

COMPLESSITÀ' QUICKSORT CASO MEDIO

Intuitivamente una funzione che descrive il comportamento medio di quicksort è data da una media di tutte le funzioni che descrivono quicksort. Per sequenze di lunghezza n abbiamo $n!$ possibili input da ordinare. A ognuna di queste $n!$ combinazioni è associato un albero di ricorsione generato da quick sort per l'ordinamento.

Poste queste premesse possiamo dire che quindi il tempo medio è

$$T_H(n) = \frac{1}{n!} \sum_{i=1}^{n!} T_H^i(n)$$

Orviamente noi non sappiamo da quest'espressione T_H^i che valori assume. Per studiare queste T_H^i (che possono essere visti come alberi perché quicksort è ricorsivo) considereremo sequenze in cui non ci sono duplicati. A ciò dobbiamo implementare il concetto di rango di un elemento

RANGO DI UN ELEMENTO

Definiamo **RANGO** di $x \in A$ sequenza, il numero di elementi di A minori o uguali di x .

Si osservi che il rango di x in una sequenza di n numeri assume i valori compresi tra 1 e n .

OSS

Il range di un elemento definisce univocamente la partizione

DIMOSTRAZIONE

Supponendo di partizionare rispetto a x , studiamo vari casi.

● RANGO 1

L'unico partizionamento possibile è quello totale quindi volendo la parte di sinistra ha 1 elemento

● RANGO 2

PARTIZIONA in questo caso scambia il pivot con l'unico elemento minore o uguale di x , chiamiamolo y

x	$>$	y	$>$
$i \uparrow$		j	

y si troverà nella partizione di sinistra, x in quella di destra, i e j si incroceranno e quindi avremo 1 elemento nella parte di sinistra (y) e i restanti a destra

● RANGO 3 (generalmente > 2)

PARTIZIONA agisce sempre allo stesso modo ovviamente

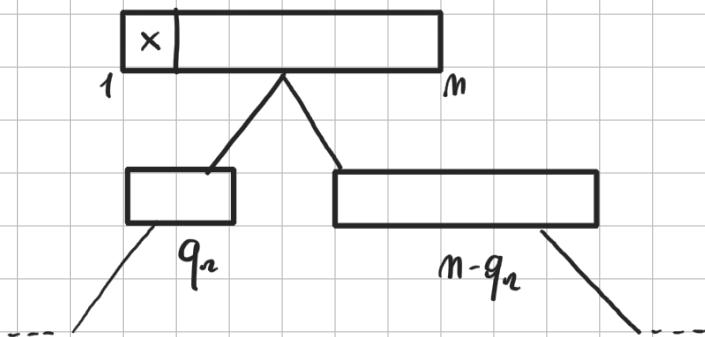
x	$>$	y	$>$	z
$i \uparrow$		$j \leftarrow j$		$z, y \leq x$

i e j vengono scambiati, i incrementa di 1 e si ferma, j decrementa fino a trovare y che combinerà con l'elemento adiacente a x , dopodiché si incroceranno gli indici. Abbiamo a sinistra 2 elementi.

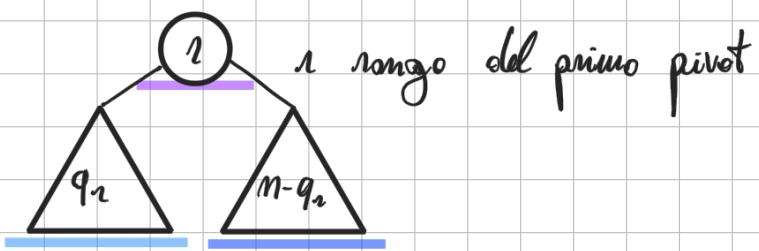
Questo discorso si può estendere per infiniti numeri, per come è definito l'algoritmo di partizionamento se il rango di una sequenza è r , la partizione di sinistra sarà sempre lunga 2^{r-1} che chiameremo q_r (fatta eccezione per $r=1$), quindi abbiamo implicitamente una funzione che definisce il partizionamento dato r

ALBERI "GENERATI" DAL RANGO

Data una sequenza di n elementi, sia r il rango di x avremo la seguente partizione



Questo significa che abbiamo, tramite partizione e le sue chiamate ricorsive, costituito un albero del seguente tipo



Secondo questo principio, avendo il rango sulla prima chiamata, suddivideremo le $m!$ possibili combinazioni (partizioni) in classi di alberi.

