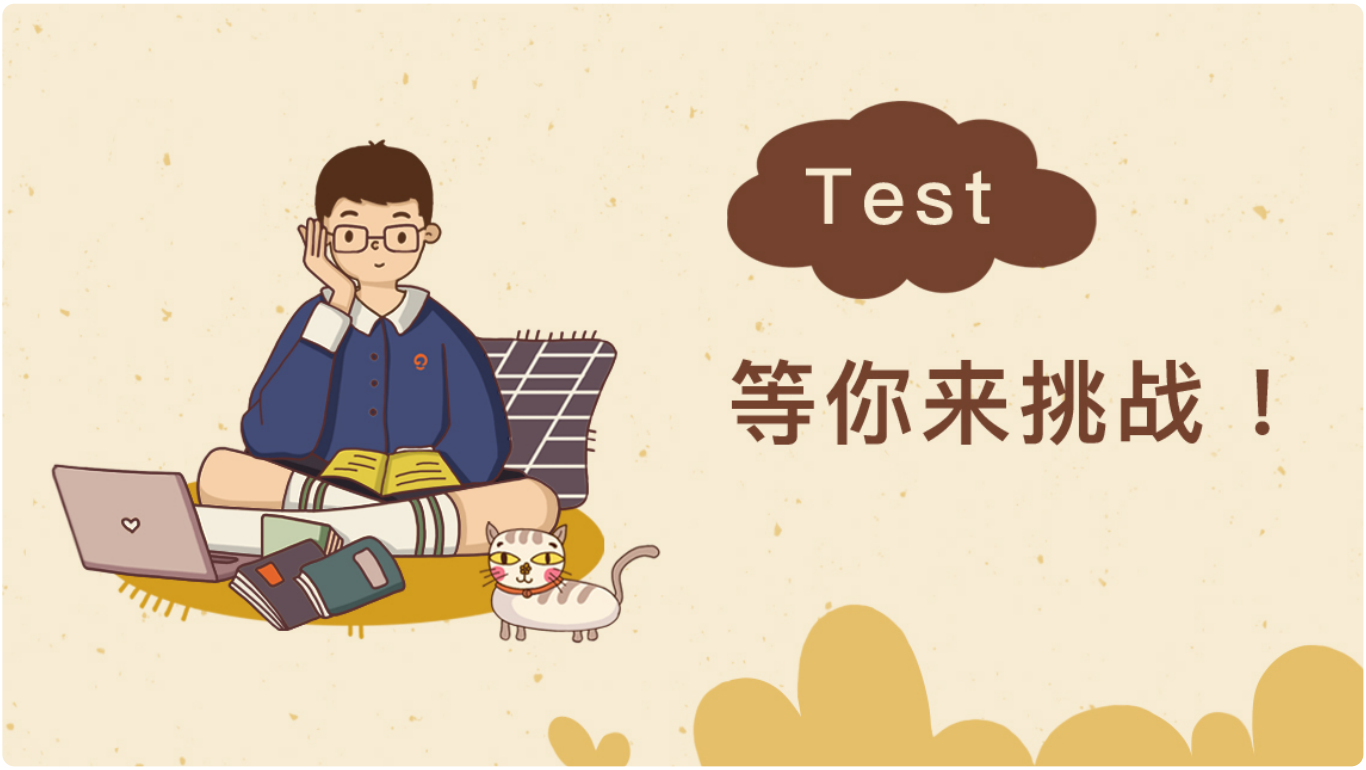
测一测 | 检索算法基础，你掌握了多少？

2020-04-03 陈东

检索技术核心20讲 进入课程 >



讲述：陈东
时长 01:18　大小 1.21M

你好，我是陈东。欢迎来到基础技术篇的测试环节！

经过这几篇的学习，检索相关的基础数据结构和算法，你掌握了多少呢？为了帮助你巩固和复习之前讲到的知识，我精心设计了一套测试题，希望能帮你巩固所学，温故知新。

在这套测试题中，有 20 道选择题，每道题 5 分，满分为 100。这是我们这套测试题最核心的部分。建议你花上 30 分钟，好好完成这套题目。

最后呢，我还为你准备了一道主观题，这道题为选做。 如果你对自己有更高的要求，我希望你可以认真思考一下，然后把你的思考过程和最终答案都写在留言区，我们一起探讨。因

为主观题考察的是你的设计能力，所以你可以多思考几天。我会在下周三把解题思路放到评论区置顶，到时，记得来看啊!

