作业 5

SA20225085 朱志儒

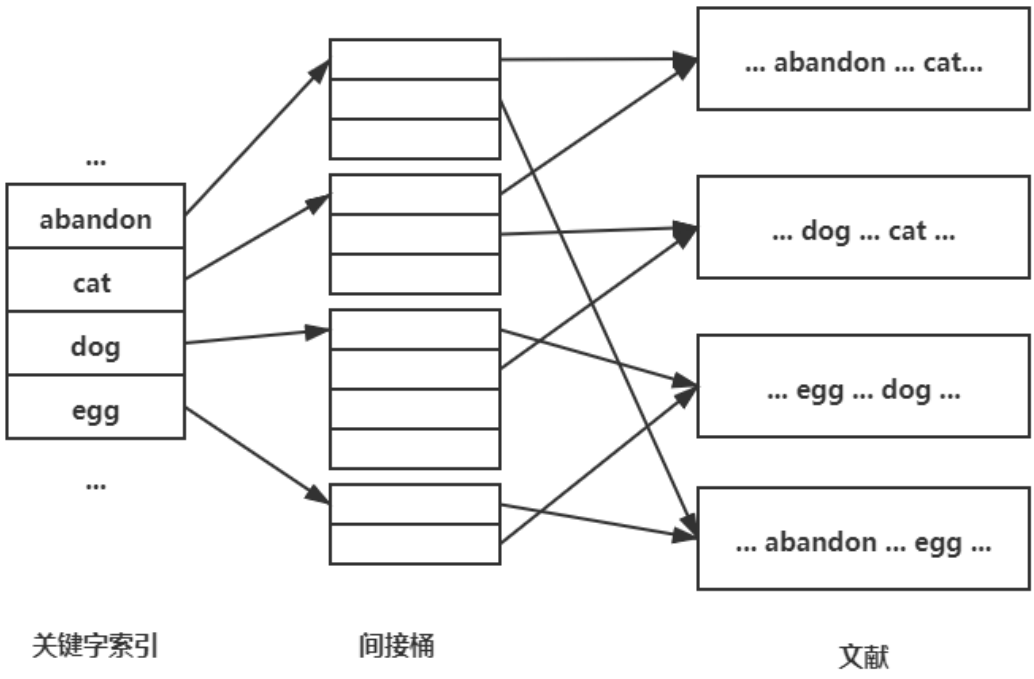
1、 假设我们在数据库中设计了如下基本表来存储文献：paper(id: int, title: varchar(200), abstract: varchar(1000))。最常见的文献查询可以描述为“查询 title 中同时包含给定关键词的文献”，关键词可以是一个，也可以是多个。请回答下面问题（假设所有文献都是英文文献）：

1) 假如在 title 上创建了 B+-tree 索引，能不能提高此查询的效率（须解释理由）？

2) 由于文献 title 的关键词中存在很多重复词语，因此上述文献查询可以借鉴我们课上讲述的支持重复键值的辅助索引技术来进一步优化。请基于此思想画出一种优化的索引结构，简要说明该索引上的记录插入过程以及文献查询过程。

解：（1）在 title 上创建 B+-tree 索引并不能提高此查询的效率，因为 B+-tree 是依据整个 title 的字典序构建的，而最常见的文献查询是依据给定关键词而不是整个 title 进行的，若有多个文献的 title 匹配，可能需要遍历整个 B+-tree，因而不能提高查询效率。

（2）**索引结构：**



插入过程：从 title 中提取关键词，如果该关键词不在关键词索引中，则在关键词索引中加入该关键词索引，新索引指向一个新间接桶，在新间接桶中加入指向该文献的指针。如果该关键词在关键词索引中，则依据该索引找到间接桶，在间接桶中加入指向该文献的指针。

查询过程：如果对单个关键词进行查询，则在关键词索引中搜索该关键词，从而得到该关键词所指向的间接桶，桶中所指向的文献均是包含该关键词的文献。

如果对多个关键词（例如，X 和 Y）进行查询，则先在关键词索引中找到 X 所指向的间接桶，将桶内指向的文献作为一个集合S_X，再在关键词索引中找到 Y 所指向的间接桶，将桶内指向的文献作为一个集合S_Y，则两个集合S_X, S_Y的交集所对应的文献就是包含关键词 X 和 Y 的文献。

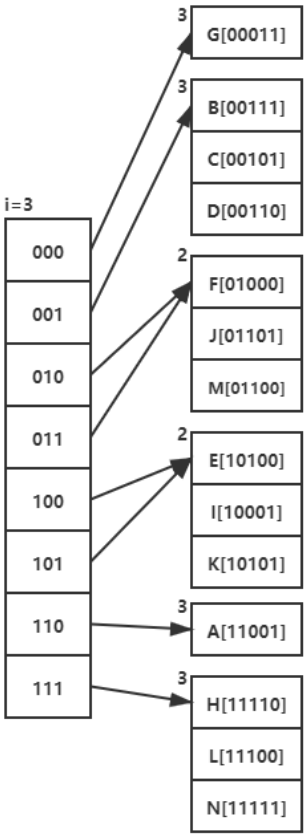
2. 假设有如下的键值，现用 5 位二进制序列来表示每个键值的 hash 值。回答问题：

