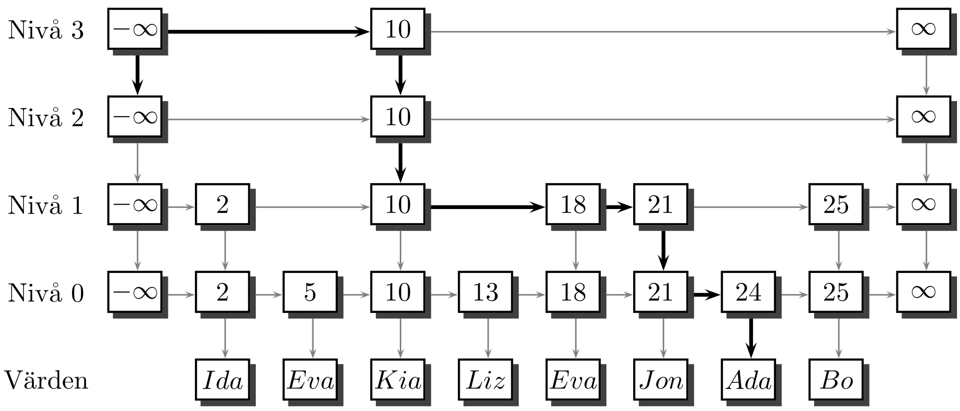
Skiplistor

Det här projektet handlar om skiplistor, en datastruktur för uppslagsverk som gör det möjligt att snabbt kunna hitta en nyckel (t.ex. ett telefonnummer) och därmed dess värde (t.ex. telefonabbonentens namn). Den fyller ungefär samma funktion som t.ex. klassen *TreeMap* i Java (som dock bygger på en helt annan datastruktur).

En ordentlig beskrivning av hur skiplistor fungerar hittar ni i avsnitt 8.6 i *Data Structures and Algorithms in Java* av Goodrich and Tamassia, som är kurslitteratur på Datastrukturer. Chalmersbiblioteket har boken i elektronisk form. (Välj Books24x7 BusinessPro bland chalmersbibliotekets E-boksamlingar eller gå in i Chans). Nedan följer en kortfattad beskrivning av hur listan ser ut, mest för att introducera terminologin som används här.



Figur 1: Exempel på skiplista med nycklarna 2,5,10,13,18,21,24,25

Antag att vi har n stycken tal (nycklar) vi vill lagra i en skiplista. Ordna talen och skapa en lista med n noder, där varje nod innehåller ett av talen. Detta är nivå 0 i skiplistan. Singla en slant för varje nod: blir det krona så skapas en ny nod på nivå 1 som innehåller samma tal, dvs med sannolikhet 1/2 kopieras en nod upp till nivå 1, oberoende av vad som sker med övriga noder. Fortsätt sedan på samma sätt: varje nod på nivå i får följa med till nivå $i+1$ med sannolikhet 1/2. Varje nivå ska dessutom börja med $-\infty$ och sluta med $+\infty$. Figur 1 visar *ett* av många möjliga *utfall* då talen $2, 5, 10, 13, 18, 21, 24, 25$ har stoppats in i en skiplista.

1. Antal noder per nivå (vecka 1)

   (a) Vad är sannolikheten att exempelvis det första talet når upp till nivå i (eller högre)?

   (b) Om man vet att k stycken tal når nivå i, på hur många sätt kan man då välja ut dessa tal?

   (c) Vad är sannolikheten att det finns exakt k stycken noder på nivå i, där $k = 0, 1, \ldots, n$ och $i = 0, 1, \ldots$, om man inte räknar med ändnoderna $-\infty$ och $+\infty$?

2. Värsta fallet (vecka 1)

   Det värsta som kan inträffa i en skiplista, med avseende på tidsåtgång vid sökning, är att alla tal når upp till samma nivå.

   (a) Vad är sannolikheten att exempelvis det första talet når upp till nivå i men inte högre?

   (b) Vad är sannolikheten att alla n talen når nivå i men ej $i + 1$, där $i = 0, 1, 2, \ldots$?