还等什么，点击下面按钮开始测试吧!

戳此答题

主观题

假设有一个员工管理系统，它存储了用户的 ID、姓名、所属部门等信息。如果我们需要它支持以下查询能力：

1. 根据员工 ID 查找员工信息，并支持 ID 的范围查询；

2. 根据姓名查询员工信息；

3. 根据部门查询部门里有哪些员工。

那使用我们在基础篇中学习到的知识，你会怎么设计和实现这些功能呢？（小提示：你可以先想一下，这个员工管理系统是怎么存储员工信息的，然后再来设计这些功能）

上一篇　特别加餐 | 倒排检索加速（二）：如何对联合查询进行加速?

精选留言 (9)

写留言

陈东 置顶
2020-04-08

这是一道开放的设计题，并没有标准答案，但是，我会给你一个参考的解答思路。你可以和你自己的方案进行对比，看看有哪些相同或者不同的地方，这些地方是否合理。下面是具体的解答思路。

这道题中其实有一个隐含的问题：员工信息应该如何存储？由于员工名单本身就是一个...
展开

一步
2020-04-03

一般是这样存储的：
有个用户表：用户Id，用户名称 ,部门id
id, name, departmentId
还有个部门表：部门ID，部门名称

id, name...

展开

作者回复: 第一个问题分析得很好！你基本解释清楚了数据库是如何满足第一个问题的。当然，如果只使用基础篇学到的知识，数组就可以了。

第二个问题你考虑到了模糊查询问题，所以以字为单位建立倒排索引，这也是很好的思考！

1

密码123456

2020-04-07

我以为，我理解了。到了做题的时候，发现我错了。原来我并不是特别理解。

作者回复: 题目都附了讲解，可以再看一看

明翼

2020-04-04

我先不看别人答案给出我的理解，首先员工增减很少，查询更多又要求可以查询id范围，所以我想员工以数组存储，数组是有序的就可以根据ID做二分查找和范围遍历。部门里面有个员工的指针可以指向最小id员工，每个员工也有指针指向下一个ID的员工，将整个部门员工串起来，至于这个链为什么有许，我想如果一个部门员工很多可以改成多层链表。另外建立一个以姓名为key的值为id的hashmap.可以满足根据姓名查询到id.再根据ID做...

展开

作者回复: 很好！有自己的理解和设计。能针对自己理解的场景给出合理的数据结构和解决方案。

最后一句其实很实在，的确如果数据量不大的话，简单粗暴也不失为一种解决方案。

刘凯

2020-04-03

老师，无形的回答中，用户索引文件存成文件是什么样的，我可以理解用数组存用户ID个用用户表行号，第一问？是不是数组的长度等于最大的用户ID，假如ID是整型，数组的下表就是用户ID，这样费时就是O(1) 找到行号？第二问？假如我理解对了，这个数组一什么格式保存到文件中。第三问，他留言中姓名的倒排索引倒排的什么，是怎么保存的，第四个问题，他说的部门倒排怎么实现的，内存中用的什么技术保存的，持久话到文件是怎...

展开

作者回复: 对于无形的回答，我从两方面给你分析:
1.内存中的数据结构是怎么样的
2.磁盘中的文件是怎么样的（其实是怎么将内存数据持久化到磁盘）
由于我们在基础篇中没有讲到磁盘（我在进阶篇会讲到），因此我出这道题的目的是只要考虑内存中的数据结构就好了。不过既然你们说到了持久化，我就一起聊聊。
对于题目基础，他做了一个预处理:假设所有用户信息存在一个“用户文件”中。每一行就是一个用户信息。
对于第一个问题，他是用有序数组实现的，数组中的单元为（用户ID，该信息在文件中的行号），这样可以支持ID查找和范围查找。如果要将这个数组持久化到磁盘，其实可以有很多种处理方式（比如二进制写入磁盘数据块；或者简单点理解，也可以每个数组元素写一行；或者每个元素不就是存着ID和行号么？这些都是数字，你就在文件里写入这些数字，用空格和逗号隔开就可以了；）。这个文件，就叫做“用户索引文件”。
第二个问题他是使用倒排索引完成。倒排索引的key是员工姓名，posting list是员工ID的列表（因为员工可能重名），可以用数组或链表实现。倒排索引也可以保存为一个文件，你可以文件的每一行保存（key+空格+ ID列表）就好了。ID列表中可以用逗号分隔ID。这个就是他说的“倒排索引文件”。
第三个问题他也是用倒排索引，以部门ID或部门名字为key，以员工ID列表为posting list就好了。持久化和第二个问题的方法一样。
因此，在磁盘中，一共会有四个文件，分别是“用户文件”，“用户索引文件”，“姓名倒排索引文件”，“部门倒排索引文件”。这就是他的持久化方案。

