N., Singh, A., Keshavan, R. H., Vu, T., Heldt, L., Hong, L., Tay, Y., Tran, V. Q., Samost, J., Kula, M., Chi, E. H., & Sathiamoorthy, M. (2023). Recommender Systems with Generative Retrieval. http://arxiv.org/abs/2305.05065

Sun, W., Yan, L., Chen, Z., Wang, S., Zhu, H., Ren, P., Chen, Z., Yin, D., de Rijke, M., & Ren, Z. (2023). Learning to Tokenize for Generative Retrieval. Http://Arxiv.Org/Abs/2304.04171. http://arxiv.org/abs/2304.04171

Tay, Y., Tran, V. Q., Dehghani, M., Ni, J., Bahri, D., Mehta, H., Qin, Z., Hui, K., Zhao, Z., Gupta, J., Schuster, T., Cohen, W. W., & Metzler, D. (2022). Transformer Memory as a Differentiable Search Index. Neurips.

Zheng, B., Hou, Y., Lu, H., Chen, Y., Zhao, W. X., Chen, M., & Wen, J.-R. (2024). Adapting Large Language Models by Integrating Collaborative Semantics for Recommendation. ICDE. http://arxiv.org/abs/2311.09049