

2009 年度 冬学期 計算数学 II レポート問題

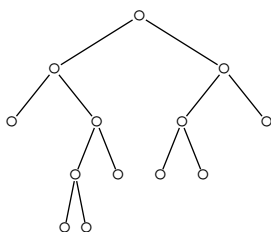
二分探索木の分析においては、 n 個のキーがランダムに与えられたとして、構成や探索に要する時間の平均を調べた。平均を取るときランダムに与えられるのは、あくまでも n 個のキーであって、二分木がランダムに与えられるわけではない。この違いを理解するために、後者の場合、すなわち木がランダムに与えられる場合を分析してみよう。

その前に 二分探索木の場合を少し思い出そう。時間を計るかわりに 比較の回数を数えることにすると、ある二分探索木 T を構成するのに要する比較の回数は、次のように与えられる。いま、 $d(v)$ でノード v の深さを表すとする。すなわち、 v がルートするとき、 $d(v) = 0$ となり、ルートのすぐ下の子ノード v の場合 $d(v) = 1$ となり、以下順次 $d(v) = 2, 3, \dots$ となる。言い方を変えれば、 $d(v)$ はルートからノード v までたどったとき、そのパスに現れるエッジの数のことである。

(1) 二分探索木 T を構成するのに要する比較の回数は $\sum_v d(v)$ で与えられる。ここで、 v は T の (葉以外の) 内部ノード全体の上を走る。このことを、直感的でかまわないので、説明せよ。ただし、二分探索木において、キーは内部ノードのみに格納されるものとし、葉にはキーは存在しないとしている。

つまり、二分探索木の構成に要する比較の回数の平均は、 $\sum_v d(v)$ の平均として求めることができたのである。ただし、平均を求める際には、 n 個のキーがランダムに与えられたとして計算したのだった。

以下では対照的に、二分木がランダムに与えられたとして、 $\sum_v d(v)$ の平均を求めてみる。まずは、 n 個の内部ノードをもつ二分木がいくつあるかを求めよう。ここで二分木とは



のような構造である。この例の場合、6 個の内部ノードを持つ。

(2) n 個の内部ノードを持つ二分木の総数を C_n と書くことにする。 $C_0 = 1$ である (葉のみをもつ二分木が一つ)。母関数 $h(x) = \sum_{n=0}^{\infty} C_n z^n$ を考える。数え上げ組み合わせ論の知識を用いると、 $h(x)$ は二次方程式 $y = xy^2 + 1$ の

解 $y = h(x)$ で与えられることが知られている. この事実を用いて,

$$C_n = \frac{1}{n+1} \binom{2n}{n}$$

となることを示せ. C_n は Catalan 数と呼ばれている.

さて, n 個の内部ノードをもつ二分木 C_n 個がランダムに与えられたとき, ルートの左の部分木が k 個の内部ノードをもち, ルートの右の部分木が l 個の内部ノードをもつ確率は $C_k C_l / C_n$ である. ただし当然 $k + l = n - 1$ でなければならない.

n 個の内部ノードをもつ二分木がランダムに与えられたとき, $\sum_v d(v)$ の平均を確率変数 X_n で表すことにする. ただし, v は前と同じように内部ノード全体を動くとしている.

(3) 確率変数 X_n に対する確率母関数 $\sum_{k=0}^{\infty} \Pr\{X_n = k\} x^k$ を $f_n(x)$ と置くと, $n \geq 1$ ならば再帰方程式

$$f_n(x) = \sum_{k+l=n-1} \frac{C_k C_l}{C_n} x^{n-1} f_k(x) f_l(x)$$

を満たす. $n = 0$ のときは, $f_0(x) = 1$ である. この再帰方程式の導出方法を説明せよ.

(4) さらに変数 z を用意して, $g(x, z) = \sum_{n=0}^{\infty} C_n f_n(x) z^n$ と置く. このとき, $g(1, z)$ を初等関数として表せ. ヒント: $f_n(1)$ の値は?

(5) $g(x, z) = 1 + z \cdot g(x, xz)^2$ が成り立つことを示せ. これを用いて,

$$g_x(1, z) = \frac{1 - 3z + (z-1)\sqrt{1-4z}}{z(1-4z)}$$

を示せ. さらに z の冪級数に展開することで

$$f'_n(1) = \frac{4^n}{C_n} - 3n - 1$$

を示せ.

(6) n 個の内部ノードをもつ二分木がランダムに与えられたとき, $\sum_v d(v)$ の平均 $E[X_n]$ を, $(n$ の初等関数) $+ o(1)$ の形で求めよ. ヒント: Catalan 数 C_n は Stirling の公式

$$n! = \frac{\sqrt{2\pi n} n^n}{e^n} \left(1 + \frac{1}{12n} + O(n^{-2}) \right)$$

を用いて評価.

(7) 上の結果のオーダーを, n 個のランダムなキーから二分探索木を構成する場合と比較せよ. その差異はどうして生じるのか.

二分探索木において, 構成に要する時間が分かると, 探索に要する時間 ($\sim$ 比較回数) も分かるのであった. 具体的には, 構成に要する比較回数の平均を $E(n)$ とすると, 探索にかかる平均の比較回数は, $E(n)/n + 1$ で求めることができる (探索に成功する場合). ランダムな n 個のキーから二分探索木を構成したときには, $E(n) = O(n \log n)$ だったので, 探索時間は $O(\log n)$ となる. しかし, もし $E(n)$ が今回の課題で求めた平均 $E[X_n]$ だったらどうか. 探索時間が $O(\log n)$ となるのは, 二分木だから当たり前というようなものではないことが, 分かるであろう.