A Hybrid Approach of Statistics and Embeddings for Multilingual and Multi-Locale Recommendation

Team: gpt_bot

Weijia Zhang, Jin Zhan, Zhongshan Huang, Lu Wang, Qiang Wang

Background

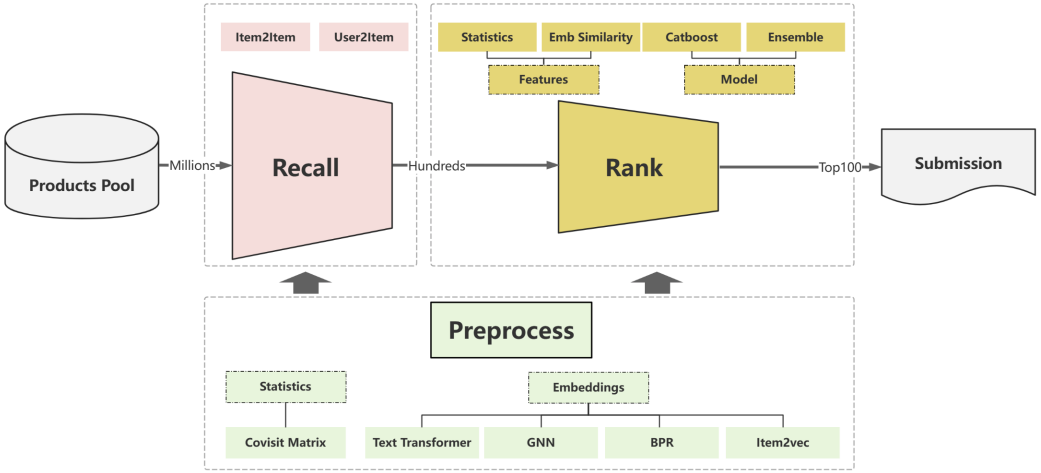
To encourage the development of multilingual recommendation systems, Amazon published a multilingual and multi-locale shopping session dataset, and KDD Cup 2023 challenge on Multilingual Session Recommendation Challenge was hosted based on this dataset.

There are 3 tasks in this Challenge:

1. Next Product Recommendation
2. Next Product Recommendation for Underrepresented Languages and Locales;
3. Next Product Title Generation

Method

As depicted in the figure, our solution is composed of two main stages: recall and rank. Although these stages are distinct, they share a significant amount of data. Therefore, we first generate multiple matrices and various embeddings through covisit statistics, text transformer, GNN(ProNE), BPR and item2vec in the preprocessing. These preprocessed results serve as the foundation for both recalling and ranking.



The first stage is candidate item recalling, which involves retrieving a set of potential products that each user may be interested in from the entire pool of products. This is achieved by utilizing the preprocessed matrices and embeddings. Approximately 200 items are retrieved for each user.

The second stage is candidate item ranking, where we predict the probability of a user engaging with each retrieved item. The ranking model is an ensemble of two Catboost models, which involve many statistical features and embedding similarity features.

Experiment Results

Finally, we achieved 4th place in Task1 and 3rd place in Task2. The metrics(MRR@100) of each single model and the ensemble model are listed in the following table.

Models	Task1	Task2
A	0.4030	0.4580
B	0.3968	0.4468
Ensemble	0.4047	0.4601