

## 数据库调优课程研究报告之索引研究

|  |  |
| --- | --- |
| 院 系： | 软件学院 |
| 组 长 学 号： | 19212010054 |
| 组 长： | 吴俊 |
| 小 组 成 员： | 吴俊 |
| 完 成 日 期： | 2019年12 月 9 日 |

## 目录

[数据库调优课程研究报告之索引研究 1](#_Toc23778)

[目录 1](#_Toc30451)

[图目录 2](#_Toc12898)

[表目录 3](#_Toc18778)

[一、 系统概述 4](#_Toc13845)

[二、 系统平台 5](#_Toc29358)

[三、 系统数据库设计 6](#_Toc11638)

[四、 系统索引算法 11](#_Toc147)

[4.1 线性哈希(Linear Hash) 11](#_Toc15693)

[4.2 扩展哈希(Extensible Hash) 13](#_Toc25503)

[4.3 B+树索引(B+ Tree Index) 14](#_Toc7962)

[五、 实验设置 15](#_Toc21898)

[六、 实验过程 16](#_Toc23299)

[七、 结论 17](#_Toc26092)

## 图目录

[图 三-1 总体E-R图 6](#_Toc22224)

[图 三-2 图书E-R图 7](#_Toc12633)

[图 三-3 学生E-R图 7](#_Toc7112)

[图 三-4 教师E-R图 8](#_Toc6842)

[图 四-1 线性哈希算法数据插入流程图 11](#_Toc18170)

[图 四-2 线性哈希算法数据查询流程图 12](#_Toc2486)

[图 四-3 线性哈希算法数据删除流程图 12](#_Toc3853)

[图 四-4 扩展哈希算法数据插入流程图 13](#_Toc19039)

[图 四-5 扩展哈希算法数据查询流程图 14](#_Toc451)

[图 四-6 扩展哈希算法数据删除流程图 14](#_Toc17039)

## 表目录

[表 三-1 图书信息表 8](#_Toc2500)

[表 三-2 学生信息表 9](#_Toc11358)

[表 三-3 教师信息表 9](#_Toc29782)

[表 三-4 图书借阅表 9](#_Toc17866)

[表 三-5 图书归还表 10](#_Toc4855)

# 系统概述

本文设计实现了一个比较不同索引算法性能特点的系统。该系统是由Java开发的Spring Boot Web应用程序，其中线性哈希算法(Linear Hash)，扩展哈希算法(Extensible Hash)以及B+树索引算法(B+Tree Index)均由Java实现，系统实现了每种索引算法的增加、删除和查询数据的方法；系统模拟生成的数据均以设计的表名作为文件名，以csv格式保存在文件中，每个文件代表一张数据库表，文件中的每一行代表一条数据记录，每条数据记录所在的行代表其在磁盘上的位置，以这种方式模拟一个数据库；系统以图表的形式可视化实验过程中不同索引算法在不同实验条件下的性能，不仅可以看出不同索引算法随实验条件改变而变化的整体趋势，而且可以对比索引算法之间的性能差异。

本文设计了三种不同的实验方案，分别是：探索索引种类与数据量的关系，探索索引种类与索引字段类型的关系以及探索不同索引种类的查询范围。在探索索引种类与数据量的关系实验中，系统保持索引字段不变，观察随着插入或删除的数据量的变化，不同索引算法插入，查询和删除数据的用时变化；在探索索引种类与索引字段类型的关系实验中，系统保持每次插入的数据量不变，观察不同索引算法在索引字段变化时的表现；在探索不同索引种类的查询范围实验中，系统保持索引字段和数据量不变，改变查询要求，观察不同索引算法能否完成任务。

