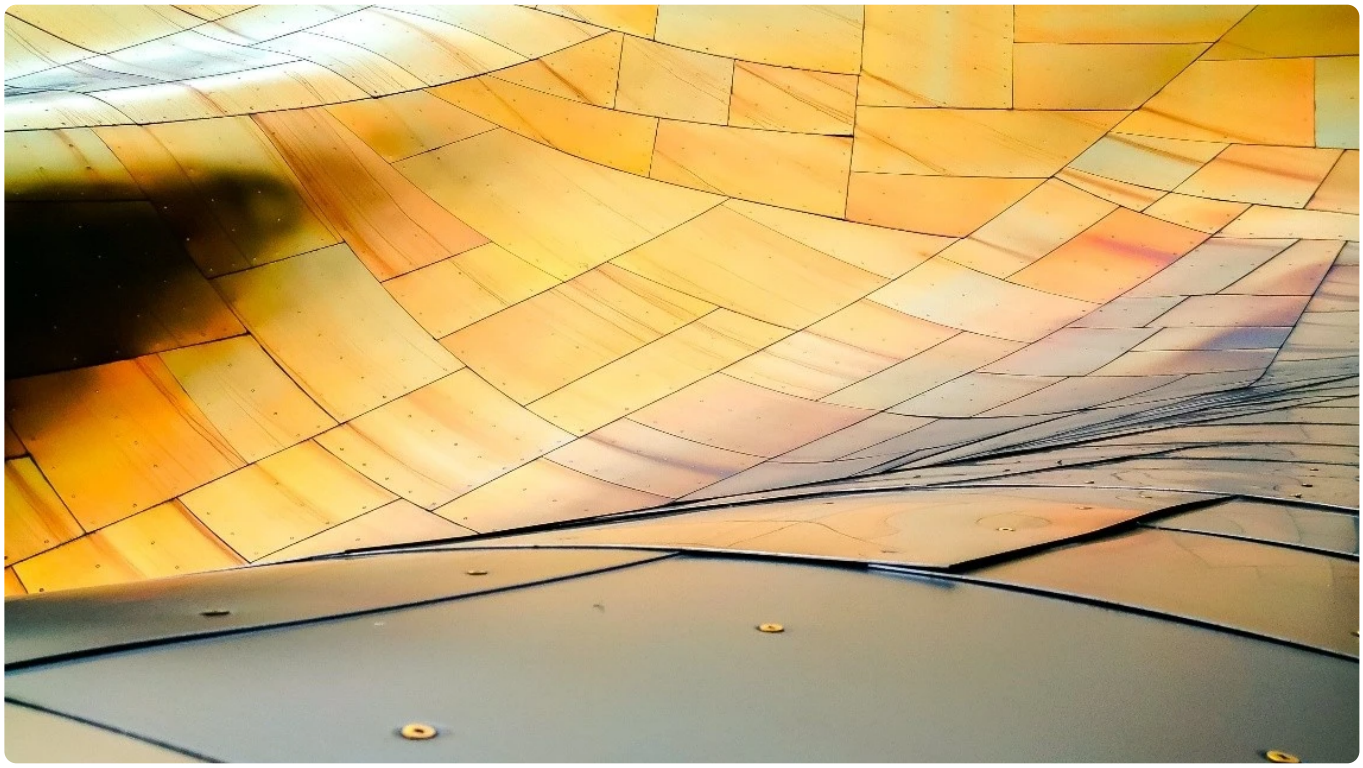
# 加餐1 | 选择和努力同样重要：聊聊如何读论文和选论文

2021-10-18 徐文浩

《大数据经典论文解读》 课程介绍 >



**讲述：徐文浩**

时长 12:39 大小 11.60M

你好，我是徐文浩。在课程刚上线的时候，就有不少同学留言，想要我给出一些应该怎么读论文，以及选择读什么论文的建议。所以今天我们就通过加餐环节，来聊聊这个话题。

在很多传统的科学领域里，比如数学、物理领域，阅读论文往往是科研需要。不过在计算机这个领域之中，即使你不做研究，想要成为一名优秀的工程师，读论文也是必不可少的一个环节。

因为计算机本身是一个非常重视工程的学科，理论到实践的路径很短，往往密不可分。而论文往往就是最新的工程进展，相信你在过去几篇的论文解读过程中，已经感受到这一了。想要能够站在工程研发的第一线，除了读一读这些论文，没有其他太好的资料。从这个角度来说，这些论文**非读不可**。

读论文的方法，其实网上也有不少人分享过。不过，作为工程师来说，我们跟做科研读论文的目标和方法，还是有不少差异的。

两者的共通之处，是都需要通过阅读最新的论文，来理解整个领域最新的进展。但是，科研人员注重在特定领域的深入了解，追求深入，想要寻求新的想法、突破，找到新的研究方向。而作为工程师，更多是学习理解，并和现有的工程系统加以联系，看看是否可以把整个领域最新的进展和自己的工作结合起来。