   (c) Vad är sannolikheten att alla n talen når upp till samma nivå?

   Ledtråd: Om alla når upp till samma nivå, betyder det att ingen når nivå 1, eller att alla når nivå 1 men ej 2, eller att alla når nivå 2 men ej 3, eller att alla når nivå 3

men ej 4, eller ...
Utnyttja (a) och att ”eller = $\cup$”.

3. Minnesutrymme (vecka 2)

   (a) Låt X beteckna antalet noder på nivå i. I uppgift 1 har ni räknat på sannolikheten att $X = k$. Vad kallas den stokastiska variabeln X och vilka parametrar har den?

   (b) Vad är förväntat antal noder på nivå i, $i = 0, 1, \ldots$, även här exklusive ändnoder? Ledtråd: Utnyttja väntevärdet i binomialfördelningen.

   (c) Vad är förväntat totalantal noder i skiplistan, exklusive ändnoder?

4. Skiplistans höjd (vecka 2)

   (a) Låt Y beteckna höjden (dvs översta nivån) på första talets stapel. I uppgift 2a har ni räknat på sannolikheten att $Y = i$. Vad kallas den stokastiska variabeln $X = Y + 1$ och vilken är dess parameter?

   (b) Förklara varför höjden i en skiplista som består av n stycken tal, H_n, kan beskrivas som $H_n = \max\{X_1, \ldots, X_n\} - 1$, där $X_1, \ldots, X_n$ är oberoende och geometriskt fördelade stokastiska variabler med parameter 1/2.

   (c) Bestäm fördelningsfunktionen för H_n.

   (d) Bestäm frekvensfunktionen för H_n. Utnyttja att $f(k) = F(k) - F(k-1)$ för heltalsvärda stokastiska variabler.

   (e) Använd dator för att approximativt beräkna $E[H_{2^i}]$, $i = 1, 2, ..., 7$. Ser du något mönster? Utnyttja detta för att motivera att $E[H_n] \approx \log_2 n$. (Låt $n = 2^i$ så att $i = \log_2 n$.)

   Ett explicit uttryck för $E[H_n]$ för godtyckligt n finns inte. Se dock gärna Theorem 1 på http://math.sun.ac.za/~prodinger/postscriptfiles/florenz.ps, där ett asymptotiskt uttryck är givet.

5. Tidsåtgång för sökning (frivillig)

   När vi söker efter en nyckel (och dess värde) så börjar vi alltid på $-\infty$ längst upp i vänstra hörnet. Därefter följer vi pekarna rakt åt höger så långt det går utan att gå förbi det tal vi söker. När vi inte kan gå åt höger längre för att nästa tal är större än det sökta så går vi istället ner en nivå. De tjocka pilarna i figur 1 visar sökvägen för talet 24.

   (a) Låt Z_i vara antalet noder vi besöker vi på varje nivå? Vad är $E[Z_i]$?

   Ledtråd: Tänk dig att du följer pilarna baklänges på nivå i. Då kommer vi gå åt vänster på samma nivå ända tills vi stöter på ett tal som finns på nivå i och dessutom på nivå i+1. För varje tal på nivå i (inklusive det vi börjar på) är sannolikheten $\frac{1}{2}$ att det finns på nivå i+1 också (detta händer då det blev krona i motsvarande slantsingling). Z_i är alltså antalet oberoende slantsinglingar som behövs innan en krona dyker upp!

   (b) Försök argumentera med hjälp av 4e och 5a ovan för att det förväntade antalet noder som vi besöker är $\approx 2\log_2 n$.

   Obs: Informationen från 4e och 5a räcker inte för att göra denna argumentationen helt strikt, eftersom H_n och Z_i inte är oberoende.

   Av detta följer att tidsåtgången vid sökning asymptotiskt är $O(\log_2 n)$. Tidsåtgången för insättningen och borttagning av tal är av samma storleksordning, men det går vi ej närmre in på här.

Sista inlämningsdag: Mån 12/9.