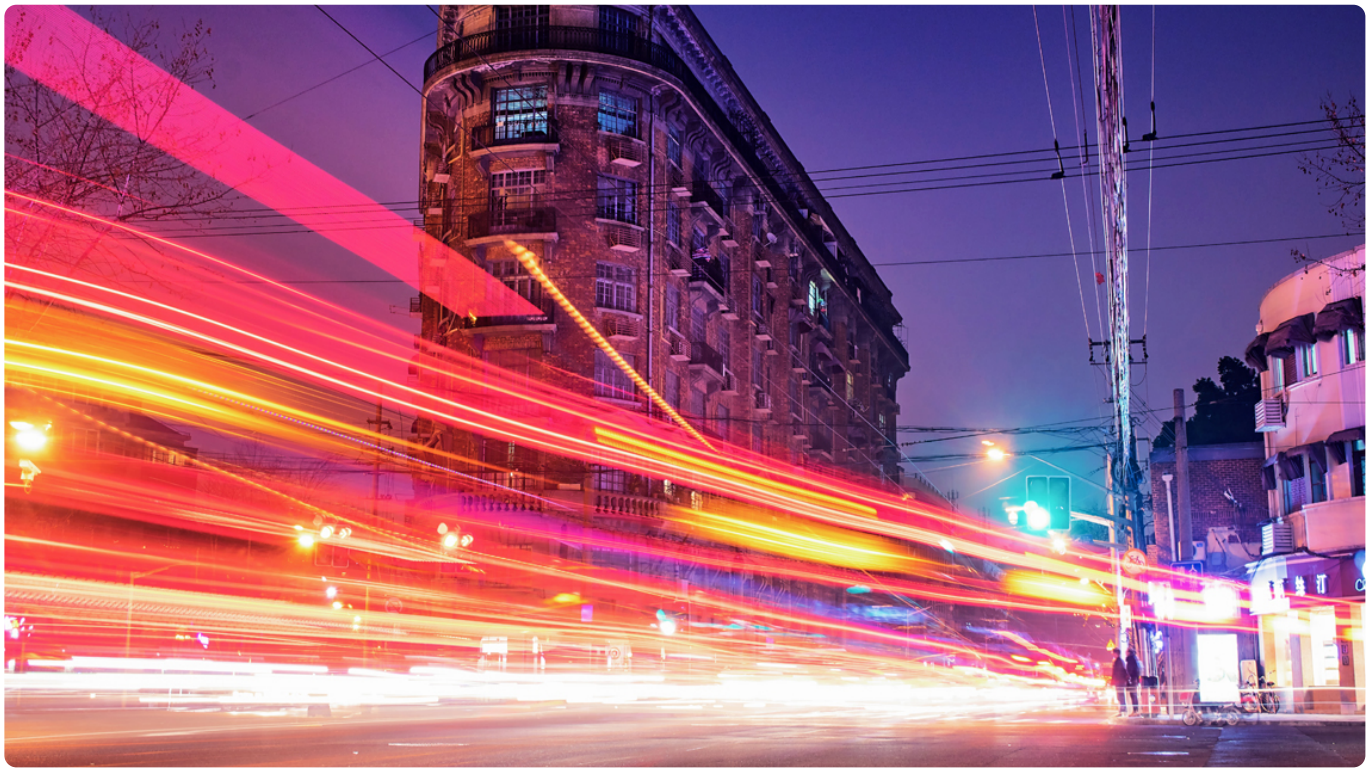
开篇词 | 学会检索，快人一步！

2020-03-23 陈东

检索技术核心20讲 进入课程 >



讲述：陈东

时长 13:39 大小 12.51M

你好，我是陈东，本硕都毕业于北京大学，目前是奇虎 360 商业产品事业部资深总监。

提起检索，我真有挺多话要说，因为这是我从学生时代到职场一路走来，一直都在学习和从事的事情。

我在研究生时期就加入了北大网络实验室。这个实验室当时最有名的一个产品，就是 1997 年发布的北大天网，它是国内第一个基于网页索引的中文搜索引擎。在这里，我接触到了搜索引擎的原理，并开始研究海量数据的存储和检索。