La prima classe contiene le sequenze in cui il primo pivot ha rango 1 (quindi genera un albero che ha a sinistra un nodo e a destra $n-1$) e così via. Questi alberi nella classe avranno in comune **SOLO** il primo partizionamento. Se calcoliamo la media di tempo di ogni classe e poi la media tra tutte le classi, avremo proprio il tempo medio totale:

$$T_H(m) = \frac{1}{m} \sum_{r=1}^m T_n^{(r)}(m)$$

m classi perche' $1 \leq r \leq m$

T_H PER CLASSI $r \geq 2$

Nel caso in cui il rango è maggiore di 2 avremo una partizione in cui i sott'alberi hanno rispettivamente q_r e $m-q_r$ nodi

Quindi l'espressione sarà:

$$T_H(n) = T_H(q_1) + T_H(n-q_1) + \Theta_m$$

↓ partizione 1x ↑ partizione dx ← tempo della radice, ovvero partizione su m

Noi però non conosciamo i ranghi delle sottosequenze, quindi ci riportiamo al caso generale.

Si noti che l'espressione generale è definita come

$$T_H(n) = \frac{1}{m} \sum_{i=1}^m T_H^i(n)$$

mentre il tempo medio per classe è

$$T_H(n) = T_H(q_1) + T_H(n-q_1) + \Theta_m$$

Possiamo notare come abbiamo definito due funzioni mutuamente ricorsive (ben fondate comunque perché le chiamate sono fatte su input strettamente minore, arrivando al caso base prima o poi.)

Sostituendo $T_H^i(n)$ in $T_H(n)$ avremo:

$$T_H(n) = \frac{1}{m} \sum_{i=1}^m (T_H(q_1) + T_H(n-q_1) + \Theta(m))$$

(è s'inteso che sia un'equazione di ricorrenza in cui il caso base è per $n=1$)

FORMA CHIUSA DI T_H

Dimostriamo col secondo principio d'induzione in cui il caso base

$$\text{e } r=1 \text{ e } P(m) = T_H(m) = \frac{1}{m} \sum_{i=1}^m (T_H(q_i) + T_H(m-q_i) + \Theta(m))$$

Isoliamo il caso $r=1$ e riscriviamo l'equazione

$$T_H = \frac{1}{m} \left(T_H(q_1) + T_H(m-q_1) + \Theta(m) + \sum_{i=2}^m (T_H(q_i) + T_H(m-q_i) + \Theta_m) \right)$$

Si noti che $q_1=1$ e $q_2=r-1$, avremo

$$T_H = \frac{1}{m} \left(T_H(1) + T_H(m-1) + \Theta(m) + \sum_{i=2}^m (T_H(i-1) + T_H(m-(i-1)) + \Theta_m) \right)$$

\downarrow \downarrow \downarrow
 $\Theta(1)$ $O(m^2)$ Θ_m
→ ↔ ↓
 $\frac{1}{m} \cdot O(m^2) = O(m)$

(il caso peggiore è $O(m^2)$)

$(r-1)$ varia tra 1 e $m-1$, sostituiamo gli indici in modo equivalente

$$T_H = O(m) + \frac{1}{m} \left(\sum_{q=1}^{m-1} (T_H(q) + T(m-q) + \underline{\Theta(m)}) \right)$$

$\Theta(m)$ è una costante che possiamo portare fuori diventando $\Theta(m^2) \cdot \frac{1}{m} = \Theta(m)$

Averemo $O(m) + \Theta(m) + \frac{1}{m} \sum_{q=1}^{m-1} \dots$, $\Theta(m)$ ha "la meglio" su $O(m)$

perché $O(m)$ non può essere più di m mentre $\Theta(m)$ è m .