A [11001] B [00111] C [00101] D [00110] E [10100] F [01000] G [00011] H [11110] I [10001] J [01101] K [10101] L [11100] M [01100] N [11111]

1）如果将上述键值按 A 到 N 的顺序插入到可扩展散列索引中，若每个桶大小为一个磁盘块，每个磁盘块最多可容纳 3 个键值，且初始时散列索引为空，则全部键值插入完成后该散列索引中共有几个桶？并请写出键值 E 所在的桶中的全部键值。

2）前一问题中，如果换成线性散列索引，其余假设不变，同时假设只有当插入新键值后空间利用率大于 80%时才增加新的桶，则全部键值按序插入完成后该散列索引中共有几个桶？并请写出键值 B 所在的桶中的全部键值（包括溢出块中的键值）。

解：（1）全部键值插入后：



显然，该散列索引中共有 6 个桶，键值 E 所在的桶中的全部键值为 E、I、K。

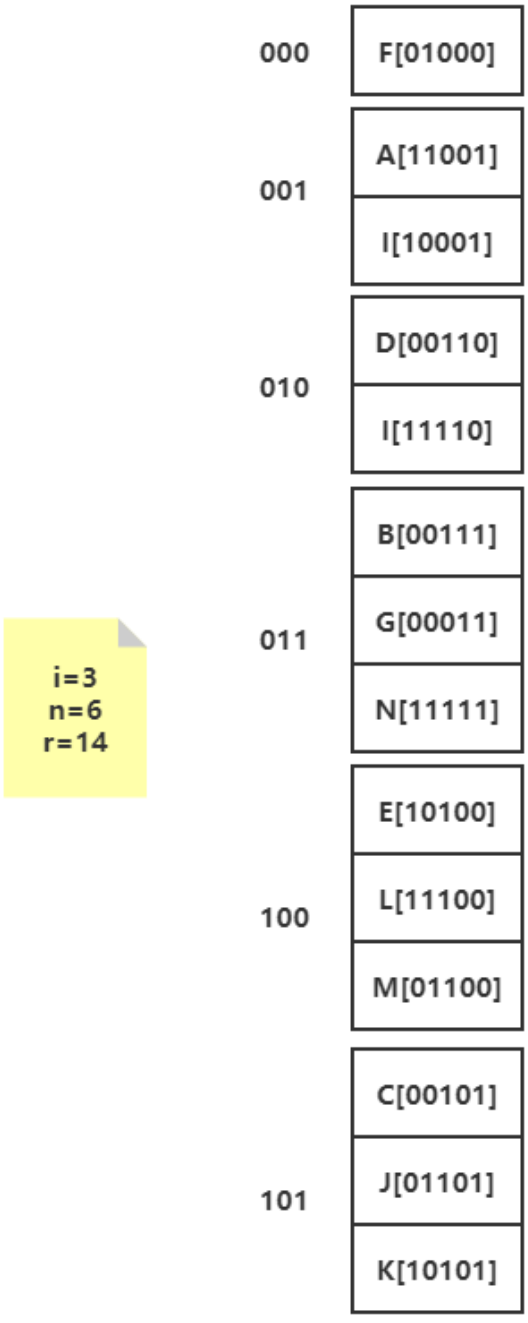
（2）由题可知：

$$\frac{r}{3n} < 0.8$$

即

$$\frac{r}{n} < 2.4$$

全部键值插入后：



显然，该散列索引中共有 6 个桶，键值 B 所在的桶中的全部键值为 B、G、N。

3、 对于 B+树，假设有以下的参数：

参数	含义	参数	含义
N	记录数	S	读取一个磁盘块时的寻道时间
n	B+树的阶，即节点能容纳的键数	T	读取一个磁盘块时的传输时间
R	读取一个磁盘块时的旋转延迟	m	在内存的 m 条记录中查找 1 条记录的时间（线性查找）

假设所有磁盘块都不在内存中。现在我们考虑一种压缩 B+树，即对 B+树的节点键值进行压缩存储。假设每个节点中的键值压缩 1 倍，即每个节点在满的情况下可压缩存储 2n 个压缩前的键值和 2n+1 个指针。额外代价是记录读入内存后必须解压，设每个压缩键值的内存解压时间为 c。给定 N 条记录，现使用压缩 B+树进行索引，请问在一棵满的 n 阶压缩 B+树中查找给定记录地址的时间是多少？（使用表格中的参数表示，n＋1 或 n−1 可近似表示为 n）？

解：压缩 B+-tree 的树高为：

$$\mathrm{h} = \log_{2\mathrm{n}} N$$

读取一个磁盘块的 IO 时间为：

$$\mathrm{S} + \mathrm{R} + \mathrm{T}$$

读取一个节点并查找一条记录的时间为：

$$\mathrm{S} + \mathrm{R} + \mathrm{T} + 2\mathrm{nc} + 2\mathrm{n}$$

查找的总时间为：

$$(S + R + T + 2nc + 2n)\log_{2n} N$$