物联网信息处理实验三

实验要求：

使用自己所擅长的程序语言编写 Flajolet-Martin 算法(估计数据流中独一无二的元素的个数，即 the number of unique elements seen so far)(其实书本上所描述的算法只是原始的 Flajolet-Martin 算法的简化版本，想详细了解原始 FM 算法的同学，请参看课本 125 页的文献[4])。

实验步骤：

1：以 stream_for_fm.txt 文件为自己所写程序的输入，读取文件中数据(数值范围是 1-2^{25})；

2：请编写一个精确算法，来计算整个文件 stream_for_fm.txt 中有多少个不同的元素(number of unique elements)；[可以通过有序链表来实现，共有 106862 个不同元素]；

3：使用哈希函数：h(x) = a*x + b，其中 a,b 为从整数 1--2^{25}+1 中随机选取的两个整数，x 为 stream_for_fm.txt 文件中的一个数，则此时对于元素 x 来说，a 为 h(x)的二进制形式中尾部的 0 的个数；计算整个文件处理完毕后的 R 值（最大的 a 的值，既所有元素的哈希值中最大的尾长）。输出 2^R 作为元素的个数的估计；

4：使用书上 110 页（组合估计）的技巧，估计元素元素个数。假设有 m 分组，每个分组包含 L 个哈希函数，共有 m*L 个哈希函数，对于每个分组中的每个哈希函数，计算其在文件处理完毕后的 R 值，令其为 Rv[i,j]，则分组 i 中的平局估计值为 R_average[i]=(R[i,1]+R[i,2]+…+R[i,L])；所有分组都计算完毕后，我们可

以得到数组 R_average[1], R_average[2],…,R[m]，对其进行排序后，求出中位数 R_median，令 2^{R_median} 做为最终的元素个数的估计。

5：设真实的元素个数为 N，令 m=1,L=1，重复实验步骤 4，20 次，可以得到 20 个元素个数的估计值 $N_1=2^{R\_median[1]}$，计算平均误差 error_sum $=\{[(N_1-N)^2+(N_2-N)^2+(N_3-N)^2+\ldots++(N_{20}-N)^2]/N\}^{0.5}$;

6：令 m=4,L=4，重复步骤，比较当 m=1,L=1 时的平均误差与 m=4,L=4 的平均误差。

7： 编写 loglog 算法。

实验考察要求：

1：分析并讲解自己所编写程序；

2：分析 m=1,l=1 时和 m=4,l=4 时，平均误差有区别的原因；

3：编写 loglog 算法，比较其与 FM 算法的差异；

4：不允许组队，如果在考察中，发现两人的程序完全相同，则两人本次的实验成绩为 0；

5：本次实验 4 月 29 号检查。

Loglog 算法

1：Initialize M[1],M[2],…M[m] to 0；

2: for each element x read from the file “stream_for_fm.txt”, do the following 3-6:

3：Let h(x) represent the hash value (in binary form) of the element x, let p(y) be the rank of first 1-bit from the right in y（for example, if

y=1100 then p(y)=3, if y=1111, then p(y)=1）;

4: set j = h(x)%m + 1; //*treat h(x) as an integer, then get the bucket id*//

5: set w = floor(h(x)/m); //*remove the $\log_2(m)$ least significant bits in h(x)*//

6: set M[j] = max(M[j],p(w));

7: set s = (M[1]+M[2]+…+M[m])/m; and return E = 0.39701*m*2^s as the estimate for the number of distinct elements in the file;