第12章 数据挖掘（关联规则与分类）

在某超市的销售数据中，发现有如下频繁项集：

{牛奶}，支持度为60%

{面包}，支持度为70%

{牛奶, 面包}，支持度为50%

1. 请写出从中可以导出的所有非平凡的关联规则。
2. 计算每条规则的支持度与置信度。
3. 设最小支持度阈值为40%，最小置信度为70%，指出哪些规则是强关联规则。

第13章 缓冲区管理（替换策略与优化）

考虑以下两种缓冲区替换策略：

LRU（最近最少使用）

基于块访问代价和引用率的替换策略（引用代价比：Cost/Frequency）

（1）请比较它们在以下两个场景下的适用性：

场景A：块大小均等，读取成本相同；

场景B：块大小和读取代价不同（如视频流或网页）。

（2）对于场景B，请给出你设计的替换策略，并说明你考虑了哪些权衡因素。

第十四章 索引

简答题

1. Explain the main difference between a dense index and a sparse index.：

→ 稠密索引为数据文件中的每一个搜索键值都建立一条索引项，而稀疏索引只为部分键值建立索引项，通常是每个数据块一个索引项。

2. Why is a hash index not suitable for range queries?

→ 因为哈希过程打乱了键值的顺序，相关的键值可能被存储在完全不同的哈希桶中，因此无高效地进行范围查询。

3. Describe one situation where using a clustered index is preferable to a non-clustered index.

→ 在执行范围查询或查询经常返回按索引字段排序的大量数据时（例如按入职日期查询所有员工），使用聚集索引比非聚集索引更合适。

4. What happens during an insertion in a B+ tree when a node is full?

→ 当B+树中的某个节点已满时，系统会将该节点分裂，并将中间键提升到父节点。如果父节点也已满，则该操作可能会向上层级级联。

大题

1 B+ 树索引插入与删除分析题

给定一个空的 B+ 树，其阶为 d = 2（即每个内部节点最多有 4 个子节点），请依次插入如下关键字序列：（10, 20, 5, 6, 12, 30, 7, 17）

1. 请画出插入完所有关键字后的 B+ 树结构图。

2. 接着删除关键字 `6` 和 `12`，说明每一步的合并或借用过程，并画出最终的树。

3. 简述 B+ 树相比普通 B 树的优势。

2 索引类型选择与优化策略

假设你是数据库管理员，有一个学生表 `Student(SID, Name, Age, GPA, Major)`，其中数据量为 100 万条记录。你需要为以下查询提供索引优化建议：

1. 查询语句一：

SELECT \* FROM Student WHERE SID = '20251234';

2. 查询语句二：

SELECT \* FROM Student WHERE Age BETWEEN 18 AND 22;

3. 查询语句三：

SELECT \* FROM Student WHERE Name LIKE 'Li%';

\* 应该选择哪种索引（聚集/非聚集、稠密/稀疏、B+ 树/哈希）？并说明原因。

\* 是否能为该表建立多个聚集索引？为什么？

\* 说明聚集索引与辅助索引在维护开销方面的差异。

3

1. 解释哈希索引的基本结构，包括哈希函数、哈希桶（bucket）及溢出处理方法。

2. 为什么哈希索引不适合执行以下查询？请说明理由：

SELECT \* FROM Customer WHERE CustomerID BETWEEN 1000 AND 2000;

4 插入、删除、更新操作如何影响索引结构（B+ 树与哈希分别讨论）？

一、简答题（每题4分，共40分）

1. 简述聚集索引与辅助索引的主要区别。
2. 为什么稀疏索引只能用于有序文件？
3. 什么时候选择使用稠密索引比使用稀疏索引更合适？请说明原因。
4. B树与B+树在结构和查找方面有何不同？
5. 为什么在数据库系统中B+树比B树更常用？
6. 位图索引通常用于什么类型的列？它的优缺点有哪些？
7. 哈希索引适合处理哪类查询？与B+树相比，它的局限性有哪些？
8. 同一张关系表是否可以存在两个不同搜索键的聚集索引？请解释原因。
9. 简述多级索引（multi-level index）的原理与优点。
10. 简述可扩展哈希（extendible hashing）在发生桶溢出时的处理策略。

二、大题（共60分）

1. 【B+树构建与操作】（20分）

已知一个阶数为2的B+树（即每个节点最多4个子节点），请按如下顺序插入键值：

10，20，5，6，12，30，7，17

（1）请画出每一步插入后的B+树变化过程；（10分）

（2）最终构建完成后的B+树结构图；（5分）

（3）在该树中删除键值 6，画出删除后的B+树，并说明处理过程。（5分）

2. 【索引选择分析】（20分）

某张学生表 Student(sid, name, gender, age, major) 中，有如下查询需求：

查询性别为女的所有学生；

查询学号为某个特定值的学生记录；

查询所有年龄在20岁以上的学生；

按专业字段分组统计人数。

（1）分别为上述四个查询推荐合适的索引类型（位图索引、哈希索引、B+树索引、聚集/辅助索引），并说明理由。（每问5分）

3. 【多级索引与磁盘访问分析】（20分）

某数据文件有100万个记录，页大小为4KB，每页可以存放100条记录。

（1）如果为该数据文件创建一个稠密索引，索引项大小为20字节，估计索引文件大小为多少？（5分）

（2）如果使用两级稀疏索引，并且每一页存放200个索引项，估算最多需要访问多少个磁盘块来定位某条记录？（10分）

（3）相比多级索引，使用B+树有何优点？（5分）

一、B+树大题（20分）

某数据库系统中使用阶为2的B+树（每个非根内部节点最多含4个孩子）。请按以下顺序插入键值：15, 5, 25, 10, 30, 20, 35, 40, 1

1. 每次插入后若发生节点分裂，请说明处理过程并画图展示变化。（10分）
2. 插入完成后画出最终的B+树结构图。（5分）
3. 若要查找范围 [10, 30] 内的所有键值，请说明查找过程。（5分）

二、B树大题（20分）

设一棵B树的最小度数 t = 2（即每个节点最多存储 2t-1=3 个键），请按顺序插入下列键值：

20, 10, 30, 5, 15, 25, 35, 40

1. 插入过程中若发生节点分裂，请说明中间键如何上升并画出每步结构。（10分）
2. 插入完成后，删除键值 25，并说明删除过程中的合并或借位操作。（10分）

三、位图索引大题（20分）

某公司员工表包含如下字段：Emp(id, name, gender, dept, salary\_grade)。其中：

gender 只有 M/F 两种取值；

dept 有 A、B、C 三种部门；

salary\_grade 共有 4 档。

数据库管理员为 gender 和 dept 建立了位图索引。

1. 请为 gender 和 dept 字段构造合理的位图结构示意图（以10条记录为例）。
2. 说明如何通过位图索引高效处理如下SQL查询：

SELECT \* FROM Emp WHERE gender = 'F' AND dept = 'B';

1. 位图索引有哪些优势与局限？适用于什么类型的数据？