FECHA: / /

Bloom Filters

Suponga una BD $S = \{s_1, \ldots, s_m\} \subseteq D$

tal que $|D| \leq n$ y $D = \{d_1, \ldots, d_n\}$

Para guardar S podemos usar un bitmap

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

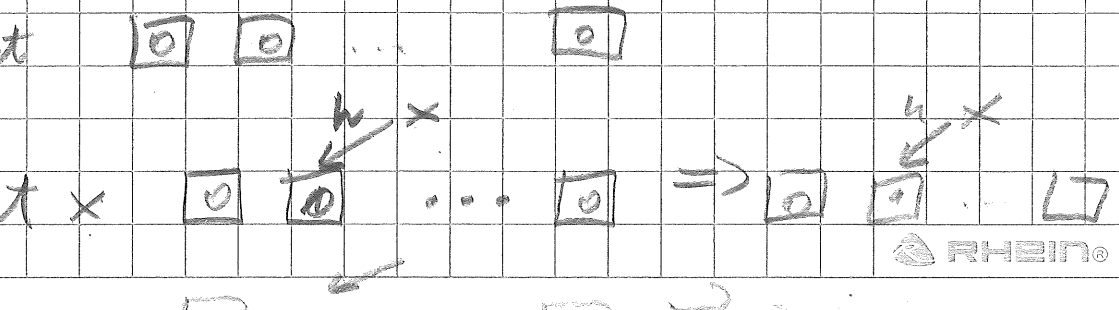
Dado cualquier objeto $d \in D$:

- ¿Cómo veo si $d \in S$?
- ¿Cómo inserto d en S?

¿Cuál es el problema de esta estructura B?

Ejemplos: - D son documentos
- " son palabras o numeros de 32 bits
- etc.

Solución: usar un bitmap B de tamaño n con $n < |D|$ y una función de hash $h: D \to \{1, \ldots, n\}$



FECHA: / /

Para cualquier objeto $d \in D$:

¿Cómo verificamos si $d \in S$?

- Si $B[h(d)] = 0 \Rightarrow d \notin S$ ✓
- Si $B[h(d)] = 1 \longrightarrow d \in S$ ✓

tenemos falsos positivos $\longleftarrow d \notin S$ ✓

Dados un $d \notin S$ y un $\varepsilon \in [0, 1]$

¿de qué tamaño escogemos n tal que:

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

Sea $d \notin S$ y suponemos que $h(d)$ es un número aleatorio. Por ejemplo:

$$P(B[i] = 1) < \varepsilon$$

¿Cuál es la probabilidad que la i-ésima celda sea 0?

RHEIN®

FECHA: / /

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

¿Cuál es la probabilidad que la i-ésima celda sea 0 después de m inserciones?

$$P\left(\begin{array}{l}\text{i-ésima celda sea 0}\\ \text{después de m-inserciones}\end{array}\right) = \left(1 - \frac{1}{n}\right)^m$$

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

Por lo tanto:

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

$$\Rightarrow \quad 1 - \varepsilon \leq \left(1 - \frac{1}{n}\right)^m \simeq e^{-\frac{m}{n}} \qquad \left[\left(1 - \frac{1}{n}\right)^n \underset{n\to\infty}{\to} e^{-1}\right]$$

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

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

$$\therefore \left[\frac{m}{-\ln(1-\varepsilon)} \leq n\right] \Rightarrow \frac{1}{-\ln(1-\varepsilon)} \leq \frac{n}{m} = c$$

proporción de bits extra por objeto

$$-\ln(1-\varepsilon)$$

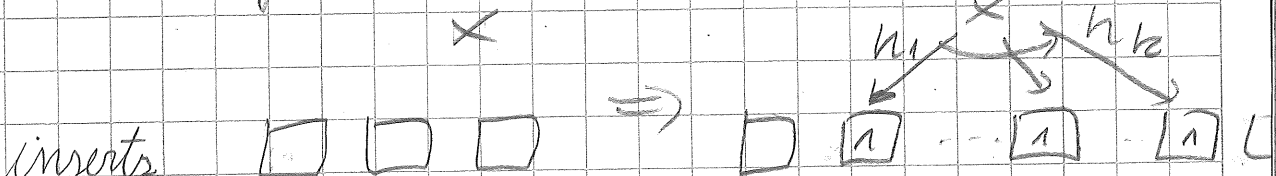
FECHA: / /

$\varepsilon = 0.1 \longrightarrow n \geq 9.9$ hm

$\varepsilon = 0.01 \longrightarrow n \geq 99.9$ m

$\varepsilon = 0.001 \longrightarrow n \geq 999.9$ m

¿Cómo mejoran esta estructura? → Open fallo



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

- Si $\exists i \; B(h_i(d)) = 0 \Rightarrow d \notin S$ ✓
- Si $\forall i \; B(h_i(d)) = 1 \Rightarrow d \in S$ ✓

