《高级算法》课程

实验教学指导书

课程编号：CS32212P

注：1. 前 4 个实验为基础性实验
2. 后两个实验为系统性实验
3. 《高级算法》教学大纲可能变动，担任实验教学的教师或助教应根据教学大纲内容调整本指导书，明确每个实验的目的、内容、方法、验收等环节的要点

实验一

实验名称：用 minHash 实现集合的相似性连接

实验学时：3 学时

一、 实验目的

1. 理解随机算法的概念
2. 体会随机算法的简洁性
3. 体会随机算法的高效性
4. 观察 Hash 函数个数对算法性能的影响
5. 根据实验结果，得出最佳的经验参数设置
6. 规范撰写实验报告

二、 实验内容

分别实现两重循环算法和 minHash 方法求解如下计算问题，利用教师发布在教学 QQ 群中的三个公开实验数据集(AOL, DILICIOUS, LINUX)开展对比实验和扩展性实验，观察 Hash 函数个数对算法性能的影响，并根据实验结果讨论经验参数设置办法。

输入：集族 $R=\{r_1,r_2,\ldots,r_n\}$,实数 $c\in(0,1]$

输出：$\{<r,s>|\ r,s\in R$ 且$|r\cap s|\geq c\}$

三、 实验步骤

1. 实现 Naïve 算法(两重循环计算交集大小)
2. 实现 minHash 算法求解问题，比较两种算法的计算结果是否一致，并解释观察到的现象
3. 设置不用的 Hash 函数个数，比较返回结果的差别和运行效率的差别，总结得出 Hash 函数个数的最佳经验设置公式
4. 更换数据集，验证上述公式的一致性
5. 抽取数据集中不同比例的数据子集开展实验，讨论方法的扩展性
6. 撰写实验报告

四、 数据集

1. 逻辑上每个集合形如 $r_i=\{e_{i1},e_{i2},\ldots,e_{ij}\}$
2. 物理上每个集合在输入文件中被存为若干行(可能不相邻)，每行形如“i,j”，表明第 i 个集合包含了元素 j
3. 在内存中，建议将每个集合 r_i 组织为一个有序数组 $R[i]$，其中 R[i][0]存储集合中的元素个数，$R[i][j]$存储该集合包含的具体元素

五、 要求

真个工程应包含数据加载、Naïve 方法、MinHash 方法等模块，实验在 Main 函数中通过调用相应方法进行。各模块应该独立实现在源代码文件中；数据加载能够通过包含头文件实现代码重用。

六、 验收要点

1. 实现过程是否规范、完整
2. 要求的实验环节是否达成
3. 有无新颖点
4. 实验报告是否规范

实验二（二选一）

选择之一

实验名称：比较三种中位数选取算法的效率
实验学时：3 学时

一、 **实验目的**

1. 通过比较三种中位数选取算法的效率进一步理解随机算法的概念、简洁性和高效性
2. 理解随机算法参数设置的效率和复杂性分析结果的有效性
3. 规范撰写实验报告

二、 **实验内容**

分别三种算法求解如下计算问题。方法 1：先排序后直接抽取；方法 2：《算法设计与分析》第 3 章讲授的线性时间中位数选取算法；方法 3：本课程第 2 章讲授的 lazySelect 随机算法。

输入：实数集合 $R=\{r_1,r_2,\ldots,r_n\}$,正整数 $k\in\{1,2,\ldots,n\}$

输出：$\min(R,k)$—R 中第 k 小的元素

三、 **实验步骤**

1. 实现先排序后直接抽取算法
2. 实现线性时间选取算法
3. lazySelect 随机算法
4. 随机产生服从均匀分布、正态分布、Zipff 分布的数据和均匀选取的 k，开展实验，比较三种算法的性能和扩展性
5. 撰写实验报告

四、 **要求**

真个工程应包含数据生成、排序选择方法、线性时间选取算法，lazySelect 方法，实验在 Main 函数中通过调用相应方法进行。各模块应该独立实现在源代码文件中

五、 **验收要点**

1. 实现过程是否规范、完整
2. 要求的实验环节是否达成
3. 有无新颖点
4. 实验报告是否规范

选择之二

实验名称：QuickSort 算法的再探讨

一、 **实验目的**

1. 进一步理解 QuickSort，找出其问题、改进其代码
2. 理解随机算法复杂性分析结果的有效性
3. 规范撰写实验报告

二、 **实验内容**

严格按照《算法导论》上的伪代码实现 QuickSort 算法，在不同数据集上运行算法、观察实验现象、解释实验现象、改进 quickSort 算法的代码。

输入：整数数组 $A[0:n-1]$

输出：排序后的数组 $A[0:n-1]$

```
QuickSort (A, p, r)
    if p < r
        q = Rand_Partition (A, p, r)
        QuickSort (A, p, q-1)
        QuickSort (A, q+1, r)
Rand_Partition (A, p, r)
    i = Random (p, r)
    exchange A[r] with A[i]
    x = A[r]
    i = p – 1
    for j = p to r – 1
        if A[j] <= x
        i = i + 1
        exchange A[i] with A[j]
    exchange A[i+1] with A[r]
    return i + 1
```