# 系统平台

系统开发语言：Java

系统开发工具：Jetbrain IntelliJ Idea

系统开发框架：Spring Boot，Echarts

系统开发环境：

Windows 10 专业版 Intel(R) Core(TM)i5-8250U CPU @1.60GHz 1.80GHz

16.0GB内存，64位操作系统

系统运行环境：

Windows 10 专业版 Intel(R) Core(TM)i5-8250U CPU @1.60GHz 1.80GHz

16.0GB内存，64位操作系统

# 系统数据库设计

本系统设计的关系数据库应用场景是大学图书馆的图书管理系统。系统共设计了五张数据表，分别是图书信息表(Book)，学生信息表(Student)，教师信息表(Teacher)，图书借阅表(BookBorrow)以及图书归还表(BookReturn)。为了方便比较不同索引算法的性能，系统并未将这些表设计到数据库中，而是保存为多个csv文件，一张表对应一个文件名与表名相同的csv文件，数据保存在文件中，文件中的一行代表一条数据记录，数据记录所在的行代表其磁盘地址，以这种方式模拟一个图书管理系统的数据库。

如下为该数据库的E-R图：

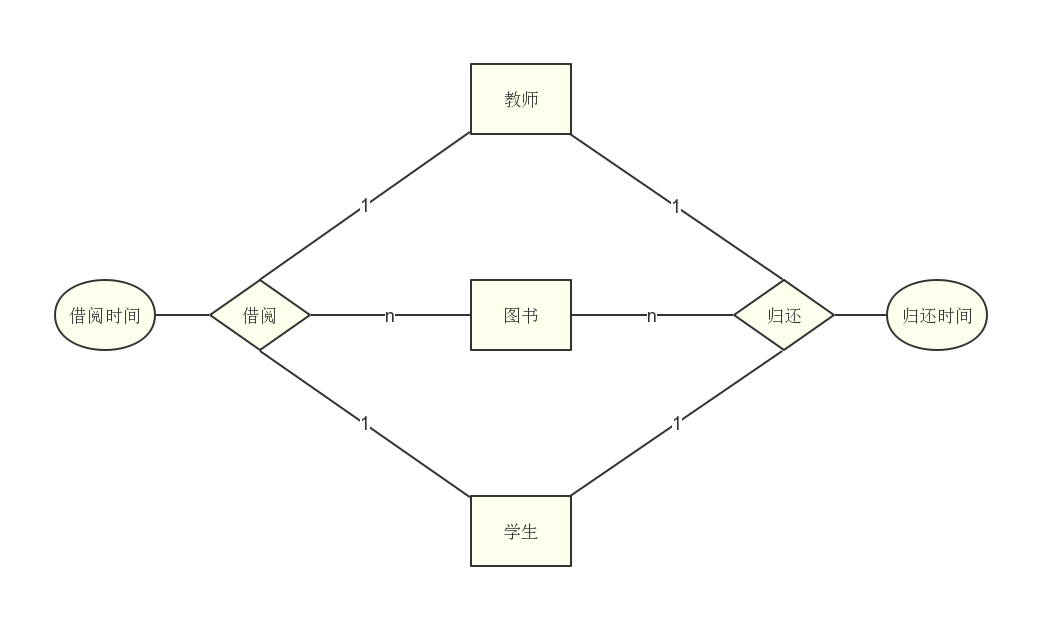


图 三-1总体E-R图

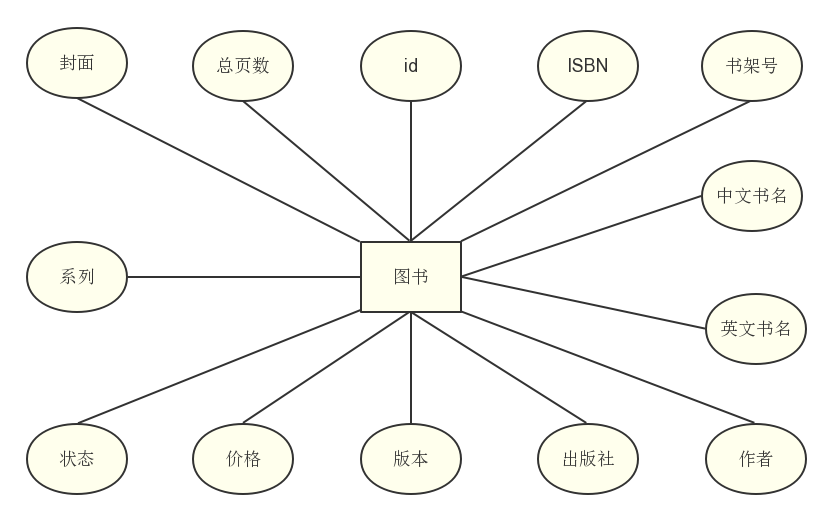


图 三-2图书E-R图

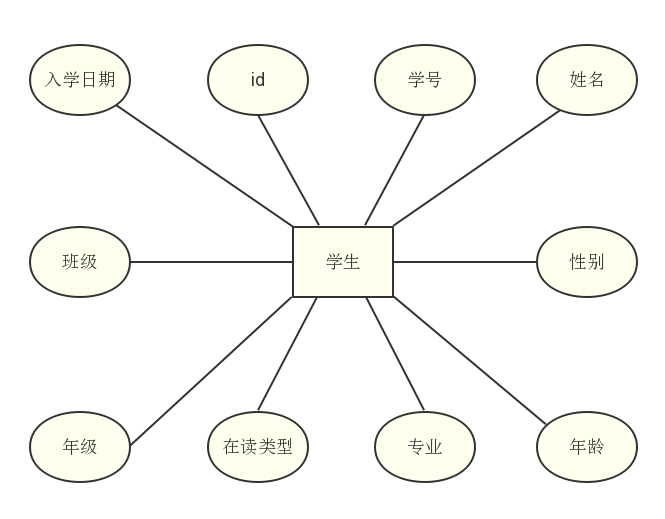


图 三-3学生E-R图

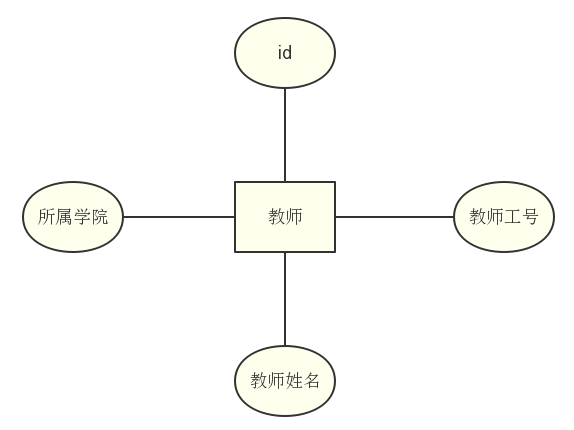


图 三-4教师E-R图

数据表设计如下：

表 三-1图书信息表

|  |  |  |  |
| --- | --- | --- | --- |
| 字段 | 类型 | 约束 | 描述 |
| id | int | primary key | 主键，记录唯一标识 |
| book\_isbn | varchar(20) | not null | 图书ISBN编号 |
| book\_bookcase\_no | varchar(20) | not null | 图书书架号 |
| book\_name\_cn | varchar(30) | not null | 中文图书名 |
| book\_name\_en | varchar(30) | not null | 英文图书名 |
| book\_author | varchar(20) | not null | 图书作者 |
| book\_press | varchar(20) | not null | 出版社 |
| book\_version | int | not null | 图书版本 |
| book\_price | float | not null | 图书单价 |
| book\_series | varchar(30) |  | 图书所属系列 |
| book\_cover | varchar(30) |  | 图书封面 |
| book\_total\_page | int | not null | 图书总页数 |
| book\_status | int | not null | 图书状态，0表示未借出，1表示已借出 |

