

Pre-rank Model

Monday, April 1, 2024

12:19 PM

Pre-rank

score ~ 1000 items

single inference cost is small

OK to have low accuracy

Rank

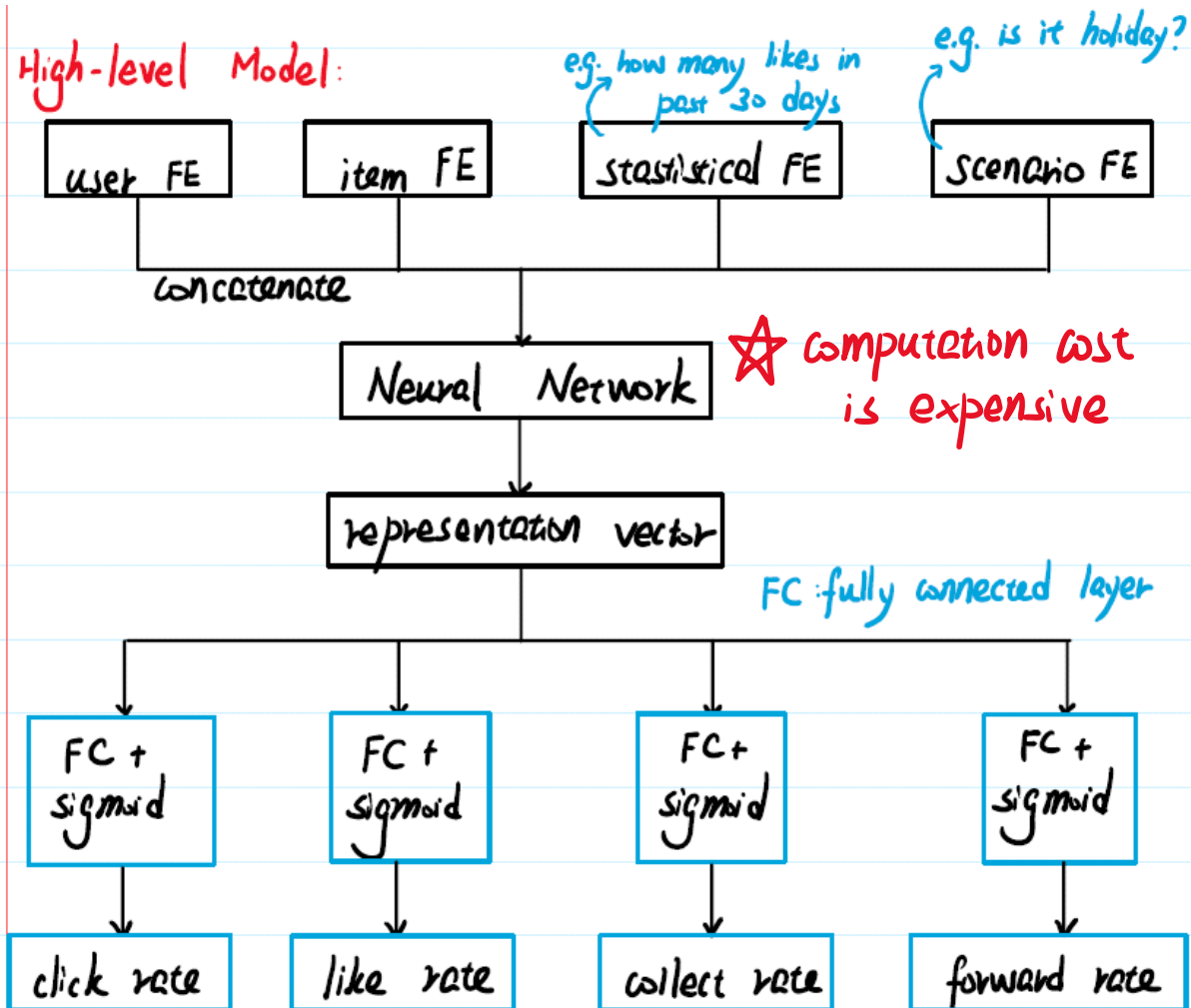
score ~ 100 items

cost is large

need high accuracy

Ranking Model Review:

High-level Model:

