Filtros de Bloom

Certo.

Para essa ideia fazer sentido, é preciso que o acesso ao filtro seja muito rápido.

E a melhor maneira de conseguir isso é usar um filtro imperfeito.

Quer dizer, vão haver casos em que o filtro vai dizer SIM mas o elemento não está na base de dados.

Nesses casos, a base de dados terá que ser consultada para verificar que o elemento realmente não está lá.

Por outro lado, quando o filtro responde NÃO é porque o elemento não está lá mesmo.

Legal.

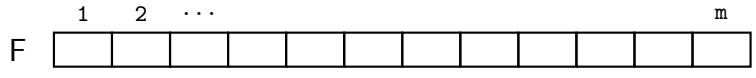
Mas como é que a coisa funciona?

Bom, a ideia é simples.

Quer dizer, nós vamos usar k funções de dispersão

$$\mathsf{h}_1(.), \quad \mathsf{h}_2(.), \quad \ldots, \quad \mathsf{h}_k(.)$$

E um vetor de bits com tamnanho m, que vai fazer o papel de filtro

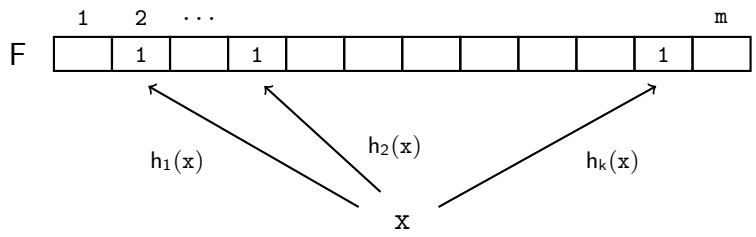


Agora imagine que nós fazemos a inserção do elemento `x` na base de dados.

Daí, para registrar esse fato no filtro, nós calculamos as posições,

$$\mathtt{i}_1 \leftarrow \mathsf{h}_1(\mathtt{x}), \qquad \mathtt{i}_2 \leftarrow \mathsf{h}_2(\mathtt{x}), \qquad \ldots \qquad \mathtt{i}_k \leftarrow \mathsf{h}_k(\mathtt{x}),$$

e ajustamos os bits nas posições $\mathtt{i}_1, \ldots, \mathtt{i}_k$ para o valor 1 — (*no início, todos os bits tem o valor 0*)



E a mesma coisa acontece com todos os elementos que são inseridos na base de dados.

Agora, imagine que ocorre uma consulta pelo elemento `z`.

Daí, nós calculamos as posições

$$\mathtt{j}_1 \leftarrow \mathsf{h}_1(\mathtt{z}), \qquad \mathtt{j}_2 \leftarrow \mathsf{h}_2(\mathtt{z}), \qquad \ldots \qquad \mathtt{j}_k \leftarrow \mathsf{h}_k(\mathtt{z}),$$

e examinamos os bits nessas posições do filtro.

Caso ao menos um desses bits seja igual a zero, nós temos a certeza de que o elemento `z` não está na base de dados — (*porque?*)

Por outro lado, se todos os bits são 1, então existem dois casos

1. pode ser que o elemento `z` tenha sido inserido na base de dados
   — (*o que fez com que os bits recebessem o valor 1*)
2. ou pode ser que `z` não esteja lá, e que os bits tenham recebido o valor 1
   na inserção de outros elementos

E é assim que surge a possibilidade de um falso positivo.