**《高级算法》课程**

**实验教学指导书**

**课程编号：**CS33604

注：1. 第1,4个实验难度相对高，对非计算机科学与技术专业学生提供选择性选项

2. 第2,3个实验为基础性实验

3. 所有实验课下完成，助教和教师课堂验收实验成果，抽查代码

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**：2学时

1. **实验目的**
2. 理解随机算法的概念
3. 体会随机算法的简洁性
4. 体会随机算法的高效性
5. 观察Hash函数个数对算法性能的影响
6. 根据实验结果，得出最佳的经验参数设置
7. 规范撰写实验报告
8. **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,…,*rn*},实数c∈(0,1]

**输出：{<*r*,*s*>| *r*,*s***∈*R*且|*r*∩*s*|*c***}**

1. **实验步骤**
2. 实现Naïve算法(两重循环计算交集大小)
3. 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
4. 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
5. 更换数据集，验证上述公式的一致性
6. 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
7. 撰写实验报告
8. **数据集**
9. 逻辑上每个集合形如*ri*={*ei*1,*ei*2,…,*eij*}
10. 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*
11. 在内存中，建议将每个集合*ri*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
12. **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。

1. **验收要点**
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

**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**：2学时

1. **实验目的**
2. 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
3. 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
4. 规范撰写实验报告
5. **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,…,*rn*},正整数*k*∈{1,2,…,*n*}

**输出：**min(*R*,*k*)—*R*中第*k*小的元素

1. **实验步骤**
2. 实现先排序后直接抽取算法
3. 实现线性时间选取算法
4. lazySelect随机算法
5. 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
6. 变换随机算法中的关键参数3/4，重复实验，观察算法性能变化
7. 撰写实验报告
8. **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

1. **验收要点**
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

**选择之二(选择此项实验，成绩最高分为8分)**

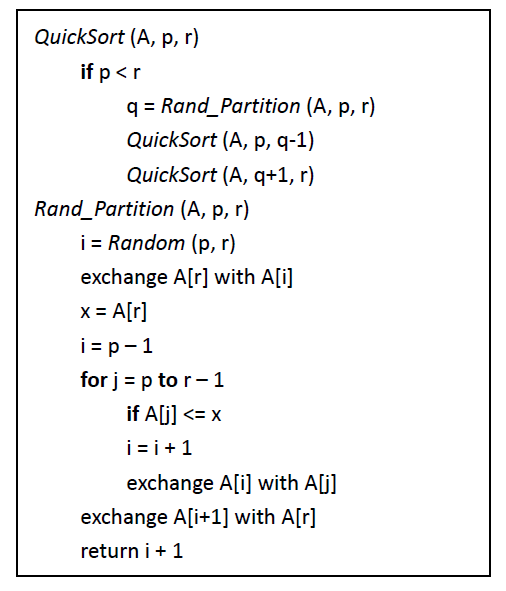
**实验名称：**QuickSort算法的再探讨

1. **实验目的**
2. 进一步理解QuickSort，找出其问题、改进其代码
3. 理解随机算法复杂性分析结果的有效性
4. 规范撰写实验报告
5. **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]



1. **实验步骤**
2. 严格按照上述伪代码实现QuickSort算法
3. 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
4. 找出原因，解释观察到的实验现象
5. 写出改进后的算法伪代码，重复实验，检查现象是否消失
6. 调用标准库中的QuickSort算法，检查它是否会发生实验中的现象
7. 撰写实验报告
8. **要求**

严格按照实验步骤进行，要求代码规范，思想清晰

1. **验收要点**
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

**实验三**

**实验名称：**最小生成树期望权值计算

**实验学时**：2学时

1. **实验目的**
2. 理解蒙特卡罗算法的过程和效用
3. 体会由经验公式得到理论公式的过程
4. 规范撰写实验报告
5. **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。

1. **实验步骤**
2. 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)

1. 调用第1步实现的算法，实现对*n*顶点图的均匀抽样；
2. 在抽样样本上计算最小生成树并计算其权值的数学期望
3. 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；
4. 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
5. 尝试用理论分析解释实验结果
6. 撰写实验报告
7. **要求**
8. 严格按照实验步骤中的要求完成实验
9. **验收要点**
10. 实现过程是否规范、完整
11. 要求的实验环节是否达成
12. 有无新颖点
13. 实验报告是否规范
14. 只完成了实验并撰写实验报告，最高分为8分
15. 分析实验结果，整理规律并进行表述，在8分的基础上加分

**实验四**

**实验名称：**大数据抽样

**实验学时**：2学时

1. **实验目的**
2. 进一步理解抽样过程
3. 学习和掌握一种大数据抽样方法
4. 规范撰写实验报告
5. **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

1. **实验步骤**
2. 阅读文献，理解论文中的抽样过程，并理解其分析结论
3. 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；
4. 撰写实验报告
5. **要求**
6. 论文阅读过程中，允许大家广泛开展讨论和交流；
7. 方法实现过程中，允许大家广泛开展讨论和交流；
8. 拒绝相互之间拷贝代码，一经查实，本实验按0分计算；
9. 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

**替代方案（计算机科学与技术专业学生不得选择替代方案）**

**实验名称：**BloomFilter的抽象数据类型实现

1. **实验目的**
2. 理解和体会随机数据结构在数据内存管理中的效用
3. 理解用期望复杂性表示的计算效率结果
4. 理解随机数据结构支持确定型算法的效果
5. 规范撰写实验报告
6. **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。

1. **实验步骤**
2. 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
3. 将集合存储为整数数组，支持元素插入、元素查找；
4. 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）
5. 比较两种方案在插入、查找等操作在两种实现方式的性能
6. 思考你实现的BloomFilter能否有效支持集合元素的删除操作
7. 撰写实验报告
8. **要求**
9. 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
10. 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
11. **验收要点**
12. 实现过程是否规范、完整
13. 要求的实验环节是否达成
14. 有无新颖点
15. 实验报告是否规范