后来，我又参加了 IBM “天才孵化计划”，参与了一个基于地理位置的图片社交创新项目。在这里，我遇到了一个特别棘手的挑战，就是要基于每个用户的位置和他们发过的图片

进行好友推荐。在这个过程中，我要不断地解决各种检索问题，包括地理位置检索、图片检索、相似用户检索等等。

幸运的是，这些问题都围绕着一个核心问题，也就是如何高效检索。所以，我还是在短时间内很好地完成了这个任务。解决棘手的问题，给我带来了成就感。同时，这也让我产生了一个想法，我想要深入去研究“如何在不同的应用中做好检索”。

于是，毕业后，我先后加入了微软广告技术团队、创业公司聚效广告和奇虎 360。尽管这些都属于互联网广告行业，但它们的背后都离不开检索知识和技术的支持。而这么多年打造高性能广告引擎的工作经历，让我对于如何做好检索，以及如何使用检索技术支持业务，有了深入的思考和理解。

为什么要学习检索技术？

说到这，我想问你一个问题：你理解的检索技术是什么？说实话，在专栏刚开始筹备的时候，我就问过很多人这个问题，不少人的第一反应就是：检索就是搜索引擎或者数据库吧。实际上，检索技术的覆盖范围远不止这些。

不管是在数据库和搜索引擎里，还是在新闻推荐平台、电商平台、生活服务平台中，如果我们把这些业务场景抽象出来，它们的本质其实都是在海量的信息中，快速筛选出我们需要的内容或服务，而这都和检索技术紧密相关。也就是说，即便业务形态不同，但是在这些平台的架构设计中，也都有着相似的检索模块实现。

那检索技术到底是什么呢？我们可以用一句话来概括**检索技术：它是更底层的通用技术，它研究的是如何将我们所需的数据高效地取出来。**而如何打造高性能的检索引擎，是这些业务都需要面对的核心问题。甚至我们可以说，提供“更快”“更好”的检索服务，其实就是这些公司的核心竞争力之一。

除了解决业务场景，在我们的实际工作中，“如何快速检索数据”也是最常见的需求。

比如说，无论你是从事底层架构开发，还是业务开发，我相信你都有可能会面对“为啥我的程序运行得这么慢”的问题。比如说，我们打开一些网站或应用的时候，常常会看到“loading”或“数据加载中”的提示，这很大可能是因为，程序从数据库中检索相关数据比较缓慢。如果我们能合理地使用数据库的索引功能，往往能提升好几倍的加载速度，从而大大减少用户的等待时间。

再比如说，在具体写代码的时候，我们会用到一些系统提供的容器，比如 ArrayList、LinkedList。尽管这些容器都提供了“查询是否包含某个元素”的功能，但是它们的性能相差甚远，如果不了解这些容器的检索原理，我们就有可能因为使用不当而导致程序运行缓慢。因此，**了解和使用合适的检索技术，往往能有效提升整个程序的执行效率。**

刚才举的例子都很具体，那我们再从更高的视角，来看看当下我们所处的这个时代。随着 5G 等新技术的普及，我们收集和存储的数据会越来越多。毫无疑问，在这样一个信息爆炸甚至过剩的时代，如何对信息进行高效检索，也将是这个时代必不可少的技能之一。再说回到我们自己的 IT 行业，技术变化之快，我就不必多说了。我相信，掌握好检索技术，能帮助你在行业的快速变化和发展中找到新的机会，让你有更多的机会进入更好的平台，施展自己的才华。

为什么检索技术难学?

