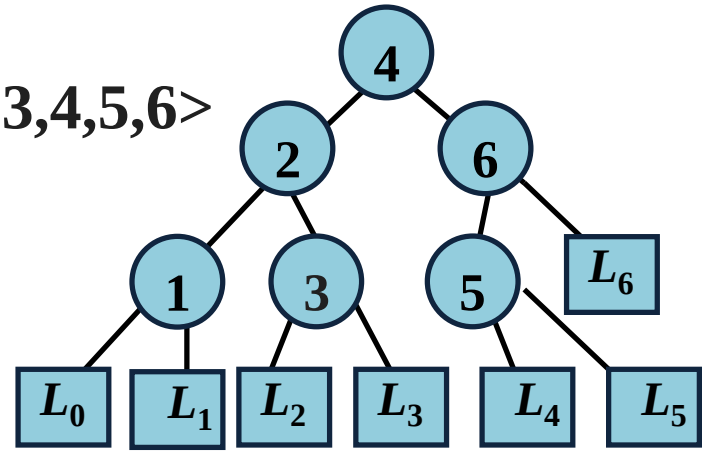
最优二叉检索树

二叉检索树

集合S为排序的 n个元素，$x_1<x_2<\ldots<x_n$，将这些元素存储在一棵二叉树的结点上，以查找 x 是否在这些数中．如果 x 不在，确定 x 在那个空隙（方结点）．

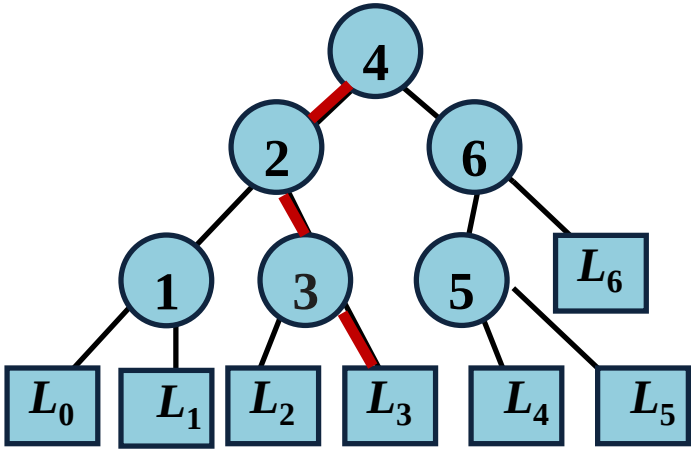
实例：

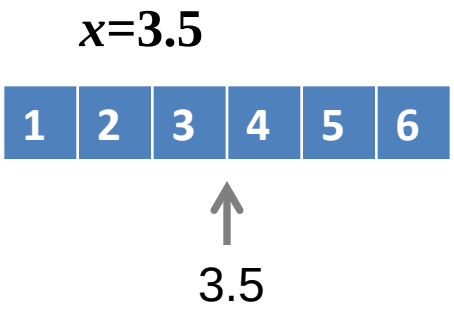
S=<1,2,3,4,5,6>



二叉树的检索方法

1. 初始，x与根元素比较；
2. x < 根元素，递归进入左子树；
3. x > 根元素，递归进入右子树；
4. x = 根元素，算法停止，输出 x；
5. x 到叶结点算法停止，输出 x不在数组.





数据元素存取概率分布

空隙:

$(x_0, x_1), (x_1, x_2), \ldots, (x_{n-1}, x_n), (x_n, x_{n+1})$,

$x_0 = -\infty, \quad x_{n+1} = +\infty$

给定序列 $S = < x_1, x_2, \ldots, x_n >$,

x 在 x_i 的概率为 b_i,

x 在 (x_i, x_{i+1}) 的概率为 a_i,

S 的存取概率分布如下:

