

## Bloom Filters

Supongamos una BD  $S = \{s_1, \dots, s_m\} \subseteq D$

tal que  $|D| \leq N$  y  $D = \{d_1, \dots, d_N\}$

Para guardar  $S$  podemos usar un bitmap

$B \in \{0, 1\}^N$  tal que  $B[i] = 1$  si  $d_i \in S$

Dado cualquier objeto  $d \in D$ :

- ¿cómo vemos si  $d \in S$ ?
- ¿cómo insertamos  $d$  en  $S$ ?

¿Cuál es el problema de esta estructura  $B$ ?

Ejemplo: -  $D$  son documentos  
 - " son palabras o números de 32 bits  
 - etc.

Solución: usar un bitmap  $B$  de tamaño  $n$   
 con  $n \leq |D|$  y una función de  
 hash  $h: D \rightarrow \{1, \dots, n\}$

Start 0 0 ... 0

insert  $x$  0 0 ... 0  $\Rightarrow$  0 1 ... 0



Para cualquier objeto  $d \in D$ :

¿cómo verificamos si  $d \in S$ ?

- si  $B[h(d)] = 0 \Rightarrow d \notin S \checkmark$
- si  $B[h(d)] = 1 \begin{cases} \rightarrow d \in S \checkmark \\ \rightarrow d \notin S \checkmark \end{cases}$   
tenemos falso positivo.

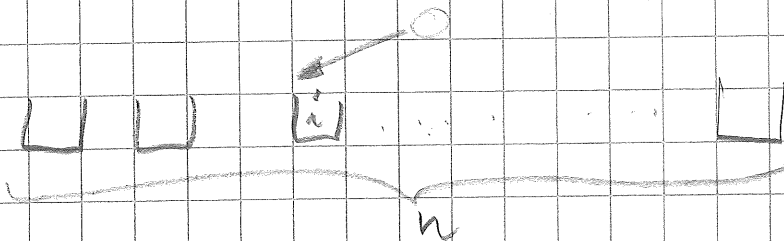
Dado, un  $d \notin S$  y un  $\epsilon \in [0, 1]$

¿de qué tomamos elegimos  $n$  tal que:

$$P(B[h(d)] = 1) < \epsilon \quad ?$$

Sea  $d \notin S$  y suponemos que  $h(d)$  es un proceso estocástico. Por estimar:

$$P(B[h] = 1) < \epsilon$$



¿cuál es la probabilidad que lo  $i$ -ésimo objeto sea 0?

$$P(i\text{-ésimo objeto sea } 0) = \left(1 - \frac{1}{n}\right)$$

¿Cuál es la probabilidad que el  $i$ -ésimo objeto sea 0 después de  $m$  inserciones?

$$P(i\text{-ésimo objeto sea } 0 \text{ después de } m\text{-inserciones}) = \left(1 - \frac{1}{n}\right)^m$$

$$P(B[i]=0) = \left(1 - \frac{1}{n}\right)^m$$

Por lo tanto:


$$P(B[i]=1) = 1 - \left(1 - \frac{1}{n}\right)^m \leq \epsilon$$

$$\Rightarrow 1 - \epsilon \leq \left(1 - \frac{1}{n}\right)^m \approx e^{-\frac{m}{n}} \quad \left[ \left(1 - \frac{1}{n}\right)^n \xrightarrow{n \rightarrow \infty} e^{-1} \right]$$

Despejando:  $\ln(1 - \epsilon) \leq -\frac{m}{n}$

$$\frac{m}{n} \leq -\ln(1 - \epsilon) = \ln\left(\frac{1}{1 - \epsilon}\right)$$

$$\therefore \boxed{\frac{m}{-\ln(1 - \epsilon)} \leq n} \Rightarrow \frac{1}{-\ln(1 - \epsilon)} \leq \frac{n}{m} = c$$

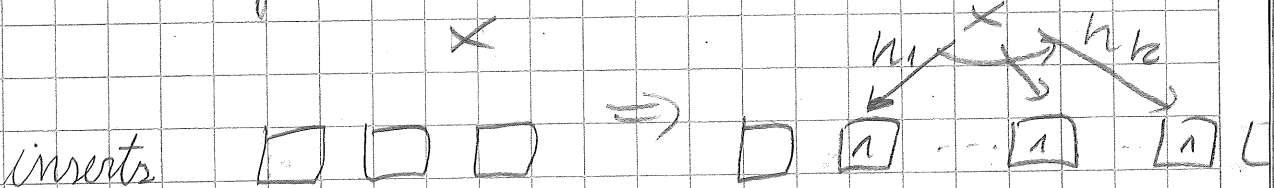
proporción  
de bits extra por  
objeto. 

$$\epsilon = 0.1 \rightarrow n \geq 9.9m$$

$$\epsilon = 0.01 \rightarrow n \geq 99.9m$$

$$\epsilon = 0.001 \rightarrow n \geq 999.9m$$

¿cómo mejorar esta estructura?  $\rightarrow$  bloom filter



Para cualquier objeto  $d$ : ¿cómo verificar si  $d \in S$ ?