说了这么多检索的好处，那说到底，我们怎么才能学好检索技术呢？我曾经想过整理一些材料，给我自己团队的成员进行培训，帮助他们更好地完成工作，但我发现很难找到理想的教材。

我认真了解并且分析之后发现，这主要有两方面原因。

一方面，经典教材大都太过理论化，和实际工作结合不紧密，学习难度太大。比如，《现代信息检索》和《信息检索导论》就是两本非常经典的书籍，但是大部分人都反馈读起它们来比较吃力，书中的内容组织和例子都和我们的实际工作有比较大的差距。

另一方面，和实际工作结合的教材，往往都是从某一行业的视角出发，全面介绍这个行业方方面面的所有技术，而不是专注于某一个基础技术。

比如说，数据库方面的教材，一般会介绍关系模型、SQL 语句、事务处理等内容；搜索引擎方面的教材会介绍爬虫系统、文本挖掘、自然语言处理、网页链接分析等内容，真正涉及检索技术的篇幅并不多。如果以这些书籍为教材，我们根本无法聚焦到检索技术的学习上，难以快速、系统地掌握这门实用的知识。

总结来说，不管是经典教材，还是和实际结合的教材，对读者的知识储备要求都比较高，那无形之中就建立了一个“高门槛”，很多想学习检索技术的人都是被这样的“高门槛”拦在门外的。

于是，我就开始想，那我是不是可以结合我这么多年对检索技术的理解，将我自己多年从事相关工作的经验总结出来呢？于是，经过近半年的精心准备，《检索技术核心 20 讲》这个专栏就诞生了。

专栏是如何设计的？

如果用一句话来概括一下这个专栏的交付目标，那就是，**我想通过这个专栏系统地梳理检索技术的知识，去除冗杂的知识旁支，聚焦于最通用、最核心的检索技术，帮助更多有学习热情、有工作需求的工程师找到学习检索的方法，快速入门、积累经验，解决实际工作中的检索问题。**

所以，对于这个专栏的内容设计，我给自己定了这么几个目标：

1. **聚焦核心知识，帮你全面了解检索技术。**我会将我了解的不同行业、不同系统的检索技术进行提炼，帮助你掌握检索技术的核心知识。我也会将不同行业的检索相关的知识进行体系化的梳理，帮助你更好地将各种检索技术进行横向比较，融会贯通，从而构建起自己的检索知识体系。
2. **注重实用性，帮你解决实际工作中的问题。**我会通过工业界中的实际案例，来详细讲解不同行业会用到的检索技术。这些案例覆盖了多个应用场景和环节，能够解决现阶段你工作中遇到的大部分检索问题，让你能够学以致用。并且，我也不会单纯地讲案例，我更多的会引导你通过这个案例进行更深入的思考，让你在之后的学习和工作中能做到举一反三，解决更多实际问题。
3. **破除“高门槛”，帮你提高学习效率。**首先，学习这个专栏对你基础知识的要求不多，只要你熟悉数组和链表，知道怎么评估时间代价，你就可以学习这个专栏。并且，为了避免枯燥的原理分析，我会少用甚至是不用公式，更多地使用具体的例子以及大量的配图，来帮助你理解检索相关的知识。除此之外，对于每一讲的知识点，我都做了合理的分配和设计。虽然知识的深度在逐步增加，但是跟着我的节奏，相信你依然可以很容易理解和吸收它们。

基于这几个目标，我把整个专栏的核心内容分成三大部分：基础技术篇、进阶实战篇和系统案例篇。

在**基础技术篇**，我会以常见但是核心的数据结构和检索算法作为入门，开启整个专栏的讲解。如果你经验尚浅，那这部分内容可以帮助你打好扎实的基础；如果你有一些实战经验，

那这部分内容能让你站在检索技术的角度，重新审视之前熟悉的数据结构和算法，帮助你构建自己的检索知识体系。

在**进阶实战篇**，我会结合工业界的实际应用场景，更深入地介绍一些高级检索技术，总结一些架构设计的思想，让你能学习到许多工业界的实用且有技术深度的解决方案。如果能深入理解并掌握这部分内容，我相信你会成为各种行业的优秀工程师。

