Item Collaborative Filtering

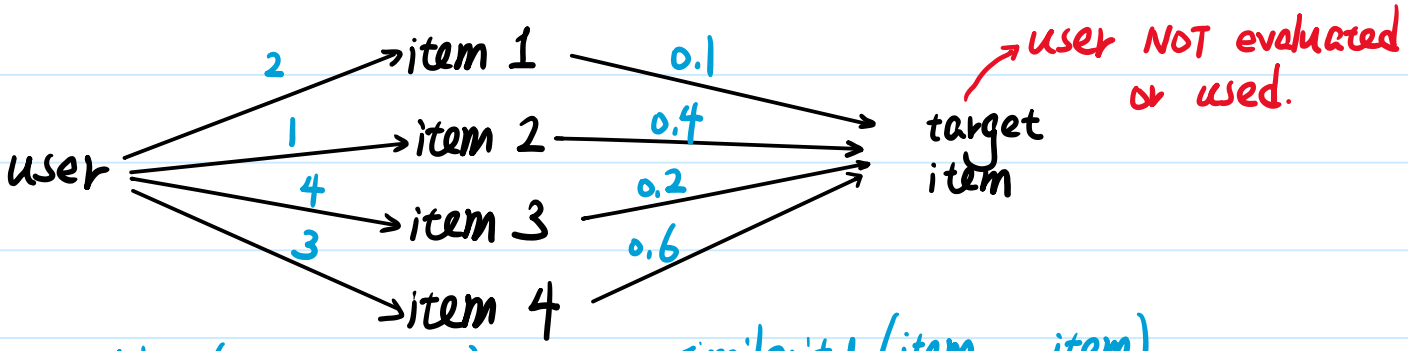
Monday, February 26, 2024 7:40 AM

Example:

I have watched "John Wick I".
I have NOT watched "John Wick II"
→ Recommend me: "John Wick II"

because "Joh Wick I" and "John Wick II" are similar.

How does the system know two items are similar?



like (user, $item_j$) similarity ($item_j$, item)

estimated like to target item:

$$\Sigma \text{ like} \cdot \text{similarity} = 2 \times 0.1 + \cdots + 3 \times 0.6 = 3.2$$

Calculate similarity:

user set W_1 like item i_1

user set W_2 like item i_2

$$\text{similarity}(i_1, i_2) = \frac{|W_1 \cap W_2|}{\sqrt{|W_1| \cdot |W_2|}} \in [0, 1]$$

# of users like both i_1 and i_2.

Item Collaborative Filtering

Monday, February 26, 2024 7:40 AM

if we include "how much" users like items i_1, i_2:

$$\text{similarity}(i_1, i_2) = \frac{\sum_{v \in w_1 \cap w_2} \text{like}(v, i_1) \cdot \text{like}(v, i_2)}{\sqrt{\sum_{u_1 \in w_1} \text{like}^2(u_1, i_1)} \cdot \sqrt{\sum_{u_2 \in w_2} \text{like}^2(u_2, i_2)}}$$

(cosine similarity)

$\text{like}(u_i, i_1) = [4, 1, 4, 3, \cdots\cdots 4]$

user #3 rating on item i_1

How item CF is used in candidate retrieval?

Offline calculation:

① index from "user ID" to "their interacted items"

② index from "item" to "item" similarity score.

Example:

① user #1 (user ID) ⟶ item #1 (item ID): 4 (how much the user likes); item #7: 3; ····

② item #1 (item ID) ⟶ item #7 (item ID): 0.5 (similarity score between two items); item #9: 0.2; ····

Item Collaborative Filtering

Monday, February 26, 2024 7:40 AM

Online retrieval:

① Given user ID, get their recently interacted N items using "user → item" index
   ↳ named "last N"

② for each item in "last N", get their top k similar items using "item → item" index
   ↳ total # of items: N · K

③ use equation: Σ like · similarity to estimate like score for each item in N · K items
   ↳ M retrieval channels in parallel

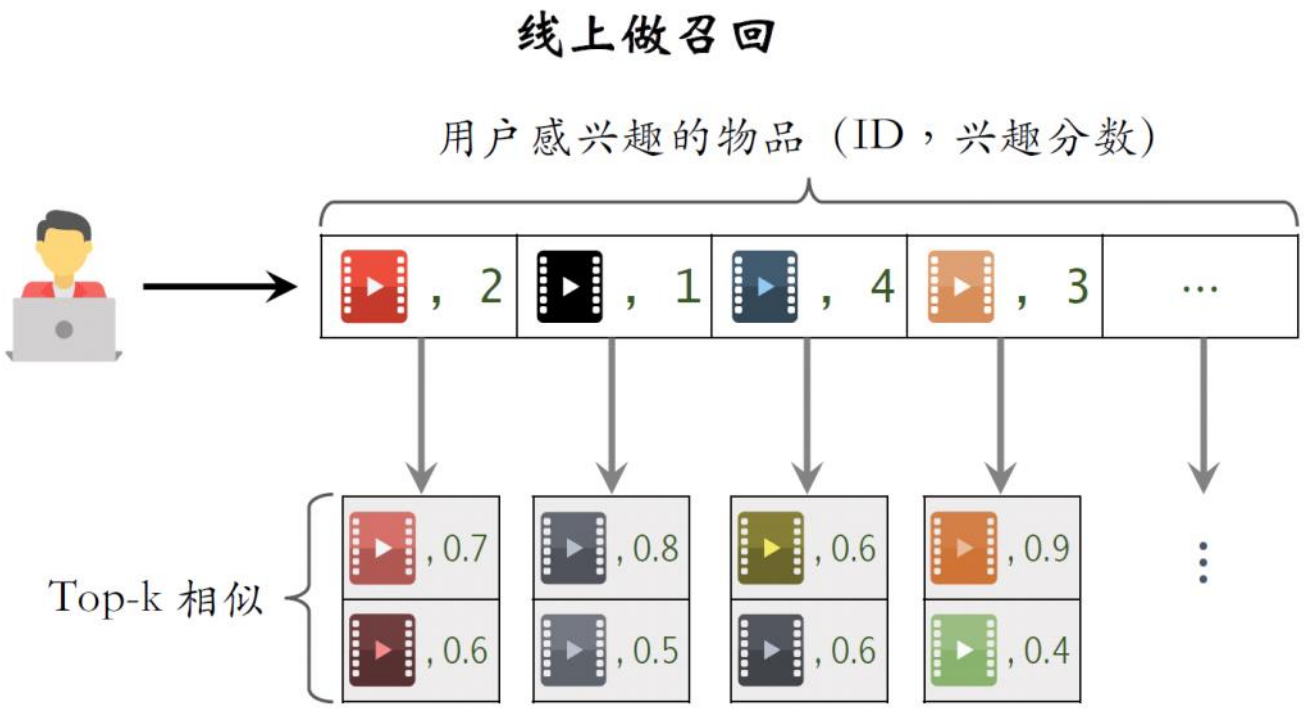
④ return top 100 items as the retrieval.
   ↳ total # of retrieval results: 100 · M

Why use index?

if No index, we need to calculate like score for millions of items. → heavy computation
using index relies on offline calculation; but online calculation is light.

Item CF

Monday, February 26, 2024 8:35 AM



(picture from Shusen Wang on Youtube/Bilibili)