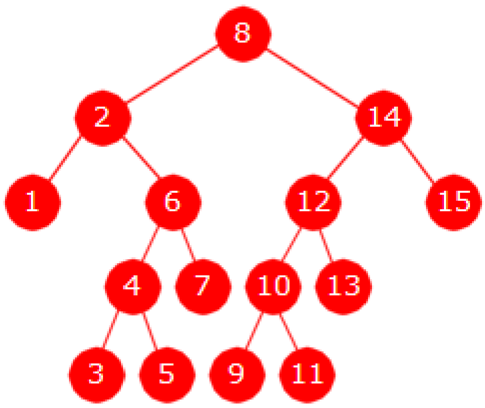
Algoritmer og datastrukturer – Avsnitt 5.2.15 – Algoritmeanalyse

Avsnitt 5.2.15.5 - Gjennomsnittlig avstand mellom to «naboer» i inorden i et binært søketre med *n* forskjellige verdier

Hver permutasjon av tallene fra 1 til n definerer et binært søketre, dvs. det treet vi får ved å legge inn verdiene i den rekkefølgen de har i permutasjonen. Siden det er $n!$ forskjellige permutasjoner, blir det $n!$ binære søketrær. Vi antar at alle har samme sannsynlighet.

La p og q være to noder i et binært søketre slik at p kommer rett etter q i inorden. Vi sier at de to er «naboer» i inorden. Da må enten p ligge i det høyre subtreet til q eller så må q ligge i et venstre subtreet til p. I begge tilfellene kan vi si at det er en bestemt avstand mellom dem. Det er antall kanter på veien fra den ene ned til den andre.



Figur 1 : *Et binært søketre*

I *Figur* 1 ser vi at avstanden mellom nodene 1 og 2 er 1, avstanden mellom 2 og 3 er 3, den mellom 3 og 4 er 1, den mellom 4 og 5 er også 1, den mellom 5 og 6 er 2, osv. Den korteste avstanden mellom to «naboer» er alltid 1. Men den lengste avstanden vil variere fra tre til tre. I treet i *Figur* 1 er det 4 som er størst og det er avstanden mellom 8 og 9.

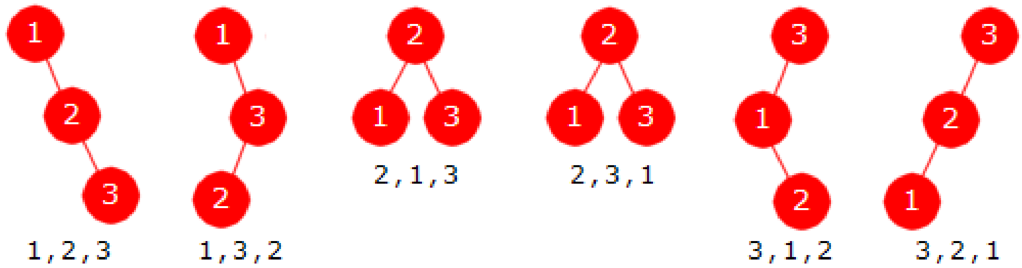
Vi kan finne summen S av alle naboavstandene. For treet i *Figur* 1 får vi:

(1) $S = 1+3+1+1+2+1+3+4+1+1+2+1+2+1 = 24$

La S_n være den gjennomsnittlige avstandssummen for binære søketrær med n forskjellige verdier. I treet i *Figur* 1 er $n = 15$. Spørsmålet er hvor typisk dette treet er. Hvordan vil S_{15} være? Større, lik eller mindre enn 24? Det viser seg at S_{15} = 23,4. Men det er jo det vi skal komme frem til i dette notatet.

La $A_n = \dfrac{S_n}{n-1}$ være den gjennomsnittlige naboavstanden. Treet i *Figur* 1 har en gjennomsnittlig naboavstand på 24/(15 – 1) = 1,71.

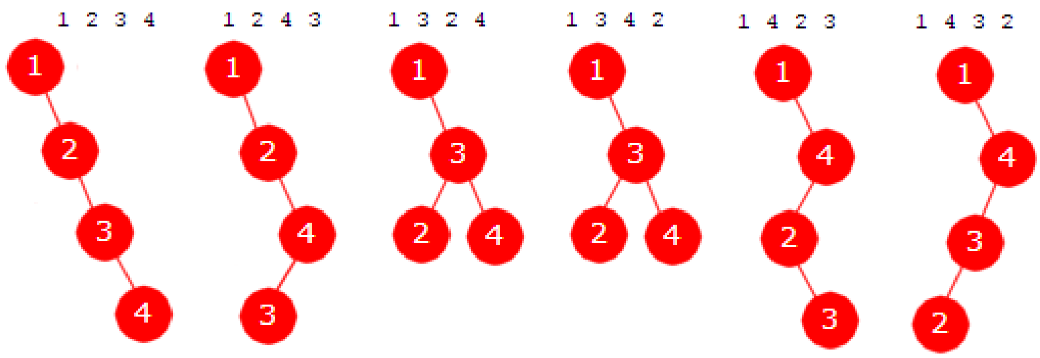
Et binærtre med bare én node har ingen naboer og dermed ingen naboavstander. Men vi sier likevel at $S_1 = 0$. Det er to forskjellige binære søketrær med to verdier, men i begge er det kun ett nabopar og mellom dem er avstanden 1. Dermed blir $S_2 = 1$. For tilfellet $n = 3$ blir det seks trær siden tre tall kan permuteres på seks forskjellige måter:



Figur 2 : *Det er seks trær for* $n = 3$

Avstandssummene for de seks trærne er henholdsvis 2, 3, 2, 2, 3 og 2. Dermed blir $S_3 = \frac{14}{6} = \frac{7}{3} = 2{,}333\ldots$

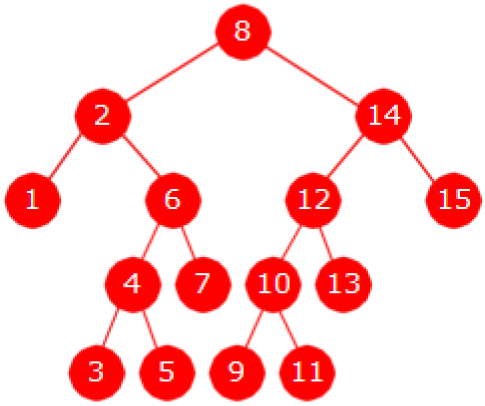
Det kunne også være interessant å finne S_4. Vi skal finne en generell formel for S_n og da er det fint å kunne teste formelen for de kjente tilfellene.



Figur 3 : *Trærne for de seks permutasjonene av* 1 *til* 4 *der* 1 *står først*

For $n = 4$ er det 24 trær. *Figur* 3 viser de seks trærne gitt av de seks permutasjonene av 1 til 4 der 1 står først. For disse er avstandssummene henholdsvis 3, 4, 4, 4, 5 og 5. Til sammen 25. På grunn av symmetrien blir det også 25 for de seks trærne gitt av de seks permutasjonene av 1 til 4 der 4 står først. Tegner en opp de seks trærne gitt ved at 2 står først ser en fort at der blir det sammenlagt 21 og det samme for de der 3 står først. Til sammen 25 + 21 + 21 + 25 = 92. Dermed blir $S_4 = \frac{92}{24} = \frac{23}{6} = 3{,}8333\ldots$.

La T være et binært søketre gitt av en permutasjon av tallene fra 1 til n der tallet k står først. Da vil det være $k-1$ noder i rotnodens venstre subtre og $n-k$ noder i rotnodens høyre subtre. Hvis vi kjenner avstandssummene for de to subtrærne, kan vi da finne avstandssummen for hele treet?



Figur 4 : *Binært søketre med* 15 *verdier*

Permutasjonen 8, 2, 14, 1, 6, 12, 15, 4, 7, 10, 13, 3, 5, 9, 11 er en av dem som gir treet i *Figur* 4. Siden 8 står først blir den rotnode. Det svarer til $n = 15$ og $k = 8$. Avstandssummene for venstre og høyre subtre er 9 og 8. I tillegg må vi ha avstanden fra 7 til 8 og den fra 8 til 9. Generelt betyr det avstanden fra den siste i inorden i det venstre subtreet til rotnoden og avstanden fra rotnoden til den første i inorden i det høyre subtreet. Vi får derfor flg. sum for treet i *Figur* 4:

(2) $$S = 9 + 3 + 4 + 8 = 24$$

Dette kan generaliseres. Den gjennomsnittlige avstanden mellom rotnoden i et binært søketre med til n verdier og noden som kommer først i inorden, er gitt ved til $H_n - 1$. Se punkt 4 i *Avsnitt* 5.2.15. På grunn av symmetrien er det den samme gjennomsnittlige avstanden mellom rotnoden og den siste i inorden. Hvis vi har et binært søketre T gitt av en permutasjon av tallene fra 1 til n der tallet k, $1 < k < n$ står først, vil vi derfor i gjennomsnitt for få flg. avstandssum:

(3) $$S = S_{k-1} + H_{k-1} - 1 + 1 + H_{n-k} - 1 + 1 + S_{n-k} = S_{k-1} + H_{k-1} + H_{n-k} + S_{n-k}$$

Vi får $H_{k-1} - 1 + 1$ fordi $H_{k-1} - 1$ er avstanden mellom rotnoden i det venstre subtreet og den siste i inorden i det treet og 1 er avstanden fra rotnoden i subtreet til rotnoden i T. På samme måte er det for $H_{n-k} - 1 + 1$.