2

无形
2020-04-03

需要三个功能，一个是用户基本信息的存储，按ID查找用户信息并支持单位查找，根据姓名和部门查找
以文件实现为例，创建三个文件
1.用户文件
2.用户索引文件...
展开

作者回复: 考虑得很全面！从存储到检索都描述得很清晰。而且还考虑到了姓名的模糊查询。

范闲
2020-04-03

1.员工工号一般都是从零开始增长，可以使用vector。支持随时范围查询和直接索引。新员

工直接pushback，如果有人离职的话就需要数据搬移。
2.姓名可以正派索引，一个姓名可能是不同员工，所以用链式hash不错.负载因子过高的时候可以渐进式rehash
3.部门隐含了层级关系，大部门可能包含子部门。可以用跳表或者多个hash
展开

作者回复: 思考得很细致！整体思路很清晰，还考虑到了部门隐含的层级关系。

峰
2020-04-03

先给偶的答案

一条数据记录大致就是id ，员工姓名，部门，其他信息。。。

由于要支持id范围查询，记录要按id排序，然后id的点查以及考虑到会有更新操作，所以...
展开

作者回复: 你的思考已经非常深入了！我就不点评你的题目答案了，和你聊聊你的问题吧。
1. 是否要学习硬件知识?
随着学科的精细化分工，知识也变得越来越细化，全才是非常少见的。更多的时候，我们是有重点地选择某部分知识进行钻研，然后对于其他领域进行一些了解。
比如说，检索技术的知识导图中，你会看到我就划分了存储介质层，数据结构和算法层，检索专业领域层，还有应用层。对于大部分软件开发工程师而言，对于存储介质，做到了解即可。了解的目的，是要能选择合适的技术方案来搭建对应的系统。
幸运的是，硬件革命性地发展并没有那么快，现在我们常用的存储介质，其实就是内存，磁盘，还有SSD。因此，只要稍微花一些时间，了解一下它们的特点，就能在很长时间内帮助你做合适的设计和决策。
当然，如果能更深入地了解硬件知识，做到软硬件通吃，那么这样的人才，就有可能做出一些突破性的成果。

2. 是否可以有抽象的中间层?
实际上，知识的发展体系，就是一个逐步抽象中间层的过程。比如说，从汇编语言到高级语言，就是一个典型的例子。
对于存储和检索也一样。比如说数据库，其实就是使用SQL语句屏蔽了很多技术细节，至于数据是用倒排索引进行检索的，还是B+树，这些细节都已经帮你屏蔽掉了。
包括现在也有人在研究如何统一关系型数据库和NoSQL，用统一的存储和查询机制来解决。
还有云存储，云计算，本质也是将后端的细节屏蔽掉，让使用者只需要调用put和get就可以获得数据，至于后面是什么存储介质，什么数据结构，并不用关心。

让平台越来越智能化和傻瓜化，搭建越来越多便捷的中间层和平台，就是大量工程师在持续进行的工作。因此，如果我们能具备这样的能力，那么就能在这样的浪潮中找到属于自己的机会。

pedro
2020-04-03

用户ID、姓名、部门是一个存储单位，其中以 ID 作为主键建立有序正排索引，支持范围查询，以姓名建立正排索引支持从姓名查询员工信息，部门建立倒排索引，典型的一个部门多员工的情况，从部门可以查到多个员工。

展开

作者回复: 整体思路没问题。不过可以细化一下，用什么数据结构？怎么支持范围查询？