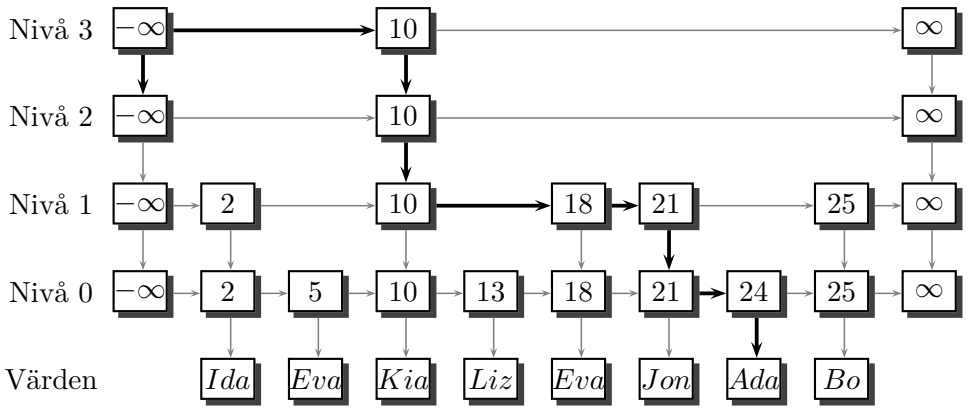
Skiplistor

Inlämningsuppgift nr 1 i MSG810 hösten 2018

Det här projektet handlar om skiplistor, en datastruktur som används för snabbare sökning i listor för att snabbt hitta en nyckel (t.ex. telefonnummer eller uppslagsord) och visa dess värde (t.ex. telefonabonnentens namn eller förklaringen på ordet). Bra beskrivningar på skiplistor finns på nätet, t.ex. på Wikipedia eller genom att bara Googla på 'skip lists'. En ordentlig beskrivning hittar du annars i avsnitt 8.6 i *Data Structures and Algorithms in Java* av Goodrich & Tamassia, som finns att låna på Chalmers-biblioteket. Nedan följer en kortfattad beskrivning av hur listan konstrueras och ser ut, mest för att introducera terminologin.



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

Antag att vi har n stycken tal (nycklar) som 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. För att skapa nästa nivå singlar du sedan slant för varje nod: blir det krona så skapas samma nod på nivå 1 med samma tal, och blir det klave skippas den noden på den nivån. Dvs med sannolikhet 1/2 kopieras noden upp till nivå 1. Fortsätt sedan på samma sätt: varje nod på nivå i får följa med till nivå $i+1$ med sannolikhet 1/2. Proceduren fortsätter tills nästa nivå blir tom. Varje nivå ska dessutom börja med $-\infty$ och sluta med $+\infty$. Figur 1 ger ett exempel på ett möjligt utfall av alla tänkbara skiplistor för talen 2, 5, 10, 13, 18, 21, 24, 25.

Uppgift 1-2 kan ni klara första veckan av kursen medan uppgift 3-5 kräver kunskaper om stokastiska variabler och görs lämpligen vecka 2.

Det är viktigt att ni motiverar era svar. Ta med era uträkningar, skriv tydligt om händelser är oberoende, disjunkta osv.

Frågor relaterade till vecka 1:

1. Antal noder per nivå:
   (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 $k = 0, 1, ..., n$ och $i = 0, 1, ...$, om man inte räknar med ändnoderna $-\infty$ och $+\infty$?

2. Värsta fallet.

   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 $i = 0, 1, 2, ...$?

   (c) Vad är sannolikheten att alla n talen når upp till samma nivå? Förenkla uttrycket så långt det går.

   Ledtråd: Om alla når upp till samma nivå, betyder det att ingen når nivå 1, eller att alla når nivå 1 men inte 2, eller att alla når nivå 2 men inte 3, osv. Utnyttja (a) och att "eller $= \cup$ (union)".

Frågor relaterade till vecka 2:

3. Minnesutrymme.

   (a) Låt X beteckna antalet noder på nivå i. I uppgift 1 har ni räknat på sannolikheten att $X = k$. Vilken fördelning har den stokastiska variabeln X och varför? Vilka parametrar har fördelningen?

   (b) Vad är förväntat antal noder på nivå i för $i = 0, 1, 2, ...$, exklusive ändnoder?

   Ledtråd: Utnyttja väntevärdet i X's fördelning.

   (c) Vad är förväntat totalantal noder i skiplistan, exklusive ändnoder? Förenkla uttrycket så långt det går.

4. Skiplistans höjd.

   (a) Låt Y beteckna höjden, dvs översta nivån, på första talets stapel (i exemplet i figur 1 har första talets stapel höjd 1). I uppgift 2(a) har ni räknat på sannolikheten att $Y = i$. Vilken fördelning har den stokastiska variabeln $X = Y + 1$ och varför? Vilken parameter har fördelningen?

   (b) Förklara varför höjden H_n i en skiplista som består av n stycken tal kan beskrivas som $H_n = \max\{X_1, ..., X_n\} - 1$, där $X_1, ..., X_n$ är oberoende stokastiska variabler med samma fördelning som i uppgift 4(a)?

   (c) Bestäm fördelningsfunktionen F_{H_n} för H_n.

   Ledtråd: $\max\{X_1, ..., X_n\} \leq k$ är ekvivalent med att $X_1 \leq k, \ldots, X_n \leq k$.

   (d) Bestäm frekvensfunktionen f_{H_n} för H_n.

   Ledtråd: För heltalsvärda stokastiska variabler gäller att $f(k) = F(k) - F(k-1)$ om k är ett heltal.

   (e) Använd formeln $E[H_n] = \sum_{k=1}^{\infty} k f_{H_n}(k)$ och en dator för att approximativt beräkna väntevärdet $E[H_{2^m}]$, för $m = 1, 2, ..., 7$. Ser du något mönster? Utnyttja detta för att motivera att $E[H_n] \approx \log_2 n$. (Låt $n = 2^m$ så att $m = \log_2 n$.)

   (Ett explicit uttryck för $E[H_n]$ för godtyckligt n finns inte, utan enbart axymptotiska uttryck.)

Frivilliga frågor:

5. När vi söker efter en nyckel (och dess värde) så börjar vi alltid på $-\infty$ längst upp i vänstra hörnet. Därefter följer vi pilarna rakt åt höger så långt det går utan att gå förbi det tal vi söker. När vi inte kan gå åt höger längre utan att nästa tal blir 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 på varje nivå. Vad är väntevärdet $E[Z_i]$?

   Ledtråd: Tänk dig att vi följer pilarna baklänges på nivå i. Då kommer vi att 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 sannolkheten $1/2$ att talet finns på nivå $i + 1$ också (om 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 4(e) och 5(a) ovan för att det förväntade antalet noder som vi besöker är $\approx 2 \log_2 n$.

   Obs: Informationen från 4(e) och 5(a) räcker inte för att göra denna argumentation 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 inte in närmre på här.