Exposure Filter and Bloom Filter

Friday, March 22, 2024 8:40 AM

Background:

① if users watched an item before, do not recommend the same item again.

② for every user, track the watching history

③ for every item generated from retrieval exclude those included in watching history

④ user watched n items and retrieval generates r items to check if the item exist in watching history

time complexity $O(r \cdot n)$

Brute force is slow

Bloom Filter

① a data structure to determine if an item ID is in the set of watched items

② if NO: the item must NOT be in the set

③ if YES: the item is LIKELY in the set

(it may put an item that was NOT watched into "watched" by mistake)

Exposure Filter and Bloom Filter

Friday, March 22, 2024 8:40 AM

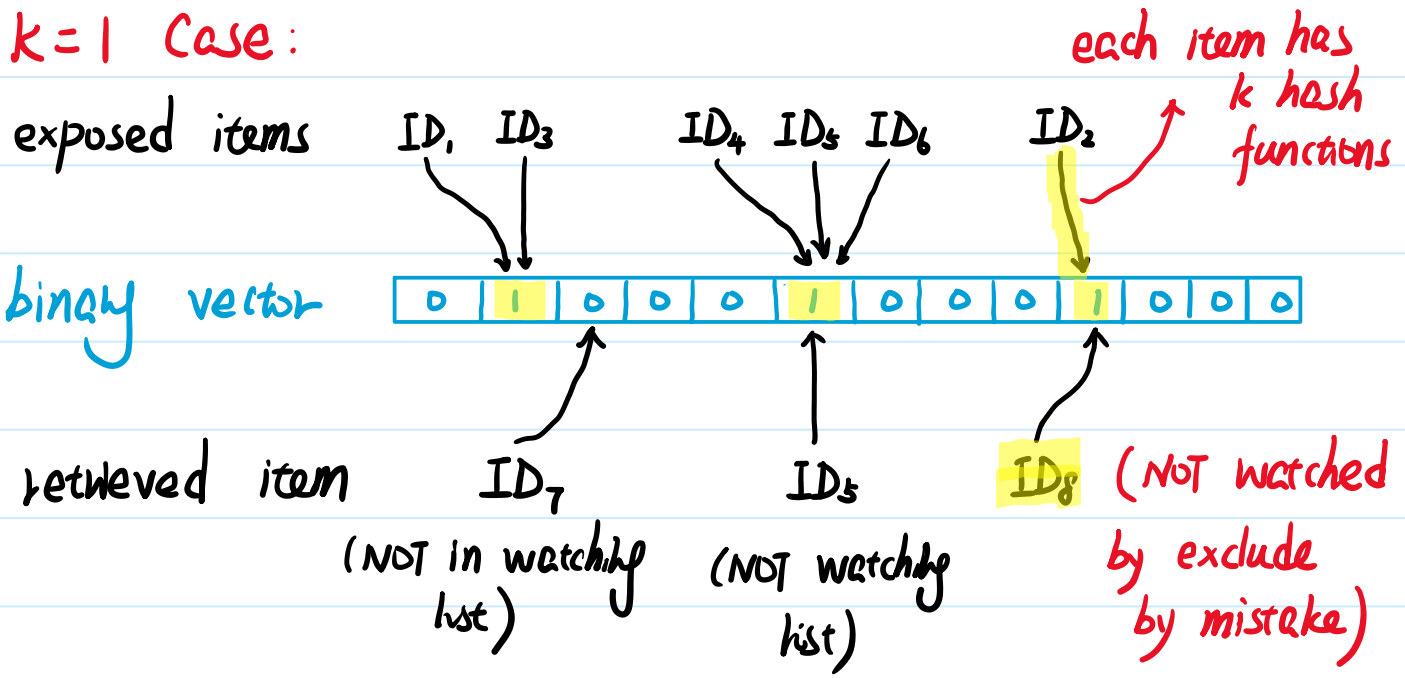
Bloom Filter:

① item ⟶ binary vector with m dimension

② each user has a "watched item set". stored as a vector with m bit space

③ Bloom filter has k hash functions that map each item ID to integers between 0 and $m-1$.

