



下载APP



加餐2 | 设置你的学习“母题”：如何选择阅读材料？

2021-12-31 徐文浩

《大数据经典论文解读》

课程介绍 >



讲述：徐文浩

时长 10:02 大小 9.19M



你好，我是徐文浩。到目前为止，我们已经一起学习了大量的论文。而在之前，我也在课程里和你聊过阅读一篇论文的办法。不过，计算机领域的论文浩如烟海，而且每年也不断地有大量的新论文出现。那么，**作为一个工程师，我们该选择哪些论文去研究学习呢？**有不少同学，也在课程刚开始的时候，提出了这样的疑问。

那么，今天我们就通过加餐环节，来聊一聊这个问题。

领资料



我们这一讲的目标，更多针对的是之前没有太多通过研读论文，来拓展自己在技术领域的“宽度”的同学们。如果你已经深入研究某一个领域，并且经常需要精读最新的论文了解学术界的进展，那么这一讲的内容就并不是针对你的学习场景的。



选择学习的“母题”

我们之前说过，作为一个工程师，想要通过论文来学习，大部分情况下的目标往往不是“深”，而是“广”。我们想要的是，能够把领域里的最新进展和自己的工作结合起来。所以，在寻找具体要学习什么论文之前，我们先要确定自己学习研究的“母题”。

“母题”的选择，一方面要和你的工作结合起来，否则即使你花时间、花精力把论文搞懂了，但是没有实践应用的机会，往往过不了多久就忘了。另一方面，又需要有一定的宽泛性，否则往往只是在一些实现细节里面打转，起不到开阔眼界、把各个领域联系起来的作用。

那么，最合理的“母题”的范围，最好是一个你需要解决的“核心问题”，而不是某一个特定的关键字或者某一门特定的课程。比如，我们这个课程里选择的论文，就是围绕着“大数据”这个母题的。**更准确地来说，我们所有选择的论文，都是围绕着“工业界的大数据系统”这个母题的。**

因为，我自己过去所需要解决的问题，就是如何在工程上能够及时处理海量的互联网数据。所以，无论是 MapReduce 这样的大数据批处理系统、Bigtable 这样的分布式 KV 数据库，还是 MillWheel 这样的近实时的流式数据处理系统，都会为我所要解决的问题提供价值。

而在选择论文的时候，我们选择的每一篇论文，都对应着一个实际在工业界中得到长期应用的系统。

如果不围绕我们关注的问题，那我们可以挑选的论文范围一下子会变得非常大。比如，光一个分布式共识问题，我们就有 Paxos、Multi-Paxos、ZAB、Raft、Viewstamped Replication 等问题可以去讲。但是，对于实际的大数据系统应用来说，对 Paxos 有认识，能理解掌握 Raft 算法，对于大数据系统层面的开发，也就足够了。

如果你漫无边际地去找所有和“大数据”这个关键字沾边的论文，那很有可能读了半天，什么具体问题都解决不了。这样，你在学习的过程中会缺少“正反馈”，很难坚持下去。

同时，我也不建议你把你关注的“母题”设置得过于狭隘。在“大数据”领域，你当然可以只去关注“分布式 KV 数据库”这样一个细分的领域。光这个领域，从 Bigtable 开始，你就有 HBase、Cassandra、Dynamo、Voldemort，以及许许多多别的论文和系统可以研究学习。

如果你是专门做分布式 KV 数据库的研发，这些论文你自然需要深入阅读。但是如果只局限在这些内容上，会让你错过整个大领域的发展，无论是 Google Spanner 带来的强一致的分布式关系数据库，还是这两年随着 Dataflow 模型的完善，开始逐步站上历史舞台的实时数据库，你都很难仅仅通过分布式 KV 数据库的论文了解到。

Survey 和书籍是最好的起点

在选定了你要学习的“母题”之后，我给你第一步的建议是，先别着急找一篇某个系统、算法的具体论文来读，而是先从课程、书籍、Survey 入手，对整个领域有一个概括性的认识。

我之前就说过，“论文”因为它本身是发表给业内的其他同行审阅的，往往会“微言大义”，有很多前置知识往往是一笔带过的。在这种情况下，你直接拿一篇论文来“啃”，往往会因为缺少背景知识，让整个学习过程变得效率低下。

所以，你最好是能够找到自己所关注领域的“历史脉络”，而这个历史脉络，往往可以在 Survey 性质的论文、经典的课程和书籍中找到。比如，在我们的这个大数据论文的课程里，我就推荐过《Streaming Systems》《数据密集型系统设计》，以及《Big Data : A Survey》这篇论文。

然后，在深入阅读某一篇论文之前，你可以先快速浏览一下这些阅读材料，确保自己对整个大数据领域有一个整体性的认识，这会大大节约你后续学习的时间。

一方面，这些材料往往更成体系，并且对很多知识点都给出了具体的解释、分类和对比，这样你在后续阅读论文的时候，至少不会遇到太多让你有陌生感的名词。另一方面，在遇到论文里的有些内容，你觉得似是而非、不太理解的时候，你也可以回到这些书籍和课程，做一些对照。