如果说科研人员追求的是深度，要深入一个领域；那么工程师更多是希望有一定的广度，能够把各个领域联系起来。

所以，对于工程师来说，怎么读论文的确是一个很有价值的话题。今天我就来分享一下，我对如何读论文与选论文的经验和方法。这个话题我会分成两个部分：

**第一部分，是微观层面，针对单篇论文，该如何阅读、理解、消化。**

**第二部分，是宏观层面，即如何去找自己该读哪些论文，怎么有节奏地建立学习路径。**

这节课，我会重点讲一下微观层面，怎么阅读单篇论文，希望你在学完之后，一方面能够学会用系统化、结构化的方法阅读论文学习；另一方面，你也能够找到适合自己的学习方法。后面我会再花一节课的时间，来讲一讲怎么收集论文的阅读材料，通过论文持续性地学习、研究和进步。

## 结构化阅读：如何阅读一篇论文？

我们先来聊聊，如果想要去读一篇论文，应该怎么入手。

读论文不是读小说，不是平铺直叙，从头到尾看一遍就好。而且，有些论文你可能不只要读一遍。所以一个合理的方式，就是**根据不同的重点，分几次读整篇论文**。

### 先读摘要、概述和结论

在开始阅读一篇论文的时候，我认为比较合适的方式，是先去读一下论文的标题、摘要（Abstract）、概述（Introduction）以及对应的结论（Conclusions）部分。**快速阅读这一部分，可以帮助我们确认两个关键点**，一个是知道这篇论文究竟要解决什么样的问题、

解决得怎么样，另一个是理解论文的整体结构，这样后续我们想要深入阅读，也可以做到有的放矢。

就以我们之前讲解过的 Bigtable 为例，在论文的摘要部分，就能一下子吸引当时要做海量数据处理和服务的工程师。它的原文是这么写的：“Bigtable 是一个用来管理结构化数据的分布式存储系统，被设计能伸缩到一个非常大的尺度：PB 级别的数据会分布到上千普通的服务器上。”

> Bigtable is a distributed storage system for managing structure data that is designed to scale to a very large size: petabytes of data across thousands of commodity servers.

而在后面的概述部分，我们能够看到，Bigtable 是一个泛用的系统而不是针对某个特定产品的专用系统，但它并不支持 SQL 模型。在结论部分，我们又能看到 Google 已经开始尝试在 Bigtable 中实现更多的新功能，比如二级索引、跨数据中心的多主数据复制。

> Infrastructure for building cross-data-center replicated Bigtables with multiple master replica.

这些对于论文整体性的介绍内容，可以让我们理解整篇论文的主题和目标，帮助我们决定是不是要深入阅读这篇论文。比如 Bigtable 这样的描述，对于想要做分布式数据库，或者面临 MySQL 集群的可伸缩性和可维护性的挑战的工程师来说，就一定要读下去。

而有些论文，可能内容同样非常好，非常有价值，但是可能和你最近要学习研究的内容关系不大，不一定非常感兴趣，那么你完全可以先放到一边，以后再读。

## 快速阅读全文，做好问题笔记

在确定要读整篇论文之后，我建议你先用比较快速的办法读一遍论文全文。在这个过程中，你会完全对论文讲解的内容有一个基本的理解。

这里我要强调一点，**想要读好论文，一定要做笔记**。论文和我们平时学习的教科书或者是编程语言书不一样，它不是以“教育”为目的，而是以发表创新的研究结果为目的。所以，论文里往往容易有类似于“微言大义”的情况出现。发表的论文的目标读者，往往是

这个研究领域的科研人员或者专家。所以它会假设读者已经有充分的前置知识，很多内容往往就两三句话，甚至是一个关键词。

以我们前面解读过的 GFS 的论文为例，在课程中我专门花了一讲时间重点讲解了数据传输的过程，特别是其中的流水线（Pipeline）式传输的含义。但是在论文原文里，这部分内容不过是小小的一段而已。

所以，你第一遍快速浏览论文的时候，很多细节不清楚是非常正常的。在这个过程中，你只要做好笔记，了解论文里讲解的系统和知识的整体架构就好了。对于很多知识的具体细节，你可以先记录下来，后面再深入研究。

还是以 Bigtable 的论文为例，快速浏览完这篇论文，你对 Bigtable 论文里的数据模型（Section 2）、API（Section 3）、系统组成部分（Section 4）、具体实现（Section 5），进一步完善（Section 6）以及实际的性能评估（Section 7）就有了一个基本概念。

