Bridging Search and Recommendation with Generative Retrieval

October 21, 2024
Published by Gustavo Penha, Ali Vardasbi, Enrico Palumbo, Marco de Nadai, Hugues Bouchard

Search engines and recommendation systems use different signals to represent users and catalog items, catering to their distinct tasks. In search engines, explicit user signals such as natural language queries are central, while recommendation systems rely more heavily on historical sequences of user interaction data to model user preferences. Although they often coexist on many industrial online platforms, these systems typically operate with separate models and input features. However, we hypothesize that search data could enhance recommendation systems, and vice versa, as the tasks capture different user behaviors and contain complementary information (content-based and collaborative-filtering-based) to represent items in the catalog.

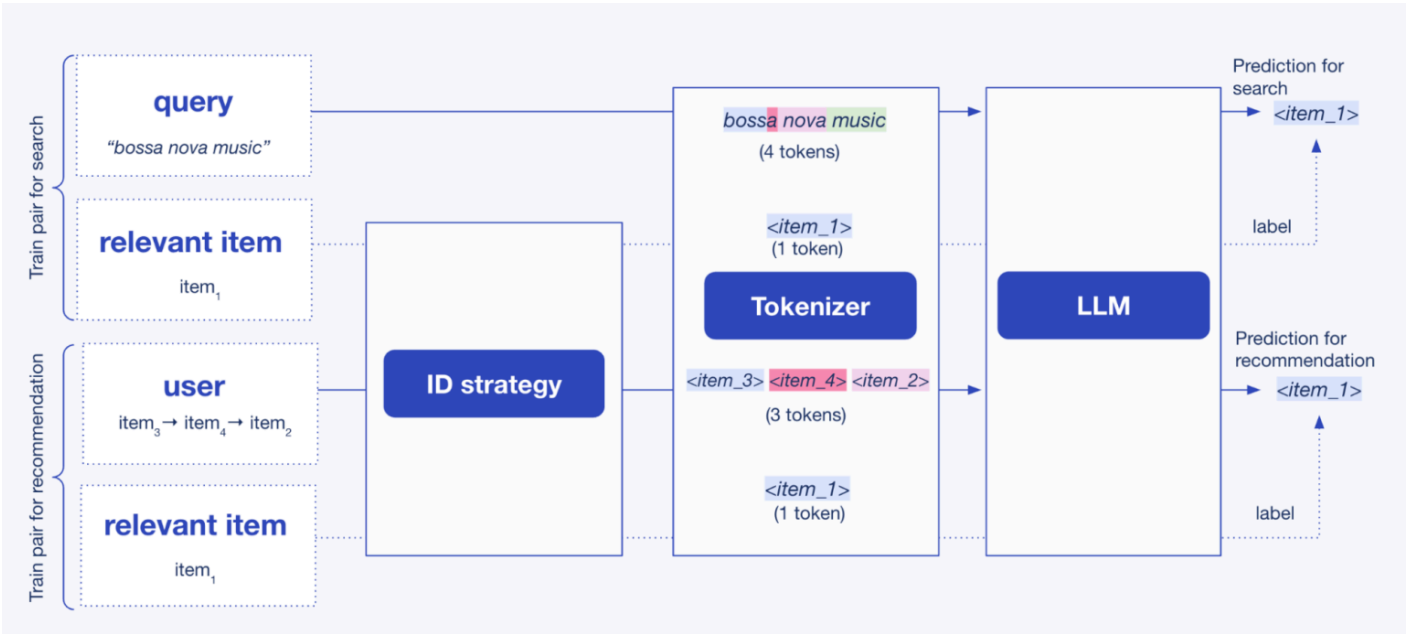
Generative retrieval for search and recommendation is a promising new paradigm to retrieve items. It offers a compelling way to unify multiple tasks within a single

recommendation. Our experiments, conducted using both simulated and real-world datasets, demonstrate that a joint generative model can indeed outperform task-specific approaches. A key finding of our analysis is that the regularization effect on the item's latent representation plays a significant role in achieving performance gains.

In this post we explore how the generative model retrieves items for both search and recommendation tasks. We also present our main hypotheses, summarize the results, and offer our conclusions.

Joint Generative Model

Generative retrieval models directly predict item identifiers for a given user or query. Unlike dense retrieval approaches, such as bi-encoder and two-tower models, that represent queries, users and items as embeddings in a shared space before applying nearest neighbor search, generative retrieval models learn a mapping that connects inputs to item IDs. This enables retrieval using LLMs and plays an important role in unifying multiple information retrieval tasks within a single model. In this paper, we focus specifically on applying this approach to search and recommendation tasks.

