

Análise

Certo.

Mas essa ideia só tem utilidade se a probabilidade de falso positivo for bem pequena.

E é isso o que nós vamos verificar agora.

Como usual, nós vamos assumir que as funções de dispersão tem comportamento aleatório.

Daí que, a probabilidade de que um certo i seja selecionado pela função de dispersão $h_j(.)$ no momento de uma inserção é de $\frac{1}{m}$.

A probabilidade de que ele não seja selecionado é de $1 - \frac{1}{m}$.

E a probabilidade de que nenhuma função selecione o bit i é de

$$\left(1 - \frac{1}{m}\right)^k$$

A esperteza nesse ponto é reescrever a expressão assim

$$\left(1 - \frac{1}{m}\right)^k = \left[\left(1 - \frac{1}{m}\right)^m\right]^{k/m} \simeq e^{-k/m}$$

Legal.

Agora imagine que existem n elementos na base de dados.

Então, a probabilidade de que o bit i ainda mantenha o valor 0 após as inserções é de

$$\left(1 - \frac{1}{m}\right)^{kn} \simeq e^{-kn/m}$$

E a probabilidade de que ele tenha o valor 1 é de

$$1 - e^{-kn/m}$$

Certo.

Agora nós já estamos prontos para calcular a probabilidade de um falso positivo.

Quer dizer, imagine que z é um elemento que não está na base de dados.

Aplicando as k funções de dispersão a esse elemento, nós obtemos as posições

$$i_1 \leftarrow h_1(z), \quad i_2 \leftarrow h_2(z), \quad \dots \quad i_k \leftarrow h_k(z),$$

O falso positivo acontece quando todas essas posições tem o valor 1.

E a probabilidade de que isso ocorra é de

$$\varepsilon = \left(1 - e^{-kn/m}\right)^k$$

Bom, aqui nós podemos fazer algumas observações simples

1. A probabilidade de falso positivo ε diminui quando m aumenta
— *(o que faz sentido)*
2. A probabilidade de falso positivo ε aumenta quando n aumenta
— *(o que também faz sentido)*

Mas, o comportamento da probabilidade ε com relação a k não é tão evidente.

E a sua análise não é tão fácil assim ...

O fato é que, a medida que k aumenta

- a probabilidade ε primeiro diminui
- e depois ela volta a aumentar ...

E o valor mínimo da probabilidade ε é atingido quando

$$k = \frac{m}{n} \ln 2$$

Quer dizer, esse é o número de funções de dispersão que nós devemos utilizar para um certo número de elementos n e um certo tamanho do filtro m .

Para essa escolha de k , a probabilidade de falso positivo é dada por

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

o que simplifica para

$$\ln \varepsilon = -\frac{m}{n} (\ln 2)^2$$

o que resulta em

$$m = -\frac{n \ln \varepsilon}{(\ln 2)^2}$$

Portanto, o número ótimo de bits por elementos (para uma certa probabilidade de falso positivo ε) é

$$\frac{m}{n} = \frac{\ln \varepsilon}{(\ln 2)^2} \simeq -1,44 \ln \varepsilon$$