Homework 2: Bloom Filter

赵楷越 522031910803

1 程序运行结果及实验结果数据

一般来说,后续的哈希函数 $H_i(x)$ 可由第一个哈希函数 $H_1(x)$ 简单变化生成。因此,在令后续哈希函数 $H_i(x) = H_1(x + (i-1) \cdot m)$ 时,控制 $\frac{m}{n}$ 与 k 的值分别进行多组实验,记录每组实验的误报率,得到程序运行结果如下图所示:

```
Ubuntu2204@horizon22:/mnt/hofs/share/hw25 g++ -o test main.cc
Ubuntu2204@horizon22:/mnt/hofs/share/hw25 ./test
When k = 1, m = 200, n = 100 False positive rate: 0.50
When k = 1, m = 300, n = 100 False positive rate: 0.25
When k = 1, m = 400, n = 100 False positive rate: 0.21
When k = 1, m = 500, n = 100 False positive rate: 0.27
When k = 2, m = 200, n = 100 False positive rate: 0.27
When k = 2, m = 300, n = 100 False positive rate: 0.26
When k = 2, m = 400, n = 100 False positive rate: 0.17
When k = 2, m = 500, n = 100 False positive rate: 0.17
When k = 3, m = 500, n = 100 False positive rate: 0.50
When k = 3, m = 300, n = 100 False positive rate: 0.50
When k = 3, m = 300, n = 100 False positive rate: 0.22
When k = 3, m = 300, n = 100 False positive rate: 0.18
When k = 3, m = 500, n = 100 False positive rate: 0.31
When k = 4, m = 200, n = 100 False positive rate: 0.32
When k = 4, m = 400, n = 100 False positive rate: 0.32
When k = 4, m = 400, n = 100 False positive rate: 0.16
When k = 4, m = 500, n = 100 False positive rate: 0.10
When k = 5, m = 300, n = 100 False positive rate: 0.10
When k = 5, m = 300, n = 100 False positive rate: 0.10
When k = 5, m = 500, n = 100 False positive rate: 0.10
When k = 5, m = 500, n = 100 False positive rate: 0.10
```

实验结果如下表所示,其中第二列的 k 代表的是理论最优值 $k = ln2 \cdot (\frac{m}{n})$ 的计算值。

m/n	k	k=1	k=2	k=3	k=4	k=5
2	1.39	0.5	0.41	0.5	0.51	0.55
3	2.08	0.35	0.26	0.22	0.32	0.29
4	2.77	0.21	0.17	0.18	0.16	0.18
5	3.46	0.27	0.16	0.07	0.1	0.08

2 对实验结果的简单分析

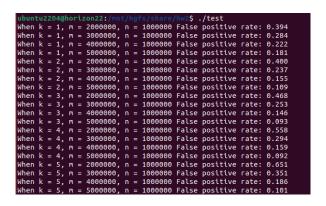
我们可以发现在不同的 $\frac{m}{n}$ 情况下,实验所得的最低误报率的 k 值与理论最优值 k 的计算值接近。比如,当 $\frac{m}{n}=2$ 时,实验所得最优值 k=2,理论最优值为 k=1.39;当 $\frac{m}{n}=3$ 时,实验所得最优值 k=3,理论最优值为 k=2.08,说明实验基本符合理论情况。

其中当 $\frac{m}{n} = 3$ 时,实验所得值与理论最优值有一定差异。简单分析可能的原因是因为:

- 1. 哈希函数取法的不同可能会导致结果的差异,本次实验中令后续哈希函数 $H_i(x) = H_1(x + (i-1) \cdot m)$,如果取了哈希函数不同,实验结果可能不同。
- 2. 本次实验只取了 100 个整数作为插入值, 100 个整数作为测试值, 实验样本数不足, 导致实验值和理论值有一定差异。因此, 我们进行了后续自定义组设置的实验。

3 进一步实验

通过调大数据规模至 n = 1000000 时,重新进行实验,让插入的数据为 0 - 999999,测试集的数据为 1000000 - 1999999,得到程序输出结果如下图所示:



将实验结果制作成为表格,如下表所示:

m/n	k	k=1	k=2	k=3	k=4	k=5
2	1.390	0.394	0.400	0.468	0.558	0.651
3	2.080	0.284	0.237	0.253	0.294	0.351
4	2.770	0.222	0.155	0.146	0.159	0.186
5	3.460	0.181	0.109	0.093	0.092	0.101

与下图中的理论值对比:

m/n	k	<i>k</i> =1	k=2	<i>k</i> =3	k=4	<i>k</i> =5	<i>k</i> =6	<i>k</i> =7	<i>k</i> =8
2	1.39	0.393	0.400						
3	2.08	0.283	0.237	0.253					
4	2.77	0.221	0.155	0.147	0.160				
5	3.46	0.181	0.109	0.092	0.092	0.101			
6	4.16	0.154	0.0804	0.0609	0.0561	0.0578	0.0638		
7	4.85	0.133	0.0618	0.0423	0.0359	0.0347	0.0364		
8	5.55	0.118	0.0489	0.0306	0.024	0.0217	0.0216	0.0229	

发现与所给理论值相差极小,证明在扩大测试数据规模之后,实验与理论效果一致。