这个时候，你的笔记更多的是一张张卡片，以及一个脑图，记录的是你疑惑的知识点，还有看完论文之后的整体概念性的理解。至于记录笔记用什么工具、什么方法并不重要，纸笔也可以，各种笔记软件也可以。不过我个人会推荐你去读一读《卡片笔记写作法》这本书，里面的方法让我受益匪浅。

## 根据笔记深入阅读，多思考为什么

在快速浏览论文之后，你又要再做一次决策，那就是是不是要深入研读论文。可能你会觉得，快速浏览一遍，有概览性的了解已经足够了。或者，有些论文中，对你来说重要的只是其中某一个知识点，其他部分你已经非常清楚了。

比如在 MapReduce 的论文中，概念逻辑其实都很简单，快速浏览一遍可能大部分知识你就已经清楚了。唯一可能比较困惑的是为什么混洗（Shuffle）的过程中需要排序，那么你只要专门研究一下这个问题就好了。

不过，在这个课程里的论文，都是非常经典的论文，是值得你对论文深入阅读的。所以，在快速浏览了论文之后，我建议你对着之前的笔记，仔细地读一遍。在这个过程中，我希望你**多问自己“为什么”，这个“为什么”会让你找到更多需要深入挖掘、深入学习的点**。

还是以我们前面讲解的 Bigtable 为例，论文里只是告诉你，引导位置（Bootstrap Location）是在 Chubby 里面，而 Tablet 的分布是直接作为一张 METADATA 表，存放在 Bigtable 自己里面。

读完论文你当然知道这个事实是这样了，但是如果你不多问自己“为什么不把这些信息放在 Master 里？”这样的问题，过一段时间，你就很容易忘记 Bigtable 的这个实现方式。也无助于你理解分布式系统应该怎么搭建，以及高可用、性能应该怎么保障这些原理性的东西。

而通过问自己为什么，自己给出答案，或者寻求到合理的答案，能够让你深入消化、理解整个论文。在这个过程中，我也建议你进一步做好笔记。你可以在原先快速浏览的笔记基础上，扩展你的笔记，把问题和答案都写下来。

## 搜索阅读辅助材料，深入弄清楚知识点

但是很多时候，光靠自己给出答案，以及根据论文里原文的描述，并不足以让你完全理解“为什么”。就像我们前面说的，论文很容易“微言大义”，所以对于很多问题的深入研究，需要你**去寻找更多的资料**。

这个时候，我建议你先拿你不理解的关键词，去 Google 里面搜索一下，然后收集一些资料对照着看。这些资料可以是其他人写的文章、PPT、视频、教程等，并且这些资料中可能还会有一些他们推荐的其他阅读资料。其实，这门课程也就是这样的资料。

还以 Bigtable 的论文为例，我们知道了 Tablet 的数据分区方式，那么想要深入学习了解，可能去读一读《数据密集型应用系统设计》里面的数据分区部分，你就能够更清楚地理解 MySQL Cluster 的分区方式、Bigtable 的分区方式，乃至 Cassandra 这样系统的分区方式的差异和优缺点了，你对于“为什么”的这些问题也就有了答案。进一步，你也可以把《Cassandra: a decentralized structured storage system》的论文或者是 PPT 拿来读一下，帮助自己学习和理解这个话题。

其实很多时候，论文里面的一些关键词，就可以帮助你理解很多系统的、数据结构的、组成原理里的各种知识，让计算机科学的各类知识内化到你头脑的知识网络中去。

## 转化成你自己的语言，组织消化论文笔记

最后一步，在你研读完了论文，觉得自己已经掌握和理解了论文之后。我建议你再做一个动作，那就是**在你之前的笔记基础之上，把论文转化成自己的理解和语言**。

其实，我来写这个课，也是这样一个过程。如果你读了论文原文，也看了课程的前几讲，相信你会发现，我并不是按照论文的 Section 一个个翻译讲解过来的，而是按照我对论文中重点的理解，拆解成不同的模块输出出来。