Alle tall k fra 1 til n har samme sannsynlighet for å stå først i en permutasjon. Hvis $k = 1$ er det ikke noe venstre subtre og hvis $k = n$ er det ikke noe høyre subtre. Dermed får vi flg. rekursjonssligning for nS_n:

(4) $$\begin{aligned} nS_n &= H_{n-1} + S_{n-1} \\ &+ S_1 + H_1 + H_{n-2} + S_{n-2} \\ &+ S_2 + H_2 + H_{n-3} + S_{n-3} \\ &\;\cdot\;\cdot\;\cdot\;\cdot\;\cdot\;\cdot\;\cdot \\ &+ S_{n-2} + H_{n-2} + H_1 + S_1 \\ &+ S_{n-1} + H_{n-1}\,. \end{aligned}$$

Hvis vi slår sammen like ledd i (4) og bruker summetegn, får vi

(5) $\quad nS_n = 2(\sum_{k=1}^{n-1}(S_k + H_k)$

Vi får et uttrykk for $(n-1)S_{n-1}$ hvis vi erstatter n med $n-1$ i (5). Tar vi så differensen mellom nS_n og $(n-1)S_{n-1}$, får vi

(6) $\quad nS - (n-1)S_{n-1} = 2(S_{n-1} + H_{n-1})$

og dermed

(7) $\quad S_n = \dfrac{(n+1)S_{n-1} + 2H_{n-1}}{n}$

Vi kan teste formelen for S_n for n = 2, 3 og 4 siden vi allerede vet svarene for dem. Vi tar som utgangspunkt at S_1 = 0.

$$S_2 = \frac{(2+1)S_1 + 2H_1}{2} = H_1 = 1 \ , \qquad S_3 = \frac{(3+1)S_2 + 2H_2}{3} = \frac{4}{3} + \frac{2}{3}\frac{3}{2} = \frac{7}{3}$$

$$S_4 = \frac{(4+1)S_3 + 2H_3}{4} = \frac{5}{4}\frac{7}{3} + \frac{2}{4}\frac{11}{6} = \frac{23}{6}$$

Vi ser at dette stemmer for n = 2, 3 og 4. Uttrykket i (7) er en 1. ordens rekursjonsligning (eller differensligning), men er ikke av en direkte kjent type. Vi må derfor løse den skrittvis. Vi finner en formel for S_{n-1} ved å erstatte n med $n-1$ i (7). Da vil S_{n-1} bli uttrykt ved S_{n-2}. Setter vi så dette uttrykket for S_{n-1} inn i (7) og ordner opp i leddene, får vi:

(8) $\quad S_n = \dfrac{n+1}{n-1}S_{n-2} + \dfrac{4}{n-1}H_{n-2} + \dfrac{2}{(n-1)n}$

Vi gjentar dette, dvs. vi finner et uttrykk for S_{n-2} ved å erstatte n med $n-2$ i (7) og sette det inn i (8). Det gir oss dette:

(9) $\quad S_n = \dfrac{n+1}{n-2}S_{n-3} + \dfrac{6}{n-2}H_{n-3} + \dfrac{4}{(n-2)(n-1)} + \dfrac{2}{(n-1)n}$

Her ser vi et mønster. Fortsetter vi vil det ende opp med S_1 og H_1 på høyre side. Men siden S_1 = 0 og H_1 = 1, får vi S_n gitt som en rekke. Setter vi opp leddene i rekken motsatt vei, blir det:

(10) $\quad S_n = \sum_{k=1}^{n-1}\dfrac{2(n-k)}{k(k+1)} = 2(n\sum_{k=1}^{n-1}\dfrac{1}{k(k+1)} - \sum_{k=1}^{n-1}\dfrac{1}{k+1})$

Rekken $\sum_{k=1}^{n-1} \frac{1}{k(k+1)}$ er kjent (se 1.9 Formelsamling) og har sum $\frac{n-1}{n}$. Summen av $\sum_{k=1}^{n-1} \frac{1}{k+1}$ blir $H_n - 1$. Dermed fører (10) til det overraskende enkle uttrykket

(11) $S_n = 2(n - H_n)$

Vi kan sette inn n = 1, 2, 3 og 4 og sammenligne med fasiten. $S_1 = 2(1-1) = 0$, $S_2 = 2(2-3/2) = 1$, $S_3 = 2(3-11/6) = 7/3$ og $S_4 = 2(4-25/12) = 23/6$.

Til slutt kan vi sette opp det vi egentlig var ute etter, dvs. den gjennomsnittlige naboavstanden A_n:

(12) $A_n = \frac{S_n}{n-1} = \frac{2}{n-1}(n - H_n)$

Vi ser at $A_n \to 2$ når $n \to \infty$. For eksempel blir $A_{15} = 1{,}67$ og $A_{100} = 1{,}92$.

n	H_n	S_n	A_n
1	1	0	…
2	3/2 = 1,5	1	1
3	11/6 = 1,833…	7/3 = 2,33…	7/6 = 1.166…
4	25/12 = 2,0833…	23/6 = 3,833…	23/18 = 1,277…
5	137/60 = 2,2833…	163/30 = 5,433…	163/120 = 1,35833…
6	147/60 = 2,45	213/30 = 7,1	213/150 = 1,42
.	.	.	.
10	2,929	14,142	1,571
15	3,318	23,364	1,669
100	5,187	189,625	1,915
1000	7,485	1985,029	1,987