- Si  $\exists i \ B(h_i(d)) = 0 \Rightarrow d \notin S \checkmark$

- Si  $\forall i \ B(h_i(d)) = 1 \Rightarrow d \in S \checkmark$

Errores positivos

Dado un dSS y un  $\epsilon \in [0, 1]$  como  
entero  $n$  y  $k$  tal que:

$$P(B[h_1(d)] = 1 \wedge \dots \wedge B[h_k(d)] = 1) < \epsilon$$

Sea  $d \notin S$  y supongamos que  $h_j(d) = i_j$

$i$  cual es la probabilidad que la  $i$ -ésima  
posición sea 0?

$$P(B[i] = 0) = \left(1 - \frac{1}{n}\right)^{k \cdot m} \approx e^{-\frac{km}{n}}$$

Sea  $p = e^{-\frac{km}{n}}$  y suponga que una fracción  
 $p$  de celdas son 0  $\Rightarrow$  la cantidad de celdas  
con 0 es  $p \cdot n$ .  
 $\hookrightarrow$  es una est. exp. de  $n$  con  $p$  como parámetro.

La probabilidad de un falso positivo es

$$P\left(\bigwedge_{j=1}^k B(i_j) = 1\right) \leq \prod_{j=1}^k (1-p) = (1-p)^k$$

~~Suposición de independencia~~  $\rightarrow P(A \cap B) = P(A) P(B|A) \rightarrow$  si  $P(B|A) = P(B)$   
es indep.

Sea  $\delta \leq (1-p)^k$ , queremos minimizar  
según  $k$ .

Suposición es usada así.

\* En esto se encuentra una compensación de fuerzas:

- más  $k$  nos da más chances de encontrar un 0.
- menor  $k$  aumenta la cantidad de 0's en B.

Porque  $\min f$ , es lo mismo que  $\min g$  con  $g = k \cdot \ln(1 - e^{-\frac{k \cdot m}{n}})$ , entonces

$$\begin{aligned} \frac{dg}{dk} &= \ln(1 - e^{-\frac{k \cdot m}{n}}) + k \cdot \frac{1}{1 - e^{-\frac{k \cdot m}{n}}} \cdot (-e^{-\frac{k \cdot m}{n}}) \cdot \left(-\frac{m}{n}\right) \\ &= \ln(1 - e^{-\frac{k \cdot m}{n}}) + \frac{k \cdot m}{n} \cdot \frac{e^{-\frac{k \cdot m}{n}}}{1 - e^{-\frac{k \cdot m}{n}}} \end{aligned}$$

Si consideramos  $x = \frac{k \cdot m}{n}$  es fácil ver que:

$$G(x) = \ln(1 - e^{-x}) + x \cdot \frac{e^{-x}}{1 - e^{-x}}$$

se hace 0 en  $\ln \mathbb{R}$  y es un mínimo global.

$$\Rightarrow \boxed{k = \ln 2 \cdot \frac{n}{m}}$$

$$\begin{aligned}
 \Rightarrow f\left(\ln 2 \cdot \frac{n}{m}\right) &= \left(1 - e^{-\ln 2 \cdot \frac{n}{m} \cdot \frac{m}{n}}\right)^{\ln 2 \cdot \frac{n}{m}} \\
 &= \left(1 - \frac{1}{2}\right)^{\ln 2 \cdot \frac{n}{m}} = \left(\frac{1}{2}\right)^{\ln(2) \cdot \frac{n}{m}} \\
 &= 0.6185 \frac{n}{m}
 \end{aligned}$$

$$\text{Si } n = c \cdot m \Rightarrow$$

$$0.6185^c \leq \varepsilon$$

$$c \geq \frac{\ln(\varepsilon)}{\ln(0.6185)}$$

$$\varepsilon = 0.1 \Rightarrow c = 4.79$$

$$\varepsilon = 0.01 \Rightarrow c = 9.58$$

$$\varepsilon = 0.001 \Rightarrow c = 14.37$$

$$\varepsilon = 0.0001 \Rightarrow c = 19.1699$$