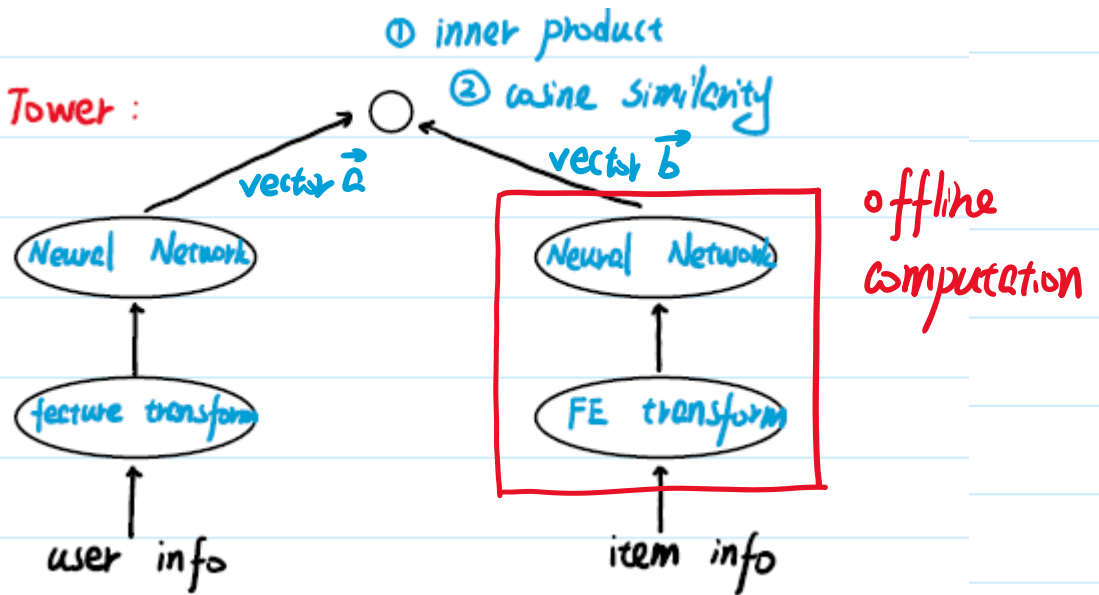


Pre-rank Model

Monday, April 1, 2024 12:19 PM

Review:

Two Tower:



Rank Model

- ① early concatenate
- ② online inference cost is large

Retrieval Model (Two Tower)

- ① late "fusion"
- ② item vector is offline computed. online inference is only for user tower

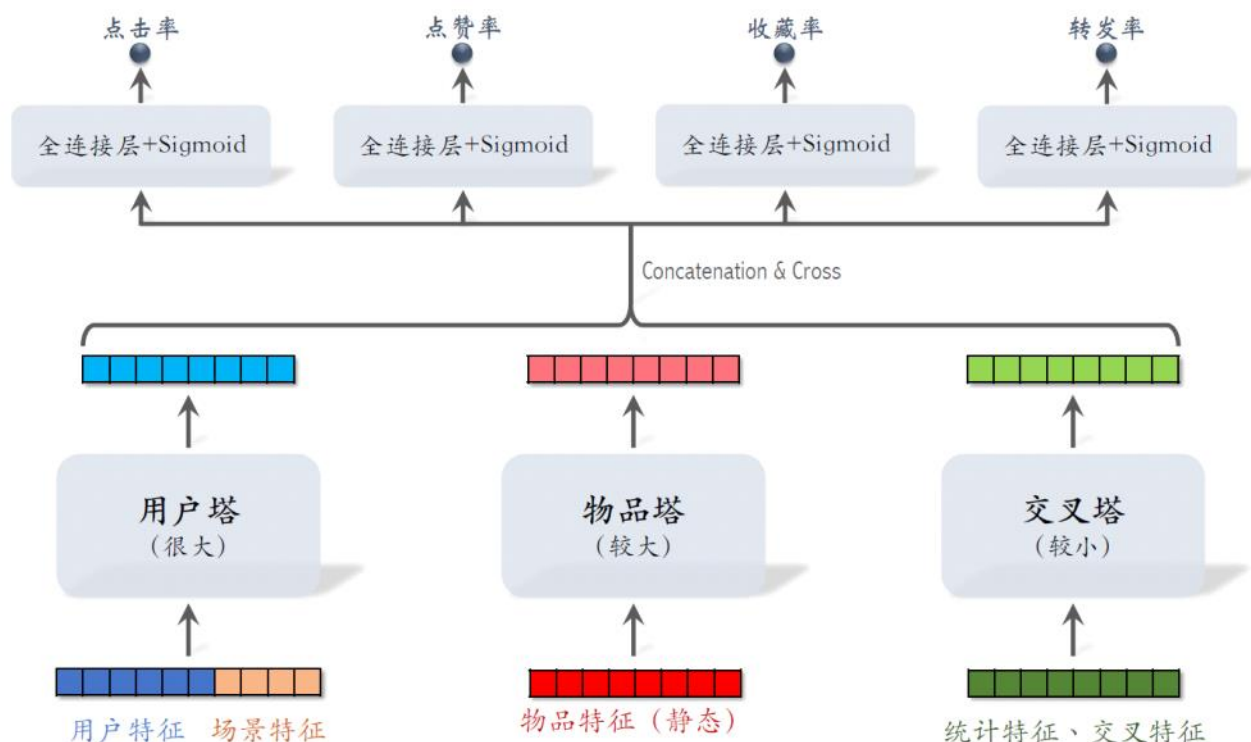
Balance between rank model and retrieval model
→ we will have "pre-rank" model.

Pre-rank Model

Monday, April 1, 2024

12:19 PM

Pre-rank Model (Three Tower Model)



(Picture from Shusen Wang on Bilibili)

User Tower:

Each time we only serve one user. so cost is ok.

Item Tower:

parameter server.

N items $\rightarrow N$ inference (use cache)
item info is static

Cross Tower:

all rely online inference (network needs to be small)

FC + sigmoid:

N times of inference (where most cost comes from)