表 三-2 学生信息表

|  |  |  |  |
| --- | --- | --- | --- |
| 字段 | 类型 | 约束 | 描述 |
| id | int | Primary key | 主键，记录唯一标识 |
| stu\_no | varchar(20) | not null | 学号 |
| stu\_name | varchar(20) | not null | 姓名 |
| stu\_gender | int |  | 性别，0表示未知，1表示男，2表示女 |
| stu\_age | int |  | 年龄 |
| stu\_major | varchar(20) |  | 专业 |
| stu\_type | int |  | 学生类型，0表示本科生，1表示硕士，2表示博士 |
| stu\_grade | int |  | 年级 |
| stu\_class | int |  | 班级 |
| stu\_admission\_date | date |  | 入学日期 |

表 三-3教师信息表

|  |  |  |  |
| --- | --- | --- | --- |
| 字段 | 类型 | 约束 | 描述 |
| id | int | primary key | 主键，记录唯一标识 |
| tea\_no | varchar(20) | not null | 教师工号 |
| tea\_name | varchar(20) | not null | 教师姓名 |
| tea\_college | varchar(30) | not null | 教师所属学院 |

表 三-4图书借阅表

|  |  |  |  |
| --- | --- | --- | --- |
| 字段 | 类型 | 约束 | 描述 |
| id | int | primary key | 主键，记录唯一标识 |
| person\_no | varchar(20) | not null | 人员编号，如学号或教师工号 |
| person\_type | int | not null | 人员类型，0表示学生，1表示教师 |
| book\_id | int | not null | 图书记录id |
| borrow\_time | datetime | not null | 借阅时间 |

表 三-5图书归还表

|  |  |  |  |
| --- | --- | --- | --- |
| 字段 | 类型 | 约束 | 描述 |
| id | int | primary key | 主键，记录唯一标识 |
| person\_no | varchar(20) | not null | 人员编号，如学号或教师工号 |
| person\_type | int | not null | 人员类型，0表示学生，1表示教师 |
| book\_id | int | not null | 图书记录id |
| return\_time | datetime | not null | 归还时间 |

# 系统索引算法

本文主要研究了三种数据库索引算法，分别为线性哈希算法，扩展哈希算法以及B+树索引算法。其中B+树索引算法是目前多种数据库中使用的最多的算法之一。前两种算法均属于哈希算法的范畴，并且是动态哈希算法，而B+树算法是由B树算法改进而来，属于树数据结构算法。在本节中，本文主要介绍三种索引算法的插入，查询和删除数据方法的实现，并用流程图的方式显示算法的原理。

## 4.1 线性哈希(Linear Hash)

线性哈希算法也叫线性散列算法，是由Witold Litwin发明并由Paul Larson推广的一种动态哈希算法。在静态哈希算法中，数据桶的个数是固定的，当发生冲突时，要么利用冲突解决策略化解冲突，要么使用开散列，将冲突的数据插入到对应数据桶的后续溢出桶中，缺点显而易见，当哈希表中的数据越来越多时，每次查询都近似于列表查找。动态哈希算法与之不同之处在于，其哈希算法是随着数据量的改变而动态变化的，并且数据桶的个数也不是固定的，而是可以动态增加的。

下面将结合流程图具体介绍线性哈希算法的数据插入，查询和删除过程。

首先是数据插入过程。

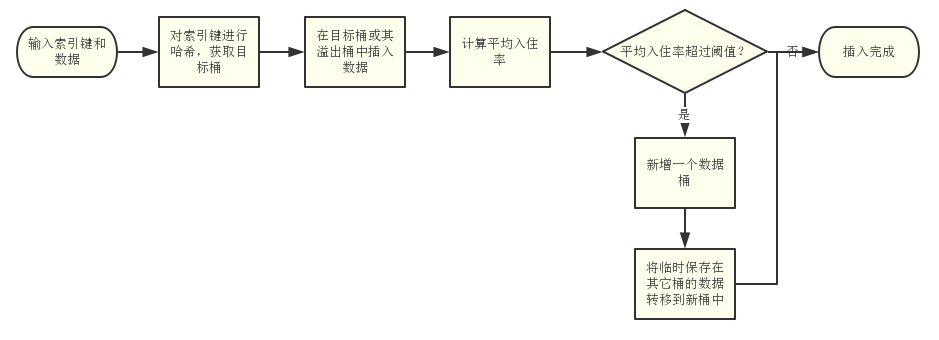


图 四-1线性哈希算法数据插入流程图

在输入索引键和数据后，首先对索引键进行哈希。这里需要注意的是，在将索引键哈希成一个整数后，需要根据目前桶的数量进行取余，取余后的值就是目标桶的下标，然后就可以获得目标桶。