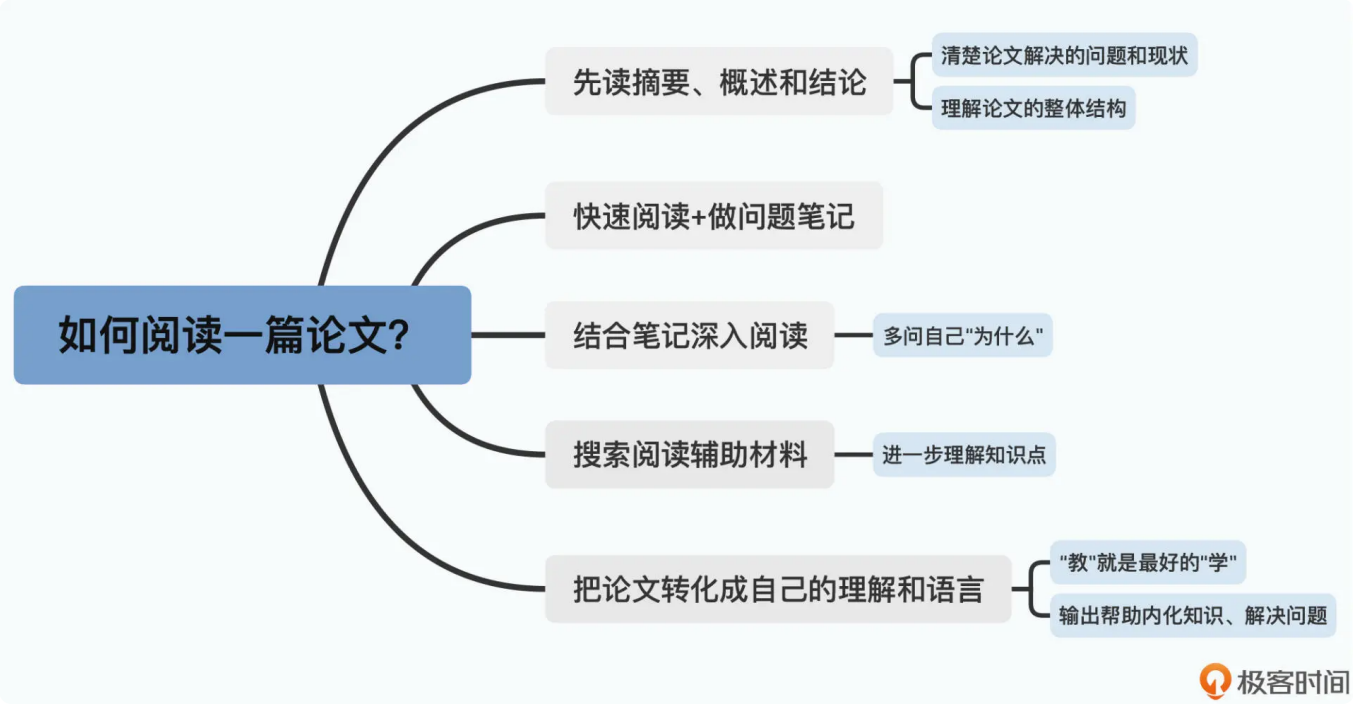
在 GFS 的讲解里，我把它拆解成了简单的设计原则、基于硬件性能的设计方法，以及分布式环境下宽松的一致性三个部分。在 Bigtable 的讲解里，我把它拆分成了具体的技术问题挑战，Bigtable 的整体架构，以及 MemTable+SSTable 的具体实现。

这都是基于我自己的思考和理解来分解论文，重新输出，而这个过程中，对我自己理解分布式系统的设计和实现也有很大的帮助。其实**“教”就是最好的“学”**，把自己理解的问题输出出去，会帮助你真正彻底理解问题，这也就是著名的“费曼学习法”。

## 小结

好了，到你总结输出的阶段，相信你对论文就已经理解得很透彻了。这个时候你会发现，其实在这个阅读论文、深入学习的过程中，你往往要读好几遍，阅读很多辅助材料，并且中间还需要不断地做笔记，而不只是简单看一遍就结束了。

第一遍，往往相当于看个目录和介绍，确定值不值得花时间读。第二遍，则是快速浏览，对论文全文有一个完整的了解。直到第三、第四遍，才是通过查询各类资料深入理解论文。而且在最后，通过整个过程中记录的笔记，整理输出，你才能够真正深刻地理解论文的内容。




希望这些经验，能够对于你未来自主学习有所帮助。

## 推荐阅读

关于如何读论文，很多业内大神也有很多非常好的经验分享。Coursera 这个线上课程平台的创始人，也是斯坦福大学的教授吴恩达老师，专门做过一个如何阅读论文的讲座。这个讲座也被斯坦福大学放上了网，我把Youtube里的链接放在了这里。

也有人根据这个讲座，写了自己的总结笔记，我也把对应的中文和英文的链接放在了这里。

## 思考题

好了，在这一节加餐的最后，我来给你留一道思考题。

在过去的几周里，我们已经讲过了好几篇论文了，你有去读过其中的论文原文吗？在读论文的过程中，你自己找到了什么帮助你有效学习和理解的办法了吗？

赞 0 提建议



上一篇 12 | 分布式锁Chubby（一）：交易之前先签合同



## 精选留言 (1)

写留言

**leslie**

2021-10-18

貌似和针对性强化是通的，基本上一篇论文/东西学下来至少3-5遍；先梗概取简单的关键点，然后通读-这个过程中必然会有一些可能还不错的东西需要扩展，把点记录到小本子或者贴个夹层进去；后面再融汇。

展开