BITTIGER

CS102 Top100高频算法设计课

第四节　大数据、位运算

左程云

版权声明

BITTIGER

Copyright Policy

All content included on the Site or third-party platforms as part of the class, such as text, graphics, logos, button icons, images, audio clips, video clips, live streams, digital downloads, data compilations, and software, is the property of BitTiger or its content suppliers and protected by copyright laws.

Any attempt to redistribute or resell BitTiger content will result in the appropriate legal action being taken.

We thank you in advance for respecting our copyrighted content.For more info see https://www.bittiger.io/termsofuse and https://www.bittiger.io/termsofservice

认识Bloom Filter

【大数据题目一】
不安全网页的黑名单包含100亿个黑名单网页，每个网页的URL最多占用64B。现在想要实现一种网页过滤系统，可以根据网页的URL判断该网页是否在黑名单上，请设计该系统。
【要求】
1．该系统允许有万分之一以下的判断失误率。
2．使用的额外空间不要超过30GB。

认识Bloom Filter

【解题点】

1，假设数据量为n，预期的失误率为p（布隆过滤器大小和每个样本的大小无关）

2，根据n和p，算出Bloom Filter一共需要多少个bit位，向上取整，记为m

3，根据m和n，算出Bloom Filter需要多少个哈希函数，向上取整，记为k

4，根据修正公式，算出真实的失误率p_true

$$m=-\frac{n\times\ln p}{(\ln 2)^2}$$

$$k=\ln 2\times\frac{m}{n}=0.7\times\frac{m}{n}$$

$$(1-e^{-\frac{nk}{m}})^k$$

只用2GB内存在20亿个整数中找到出现次数最多的数

【大数据题目二】
有一个包含20亿个全是32位整数的大文件，在其中找到出现次数最多的数。
【要求】
内存限制为2GB。

只用2GB内存在20亿个整数中找到出现次数最多的数

【解题点】
1，假设哈希表一条记录(key,value)占用8字节。2GB内存最多允许2亿+条记录
2，为了一个文件中不同数的种类不超过2亿+种，要将20亿个数分成10+个文件
3，根据哈希函数的性质，可以把20亿个数通过哈希函数与求余运算，在种数上分成10+份，而且每份之间，不会有重复的数
4，一份数就是一个文件，对每一个文件用哈希表求出现次数最多的数，并且记录下出现次数最多的数的词频。此时内存是够用的
5，10+个文件，10+个出现次数最多的数，选出现次数最多的作为结果

40亿个非负整数中找到没出现的数

【大数据题目三】
32位无符号整数的范围是0～4294967295，现在有一个正好包含40亿个无符号整数的文件，所以在整个范围中必然有没出现过的数。可以使用最多1GB的内存，怎么找到所有没出现过的数？
【大数据题目三进阶】
内存限制为10MB，但是只用找到一个没出现过的数即可。

40亿个非负整数中找到没出现的数

【原问题解题点】
1GB内存，80亿个bit，每个bit表示一个数是否出现过(0/1)，一共42亿个数，够了
【进阶问题解题点】
1，先算10MB的bitArr能给多大范围的数做精细的计数，假设范围大小为a
2，b=42亿/a，代表有多少个范围，遍历所有的数，在这个范围上做计数
3，必然会有计数不满的区间，在不满的区间上，用bitArr做精细的计数统计

找到100亿个URL中重复的URL以及搜索词汇的top *K*问题

【大数据题目四】
有一个包含100亿个URL的大文件，假设每个URL占用64B，请找出其中所有重复的URL。
【大数据题目四补充题目】
某搜索公司一天的用户搜索词汇是海量的（百亿数据量），请设计一种求出每天最热top 100词汇的可行办法。

找到100亿个URL中重复的URL以及搜索词汇的top *K*问题

【解题点】
1，哈希
2，堆

40亿个非负整数中找到出现两次的数和所有数的中位数

【大数据题目五】
32位无符号整数的范围是0～4294967295，现在有40亿个无符号整数，可以使用最多1GB的内存，找出所有出现了两次的数。
【补充题目】
可以使用最多10MB的内存，怎么找到这40亿个整数的中位数？