Falsos positivos ← $d \notin S$

RHEIN®

FECHA: / /

Dado un $d \notin S$ y un $\varepsilon \in [0,1]$ ¿cómo escoger n y k tal que:

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

Sea $d \notin S$ y suponemos que $h_i(d) = j_i$.

¿Cuál 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 supongo que una fracción p de celdas son $0 \Rightarrow$ la cantidad de celdas con 0 es $p \cdot n$.

La probabilidad de un falso positivo es

$$P\left(\bigwedge_{i=1}^{k} B(j_i) = 1\right) = \prod_{i=1}^{k} (1-p) = (1-p)^k$$

suposición de independencia → ¿es verdad?

Sea $f(k) = (1-p)^k$, queremos minimizar según k.

FECHA: / /

En esto se encuentra una compensación de fuerzas:

- mas k nos da mas chances de encontrar un 0.
- menos k aumenta la cantidad de 0's en B.

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

$$\frac{d\,g}{dk} = \ln\left(1 - e^{-\frac{k \cdot m}{n}}\right) + k \cdot \frac{1}{1 - e^{-\frac{k \cdot m}{n}}} \cdot \left(-e^{-\frac{k \cdot m}{n}}\right) \cdot \left(-\frac{m}{n}\right)$$

$$= \ln\left(1 - e^{-\frac{k \cdot m}{n}}\right) + \frac{k \cdot m}{n} \cdot \frac{e^{-\frac{k \cdot m}{n}}}{1 - e^{-\frac{k \cdot m}{n}}}$$

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

$$f_0(x) = \ln\left(1 - e^{-x}\right) + x \cdot \frac{e^{-x}}{1 - e^{-x}}$$

se hace 0 en $\ln 2$ y es un mínimo global.

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

FECHA: / /

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

$$= \left(1 - \frac{1}{2}\right)^{\ln 2 \cdot \frac{n}{m}} = \left(\left(\frac{1}{2}\right)^{\ln(2)}\right)^{\frac{n}{m}}$$

$$= 0.6185^{\frac{n}{m}}$$

Si $n = c \cdot m \Rightarrow$

$$0.6185^{c} \leq \varepsilon$$

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

$\varepsilon = 0.1 \Rightarrow c = 4.79$

$\varepsilon = 0.01 \Rightarrow c = 9.58$

$\varepsilon = 0.001 \Rightarrow c = 14.37$

$\varepsilon = 0.0001 \Rightarrow c = 19.1699$

FECHA: / /

Comentario 1: Si consideramos f como una función de p, y

$$p = e^{-k \cdot \frac{m}{n}},$$

$$\Rightarrow \ln p = -k \cdot \frac{m}{n} \Rightarrow k = -\ln(p) \cdot \frac{n}{m}$$

Se tiene que:

$$\begin{aligned} f &= (1-p)^k \\ &= (1-p)^{-\ln p \,(m/n)} \\ &= \left(\left(e^{\ln(1-p)}\right)^{-\ln p}\right)^{m/n} \\ &= \left(e^{-\ln p \cdot \ln(1-p)}\right)^{m/n} \end{aligned}$$

Por la simetría de f este se minimiza cuando:

$$p = 1/2$$

∴ Cuando $k = \ln(p) \cdot \frac{n}{m}$, el bitmap B se ve como un string aleatorio

FECHA: / /

Comentario 2: Si se tiene $S_1 = \{s_1, \ldots s_m\}$ y $S_2 = \{s'_1, \ldots, s'_n\}$ con bitmaps B_1 y B_2 (mismo largo).

¿Cómo podemos construir el bitmap de $S_1 \cup S_2$?