在获得目标桶后，需要找到数据插入的位置，首先判断目标桶是否已经装满，如果没装满则找到一个空位置插入数据，否则在其溢出桶中寻找空位置，然后插入数据。

插入数据后，需要计算平均入住率来判断是否需要新增一个桶。平均入住率的计算方法是：平均入住率=总数据量/(桶的数量 x 桶的容量)。当平均入住率超过设置的阈值后，就需要新增一个桶，然后计算本来应该插入到该新桶的数据被临时保存在哪个桶中，然后遍历该桶，重新哈希该桶中的每个元素，将哈希后应该插入到新桶里的数据转移到新桶里来。插入完毕。

接下来是数据查询过程。

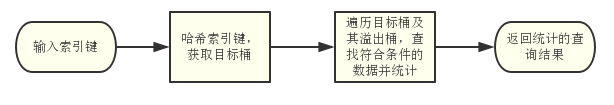


图 四-2线性哈希算法数据查询流程图

在输入索引键后，用哈希算法哈希索引键，获取目标桶的下标，进而获取目标桶。遍历目标桶及其所有溢出桶，统计符合条件的数据，在统计完成后，将统计的数据返回。

最后是数据删除过程。

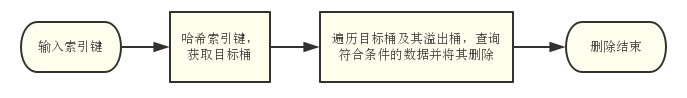


图 四-3线性哈希算法数据删除流程图

数据删除过程和查询的过程类似，只不过在找到符合条件的数据后，会将其删除。

## 4.2 扩展哈希(Extensible Hash)

扩展哈希算法与线性哈希算法类似，但也有区别。相似之处在于两种算法都用数据桶来保存数据，并且哈希函数都会变化，不同之处在于线性哈希算法的哈希函数是随着桶的数量变化而变化，而扩展哈希算法的哈希函数是随着目录的扩展而变化，并且扩展哈希中不存在溢出桶。在扩展哈希算法中，引入了目录的概念，目录中保存了数据桶的地址，不同目录中可能保存着同一个数据桶的地址。目录的每次扩展都是翻倍的，即从一个变为两个，两个变为四个，以此类推。目录的翻倍与桶的分裂都与两个概念有关，即目录的全局深度以及桶的局部深度。因为目录只有一个，因此全局深度也是唯一的，而数据桶的数量可以有很多，因此每一个桶都有其局部深度。全局深度与局部深度的关系决定了是否对目录进行扩展。

