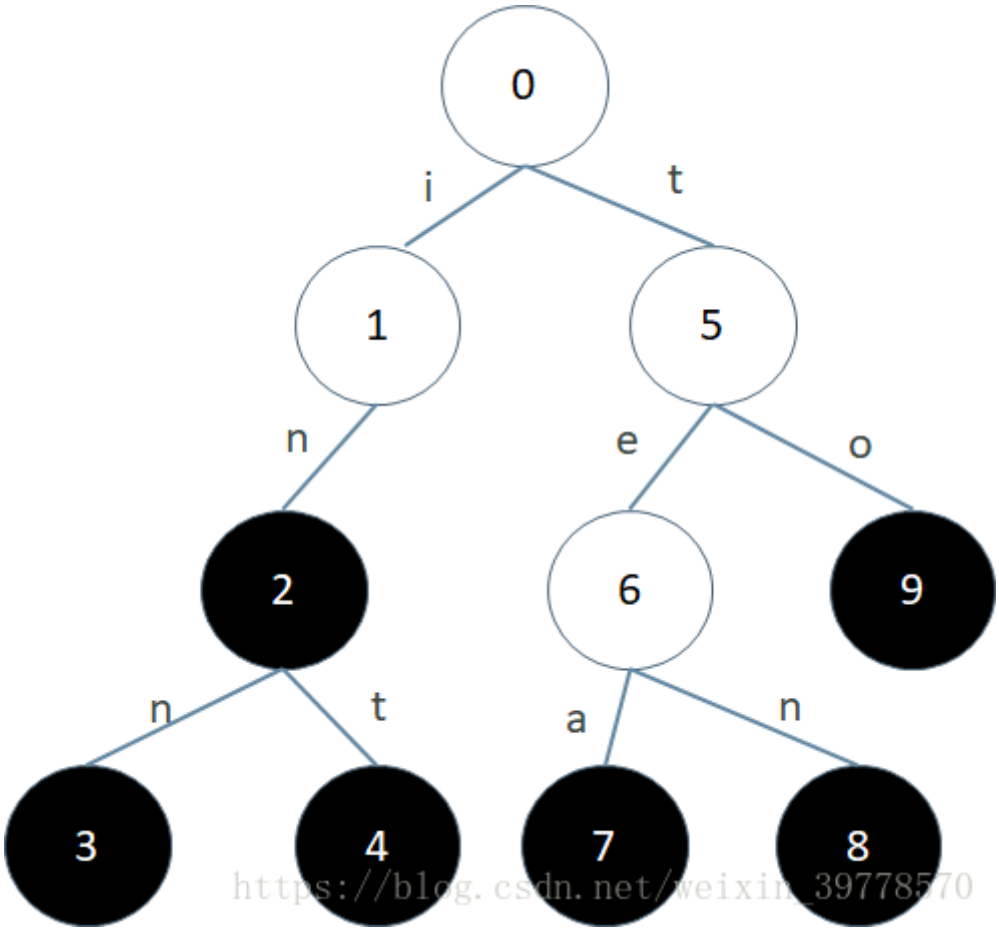
面试题

1.大量的URL中找出相同的URL

题目描述

给定 a、b 两个文件，各存放 50 亿个 URL，每个 URL 各占 64B，内存限制是 4G。请找出 a、b 两个文件共同的 URL。

- 方法1：先hash(url)%1000分成1000个约为300M的文件，然后用hash表去查找
- 方法2：前缀树：降低存储成本



2.从大量数据中找出高频词

分治+hashMap+大顶堆或者小顶堆

3.查询最热门的查询串

搜索引擎会通过日志文件把用户每次检索使用的所有查询串都记录下来，每个查询串的长度不超过 255 字节。

假设目前有 1000w 个记录（这些查询串的重复度比较高，虽然总数是 1000w，但如果除去重复后，则不超过 300w 个）。请统计最热门的 10 个查询串，要求使用的内存不能超过 1G。（一个查询串的重复度越高，说明查询它的用户越多，也就越热门。）

方法一：分治法

方法二：HashMap法

300w*(255+4)~=777M 4字节存储的是频次。

先创建map，时间复杂度是$O(N)$,

再遍历map构建10个元素的小顶堆$O(Nlog10)$

方法三：前缀树法

4.如何统计不同电话号码的个数?

