物联网信息处理课程设计指导书

下周一晚（10 月 17 日）、下下周一晚（10 月 24 日）

逸夫楼 411 进行实验指导

课程设计项目一：滑动窗口上的均匀采样算法的设计与实现

给定一个文件”data_for_programme1.txt”，其中含有10^6个数据（整数值），令$S=e_1,e_2,...,e_{10^6}$表示这些数据。给定一个滑动窗口大小$n=10$，并假设每个一秒钟到达一个数据，则在第1秒，当前滑动窗口$W=e_1$；在第 2 秒，滑动窗口$W=e_1,e_2$；在第3秒，$W=e_1,e_2,e_3$；……；在第10秒，当前滑动窗口$W=e_1,e_2,...,e_{10}$；在第$10+1$秒，当前滑动窗口$W=e_2,e_2,...,e_{10+1}$（注意，此时，元素e_1过期而从窗口中删除）；在第$10+2$秒，当前滑动窗口$W=e_3,e_4,...,e_{10+2}$；以此类推后面的情况。为方面描述，我们可以令W^t表示窗口在第t秒时所含包含的元素；则显然有：

$$W^t=\begin{cases} e_{t-n+1},e_{t-n+2},...,e_t & \text{当}t\geq n\text{时} \\ e_1,e_2,..,e_t & \text{当t<n时} \end{cases}$$

现在要求编写一个均匀采样程序，该程序有一个输入参数：$l\in\{1,2,3,4,5\}$，要求你所编写的程序维持一个大小为l样本集合A（A是一个只能存储l个整数的数组），使得在任意第$t\geq l$秒，当前窗口W^t中的元素以相等的概率被存储在A中（即被选取到A中）。所设计的程序对于所有的元素$S=e_1,e_2,...,e_{10^6}$只能读取一遍，不能将整个S中的元素都存在一个数组中在进行采样。

课程设计项目要求：

1：分析报告中，证明你所设计的算法可以满足实验要求：即在任意时刻都可以实现均匀的采样当前滑动窗口。

2：显然使用一个大小为n 的数组可以在任意时刻都完整的保存当前窗口中的所有元素，然后可以在数组上实现均匀采样l 个数据就是A 。指出这种方法在实际运用中可能缺陷。

3：输入一个时刻t，输出此时样本集A 平均值估计，比较其于此时当前窗口上的

精确平均值的差异(为得出当前窗口的精确值，你需要使用一个大小为n 的数组来保存窗口在任意时刻的状态)。

4：交正式的课程设计报告，首页上写上：你的名字，学号，课程名称，课程设计题目。

课程设计项目二：滑动窗口上的 Bloom 过滤器的设计与实现

给定一个文件”data_for_programme2.txt”，其中含有10^6个数据（整数值取值范围为$U=\{1,2,..,2*10\wedge 6\}$），令$S=e_1,e_2,...,e_{10^6}$表示这些数据。给定一个滑动窗口大小$n=10$，并假设每个一秒钟到达一个数据，则在第1秒，当前滑动窗口$W=e_1$；在第2秒，滑动窗口$W=e_1,e_2$；在第3秒，$W=e_1,e_2,e_3$；……；在第10秒，当前滑动窗口$W=e_1,e_2,...,e_{10}$；在第$10+1$秒，当前滑动窗口$W=e_2,e_2,...,e_{10+1}$（注意，此时，元素e_1过期而从窗口中删除）；在第$10+2$秒，当前滑动窗口$W=e_3,e_4,...,e_{10+2}$；以此类推后面的情况。为方面描述，我们可以令W^t表示窗口在第t秒时所含包含的元素；则显然有：$W^t=\begin{cases} e_{t-n+1},e_{t-n+2},...,e_t & \text{当}t\geq n\text{时} \\ e_1,e_2,..,e_t & \text{当t<n时} \end{cases}$

显然使用一个大小为n的数组可以完整的保存任意时刻的滑动窗口状态。先要求编写 Bloom 过滤器，使得其可以近似保存任意时刻当前窗口上元素，以近似回答任意一个查询元素是否出现在当前窗口中。更加详细的说，在任意时刻$t\geq n$你所编写的 Bloom 过滤器，未知一个 01 数组 R，基于该数组 R，给定一个查询元素$q\in U$，你的算法可以达到下面的要求：

要求 1:当q的的确确属于W^t（即当前窗口中包含元素q）时，算法需总是正确回答$q\in W^t$；

要求 2：当q的的确确不属于W^t（即当前窗口中不包含元素q）时，你的算法以一定概率可以正确的判断出$q\notin W^t$（允许一定的错误率）。

课程设计项目要求：

1：分析报告中，说明你所设计的算法可以满足实验要求：即在任意时刻都可以满足要求 1 和要求 2。

2：显然使用一个大小为n 的数组可以在任意时刻都完整的保存当前窗口中的所有元素，然后可以基于该数组准确的回答：一个查询元素q 是否属于W^t。请计算你的算法的理论错误率为多少，并与实验错误率做对比。

3：对于你的算法如何进一步改进，减少算法所使用的内存大小？

4：交正式的课程设计报告，首页上写上：你的名字，学号，课程名称，课程设计题目。