下面结合流程图来具体介绍扩展哈希的数据插入，插叙和删除过程。

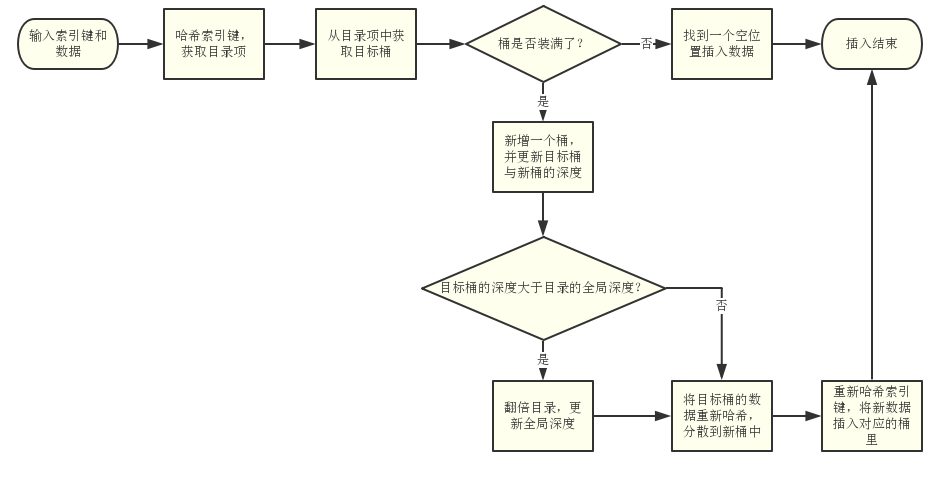


图 四-4扩展哈希算法数据插入流程图

在输入索引键和数据后，根据当前目录的全局深度对索引键进行哈希，获取目录项的下标，然后从目录项中获取目标桶的地址，进而获取目标桶。

获取目标桶后，判断桶是否已经装满，如果没有装满，在找到一个空位置插入数据，否则新建一个桶，将目标桶的局部深度加一，并赋值给新桶。然后判断更新后的桶的局部深度与目录的全局深度的大小关系，如果桶的局部深度大于目录的全局深度，则需要对目录进行翻倍。然后遍历目标桶中的数据，重新哈希，将数据分散到目标桶与新桶中，最后将新数据插入对应的数据桶。插入结束。

接下来是数据查询过程。



图 四-5扩展哈希算法数据查询流程图

在输入索引键后，对其进行哈希，获取目录项，然后从目录项中获取目标桶的地址，进而获取目标桶，最后遍历目标桶，统计符合条件的数据，遍历完成后返回统计的结果。

最后是数据删除的过程。

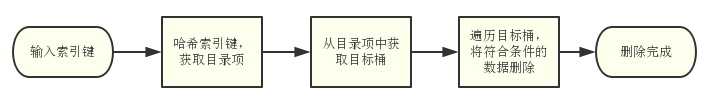


图 四-6扩展哈希算法数据删除流程图

扩展哈希算法的数据删除过程与查询类似，只不过该遍历目标桶时，如果找到了符合条件的数据，则将其删除。这里的删除过程是不完整的，完整的数据删除过程还应包含目录的缩小与桶的合并，本文考虑到目录的缩小与桶的合并需要消耗大量的计算资源，不合并的话只会对查询性能有较小的影响，因此省略了该过程。

## 4.3 B+树索引(B+ Tree Index)

B+树索引算法是由B树索引算法优化改进而来的，与B树的主要不同点在于：B+树的非叶子节点只保存键值信息，不再保存数据；数据记录都保存在叶子节点中；所有叶子节点之间都有一个链指针。B+树基于树数据结构，通常用于数据库和文件系统中。B+树的特点是能够保持数据稳定有序，其插入和修改具有较稳定的对数时间复杂度。B+树与一般二叉树的不同之处在于，B+树的元素是自底向上插入的。一颗m阶B+树的定义如下：

