FORMULE DI BASI DI DATI

22 giugno 2021

FORMULE SUI B-TREE

DEFINIZIONE DI B-TREE :
Un B-TREE è un albero bilanciato, organizzato a nodi, ogni nodo corrisponde ad un blocco dati di uno storage device

IMPORTANTE :
T(g,h) è un albero bilanciato di ordine g e altezza h. Le cardinalità/configurazioni possibili sono rappresentate nella tabella sottostante. Ricordiamo che : |sk| = numero di chiavi selezionate per nodo(nodo != radice).

Numero Massimo e minimo di chiavi per ogni nodo o radice:

CHIAVI	MIN	MAX
Radice	1	2g
Nodo	g	2g

Numero Massimo e minimo di figli per ogni nodo o radice:

FIGLI	MIN	MAX
Radice	0	2g+1
Nodo	|sk + 1|	|sk + 1|

FORMULE :

NUMERO MINIMO DI NODI(IPmin)

$$\mathrm{IP}min = 1 + 2 \cdot \sum_{i=0}^{h-2} (g+1)^i$$

NUMERO MASSIMO DI NODI(IPmax)

$$\mathrm{IP}max = \sum_{i=0}^{h-1} (2g+1)^i$$

ALTEZZA DI UN B-TREE

NK = *Numero di chiavi del B-TREE*
NK$min = 1 + g(IPmin - 1) = 2(g+1)^{h-1} - 1$
NK$max = 2g(IPmax) = (2g+1)^h - 1$
h$min = \log_{2g+1}(NK+1)$
h$max = 1 + \log_{g+1}(\frac{NK+1}{2})$

$$\text{h}min \leq \text{h} \leq \text{h}max$$

COSTO INSERIMENTO/ELIMINAZIONE DI UN NODO DA UN B-TREE(g)

CASO :	MIGLIORE	PEGGIORE	MEDIA
Inserimento	$h+1$	$3h+1$	$h+1+2/g$
Eliminazione	$h+1$	$3h$	$5h+5+3/g$

STIMA VARIABILITA' DELL'ORDINE DI UN B-TREE

k = chiave
p = *RID* = Record Identifier, ovvero puntatore a record
q = *PID* = Puntatore al nodo figlio
D = page size

$2g * len(k) + 2g * len(p) + (2g+1) * len(q) \leq D$

$$\text{g} = \frac{D - len(q)}{2(len(k) + len(p) + len(q))}$$

FORMULE SUI B+-TREE

DEFINIZIONE DI B+-TREE :
Un B+-TREE è un B-TREE in cui i record pointer sono memorizzati solo nei nodi foglia dell'albero. La struttura dei nodi foglia differisce quindi da quella dei nodi interni. Ha prestazioni migliori del B-TREE sulla ricerca sequenziale, peggiori nella ricerca per singolo valore.

ORDINE

k = chiave
q = PID = Puntatore al nodo figlio
D = page size

$2g * len(k) + (2g + 1) * len(q) \leq D$

$$\text{g} = \frac{D - len(q)}{2(len(k) + len(q))}$$

NUMERO DI FOGLIE

NR = Numero di Record
NL = Numero di foglie
u = % di utilizzo di un singolo nodo (in media è il 69%)
d = dimensione dei nodi

$$\text{NL} = \frac{NR \cdot (len(k) \cdot len(q))}{d \cdot u}$$

ORDINE DELLE FOGLIE

D = page size

$2g_{leaf} * len(k) + 2g_{leaf} * len(p) + (2g_{leaf} + 1) * len(q) \leq D$

$$g_{leaf} = \frac{D - len(q)}{2 \cdot (len(k) + len(p))}$$

Con questo risultato si è in grado di avere una stima più accurata del numero di foglie di un B-+TREE

NUMERO DI FOGLIE (stima più accurata sfruttando g_{leaf})

$$\mathrm{NL} = \frac{NR}{2g_{leaf} \cdot u}$$

ALTEZZA

Costruendo un albero in cui ciascun nodo ha il numero massimo di figli, si minimizza l'altezza del B+-TREE

$(2g + 1)^{h-1} \geq NL$
$h_{min} = 1 + \log_{2g+1} NL$

Similmente, sfruttando lo stesso ragionamento, costruendo un albero con il numero minimo possibile di figli per nodo, si massimizza l'altezza

