Linear counting

Dario Maio

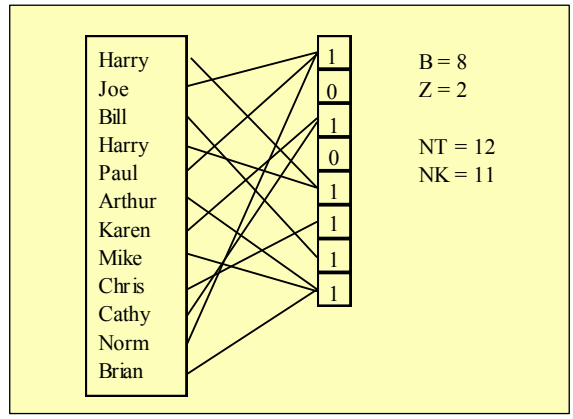
http://bias.csr.unibo.it/maio/

Il metodo "Linear counting"

- In fase di ottimizzazione si rende spesso necessario conoscere il numero di valori distinti di un attributo. Se l'attributo non è indiciato, una stima esatta (ma costosa) si può ottenere attraverso sort o hash dei valori.
- L'algoritmo di Linear counting (Whang, Vander-Zanden, Taylor 1990) è un metodo probabilistico più economico, in termini sia di tempo sia di spazio, e fornisce una buona stima del valore cercato.
- Linear counting fa uso di una **bit-map** di dimensione B e di una **funzione hash** H a valori in [0,B-1].
- La bit-map viene inizializzata ponendo tutti i bit a 0.
- Se per un valore di chiave k si ha H(k) = i, si pone a 1 l'i-mo bit.
- Dopo aver elaborato in questo modo tutte le tuple, si conta il numero, Z, di 0 presenti nella bit-map.

Esempio

- La figura mostra il caso in cui si hanno NT = 12 tuple e NK = 11 valori distinti per l'attributo in esame. La bit-map ha dimensione B = 8 e, dopo l'hashing di tutti i valori, contiene Z = 2 bit a zero.



Analisi del metodo

- Date NT tuple e NK valori di chiave, e allocando casualmente le tuple su B "bucket" sulla base del valore di chiave, il numero medio di bucket non vuoti è pari a:

$$B - Z = B \times \left(1 - \left(1 - \frac{1}{B}\right)^{NK}\right) \quad \text{ovvero} \quad Z = B \times \left(1 - \frac{1}{B}\right)^{NK}$$

- Ricordando che per B e NK "grandi" vale l'approssimazione:

$$\left(1 - \frac{1}{B}\right)^{NK} \approx e^{-\frac{NK}{B}}$$

- si ottiene:

$$Z = B \times e^{-\frac{NK}{B}}$$

- Il valore stimato di NK, indicato con NK^e, è allora : $NK^e = -B \times \ln\left(\frac{Z}{B}\right)$

Considerazioni

- Data una bit-map di B = 10^4 bit, se l'algoritmo ne pone la metà a 1 (Z = 5000), si ottiene $NK^e \approx 6932$.
- È evidentemente necessario evitare che la bit-map "saturi", ovvero che sia Z = 0. A tale proposito l'analisi probabilistica fornisce :

$$B \geq 0.2 \times NK$$

come indicazione pratica per il dimensionamento della bit-map, che si applica nel caso di NK "grande", dell'ordine almeno delle migliaia.

- Un approccio "conservativo" dimensiona la bit-map usando direttamente NT in luogo di NK nella disuguaglianza, Un approccio "aggressivo" usa un'approssimazione scadente di NK, come primo tentativo, e ripete il procedimento se la bit-map satura.