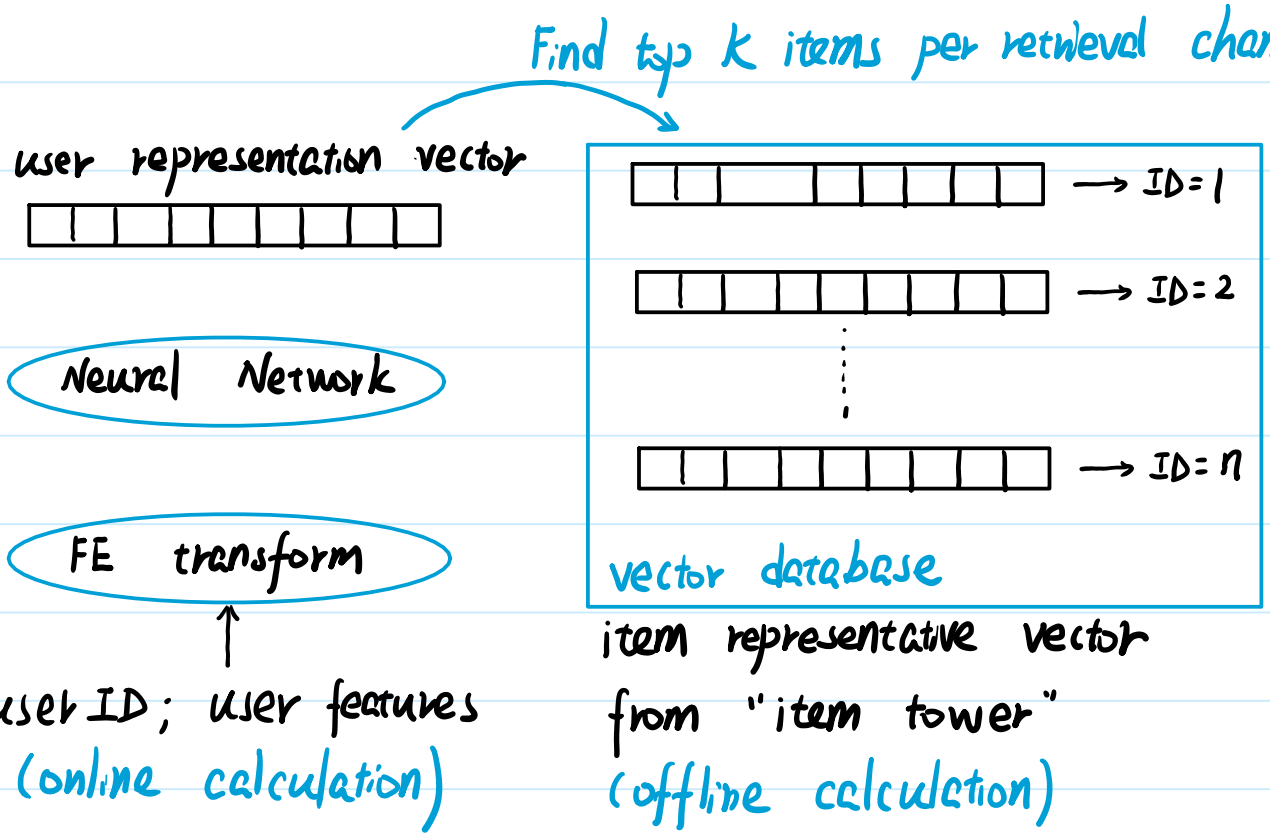
Two Tower Model Online Retrieval

Tuesday, March 12, 2024 1:15 PM

Two Tower Model Workflow:



Assumption:

① user interest changes with time

② item character remains stable.

Full Update of Model:

① based on parameters of previous model (NOT random initialize)

② use yesterday's data to train 1 epoch (shuffle data)

③ deploy new user and item tower

Two Tower Model Online Retrieval

Tuesday, March 12, 2024 1:15 PM

Incremental Update of Model:

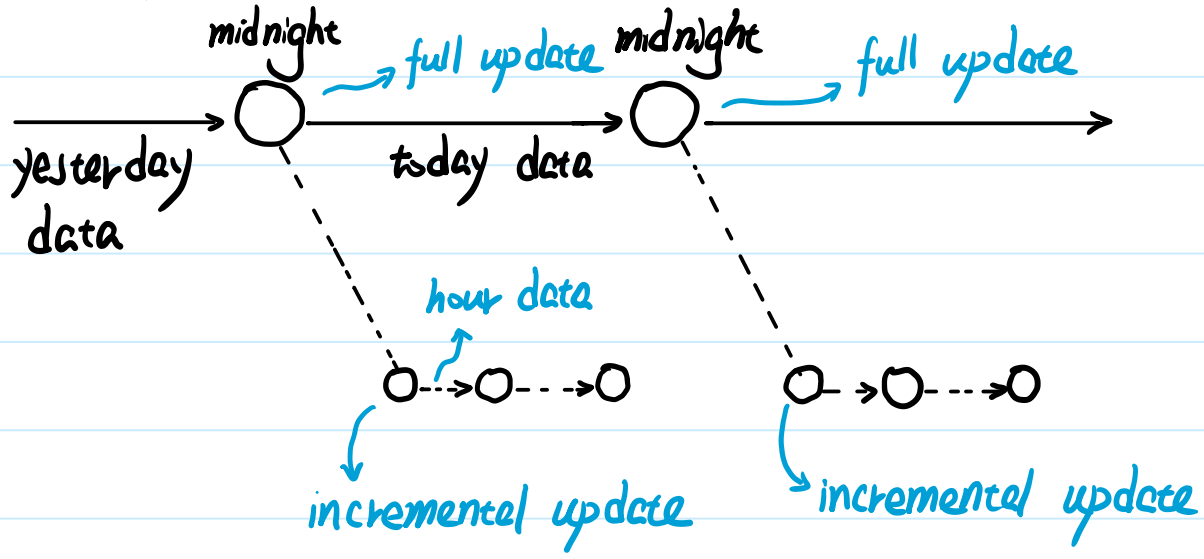
① user interest changes from hour to hour

② online data stream → TF Record file → train model

③ online learning → update embedding layer
(freeze other layers: fully connected layer)

④ deploy new user embedding layer

Full Update vs. Incremental Update:



Incremental updated model will be discarded at the end of every day.

Two Tower Model Online Retrieval

Tuesday, March 12, 2024 1:15 PM

Can we only do incremental update and remove full update ?

No: ① hourly data has bias

② full update: random shuffle data → 1 epoch

incremental update: data ordered by time

③ full update: less bias

incremental update: capture real-time interest change.