实验 11 FM Sketch 的实现

一、实验目的

通过 FM Sketch 算法，实现在误差范围内计算数据流中不同元素的个数。

二、实验内容

1、哈希函数 Produce_hash()

```
void Produce_hash(){//产生哈希函数
    cout << "process hash function...." << endl;
    for(int i = 0; i < d; i++){
        for(int j = 0; j < W; j++){//对每一个FM Sketch产生一个哈希函数的系数和常数
            a[i][j] = rand()%big_prime;
            b[i][j] = rand()%big_prime;
        }
    }
}
```

2、更新 FM Sketch 函数 FM_Sketch(int n)

```
void FM_Sketch(int n){
    for(int i = 0; i < d; i++){
        for(int j = 0; j < W; j++){
            //在第[i][j]个FM Sketch中对数字n进行哈希h(n)
            long long int num=(long long int)a[i][j] * (long long int)n%big_prime + (long long int)b[i][j];
            double x = ((double)(num%big_prime)) / ((double)(big_prime));//x的范围为[0,1]
            hashTable[i][j] = min(hashTable[i][j], x);//进行比较，选择小的
        }
    }
}
```

3、计算估计值 Get_Ans()

```
int Get_Ans(){//计算结果
    double mean[d];
    for(int i = 0; i < d; i++){
        double w_mean = 0;
        for(int j = 0; j < W; j++){
            w_mean += hashTable[i][j];
        }
        mean[i]=((double)w_mean)/((double)(W));//求每一行W个FM Sketch的平均数
    }
    sort(mean, mean+d);//排序
    int ans = (int)((double)1 + (double)1 / mean[(d-1)/2]);//求d行的中位数，估计真实值
    return ans;
}
```

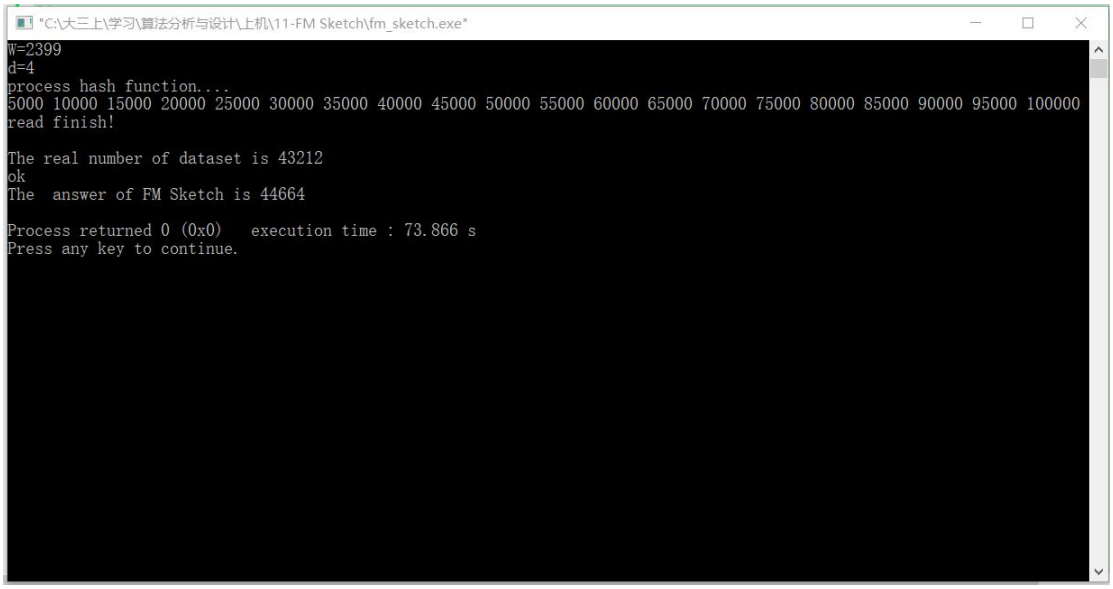
三、实验结果

真实不同元素个数为 43212 个。

1、$\varepsilon = 0.1$，$\delta = 0.01$

可以看到，在这种情况下，FM Sketch 估算的结果为 44664 个。

（44664-43212）/43212=0.03<ε，在误差允许范围内。



2、$\varepsilon = 0.08$，$\delta = 0.005$

当允许的误差稍微降一些后，FM Sketch 估算的结果为 44120 个，更接近真实值。

（44120-43212）/43212=0.02<ε，在误差允许范围内。