Genomsnittligt sökdjup i binära sökträd

Detta papper visar att binära sökträd som byggs upp av slumpmässiga data är *bra*.

Beteckningar och definitioner

Definition

Den *interna väglängden* (betecknad IPL eller I) är summan av alla noders väglängder där roten sägs ha väglängden 1, rotens barn väglängderna 2 etc.

För att definiera den *externa* väglängden kompletterar vi trädet med alla ”tomma träd” som vi kallar *externa* noder. Det nya trädet innehåller då n interna och $n+1$ externa noder (trädet är ju *fullt* dvs varje nod är antingen en intern med två barn eller ett löv utan barn).

Den *externa väglängden* (betecknad EPL eller E) definieras som summan av alla *externa* noders väglängder.

Exempel

Figuren visar hur trädets komletteras med externa noder. I trädet är $I = 11$ och $E = 22$

Ett träds väglängd är ett mått på hur ”välförgrenat” det är — för samma antal noder är ett träd med liten väglängd mer välförgrenat än ett träd med stor väglängd.

Det finns ett enkelt samband mellan den interna och den externa väglängden uttryckt i följande formel:

$$E = I + 2n + 1 \tag{1}$$

Detta kan visas med hjälp av induktion (övning!)

Användning av väglängder

Väglängderna är speciellt användbara för att uttrycka det genomsnittliga arbetet för sökning i ett binärt sökträd. Om ett träd med n noder har intern och extern väglängd I respektive E så gäller att det genomsnittliga antalet försök s för att söka en befintlig nyckel (så kallad *lyckad* sökning) är

$$s = \frac{I}{n} \tag{2}$$

Detta följer direkt av definitionen av intern väglängd: att hitta nyckeln lagrad i roten kräver ett försök, nyckelarna i rotens barn två försök etc. (Resonemanget förutsätter att alla nycklar är lika sannolika att bli sökta).

För att hitta motsvarande uttryck för *misslyckad* sökning (dvs att upptäcka att en viss nyckel inte finns i trädet) konstaterar vi att varje misslyckad sökning slutar i en extern nod. För att komma fram till den externa noden gör vi lika många försök som dess bidrag var till den externa väglängden. Om alla externa noder är lika sannolika (tänk efter vad detta egentligen innebär) så blir det genomsnittliga antalet försök u för en misslyckad sökning

$$u = \frac{E}{n+1} \tag{3}$$

(det finns ju $n+1$ externa noder).

Det är således i högsta grad intressant att titta på vad vi kan förvänta oss för väglängder i ett binärt sökträd.

Att den minsta och största möjliga väglängden är $\Theta(n \log n)$ respektive $\Theta(n^2)$ lämnas som (en ganska enkel) övning. Den svårare men intressantare frågan vad vi kan förvänta oss för väglängd i ett slumpmässigt uppbyggt träd ägnar vi nästa avsnitt åt.

Väglängd i träd uppbyggt av slumptal

Antag att vi har n unika nycklar och att ett binärt sökträd byggs upp genom att dessa nycklar successivt läggs in i trädet. Antag vidare att var och en av de $n!$ olika permutationerna av nycklarna är lika sannolik.

Påstående *Den genomsnittliga interna väglängden i de träd som genereras är*

$$1.386 n \lg n + O(n)$$

Bevis

Låt

$E(n)$ beteckna den genomsnittliga externa väglängden och

$I(n)$ beteckna den genomsnittliga interna väglängden

i de träd som skapas av alla möjliga inläggningssekvenser av de n nycklarna.

Då gäller

$$E(n) = I(n) + 2n + 1 \tag{4}$$

(Eftersom ovanstående relation gäller för varje enskilt träd måste den också gälla för medelvärdet)

Antalet nodbesök som görs vid sökning av en viss nyckel är exakt lika många som gjordes när den nyckeln lades in i trädet (om man även räknar den externa, ”tomma” noden.)

Vid det i:te inlägget så görs, i genomsnitt, $E(i-1)/i$ eftersom $i-1$ redan är lagrade och det finns i externa ”nil”-noder.

Vid första inlägget görs $E(0)/1$ besök i genomsnitt, andra inlägget $E(1)/2$ besök i genomsnitt o.s.v.

Eftersom det krävs exakt lika många nodbesök för att finna en nyckel som det behövdes för att lägga in den så erhåller vi sambandet

$$I(n) = E(0)/1 + E(1)/2 + E(2)/3 + ...E(n-1)/n \tag{5}$$

Denna differensekvation involverar både $E(n)$ och $I(n)$. Dessutom har den ”full historia” dvs det n:te värdet beror av alla föregående värden. Standardsättet att bli av med den fulla historien är att sätta upp motsvarande ekvation för $I(n-1)$:

$$I(n-1) = E(0)/1 + E(1)/2 + E(2)/3 + ... + E(n-2)/(n-1) \tag{6}$$

och sedan subtrahera (6) från (5):

$$I(n) - I(n-1) = \frac{E(n-1)}{n} \tag{7}$$

Genom att använda sambandet (4) kan vi bli av med den ena av de två obekanta. Det spelar i princip ingen roll vilken vi eliminerar men det visar sig att det blir lättare att räkna om vi använder $E(n)$. Ekvationen (7) övergår då i

$$\begin{aligned} E(n) - 2n - 1 &= E(n-1) - 2(n-1) - 1 + \frac{E(n-1)}{n} \\ E(n) &= E(n-1) + 2 + \frac{E(n-1)}{n} \\ E(n) &= 2 + \frac{n+1}{n}E(n-1) \end{aligned} \tag{8}$$

(Ekvation (8) kan kontrolleras för n=1, 2 och 3 genom räkna på alla möjliga träd av dessa storlekar.)