$P = < a_0, b_1, a_1, b_2, a_2, \ldots, b_n, a_n >$

实例

实例： $S = < 1, 2, 3, 4, 5, 6 >$

$P = < 0.04, 0.1, 0.01, 0.2, 0.05, 0.2, 0.02,$

$0.1, 0.02, 0.1, 0.07, 0.05, 0.04 >$

1, 2, 3, 4, 5, 6 检索的概率分别为：

0.1, 0.2, 0.2, 0.1, 0.1, 0.05

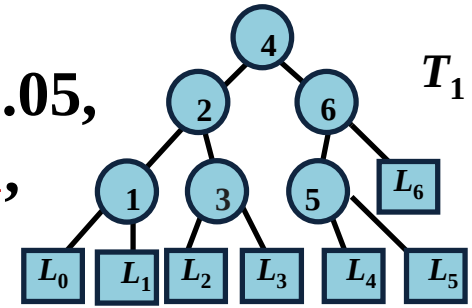
各个空隙的检索概率分别为：

0.04, 0.01, 0.05, 0.02, 0.02, 0.07, 0.04

检索数据的平均时间

S = < 1, 2, 3, 4, 5, 6 >

P = < 0.04, 0.1, 0.01, 0.2, 0.05,
0.2, 0.02, 0.1, 0.02, 0.1,
0.07, 0.05, 0.04 >



T_1

$m\ (T_1)$

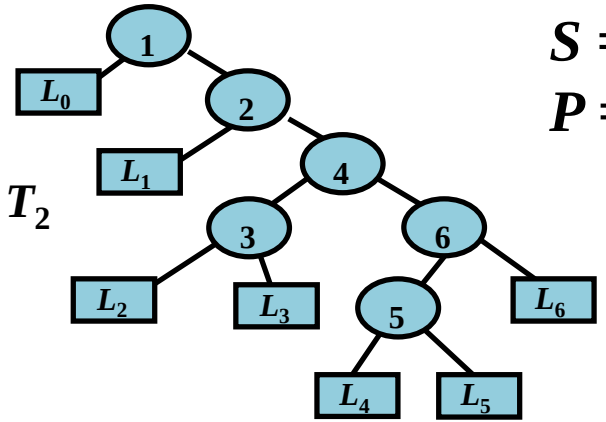
= [1×0.1+2×(0.2+0.05) +3×(0.1+0.2+0.1)]

+[3×(0.04+0.01+0.05+0.02+ 0.02+0.07)

+2×0.04]

= 1.8 + 0.71 = 2.51

检索数据的平均时间



T_2

S = < 1, 2, 3, 4, 5, 6 >

P = < 0.04, 0.1, 0.01, 0.2,
0.05, 0.2, 0.02, 0.1,
0.02, 0.1, 0.07, 0.05,
0.04 >

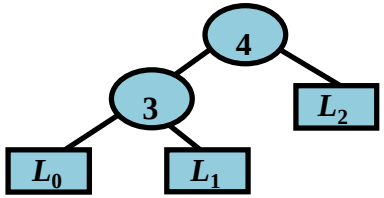
$m\ (T_2)$

= [1×0.1+2×0.2+3×0.1+4×(0.2+0.05)+5×0.1]
+ [1×0.04+2×0.01+4×(0.05+0.02+0.04)
+ 5×(0.02+0.07)]
= 2.3 + 0.95 = 3.25

平均比较次数计算

数据集 $S=<x_1, x_2,\ldots, x_n>$

存取概率分布



$P=<a_0, b_1, a_1, b_2, \ldots, a_i, b_{i+1}, \ldots, b_n, a_n>$

结点 x_i 在 T 中的深度是 $d(x_i)$, $i=1,2,\ldots,n$,

空隙 L_j 的深度为 $d(L_j)$, $j=0,1,\ldots,n$,

平均比较次数为：

$$t=\sum_{i=1}^{n} b_i(1+d(x_i))+\sum_{j=0}^{n} a_j d(L_j)$$

问题

给定数据集

$$S = < x_1, x_2, \ldots, x_n >,$$

及 S 的存取概率分布如下：

$$P = < a_0, b_1, a_1, b_2, a_2, \ldots, b_n, a_n >$$

求一棵最优的 (即平均比较次数最少的) 二分检索树.

小结

- 二叉检索树的构成
- 给定概率分布下，一棵二叉检索树的平均检索时间估计
- 什么是最优二叉检索树