数据重复问题一般用位图法

8位的电话号码，一共有10^8中可能，也就是1亿种可能。所以申请一个1亿位的空间用来保存不同电话号码，遍历每一个就行

5.从5亿个数中找到中位数

方法一：维护一个大顶堆一个小顶堆。大顶堆的最大值小于等于小顶堆的最小值，两个堆的元素个数相差不超过1，这样就能只去两个堆的堆顶元素就能得到中位数的值了。

方法二：分治，通过二进制位的最高位是0还是1，可以得到两组数，通过两组的多少和最高位的情况就能判断中位数在哪组里面。

6.topN最大的数

1.排序或者部分排序

2.分治求得最前面N个数

3.内存不够大的情况下，多次读取，因此采用分布式再汇总

4.单个机器的话就要考虑使用堆来存储，小顶堆，小于顶端元素直接舍弃；如果大于顶端元素，就去掉顶端元素，调整堆。

10亿个4字节大概是是4个G

7.40亿个数字中找是否有和给定数字相同的?

1.set/map进行映射

2.分布式读取，减少IO读取的时间

3.用bitmap，位图法，每个数只用一个位表示即可。

原来是40亿个4字节的数，现在是40亿个1位的bool数，降低了4*8=32倍，这样就能将16G的内存使用降低到500M，2G的内存条也够用。

这里要申请的是32亿个位，要把整数都覆盖了（这里要和40亿区别，40亿是个数，2^32=42亿个是int的取值范围，第32位代表int的最大值，这样就可以对应40亿甚至50亿60亿个int值了）。这样才能表示int类型的全部值。每一位用01表示就行。

4.由于申请的空间可以表示42亿个数，但是其实题目中给的是40亿，那么先对其排序，

如果数据是1 2 3 4 6 7.....这种的，那么可以用(1,4)和(6,2)来表示，这样一来，连续的数都变成了2个数表示。 来了一个新数之后，就用二分法进行查找了。

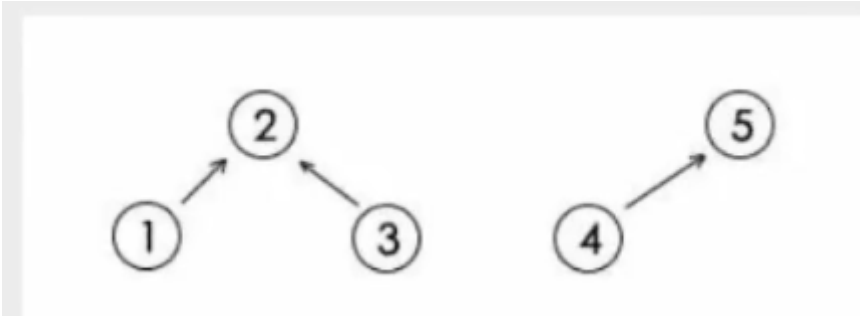
这样最多有2亿个断点，一个断点4*2=8字节，一共16亿字节，内存是1.6G，可以一次性加载处理。

8.朋友圈的个数

方法1：用bitmap来存储集合，一组对应关系就建立一个bitmap的数组，两两之间如果按位与之后不全为0就说明有交集，然后按位或一下就能够得到新的交集的bitmap