同时，在这些书籍和 Survey 里，还会给出大量引用的论文。这也就给了你一个按图索骥，寻找论文去学习研究的清单。你只需要在 [谷歌学术搜索](#) 里输入论文的名称，很容易就能找到论文的电子版。

不盲目追新，从经典开始

不过，那么多论文，我们应该从哪些论文开始读呢？

我们在学习技术的时候，往往会选择尽量去学习最新的技术。比如，现在做大数据分析，你大概率会学习使用 Spark SQL，而不是原始的 Java MapReduce。

不过，在阅读论文的时候，我给你的建议恰恰相反。**我建议你尽量去读经典的、引用数量高的论文。然后，随着技术变迁的脉络，逐渐往后读到新的论文。**

这是因为，在学习具体要使用的技术的时候，我们的目的是立刻用起来。“老”的技术往往因为有着种种的不足和缺陷，所以自然没有必要花时间去学习研究。但是，在研读论文的时候，我们的目的是搞清楚具体的技术原理和解决问题的思路。这个时候，经典论文的价值就体现出来了。

The screenshot shows the Google Scholar search interface. The search term 'millwheel' is entered in the search bar. The results page displays several academic papers. The first result is 'Millwheel: Fault-tolerant stream processing at internet scale' by T Akidau, A Balikov, K Bekiroğlu, S Chernyak, et al., published in 2013. Annotations highlight the '发表时间' (Publication Time) and '被引用数' (Citation Count) for this paper. Other results include 'The 'mill-wheel' murmur and computed tomography of intracardiac air emboli' by BJ Rubal, A Leon, BL Meyers, et al., published in 2009, and 'Loud Millwheel Murmur Presumably Caused by Air Embolism In a Patient with Pneumoperitoneum' by BO Duboczyk, published in 1954. The sidebar on the left provides filtering options like '时间不限' (No time limit), '按相关性排序' (Sort by relevance), and '包含引用' (Include citations).

在谷歌学术里，你很容易可以看到论文的引用数量和发表时间

在谷歌学术搜索里，你可以直接看到论文的引用数量和发表时间。一般来说，我建议你选取最近 5~15 年，引用数量至少在 100 以上的论文。如果是同一个主题下的论文有很多都满足这个筛选条件，那么优先选择时间早、引用数多的论文。

发表 5 年以上，引用数众多的论文，往往已经经过了时间的检验，说明它的确在这个领域里解决清楚了某一个重要的问题，或者至少是整个领域发展的一个里程碑。而超出 15 年以上的论文，时间就会有些过于久远了，往往其中的知识可以直接在教科书里就能找到，你不一定需要选择去读论文原文。

寻找相关论文，丰富问题视角

在找到对应的经典论文并研读完毕之后，你还可以进一步地阅读从这篇论文延伸出来的其他论文。一般来说，你会遇到这样几种场景：

第一种，是论文需要大量的前置知识。比如我们之前看过的 Spanner 论文里，就会引用到 Bigtable、Paxos、Chubby 等相关内容。这些对应的论文，在你阅读论文最后的参考文献部分，往往都能找到。你可以针对其中你并不理解的主题，进一步深入了解下去。

第二种，和论文相同主题下，有各色各样的其他解决方案。比如我们之前学习过的流式处理系统，除了我们看过的 S4、Storm、MillWheel 和 Dataflow 模型之外，还有 Heron、Naiad、Samza、Flink 等一系列其他的系统。这个，你就可以通过谷歌学术搜索里的相关文章，找到大量高引用的相关论文。

The screenshot shows a Google Scholar search results page for the query "Millwheel". The results list several papers related to stream processing at scale:

- Millwheel: Fault-tolerant stream processing at internet scale** by T Akidau, A Balikov, K Bekiroğlu, S Chernyak... - Proceedings of the ..., 2013 - dl.acm.org. It has 671 citations. A red box highlights the "相关文章" (Related articles) link, which is described in the text as linking to other papers on the same topic.
- Twitter heron: Stream processing at scale** by S Kulkarni, N Bhagat, M Fu, V Kedigehalli... - Proceedings of the ..., 2015 - dl.acm.org. It has 464 citations.
- Storm@ twitter** by A Toshniwal, S Taneja, A Shukla... - Proceedings of the ..., 2014 - dl.acm.org. It has 1179 citations.
- The dataflow model: a practical approach to balancing correctness, latency, and cost in massive-scale, unbounded, out-of-order data processing** by T Akidau, R Bradshaw, C Chambers, S Chernyak... - 2015 - research.google. It has 593 citations.
- Discretized streams: Fault-tolerant streaming computation at scale** by M Zaharia, T Das, H Li, T Hunter, S Shenker... - Proceedings of the ..., 2013 - dl.acm.org. It has 1222 citations.
- Naiad: a timely dataflow system** by D G Murray, F McSherry, R Isaacs, M Isard... - Proceedings of the ..., 2013 - dl.acm.org. It has 880 citations.
- S4: Distributed stream computing platform** by L Neumeyer, B Robbins, A Nair... - 2010 IEEE International ..., 2010 - ieeexplore.ieee.org. It has 1261 citations.