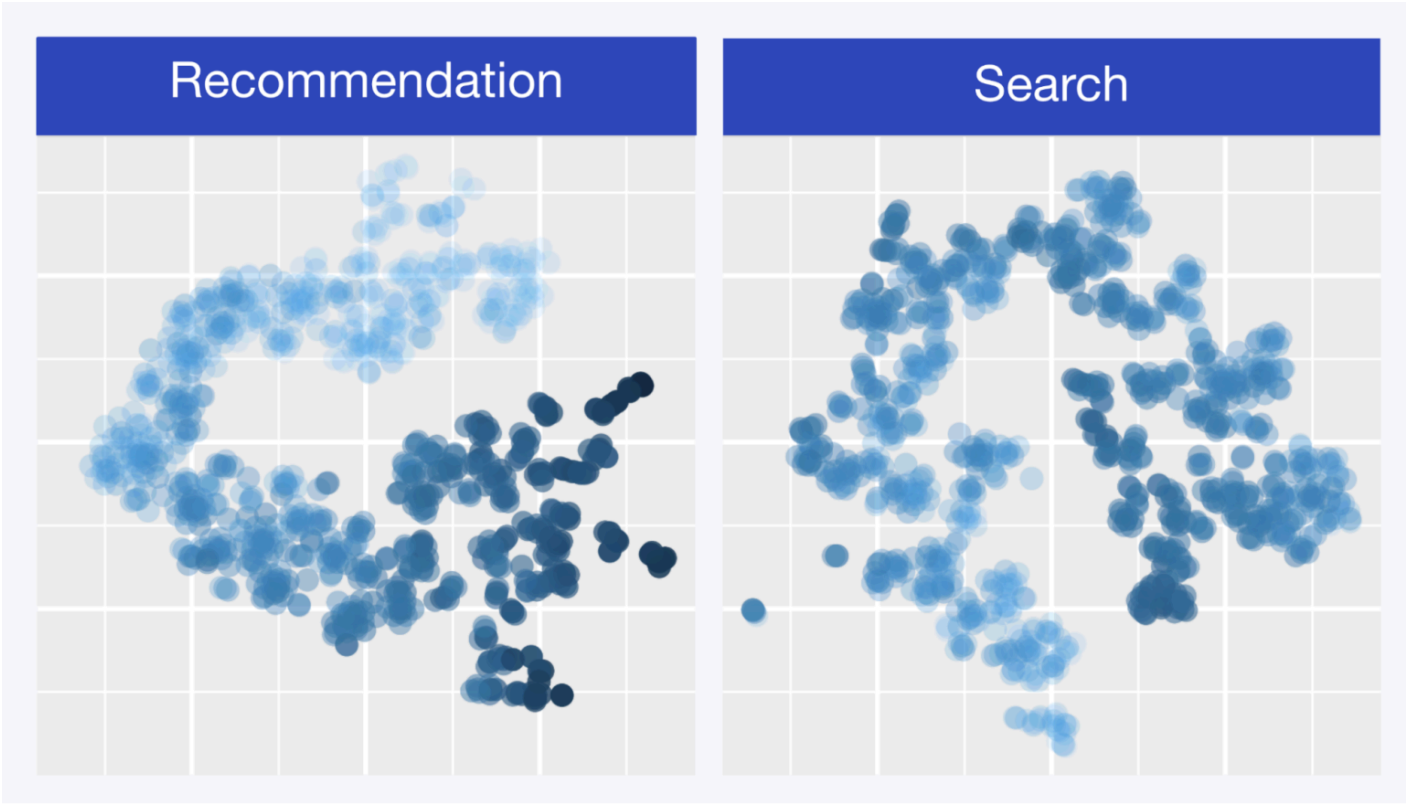


and item pairs, see bottom two boxes), to better represent each item in the catalog for each task.

We train a single generative retrieval model using both query-item pairs for search and user-item pairs for recommendation in a multi-task learning setup.

Hypotheses and Experiments

We are guided by two key hypotheses that could explain why a joint model might produce better item representations: (1) joint training regularizes the estimation of each item's popularity, and (2) joint training regularizes the item's latent representations. The motivation behind the first hypothesis is that both search and recommendation models using generative retrieval may exhibit a bias towards popular items.



t-SNE projection of the latent representations of a generative recommender (left) and generative search (right) for the MovieLens dataset. The popularity of each item (darker if the item occurs more often in the respective recommendation and search

The second hypothesis stems from the idea that search and recommendation capture different aspects of item similarity, leading to better representations. While search tends to capture content-based information, recommendation signals incorporate collaborative filtering data.

Comparison of recommendation and search signals. Items are represented by a blue dot, and connections in recommendation indicate that the item co-occurs in different user profiles, while for search it indicates that they are connected to a textual query.

Our experiments with **simulated data** show that the joint generative model outperforms task-specific models under certain conditions. For example, when testing our second hypothesis on the regularization of items latent representations, we found that improvements occur when the distribution of item co-occurrences across tasks aligns above a certain threshold.

Joint training of generative retrieval models for both recommendation and search outperforms task-specific models across **three real-world datasets,** with an average increase of 16% in R@30. Our follow-up analyses suggest that the regularization effect on item latent representations (our second hypothesis) is the main factor behind the differing predictions of the joint generative model compared to the task-specific models.

Conclusion

and recommendation tasks for generative retrieval. In future research, we plan to explore the impact of incorporating additional tasks, such as generating explanations, within a unified multi-task-learned LLM for IR, and to investigate semantic IDs in the multi-task learning framework.

For more information, please refer to our paper:
Bridging Search and Recommendation in Generative Retrieval: Does One Task Help the Other?
Gustavo Penha, Ali Vardasbi, Enrico Palumbo, Marco De Nadai, Hugues Bouchard. RecSys'24.

SHARE

CATEGORIES

Search & Recommendations

Related articles

Socially-Motivated Music Recommendation

Exploring Local Music's Place in Global Streaming

Personalizing Audiobooks and Podcasts with graph-based models

Sign up for research updates

By clicking sign up you'll receive occasional emails from Spotify. You always have the choice to adjust your interest settings or unsubscribe.

Your Email

Sign Up

Newsroom Spotify Jobs Spotify.com

Spotify R&D Engineering Spotify R&D Design

Legal Privacy Cookies About Ads