Dopo questi calcoli avremo:

$$T_H(n) = \Theta(n) + \frac{1}{m} \sum_{q=1}^{m-1} (T_H(q) + T_H(m-q))$$

Si noti che $\sum_{q=1}^{m-1} T_H(q)$ e $\sum_{q=1}^{m-1} T_H(m-q)$ (i due addendi della sommatoria) generano in totale le stesse sequenze in ordine inverso quindi dandoli sommare, ai fini della sommatoria, possiamo dire che

$$T_H(n) = \Theta(n) + \frac{1}{m} \sum_{q=1}^{m-1} 2T_H(q) \rightarrow T_H(n) = \Theta(n) + \frac{2}{m} \sum_{q=1}^{m-1} T_H(q)$$

Non vogliamo dimostrare che $T_H = \Theta(n \lg_2 n)$ poiché abbiamo già visto che nel caso migliore $T_{BS} = \Theta(n \lg_2 n)$, se il caso medio è limitato superiormente da $n \lg_2 n$ allora il caso medio sarà proprio $\Theta(n \lg_2 n)$

DIMOSTRAZIONE

Dimostriamo che $\exists c > 0 | \forall n \geq 2 \quad T_H(n) \leq c \cdot n \lg_2 n$

per induzione con base d'induzione $n=2$.

• $n=2$

$$T_H(2) = \Theta(1) + \frac{2}{2} \sum_{q=1}^1 T_H(q) = \Theta(1) + \Theta(1)$$

Questi $\Theta(1)$ che chiameremo a e b sono diversi tra loro poiché a è originata dai costi di confronto mentre b è

originata dalle diverse di partizione, quindi non le sommiamo
tra di loro. Possiamo però dire che:

$$T_H(m) \leq c \cdot n \lg_2 m \rightarrow a+b \leq c \cdot 2 \rightarrow \frac{a+b}{2} \leq c$$

Ci basterà scegliere una c positiva che rispetta la condizione per soddisfare sul caso base, a noi però serve una c unica per tutti i casi, quindi proseguiamo per vedere come individuare c

- $n > 2$

Ricordiamo che per ipotesi $(\forall q < m P(q)) \Rightarrow P(m)$ ovvero

$$\forall q < m T_H(q) \leq c \cdot q \lg_2 q$$

Questo lo possiamo sostituire nell'espressione generale per ottenere

$$T_H(m) = \Theta(m) + \frac{2}{m} \sum_{q=1}^{m-1} T_H(q) \leq \Theta(m) + \frac{2}{m} \sum_{q=1}^{m-1} (cq \lg_2 q)$$

Il problema attuale è la sommatoria. Non ci serve un'uguaglianza bensì ci basta un'approssimazione per eccesso che sia maggiore di $c \cdot q \lg_2 q$ (portiamo c fuori dalla sommatoria).

L'approssimazione in particolare è:

$$\sum_{q=1}^{m-1} q \lg_2 q \leq \frac{m^2}{2} \lg_2 m - \frac{m^2}{8}$$

Completiamo la dimostrazione grazie a quest'approximazione e
successivamente dimostriamo la validità di questa.

Raccogliamo c fuori la sommatoria e sostituiamo la relazione resta
valida per monotonicità:

$$T_H(n) \leq \Theta(n) + \frac{2c}{n} \left(\frac{n^2}{2} \lg_2 n - \frac{n^2}{8} \right) = \Theta(n) + cn \lg_2 n - \frac{cn}{4}$$

A patto di scegliere $\frac{cn}{4} > \Theta(n)$ avremo concluso perché il nostro termine,
dato alla sottrazione, sarà poco più piccolo di $cn \lg_2 n$ e per
transitività varrà proprio ciò che vorremo dimostrare.

Questo $\Theta(n)$ è principalmente il costo del partizionamento.

Questo costo è assimilabile a $n \cdot a$ con a costante (vedi sopra) perché
è lineare su n . Quindi l'espressione di $T_H(n)$ è

$$T_H(n) \leq an + cn \lg_2 n - \frac{cn}{4}$$

Quindi come detto prima se $an - \frac{cn}{4} \leq 0$ allora potremo dire
che $T_H(n) = \Theta(n \lg_2 n)$ (perché $\leq cn \lg_2 n$).