⑴每个节点至多有m-1个关键字和m个子女。

⑵除根节点外，每个节点至少有m/2个子女。

⑶叶子节点按照关键字大小从小到大顺序链接。

B+树中保存了最小的叶子节点的地址，又因为叶子节点是按关键字大小从小到大顺序链接的，因此很容易进行范围查询。

下面结合流程图来讲解B+树算法的数据插入，查询和删除过程。

首先是数据插入过程。

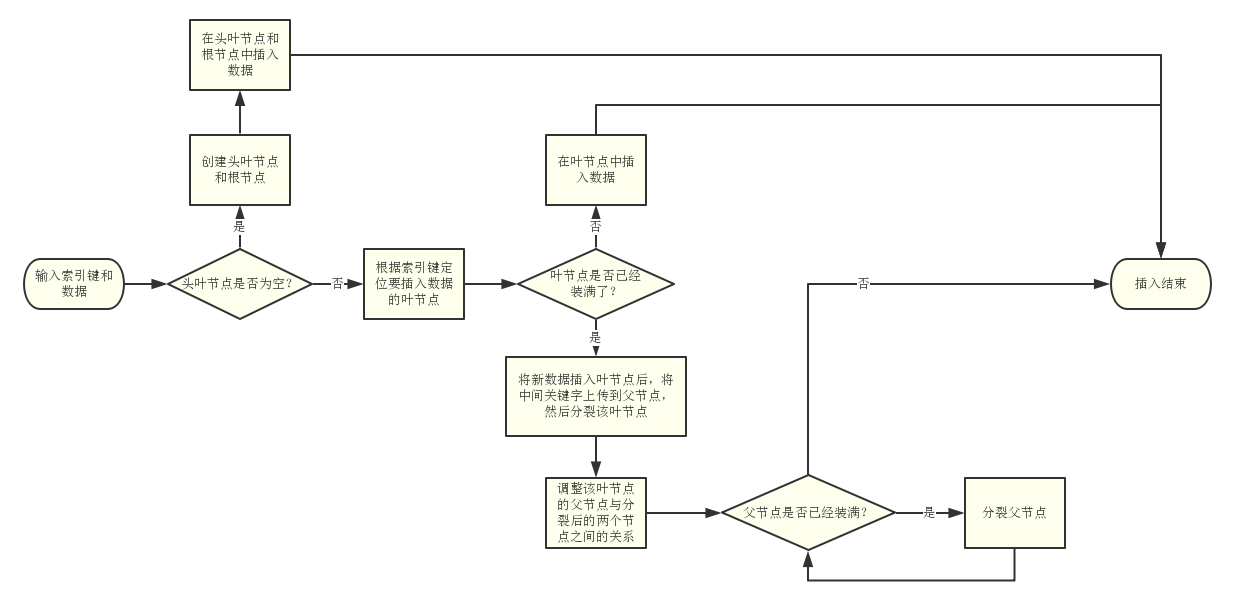


图 四-7 B+树数据插入过程流程图

在输入索引键和数据后，首先判断头叶节点是否是空的，如果是空的说明树也是空的，那么只需要创建头叶节点和根节点，然后把新数据插入到两个节点中就行了，否则需要根据输入的索引键大小关系找到适合插入数据的叶节点。

定位到目标叶节点后，判断该叶节点是否已经装满了，如果没有装满，则直接把新数据插入该叶节点，否则需要对该叶节点进行分裂。

在对叶节点进行分裂之前，先把新数据插入到叶节点中，排序，然后将位于中间的关键字上传到父节点，然后以此中间关键字为边界进行分裂，将中间关键字左边的关键字划分到左节点，将中间关键字右边（包含中间关键字）的关键字划分到右节点。

