新年的贺电 解题报告

安徽师范大学附属中学 罗哲正

1 试题来源

UOJ Goodbye Yiwei E.新年的贺电

链接：http://uoj.ac/contest/24/problem/178。

2 试题大意

猴族一共拥有 1024 台高达，每个高达有一个编号，为一个 0 到 $2^{32}-1$ 的整数。（不一定连续）

猴族一共拥有 1024 个机库，编号为 0 到 1023。每个高达都存放在某个机库中（**一个机库可能存放多个高达**）。

前线将领只知道自己管辖范围内的高达的编号而不知道位置，只有猴族首领猴腮雷清楚每个高达存放在哪个机库，所以他需要把位置信息发给前线将领。

但是，猴族的通讯技术不发达，猴族首领猴腮雷找到了你——请你设计一个通讯方式传输高达的位置。

2.1 任务描述

你需要写一个程序，实现编码和解码的功能。

2.1.1 编码

如果是编码，输入的第一行为一个字符串 “encode”。

接下来 1024 行，每行两个整数 k, v，表示编号为 k 的高达存放在编号为 v 的机库。保证 $0 \le k < 2^{32}$，$0 \le v < 1024$，**保证 k 互不相同**。

你需要输出一个 **01 串**，表示发送给前线将领的信息。

2.1.2 解码

如果是解码，输入的第一行为一个字符串 “decode”。

接下来一行，是你的程序编码出的 01 串。

接下来一行，一个正整数 Q，表示前线将领管辖了 Q 个高达。

接下来 Q 行，每行一个整数 k 表示一个高达编号。保证一定是合法的高达编号，**保证编号互不相同**。

对于每个高达编号，你需要输出一行，为它所在的机库编号。

2.2 限制与约定

如果解码出的机库编号错误，直接 0 分；

如果你的程序正常执行，设编码出的 01 串长度为 n，则你可以获得如下分数：

得分	条件	得分	条件
1	$n \le 10^5$	6	$m \le 15000$
2	$n \le 43008$	7	$n \le 14000$
3	$n \le 40000$	8	$n \le 13000$
4	$n \le 30000$	9	$n \le 12750$
5	$n \le 20000$	10	$n \le 12500$

如果有多个条件被满足，取得分最高的那一个。

对于前30%的数据，$0 \le k < 1024$。

时间限制：1s

空间限制：256MB

3 算法介绍

3.1 算法1

注意到$n \le 43008$时每个测试点可以得到2分。

观察$43008 = (32 + 10) \times 1024$，而键有32位，值有10位，所以直接把健和值编码发过去就行了，恰好需要43008个bit，每个测试点可以得到2分。

3.2 算法2

注意到有30%的数据$0 \le k < 1024$，那么我们直接传一个长度为1024的数组a_i表示键为i时的值，长度是$10 \times 1024 = 10240$个bit，可以得到30分。

配合上算法1，并使用传输字串长度来判断数据类型，可以得到44分。

3.3 算法3

$k < 1024$的时候我们只需要10240个bit即可完成传输，那么我们考虑使用hash函数，把$\left[0, 2^{32}\right)$映射到$[0, M)$中，这样就可以使用算法2解决了。

如何选取hash函数呢，我们考虑确定常数a, b使用$ak + b \mod M$取模后的值作为hash值。但是这样有一个问题，多个k可能对应着同一个hash值。对于这种情况，我们可以多随机几次，随机到不重复为止，然后把a, b, m都传过去就行了，编码长度是$64 + 10M$。

然而你会发现这样做基本没有分，因为把n个数映射到$[0, M)$中不重复的概率是$\frac{\binom{M}{n}}{M^n}$，所以要想在时限内随机到一个合理的a, b，M不能取的很小。事实上M需要与n^2同级，而这是不可接受的。

3.4 算法4

算法3的尝试虽然失败了，但是取模给了我们一个好的思路，我们考虑在模意义下传输，想到使用多项式插值和求值来完成encode和decode的过程。

选择一个模数M，然后构造多项式$f(x)$对于每一个(k, v)都有$f(k) = v$，这可以使用拉格朗日插值法来完成。这样传输的时候只需要传输M和模M意义下的$n - 1$次多项式f即可，一共要传输$(n + 1) \times L$个bit，其中L是M的二进制的长度。

但是到算法3中的问题依然存在，可能有多个键值在模M意义下相等。我们考虑多随机几次选取M使得任意两个键值在模意义下不等，M取2^20级别的质数即可，这样传输长度就是$(1024 + 1) \times 20 = 20500$，每个测试点可以得到4分。

事实上我们并不需要把M设那么大，M取18位就可以在时限内随机到可行的取值了，这样可以得到50分，结合算法1，可以得到65分。

3.5 算法5

使用插值算法已经几乎优化到极限了，我们重新开始考虑hash，有没有一种hash方式能够把n个$\left[0, 2^{32}\right)$级别的数映射到[0, 1024)中呢？

我们考虑建立一棵二叉树，叶子节点恰好有1024个，按照dfs序分别对应着0 – 1023，树上每个节点都有一个函数。对于每个键，从根开始由节点上的函数决定走向左孩子还是右孩子，走到叶子节点就能找到对应的值了。

每个节点上的随机函数很好确定，通过固定随机种子生成随机序列按照编号分配给每一个点一个随机权值，然后构造一个概率均匀的二元函数$g(k, s) \to \{0, 1\}$即可，例如$k \times s$二进制中1的个数的奇偶性。

那么需要传输的就只有树的形态，考虑使用`01`对树进行编码，定义空树的编码为`0`，非空二叉树的编码为`1`+左孩子编码+右孩子编码，这样直接dfs就可以构造出树了。

这样做树高的期望是$O(\log n)$的，树上节点个数的期望是$O(n)$的，由于常数原因，能得到50-80分。

3.6 算法6

算法6的主要弊端在于随机时往往不能均分键值。我们可以考虑在每个节点上多随机几次，选择比较均匀的一个随机函数，然后把随机次数传过去就可以了。例如我们可以在每个节点上再开两位存一下随机的次数，这样期望得分50-80。

变长编码

编码长度固定往往不是一个好的选择，因为如果设置的短了，有的时候就会溢出，而设置的足够长又会有相当多的位是浪费的。

我们可以考虑在输出一个k位二进制数的时候，先设定一个初始长度 l_0，然后传输时先输出 $k-1$ 个 `1`，再输出一个 `0`，后面紧跟着一个长度为 $l_0 + k$ 的数，这样就可以传不定长的整数了！

本题中我们要传的是随机次数，我们根据期望随机次数设定 l_0 即可。

3.7 算法7

可以发现，在键已经被分为较小的集合时，对这个集合再建立二叉树会有大量的冗余信息浪费。我们考虑当集合大小不超过8时使用算法3中提到的hash算法，一次性传输随机次数来确定位置，由于集合很小是可以在很少的次数内随机到一个可行函数的。

期望得分80-100分。

4 总结

在如今的信息学竞赛中，各种不同的新类型题目都如雨后春笋般涌现，通讯题是一个典型的例子，以往受评测系统的限制这种类型的题目无法出现在正式比赛中，而如今UOJ等允许自定义评测方式的OJ的出现使得这种类型的题目更加的容易普及。从考察点上来看，通讯类题目主要考察了信息熵方面的知识，对构造性思维也有较高的要求，是一个很有前景的学习与命题方向。