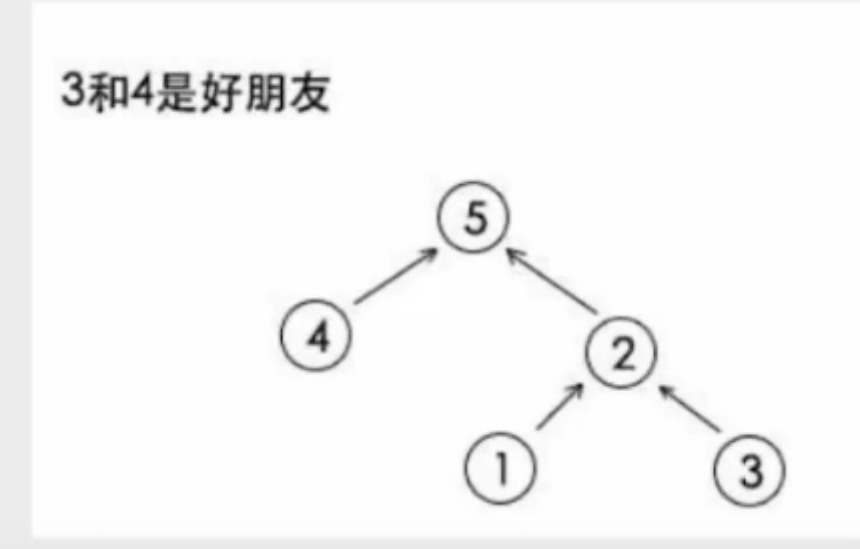
方法二：双向链表，遍历的过程中，如果存在连接就进行合并，统计链表数量就行

方法三：单向链表，反向的树，父节点指向子节点

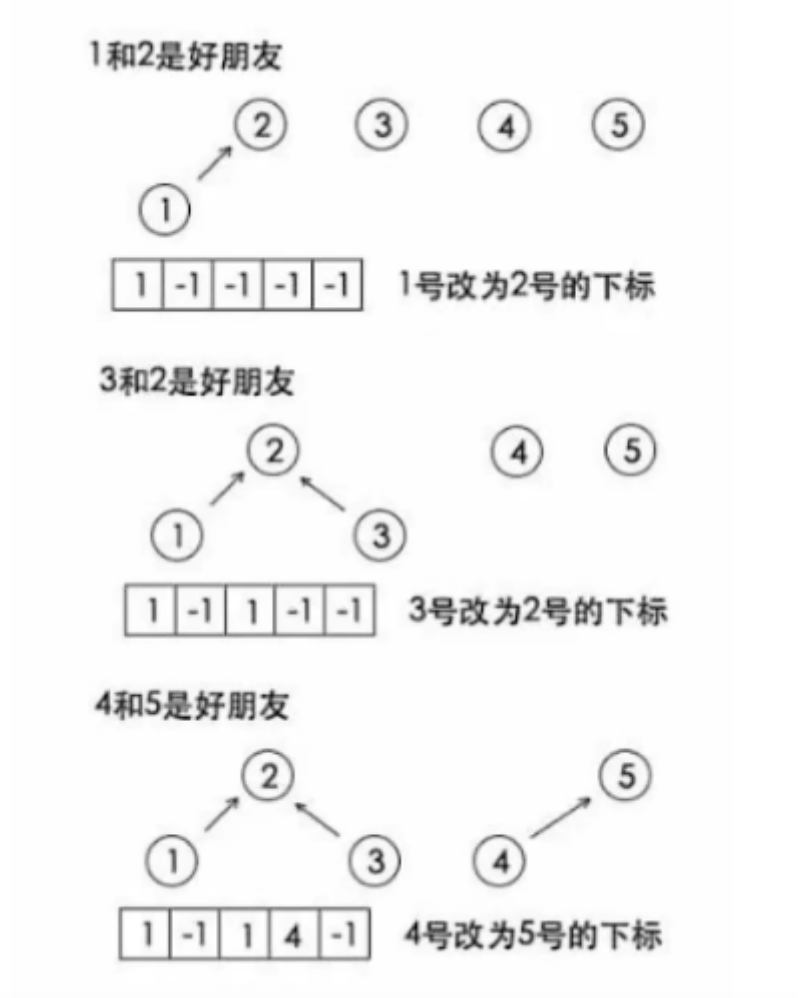


存储的方式就采用树，一对好友关系就是把两棵树树根连接起来。

如果过34是好友，那么就要把两者的根25相连



为了精简，可以将指针类型的节点改成数组存储。



最后只需要查询数组中值为-1的数量就行。

还能申请一个数组用来表示树的高度，防止树退化成链表。