Ha allora semplicemente $an - \frac{cn}{4} \leq 0 \rightarrow c \geq 4a$

Se $c \geq 4a$ allora $an - \frac{cn}{4} \leq 0$ e per transitività ovvero

$$T_1(m) \leq am + cm\lg_2 m - \frac{cm}{4} \leq cm\lg_2 m \rightarrow T(m) \leq cm\lg_2 m$$

la costante c che daremo scegliere deve rispettare 2 vincoli ovvero
 $c \geq \frac{a+b}{2}$ e $c \geq 4a$, ci basterà prendere la più grande tra le
due (che sicuramente esisterà)

Dimostriamo l'approssimazione

DIMOSTRAZIONE

Dimostriamo che $\sum_{q=1}^{m-1} q \lg_2 q \leq \frac{m^2}{2} \lg_2 m - \frac{m^2}{8}$

Studiamo la sommatoria. La forma chiusa è difficile da trovare però
possiamo trovare facilmente una quantità più grande sostituendo a
 $\lg_2 q \rightarrow \lg_2 m$ (q varia tra 1 e $m-1$ quindi sarà sicuramente minore)

$$\sum_{q=1}^{m-1} q \lg_2 q \leq \sum_{q=1}^{m-1} q \lg_2 m = \lg_2 m \sum_{q=1}^{m-1} q = \lg_2 m \left(\frac{m(m-1)}{2} \right) = \frac{m^2 \lg_2 m}{2} - \frac{m}{2} \quad \times$$

Quell' $\frac{m}{2}$ è quello che nella dimostrazione precedente aveva il
compito di "mangiare" b_m ($\frac{m^2}{8}$ che veniva moltiplicato per $\frac{2}{m}$ e poi
sottratto a b_m) ma in questa forma è troppo piccolo (perché
approssima troppo poco).

Per renderla più precisa usiamo la stessa tecnica ovvero approssimare alcuni fattori con $\lg_2 m$ e altri con qualcosa di più piccolo:

$$\sum_{q=1}^{m-1} q \lg_2 q = \sum_{q=1}^{\lceil \frac{m}{2} \rceil - 1} q \lg_2 q + \sum_{q=\lceil \frac{m}{2} \rceil}^{m-1} q \lg_2 q \quad \text{abbiamo spezzato la sommatoria}$$

Approssimo nuovamente il secondo termine con qualcosa di maggiore ($q \lg_2 m$) e il primo termine sempre qualcosa di maggiore ma più preciso

$$\sum_{q=1}^{m-1} q \lg_2 q \leq \sum_{q=1}^{\lceil \frac{m}{2} \rceil - 1} q \lg_2 \left(\frac{m}{2} \right) + \sum_{q=\lceil \frac{m}{2} \rceil}^{m-1} q \lg_2 m$$

Raccogli $\lg_2 m$ ed $\lg_2 \frac{m}{2}$ ricordando che $\lg_2 \frac{m}{2} = \lg_2 m - 1$

$$\leq (\lg_2 m - 1) \sum_{q=1}^{\lceil \frac{m}{2} \rceil - 1} q + \lg_2 m \sum_{q=\lceil \frac{m}{2} \rceil}^{m-1} q = \lg_2 m \left(\sum_{q=1}^{\lceil \frac{m}{2} \rceil - 1} q + \lg_2 m \sum_{q=\lceil \frac{m}{2} \rceil}^{m-1} q \right) - \sum_{q=1}^{\lceil \frac{m}{2} \rceil - 1} q$$

Le due sommatorie sommate danno $\sum_{q=1}^{m-1} q$, sostituendo

$$\leq \frac{m^2}{2} \lg_2 m - \frac{m}{2} \lg_2 m - \sum_{q=1}^{\lceil \frac{m}{2} \rceil - 1} q$$

Si noti che $\lceil \frac{m}{2} \rceil - 1 \geq \frac{m}{2} - 1$, se combiniamo con l'indice finale della sommatoria avremo una sommatoria in valore assoluto più piccola ma un'espressione più grande (perché la sommatoria

viene sottratta al polinomio), quindi per transitività possiamo dire:

$$\leq \frac{m^2}{2} \lg_2 m - \frac{m}{2} \lg_2 m - \sum_{q=1}^{\frac{m}{2}-1} q = \frac{m^2}{2} \lg_2 m - \frac{m}{2} \lg_2 m - \frac{\frac{m}{2}(\frac{m}{2}-1)}{2} =$$

$$= \frac{m^2}{2} \lg_2 m - \frac{m}{2} \lg_2 m - \frac{m^2}{8} + \frac{m}{4} \leq \frac{m^2}{2} \lg_2 m - \frac{m^2}{8}$$

perché $\frac{m}{4} - \frac{m \lg_2 m}{2}$ è sempre ≤ 0

Quindi abbiamo dimostrato la nostra ipotesi