Each result includes a PDF download link and a "我的个人学术档案" (My personal academic archive) button. The interface is in Chinese, with English labels for the search terms and document types.

你也可以通过谷歌学术搜索的相关文章，找到和当前阅读论文同一主题的其他论文

对于这两类论文，我建议你先尽量去读一下第一种。因为这些论文，会对于你掌握理解已经读过的论文更有帮助。而对于第二种论文，我建议你可以做一下泛读，有个了解，或者

也可以先收藏起来。这些论文，往往在你对整个领域全貌有了一定的了解之后，回过头来读，会更高效。

从工业界和学术界汲取养料

除了通过一篇论文衍生开来寻找论文，还有一个好去处，就是各大公司研究院的官方网站。无论是 Google 还是 Facebook，你都可以在他们的网站上，找到以他们公司名义发表的论文。而且，这些论文，都已经分门别类，按照研究领域和发表时间划分好了。

对于 MapReduce 这些论文，你可以在 research.google.com 里的“分布式系统与并行计算”（Distributed Systems and Parallel Computing）里面找到；对于 Hive 这样的论文，你也可以在 research.facebook.com 里的数据库（Databases）这个领域里面找到。

当然，除了 Google、Facebook、微软这样的巨头，很多互联网公司并没有一个专门的研究团队。不过，大部分公司也都会有工程团队的官方博客，比如 Twitter 工程团队的 [官方微博客](#)，也是一个值得追踪的阅读材料。

前面这些，是你从工业界里能够找到的阅读材料。还有一个办法，就是去关注一下学术界的各个会议。

就以这门课程关注的大数据领域为例，你可以去看看像 VLDB 这样每年召开的学术会议发表的论文。我们之前讲解过的 Google 的 Dataflow 的流式计算模型，就是在 [2015 年的 VLDB 会议](#) 上发表的。如果比较经典的论文你都已经读过了，你也对关注的整个领域已经有比较全面的了解了，那么追踪最新会议发表的论文就更适合你了。

小结

好了，相信通过 Survey 和书籍、论文的引用和相关论文，以及追踪各大公司研究团队的官方网站和学术会议，你不会缺少可以阅读的论文了。而通过论文发表的时间和引用数，你也可以很容易地挑选出一些经典论文先去阅读。

希望我今天介绍的这些方法和渠道，能够帮助你快速找出一些值得阅读的论文。

推荐阅读

今天的推荐阅读，我要给你推荐的不是一篇文章，而是一个软件。大部分论文你下载下来都是 PDF 的文件，我一般都会使用 [Mendeley](#) 这个软件，来收集和管理所有我要阅读的论文。并且你也可以直接使用它在论文里面做笔记，网上也有详细的[使用教程](#)。如果你未来会把阅读论文，作为你学习成长中不可或缺的一步的话，那么你可以尽早把它用起来。

思考题

最后，给你留一道小作业。根据今天我介绍的这些渠道，你能不能围绕着大数据这个主题，在我们课程的范围之外，找到一篇你觉得最值得阅读的论文推荐给大家呢？

欢迎留言说说你的方法，也欢迎把课程分享给更多的朋友。

分享给需要的人，Ta订阅后你可得 **20** 元现金奖励

 生成海报并分享

 赞 1  提建议

© 版权归极客邦科技所有，未经许可不得传播售卖。页面已增加防盗追踪，如有侵权极客邦将依法追究其法律责任。

上一篇 [加餐1 | 选择和努力同样重要：聊聊如何读论文和选论文](#)

下一篇 [加餐3 | 我该使用什么样的大数据系统？](#)

更多课程推荐

陈天 · Rust 编程第一课

实战驱动，快速上手 Rust

陈天

Tubi TV 研发副总裁



涨价倒计时

今日订阅 **¥89**，1月12日涨价至**¥199**

精选留言 (2)

写留言



在路上

2021-12-31

徐老师好，我数了下，课程进行到现在，我精读了19篇论文，泛读了11篇论文和材料，这次的加餐非常及时，因为对现在的我来讲，读论文不是难事，筛选合适的论文才困难。

如果要推荐一篇论文的话，我推荐《State Management in Apache Flink》，MillWheel那篇论文讲的是细粒度的状态管理，Flink这篇论文讲的是粗粒度的状态管理，对照来读理解...
展开



2



乐天

2022-01-03

选择母题，寻找经典，扩充背景丰富视角，从机构汲取养料



1