Ekvationen (8) är en icke-linjär differensekvation som, generellt sett, kan vara svår att lösa. I just detta fall kan dock lösningen erhållas på vårt vanliga sätt – genom att expandera:

$$\begin{aligned}
E(n) &= 2+\frac{n+1}{n}E(n-1) = \\
&= 2+\frac{n+1}{n}\left(2+\frac{n}{n-1}E(n-2)\right) = \\
&= 2+2\frac{n+1}{n}+\frac{n+1}{n-1}E(n-2) = \\
&= 2+2\frac{n+1}{n}+\frac{n+1}{n-1}\left(2+\frac{n-1}{n-2}E(n-3)\right) = \\
&= 2+2\frac{n+1}{n}+2\frac{n+1}{n-1}+\frac{n+1}{n-2}E(n-3) = \\
&\vdots \\
&= 2+2\frac{n+1}{n}+2\frac{n+1}{n-1}+2\frac{n+1}{n-2}+\cdots+\frac{n+1}{n-k}E(n-k-1) = \{k=n-2\} \\
&= 2+2\frac{n+1}{n}+2\frac{n+1}{n-1}+2\frac{n+1}{n-2}+\cdots+\frac{n+1}{2}E(1) = \\
&= 2+2(n+1)\left(\frac{1}{n}+\frac{1}{n-1}+\cdots\frac{1}{3}+\frac{1}{2}\frac{E(1)}{2}\right)
\end{aligned}$$

Eftersom $E(1)=4$ erhålles

$$\begin{aligned}
E(n) &= 2+2(n+1)\left(\frac{1}{n}+\frac{1}{n-1}+\cdots\frac{1}{3}+1\right) = \\
&= 2+2(n+1)\left(\frac{1}{n}+\frac{1}{n-1}+\cdots\frac{1}{3}+\frac{1}{2}+1-\frac{1}{2}\right) = \\
&= 2+2(n+1)\left(H(n)-\frac{1}{2}\right) = \\
&= 2+2(n+1)H(n)-(n+1) = \\
&= 2(n+1)H(n)-n+1 \qquad (9)
\end{aligned}$$

där $H(n)$ är den harmoniska summan.

$$H(n) = 1+\frac{1}{2}+\frac{1}{3}+\cdots+\frac{1}{n}$$

Observera att detta är ett exakt uttryck för $E(n)$ – vi har inte gjort några uppskattningar alls. Formeln kan således lätt kontrolleras för t.ex. n = 1, 2 och 3.

Genom att använda uppskattningen

$$H(n) = \ln n + \gamma + O(\frac{1}{n}) \tag{10}$$

så erhålles

$$\begin{aligned} E(n) &= 2(n+1)\left(\ln n + \gamma + O(\frac{1}{n})\right) - n + 1 \\ &= 2n\ln n + \ln n + 2(n+1)\left(\gamma + O(\frac{1}{n})\right) - n + 1 \\ &= 2n\ln n + O(n) \\ &= 2n\ln 2 \lg n + O(n) \\ &= 1.386 n \lg n + O(n) \end{aligned} \tag{11}$$

Om sambandet mellan intern och extern väglängd (4) användes så erhålles

$$I(n) = 1.386 n \lg n + O(n)$$

V.S.B

Observera att vi som ett delresultat har visat att lösningen till differensekvationen

$$E(n) = \frac{n+1}{n} E(n-1) + 2$$

är $O(n \log n)$ vilket är ett resultat som kan användas vid analysera quicksort.

Alternativ resonemang

Om man tycker att resonemanget som ledde fram till ekvation (5) är svårt att förstå så kan följande alternativa resonemang föras.

Antag att de nycklar vi har nycklarna $k_1 < k_2 < k_3 < \ldots < k_n$. Om vi vid uppbyggnaden av trädet börja med nyckel k_i kommer den att bli rot i trädet och vi kommer få $i-1$ noder i vänstersubträdet T_L och $n-i$ i högersubträdet T_R. Om väglängden för trädet T_L och T_R betecknas med I_L respektive I_R så blir väglängden för detta träd

$$\begin{aligned} I &= 1 + (i - 1 + I_L) + (n - i + I_R) \\ &= n + I_L + I_R \end{aligned}$$

(vi måste addera 1 för varje nod i subträden eftersom vägarna till dem först måste gå genom roten).

Om vi varje nyckel k_i är lika sannolik att bli vald som rot blir den genomsnittliga väglängden

$$\begin{aligned} I(n) &= \frac{1}{n}\sum_{i=1}^{n}\left(n + I(i-1) + I(n-i)\right) \\ &= n + \frac{2}{n}\sum_{i=0}^{n-1} I(i) \end{aligned}$$

Denna ekvation kan lösas med en liknande teknik som användes ovan eller så kan man utnyttja sambandet (1) mellan intern och extern väglängd och då får man exakt samma formel som (2).