三、 实验步骤

1. 严格按照上述伪代码实现 QuickSort 算法
2. 按要求随机生成 100 万个整数的 11 个数据集，1)数组元素各不相同的无序数组；2）一个元素占整个数组的 10%, 20%, …，100%而其他元素各不相同，运行你实现的 QuickSort，观察实验现象
3. 找出原因，解释观察到的实验现象
4. 写出改进后的算法伪代码，重复实验，检查现象是否消失
5. 调用标准库中的 QuickSort 算法，检查它是否会发生实验中的现象
6. 撰写实验报告

四、 要求

严格按照实验步骤进行，要求代码规范，思想清晰

五、 验收要点

1. 实现过程是否规范、完整
2. 要求的实验环节是否达成
3. 有无新颖点
4. 实验报告是否规范

实验三（三选一）

选择之一

实验名称：随机跳表抽象数据类型实现
实验学时：3 学时

一、 实验目的

1. 理解和体会随机数据结构在数据内存管理中的效用
2. 理解用期望复杂性表示的计算效率结果
3. 理解随机数据结构支持确定型算法的效果
4. 规范撰写实验报告

二、 实验内容

将随机跳表实现为一种抽象数据类型，并用它支持一种应用。

三、 实验步骤

1. 将随机跳表实现为一种抽象数据类型，支持结构初始化、元素插入、元素查找、元素删除、结构销毁等基本操作
2. 用随机跳表实现集族动态倒排索引：给定全集中的元素 e，返回包含 e 的所有集合编号；
3. 用传统倒排索引实现集族动态倒排索引（可以选择其他更有趣的应用）
4. 在实验 1 给出的三个数据集上比较集合插入、删除、查找等操作在两种实现方式的性能
5. 撰写实验报告

四、 要求

集合和集合中的元素都用整数编号，部分编号允许缺失不用；倒排索引需要支持集族动态变化。

五、 验收要点

1. 实现过程是否规范、完整
2. 要求的实验环节是否达成
3. 有无新颖点
4. 实验报告是否规范

选择之二

实验名称：BloomFilter 的抽象数据类型实现

一、 实验目的

1. 理解和体会随机数据结构在数据内存管理中的效用
2. 理解用期望复杂性表示的计算效率结果
3. 理解随机数据结构支持确定型算法的效果
4. 规范撰写实验报告

二、 实验内容

将随机跳表实现为一种抽象数据类型，并用它支持一种应用。

三、 实验步骤

1. 将 BloomFilter 实现为一种抽象数据类型，支持按需初始化、元素插入、元素查找、结构销毁等基本操作

2. 将集合存储为整数数组，支持元素插入、元素查找；
3. 分别用上述两种方法管理黑名单（可以选择其他更有趣的应用）
4. 比较两种方案在插入、查找等操作在两种实现方式的性能
5. 思考你实现的 BloomFilter 能否有效支持集合元素的删除操作
6. 撰写实验报告

四、 要求

1. 代码必须自己实现，不得使用网络上的代码，一旦查实，记 0 分
2. 通过简单的映射管理，黑名单中的每个元素可视为一个整数

五、 验收要点

1. 实现过程是否规范、完整
2. 要求的实验环节是否达成
3. 有无新颖点
4. 实验报告是否规范

选择之三

实验名称：用欺瞒对手技术实现集合包含连接算法

一、 实验目的

1. 进一步理解欺瞒对手技术及其与 Hash 之间的关系
2. 理解 Hash 方法与球和箱子模型之间的关系
3. 比较简单算法与随机算法之间的性能差异
4. 规范撰写实验报告

二、 实验内容

针对下面给出的计算问题实现一个基于两重循环的集合包含连接算法，实现一个基于 Hash 指纹技术的集合包含连接算法，比较算法性能。

输入：集族 $R=\{r_1,r_2,\ldots,r_n\}$

输出：$|\{\langle r,s\rangle | \ r,s\in R$ 且 $r\neq s, r\subseteq s\ \}|$

三、 实验步骤

1. 调用实验一中的数据加载方法加载数据；
2. 实现基于二分查找的包含关系验证算法；
3. 实现基于两重循环+包含验证的集合包含连接算法；
4. 实现一种数字指纹使得 $r\subseteq s \Rightarrow f(r)\,\&\, f(s)=f(r)$
5. 利用 5 中的数字指纹实现一种基于过滤+验证的集合包含连接算法；
6. 在各种参数(如指纹长度)条件下，比较两种算法的性能
7. 撰写实验报告