$2(g + 1)^{h-2} \leq NL$
$h_{max} = 1 + \log_{2g+1} NL$

$$\mathrm{h}min \leq \mathrm{h} \leq \mathrm{h}max$$

RICERCA DI VALORI

Supponendo di avere un B-+TREE con NK chiavi in NL foglie, si effettua una ricerca sequenziale di k chiavi nell'intervallo di valori compresi fra $[k_{low}, k_{high}]$

EK = Numero di chiavi all'interno dell'intervallo $[k_{low}, k_{high}]$
EL = expected leafs, ovvero è una stima del numero di foglie alla quale si deve accedere durante la ricerca

$$EL = \frac{EK \cdot NK}{NL} \text{ (proporzione -> NK : NL = EK : EL)}$$

$$\text{COSTO(Ricerca)} = 1 - h + EL$$

NUMERO DI FOGLIE DI UN SECONDARY B+-TREE

$$NL = \frac{NK \cdot len(k) + NR \cdot len(p)}{D * u}$$

ALTEZZA DI UN SECONDARY B+-TREE

$$h_{min} = 1 + \log_{2g+1} min(NL, NK)$$

$$h_{max} = 2 + \log_{g+1} \frac{min(NL, NK)}{2}$$

$$\text{h}min \leq \text{h} \leq \text{h}max$$

MODELLO DI CARDENAS

Il modello di Cardenas è utile per stimare il numero medio di pagine NP che contengono almeno uno degli ER(expected records) presi in considerazione.

Considerando i seguenti eventi :

A = "Una pagina contiene 1 degli ER record"
$\overline{\text{A}}$ = "Una pagina **non** contiene 1 degli ER record"
B = "Una pagina **non** contiene **nessuno** degli ER record "
$\overline{\text{B}}$ = "Una pagina contiene almeno 1 degli ER record"
Definiamo :

$$P(A) = \frac{1}{NP} \qquad P(\overline{\text{A}}) = 1 - \frac{1}{NP}$$

$$P(B) = P(\overline{\text{A}})^{ER} \qquad P(\overline{\text{B}}) = 1 - P(B)$$

FORMULA DI CARDENAS :

$$\phi(ER, NP) = NP \cdot P(\overline{\text{B}})$$

$$= NP * (1 - (1 - \frac{1}{NP})^{ER}) \leq MIN(ER, NP)$$

QUERY PLAN

Gestione del piano di accesso alle query, l'obiettivo e quello di scegliere il metodo più veloce ed efficiente.

Linear counting

Algoritmo che permette di stimare il numero di NK chiavi distinte di un attributo A, con A $\in$ r.
r = una qualsiasi estensione di una relazione R(T).

DATI :

BM = mappa chiave - valore
H = hash function
t_i.A = valore dell'elemento $\in$ A di una determinata tupla.
B = n. di bucket
Z = n. di elementi della bitMap uguali a 0
Nota :
Ogni elemento t_i.A sarà presente all'interno della bit-map BM seguendo la seguente logica:

if (t.A == NULL); then BM[H(t.A)] = 0;
else then BM[H($t_i.A$))] = 1;

FORMULA RISOLUTIVA :

$$NK^e = -B * \ln(\frac{Z}{B})$$

PROCEDIMENTO

Considerando l'evento A = "un bucket contiene ALMENO un valore".
dal modello di cardenas si avrà :

$$P(A) = (1 - (1 - (\frac{1}{B})^{NK}))$$
$$P(\overline{\mathrm{A}}) = (1 - (\frac{1}{B})^{NK})$$

il quale ci permetterà di formulare :

$$B - Z = B \cdot P(A) = \textit{stima del n. di bucket pieni}$$

$$Z = B \cdot P(\overline{\mathrm{A}}) = \textit{stima del n. di bucket vuoti}$$

Successivamente è importante eseguire la seguente approssimazione, applicabile solo nel caso in cui si lavori con valori molto grandi :

$$P(\overline{\mathrm{A}}) \simeq e^{-\frac{NK}{B}} = P(\underbrace{\overline{\mathrm{A}}}_{})$$

di conseguenza Z diventerà :

$$Z = B \cdot P(\underbrace{\overline{\mathrm{A}}}_{}) = B \cdot e^{-\frac{NK}{B}}$$

A questo punto, è possibile estrarre NK, il valore cercato.

$$NK^e = -B * \ln(\frac{Z}{B})$$

l'apice e aggiunto ad NK sta a significare estimated, ovvero il valore stimato, non necessariamente esatto, è stato aggiunto nel risultato per non confonderlo con la costante di Eulero.