哈爾濱工業大學

实验报告

题　目：大数据高级数据结构设计与实践作业一

专　　业　大数据科学与技术

学　　号　1170300916

姓　　名　彭钰驯

课　　程　大数据高级数据结构设计与实践

日　　期　2020-4-15

一、实验要求

任务一：

1) 生成 1 万个随机数 （srand rand 函数 随机数范围 0 到 rand_max）
2) 用 bitmap 表示这 1 万个随机数
3) 将 1 万个随机数放到 bloomfilter 容器中，错误率不高于 0.001
4) 再产生 1000 个随机数，通过 bitmap 和 bloomfilter 判断，检测 bloomfilter 的误判率

任务二：

1） 判断 file2 中 url 总数，设定错误率为 0.001；

2） 将 file2 中的 url 放入到 bloomfilter 容器中；

3） 对 file1 中的每一个 url 进行判断是否在 bloomfilter 容器中。

4） 输出所有找到的 url

二、实验环境

系统环境：Windows10

IDE：Visual Studio

三、人员安排

一人完成

四、实验过程

4.1 bitmap

所谓的 Bit-map 就是用一个 bit 位来标记某个元素对应的 Value， 而 Key 即是该元素。可以理解成一个位图是一个巨大的储存数据的桶，桶的下标表示元素，而每个桶中只保存一个比特位，若为 1 则表示该元素存在，若为 0 则表示该元素不存在。这样可以大大减少使用的空间。可以使用字符串数组来表示所有的桶，一个字符出所占空间是一个字节，8bit。所以通过给定的上下界可以确定桶的大小。除以 8 则是需要申请的字符串数组的大小。所有位均初始化为 0。

4.2 bloomfilter

结合了位图和哈希表的优点，位图的优点是节省空间，但是只能处理整型值一类的问题，无法处理字符串一类的问题。而 Hash 表却恰巧解决了位图无法解决的问题，然而 Hash 太浪费空间。bloomfilter 是一种基于二进制向量和一系列随机函数的数据结构。如果要查找某个元素 item 是否在 S 中，还是通过映射函数{f1,f2.....fk}得到 k 个值{g1,g2.....gk}。然

后再判断 array[g1],array[g2].....array[gk]是否都为 1，若全为 1，则 item 在 S 中，否则 item 不在 S 中。

五、 实验分析

任务一： 在 10000 数据量上，对布隆过滤器的参数进行设置，设置集合大小为 10000。1000 个随机数的监测，错误率为 0。

任务二： 首先浏览 file2 文件，设置布隆过滤器的集合大小为 470000.统计结果是 461281 条 url。并且将所有在 file1 的 url 和在 file2 中进行比对。若在 file2 文件中出现，则进行输出，输出至 count.txt 文件中。大概有 30000 条 url 被统计在内。