40亿个非负整数中找到出现两次的数和所有数的中位数

【原题目解题点】
1，用两个bit表示一个数出现的次数
2，超过了两次，状态就不变了
【补充题目解题点】
1，确定精细区间的大小a
2，根据a确定区间数
3，区间之间搞大计数之后，就知道中位数来自哪个区间
4，确定了来自哪个小区间，再遍历，此时在小区间上做精细计数

一致性哈希算法的基本原理

【大数据题目六】
工程师常使用服务器集群来设计和实现数据缓存，以下是常见的策略：
1．无论是添加、查询还是删除数据，都先将数据的id通过哈希函数转换成一个哈希值，记为key。
2．如果目前机器有*N*台，则计算key%N的值，这个值就是该数据所属的机器编号，无论是添加、删除还是查询操作，都只在这台机器上进行。
请分析这种缓存策略可能带来的问题，并提出改进的方案。

一致性哈希算法的基本原理

【解题点】
1，理解哈希函数的性质
2，理解一致性哈希的结构
3，理解一条数据去寻找所在机器的简单落地实现
3，理解虚拟节点技术的优化点来自哈希函数的性质

位运算题目

【位运算题目一，大俗题】
不用额外变量交换两个整数的值

位运算题目

【位运算题目二】
给定两个32位整数a和b，返回a和b中较大的。不用任何比较判断。

位运算题目

【解题点】
不会溢出的方法：
如果a的符号与b的符号不同（difSab==1,sameSab==0），则有：
　　如果a为0或正，那么b为负（sa==1,sb==0），应该返回a；
　　如果a为负，那么b为0或正（sa==0,sb==1），应该返回b。
如果a的符号与b的符号相同（difSab==0,sameSab==1），这种情况下，a-b的值绝对不会溢出：
　　如果a-b为0或正（sc==1），返回a；
　　如果a-b为负（sc==0），返回b；
综上所述，应该返回a * (difSab * sa + sameSab * sc) + b * flip(difSab * sa + sameSab * sc)。

位运算题目

【位运算题目三】
给定一个32位整数*n*，可为0，可为正，也可为负，返回该整数二进制表达中1的个数。

位运算题目

【解题点】

1，整数*n*每次进行无符号右移一位，检查最右边的bit是否为1来进行统计

2，每次进行n&=(n-1)操作，n&=(n-1)操作的实质是抹掉最右边的1，求抹掉次数

3，每次进行n&=(n-1)操作，n-=n&(~n+1)这也是移除最右侧的1的过程

4，平行法

5，MIT hackmem算法

位运算题目

【位运算题目四】
给定一个整型数组arr，其中只有一个数出现了奇数次，其他的数都出现了偶数次，打印这个数。
【进阶】
有两个数出现了奇数次，其他的数都出现了偶数次，打印这两个数

位运算题目

【原问题解题点】
整数*n*与0异或的结果是*n*，整数*n*与整数*n*异或的结果是0。所以，先申请一个整型变量，记为eO。在遍历数组的过程中，把eO和每个数异或（eO=eO^当前数），最后eO的值就是出现了奇数次的那个数
【进阶问题解题点】
1，找两个数在某个位置上不一样，假设为i位置
2，i位置其实把所有的数字分成了两堆，一堆i 位置上是1，另一堆是0
3，只把i 位置上是1的那堆数异或，就得到了其中一个数
4，另一个数简单异或得到

位运算题目

【位运算题目五】
给定一个整型数组arr和一个大于1的整数*k*。已知arr中只有1个数出现了1次，其他的数都出现了*k*次，请返回只出现了1次的数。
【要求】
时间复杂度为*O*(*N*)，额外空间复杂度为*O*(1)。

位运算题目

【解题点】
多位数异或而已

位运算题目

【位运算题目六】
只用位运算实现整数的加减乘除运算

位运算题目

1，求两个数的平均数：((x&y) + ((x^y)>>1)
2，判断奇偶，用if((a&1)== 0)代替if(a%2 == 0)
3，变符号， ~a+1，系统最小值变符号还是自己

位运算题目

刷题时看到位运算的解法出现别慌，位运算一般都是用来加速得到或者设置状态的过程，比如N皇后问题中，皇后的状态。我们会在以后的章节中讲这个题。

课程项目负责人：Catherine

邮件：weiyi@bittiger.io

左程云答疑邮箱：chengyunzuo@gmail.com

微信二维码：

关注微信，获得太阁最新信息

微信: bit_tiger

官网: BitTiger.io