四、 要求

1. 严格按照实验步骤进行，要求代码规范，思想清晰
2. 两种方法应实现在不同的源代码文件中，能通过头文件实现代码复用

五、 验收要点

1. 实现过程是否规范、完整
2. 要求的实验环节是否达成
3. 有无新颖点
4. 实验报告是否规范

实验四

实验名称：最小生成树期望权值计算

实验学时：3 学时

一、 实验目的

1. 理解蒙特卡罗算法的过程和效用
2. 体会由经验公式得到理论公式的过程
3. 规范撰写实验报告

二、 实验内容

实现一种随机图，在随机图上实现对最小生成树的抽样过程，由抽样过程实现蒙特卡罗方法计算最小生成树权值的数学期望估计，比较估计结果的准确性。

三、 实验步骤

1. 实现算法产生 n 顶点随机图的生成

   输入： n

   输出：一个 n 顶点随机图，任意两个顶点之间边的权值均匀分布于(0,1)
2. 调用第 1 步实现的算法，实现对 n 顶点图的均匀抽样；
3. 在抽样样本上计算最小生成树并计算其权值的数学期望
4. 在第 2 步第 3 步的基础上，建立 n 与最小生成树权值数学期望间的关系；
5. 对 n=16,32,64,128,256,512,1024…展开实验，考察算法运行时间的变化，并检验所建立的关系的一般性
6. 尝试用理论分析解释实验结果
7. 撰写实验报告

四、 要求

1. 严格按照实验步骤中的要求完成实验

五、 验收要点

1. 实现过程是否规范、完整
2. 要求的实验环节是否达成
3. 有无新颖点
4. 实验报告是否规范

实验五

实验名称：大数据抽样

实验学时：6 学时

一、 实验目的

1. 进一步理解抽样过程
2. 学习和掌握一种大数据抽样方法
3. 规范撰写实验报告

二、 实验内容

从下面两篇指定文献中选择一篇进行阅读，掌握其方法、实现其方法、下载论文中涉及的数据及并进行实验。

文献 1： random sample over join revisit， Sigmod 2018

文献 2： Error bounded sampling of annalistic on big sparse data, VLDB2014

文献 3： NeroCard one cardinality estimator for all tables

三、 实验步骤

1. 阅读文献，理解论文中的抽样过程，并理解其分析结论
2. 实现论文中的抽样过程，在论文中讨论的数据上开展实验，验证实验结果与论文的实验部分是否相符；
3. 撰写实验报告

四、 要求

1. 论文阅读过程中，允许大家广泛开展讨论和交流；
2. 方法实现过程中，允许大家广泛开展讨论和交流；
3. 拒绝相互之间拷贝代码，已经查实，本实验按 0 分计算；
4. 对于基础相对薄弱的同学，允许以其他实验替代；

五、验收要点

1. 实现过程是否规范、完整
2. 要求的实验环节是否达成
3. 有无新颖点
4. 实验报告是否规范

实验六

实验名称：专题随机算法实验

实验学时：6 学时

一、 实验目的

1. 检查学生知识面扩充结果
2. 训练系统实现能力和相互协作能力

二、 实验内容

本实验分组进行，每组根据本小组选择的专题随机算法，收集资料、较全面地了解其应用，并实现一种有趣的应用。

三、 实验步骤

1. 学生分组和专题确定（本课程开始时进行）；
2. 各小组成员相互协作，收集资料、整理资料、制作 ppt，选定实验内容；
3. 各组成员相互协作，实现应用；
4. 撰写实验报告

四、 要求

1. 资料收集、整理、报告情况由翻转课堂打分；
2. 各小组成员均应在系统搭建过程中发挥作用，明确分工和作用；
3. 实验验收按照分工进行考察；

五、验收要点

1. 实验内容的趣味性，有用性
2. 实现过程是否规范、完整
3. 要求的实验环节是否达成
4. 有无新颖点
5. 实验报告是否规范

替代性实验

个别同学在完成单人实验的过程中遇到极大困难时，原则上可以选择替代实验来代替指定实验。

实验一：

对于任意给定的离散概率分布 $P(x=x_i)=p_i$ 抽样存在如下两种对比性较强的两种方案：

(1) 产生(0,1)之间的随机数 r，如果 $\sum_{i\le k-1} p_i < r \le \sum_{i\le k} p_i$,则抽取 x_k 作为样本；重复该过程 n 次，得到 n 次抽样结果。

(2) 采用 Alias Table 方法：参见 https://www.keithschwarz.com/darts-dice-coins/

任务一：在某个具体应用中分别根据上述两种方案进行抽样，对比抽样方法的时间性能。

任务二：分析两种抽样方法的时间复杂度。