在**系统案例篇**，我会对当前热门的各个方向进行系统分析，比如，存储系统、搜索引擎、广告系统、推荐系统等。从中，你不仅能学到这些行业中是如何应用检索技术的，还可以了解不同行业中检索技术的共同点和不同点。这会帮助你更好地扩展自己的知识面，让你能站在架构师、甚至更高的角度去思考问题和解决问题。

通过学习这个专栏，你不仅能知道这些基础的数据结构在代码级别提升效率的方法，还能够知道在存储系统、搜索引擎、广告系统和推荐系统这些热门架构中，高效率的设计思想以及某些独特环节的技术处理方式，让你对检索技术的理解和使用都能够更上一层楼，从知道“**检索技术是什么**”，到学会“**利用检索技术解决实际问题**”，并且更深入理解“**为什么这么用**”。

专栏马上要开始更新了，这里我还想多说几句。

如果你是一个对检索完全不懂的“新人”，没关系，遇到不懂、不理解的问题，你可以随时给我留言，我会尽我所能给你解答，帮你快速成长；如果你是一个有着多年经验的高级工程师，欢迎你和我分享你工作中遇到的难题，我们共同探讨、一起成长！

最后，我想听你聊一聊，你是怎样理解检索技术的？对于这个专栏你有什么样的期待？欢迎在留言区畅所欲言，我会第一时间给你反馈！

下一篇　导读 | 三步走策略，轻松搞定检索！

精选留言 (16)

写留言

Shelly
2020-03-23

检索就像大脑的提取，我们锻炼大脑的提取速度，也要学习计算机检索的相关知识

> 作者回复: 是的！人的大脑处理数据，其实也是信息提取过程。在第五课中，你就会发现，给你一首古诗的题目，你可以快速背出这首诗；但问你有哪些古诗中包含“极”和“客”字，大脑就很难处理。这其实就会对应到正排索引和倒排索引的检索技术。

2

Kǎfkā²⁰²⁰
2020-03-23

期待

展开

2

夜空中最亮的星（华仔...
2020-03-23

第一时间订阅支持

展开

2

不记年
2020-03-23

数据检索是几乎所有业务系统中使用频率最高的部分，其性能与准确度深刻的影响着业务系统的发展，对检索系统的设计是贯穿前后端的系统性工程

展开

作者回复: 是的。任何系统基本都要支持“增删查改”。其中“查”就是数据检索。

1

无形
2020-03-23

第三时间订阅，老师好帅！！！

展开

1

rookie
2020-03-24

老师，你好，最近也在学习搜索相关的知识，有一个名词叫qanchor，一直不太了解是什么意思，想问一下陈老师您的理解

展开

作者回复: 是anchor吧?

铭毅天下（公众号）
2020-03-24

加入了 非常有必要深入学习底层检索技术了！！

展开

零下一度

2020-03-24

如何搭建公司的搜索引擎系统，希望做到业务数据不受到搜索引擎服务的影响，同时搜索引擎能比较实时提供查询统计功能。

展开

作者回复: 这个具体要看你们的“实时查询统计”的需求到底是怎么样的。如果是简单的一些固定统计，那么elastic search就可以提供；但如果是偏OLAP的灵活分析查询需求，那其实Druid和clickhouse是更合适的选择。
ps:Druid和clickhouse都是基于lsm树实现的。lsm树在进阶实战篇和系统案例篇中我都会介绍。

1

NEVER SETTLE
2020-03-24

我制定自己之后发展的领域是搜索与推荐技术领域，这个专栏再好不错了，希望可以助我一臂之力。

作者回复: 加油!

GEEKBANG_8806808
2020-03-23

不错，条理清晰，核心技术与实际应用结合紧密

展开

aoe
2020-03-23

炫酷

展开

仙女养的🐷
2020-03-23

看到就订阅了，支持

展开

汉
2020-03-23

期待

展开

五河士稻
2020-03-23

期待，加油

展开

追风的沙滩裤
2020-03-23

刚刚入职的公司就是做检索和推荐的，马上课就来了，必须订阅

1

每天晒白牙
2020-03-23

对口

展开