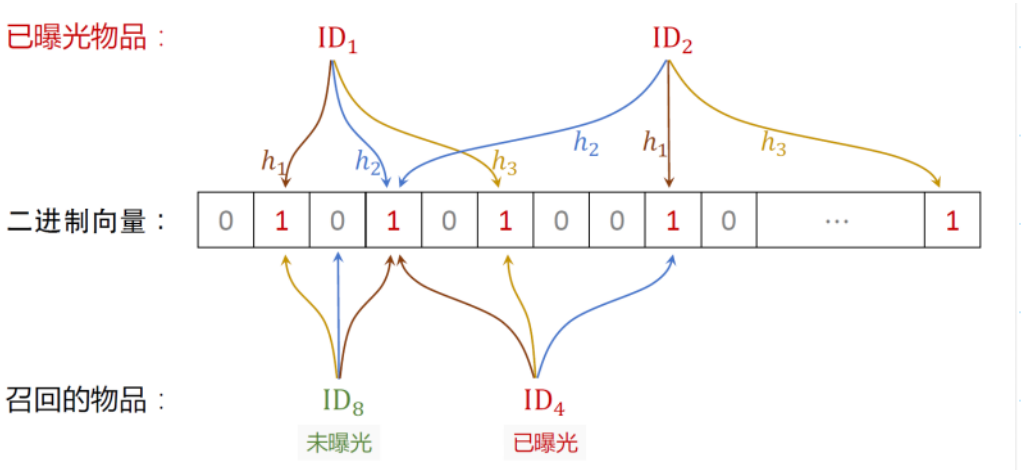
$k=1$ Case:



Exposure Filter and Bloom Filter

Friday, March 22, 2024 8:40 AM

K=3 Case:



(Picture from Shusen Wang on Youtube/Bilibili)

False Positive Probability:

① when bloom filter say "NOT in the wachlp set"
it is 100% correct

② when it say "In the watching set"
it may be wrong.

probability of being wrong: $\delta = \left[1 - \exp\left(-\frac{kn}{m}\right)\right]^k$

(a) large n → more ones in vector → easy to be wrong

(b) large m → more space in vector → hard to hash conflict

(c) k should NOT be too large or small → has optimal

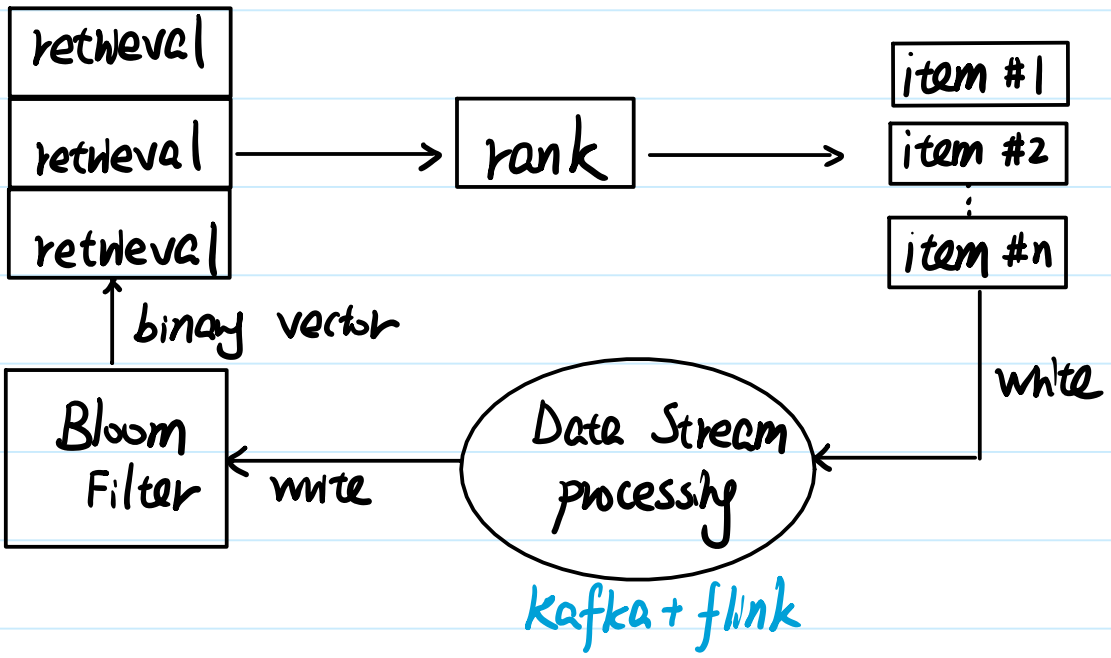
Exposure Filter and Bloom Filter

Friday, March 22, 2024 8:40 AM

Optimal parameters: (based on tolerated probability δ)

$$k = 1.44 \cdot \ln\left(\frac{1}{\delta}\right), \quad m = 2n \ln\left(\frac{1}{\delta}\right)$$

Entire Workflow



limitations of Bloom Filter:

it ONLY supports adding items; cannot delete items

(we cannot change 1 back to 0 in binary vector; because it affects other items as well)