划分完成后，需要调整原叶节点的父节点与新分裂的左右节点之间的关系，包括将左右节点的父节点指向原叶节点的父节点，以及将左节点的下一个节点指向右节点等等。

完成关系调整后，需要判断父节点是否因为子节点上传了关键字而超载了，如果没有超载，则插入完毕，否则需要递归分裂父节点，直到树结构满足要求为止。

接下来是B+树的数据查询过程。

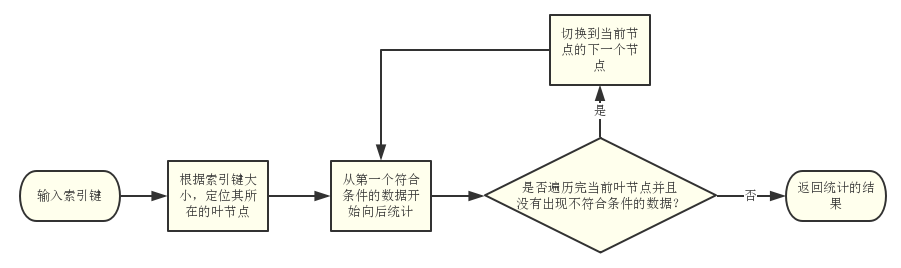


图 四-8 B+树数据查询过程流程图

在输入索引键后，首先根据索引键的大小，从头叶节点开始，找到第一个最大关键字比输入的索引键大的叶节点，因为索引键最有可能第一次出现在该叶节点中。从该叶节点的第一个关键字开始向后搜索索引键的位置，如果是等值查询，则找到第一个相等的关键字后，记录，顺序向后比较，如果遍历完当前叶节点，则进入下一个叶节点，直到出现不相等的关键字为止，返回记录的结果；如果是范围查询，同样在找到第一个比输入的最小关键字大的关键字开始，顺序向后比较并记录，如果遍历完当前叶节点，则进入下一个叶节点，直到关键字的大小比输入的最大关键字还大时结束搜索，返回记录的结果。

最后时B+树数据删除的过程。

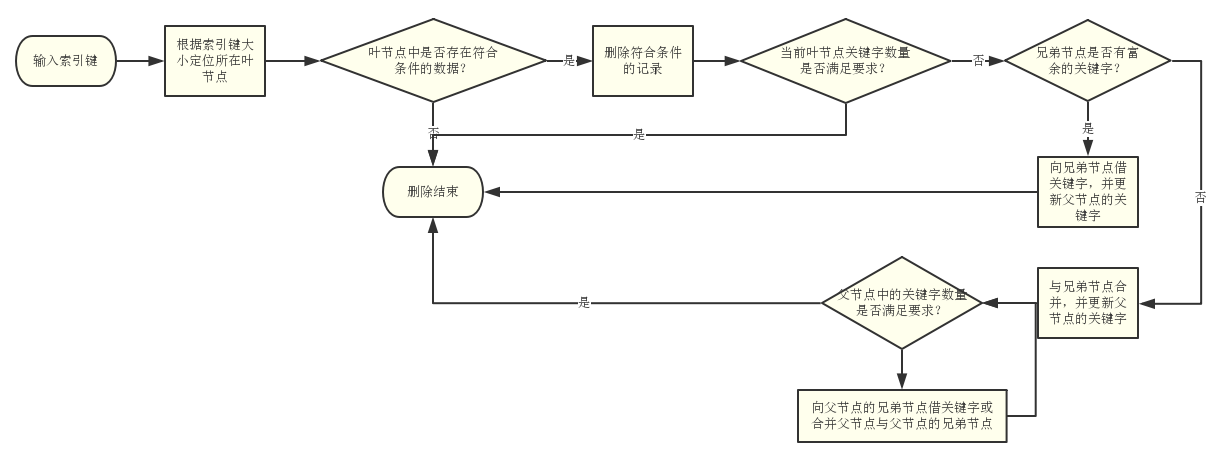


图 四-9 B+树数据删除过程流程图

在输入索引键后，首先找到索引键可能会出现的第一个叶节点，在叶节点中搜索是否存在符合要求的数据，如果没有找到符合要求的数据，则删除结束。如果找到了符合要求的数据，则将其删除，然后判断当前叶节点关键字的数量是否满足最低要求，如果满足，则删除结束，否则需要进一步操作。

如果删除数据后叶节点的关键字数量低于最低要求，则先查看其兄弟节点中是否存在富余的关键字，有富余的关键字的意思是即使从该节点中转移一部分数据，该节点的关键字数量仍然满足最低要求，如果存在富余的关键字，则向其兄弟节点借数据使当前节点达到要求，并更新父节点的关键字。如果兄弟节点没有富余的关键字，则可以将当前节点与其兄弟节点进行合并，并更新父节点的关键字。

在更新父节点的关键字后，父节点的关键字数量可能小于最低要求，这时的做法与叶节点类似，要么从其兄弟节点中借关键字，要么与其兄弟节点合并，然后向上递归，直到树结构满足要求。

# 实验设置

# 实验过程

# 结论