【团队篇】组建推荐团队及工程师的学习路径

2018-05-25 刑无刀

推荐系统三十六式 进入课程 >

讲述：黄洲君
时长 09:59 大小 5.74M

如果你是老板，或者是公司里的推荐系统包工头，那么你一定会关心：要凑齐多少人才能开始搬砖？

一个推荐系统复杂度没有上限，但是有最低标准，所以下面在估算推荐系统团队规模时，按照下限来估计，照这个方式建立的团队就叫做“有下限的团队”。

团队组建

我们先定义团队的角色，这里既然是组建“有下限团队”，当然按照能省则省的原则。

1. 算法工程师，承担的是数据科学家和机器学习工程师的双重职责，主要职责是清洗数据，训练离线推荐模型，开发算法接口，评估指标。

2. 软件开发工程师，承担算法之外的开发任务，例如数据库的搭建维护，API 接口的开发，日志的收集，在线系统的高可用等，当然“有下限的团队”可以适当简陋些，不用考虑高性能。
3. 其他非技术角色，如果是“有下限团队”的话，这一项也可以省略，如果涉及了跨部门合作，或者你不幸被老板提出各种“推得不准”的伪 Bug，那么你就需要一个这样的角色去充当工程师港湾，来阻挡外界的风雨。

接下来，估计一下每个角色的数量。这里主要是估计工程师的数量。

关于算法工程师，最低配需要 2 位，一位三年左右经验的算法工程师负责数据分析，数据清洗，推荐模型训练，评估和上线，外带一位三年以下经验的初级工程师，从中辅助分担琐碎工作。

为什么说有经验的算法工程师一位就够了，假如你使用矩阵分解作为推荐系统第一版核心算法，那么推荐使用 Quora 开源的 QMF 工具。它能在一台 32 核、244G、640G 固态硬盘的服务器上用 20 分钟完成 10 亿非零元素，千万用户和物品的矩阵分解。工具简单易用，一个有经验的工程师足够让其运转起来。

那么核心问题就是，一台机器是不是撑得起你老板的野心？我认为，撑得起，具体的估算如下。

根据我前文中对注意力的定义：内容消耗的加速度乘以内容的消耗难度。当注意力为正数时，是上马推荐系统的好时机。

因为这说明平台方已经有了注意力的原始积累，只需要加上推荐系统将它保存下来并加以扩大即可。那么组建的这个“有下限团队”最低要求就是能留住当前的注意力。

注意力为正时，每天的用户消耗内容数量应该是指数级别，比如 $f(a,t) = t^a$ (a >= 2)。其中 t 是时间，a 大于 2 时才会有正的注意力。因为它的二阶导数为：$a(a-1)t^{(a-2)}$。当然，这个数学推导不重要，只是举个例子。

每天的内容消耗量，其实就是用户产生的行为数据条数，至少是正比关系，这里从简考虑，认为二者等同。假如 a=2，t 的单位是天。那么在 t 天后，累计产生的日志数量是：

$$\sum_{i=1}^{T} i^2 = \frac{T(T+1)(2T+1)}{6}$$

现在看看，如果你公司使用的服务器和 Quora 评测 QMF 时所用服务器一样的配置，用单机运行 QMF，极限是撑多久？我简单列个方程。

$$\frac{T(T+1)(2T+1)}{6} = 6000000000$$

方程右边就是 QMF 评测时处理的 60 亿非零元素。解这个一元三次方程，得到唯一的实数解是 1441.75 天，也就是 3.9 年。

所以告诉你的老板，你一个人可以撑四年，只管定期加工资就可以了，不用加人。

那么为什么明明一位算法工程师就可以，还要外带一位，这主要是考虑，团队人才应该有梯度和备份。

关于软件开发工程师，至少需要四位，是的，你猜到了，我要证明的是需要两位，还有两位是为了人才梯度和冗余备份。分工是这样的。

1. 推荐服务输出，一位三年及以上经验的后端开发工程师，外带一位三年以下的初级工程师。负责调用推荐 RPC 服务，开发必要的过滤逻辑，填充详细字段等；
2. 反馈数据收集和管理，一位三年以上经验的运维工程师，外带一位三年以下的初级工程师，负责回收用户反馈数据，统一存储日志数据。

以上就是一个最低配推荐系统团队的配置。当然，如果能复用现有团队的部门工程师，则灵活处理。上面的估算也只是一个示例。

个人成长

下面来说说，工程师个人该如何学习和成长的问题。

推荐系统工程师和一般意义的软件工程师相比，看上去无需像 IOS 或者 Android 工程师写大量的代码；也无需像研究院的研究员那样，非得憋出漂亮的数学模型才能工作；更无需像数据分析师绘制出漂亮的图表。那推荐系统工程师的定位是什么呢？

实际上，这里说的几个“看上去无需”，并不是降低了推荐系统工程师的要求，而是提高了要求。因为你得具备三个核心素质：

1. 有较强的工程能力，能快速交付高效率少 Bug 的算法实现，虽然项目中不一定要写非常大量的代码；
2. 有较强的理论基础，能看懂最新的论文，虽然不一定要原创出漂亮的数学模型；
3. 有很好的可视化思维，能将不直观的数据规律直观地呈现出来，向非工程师解释清楚问题所在，原理所在。

首先，虽然世人目光都聚焦在高大上的推荐算法上，然而算法模块确实是容易标准化的，开源算法实现一般也能满足中小厂的第一版所需，而实现整个推荐系统的路径却不可复制，这个实现路径就是工程。

可以说，是工程能力决定了推荐系统的上限。如何提高工程能力，无他，就是反复刻意练习。但是对于入行不同年限的人来说，提高的办法则各不相同。

对于在校生，一个比较好的办法是，将看到的任何算法知识、论文或者图书，都亲手转换成代码，一个简单的算法，从你看懂到你无 Bug 地实现出来，其实还有很远的距离，在实现完成后，去阅读对应的热门开源应用，阅读它的实现方法，对照自己的，总结差距。

对于刚工作的新人，这时候你已经有一定的工程基础，并且没有太多的整块时间，那么就要好好把握工作中的项目实战。

避免重复造轮子的前提是知道有轮子，并且知道轮子好在哪，这要求你熟读现有轮子的各种，对它性能、实现方法了如指掌，如此才能在不重复造轮子的基础上安心实施拿来主义，并且可以进一步将轮子按照实际使用的所需问题进行改良。

对于工作一定年限的人，这时候你已经熟知各种轮子极其弊端，也能改进了，那么在业务逐渐增长后，需要考虑将系统中部分模块中所使用的开源加上补丁，整体升级为自研系统。这个开发可以从一些风险不高的模块着手，逐渐锻炼。

上述三个大的阶段，比较粗略。但是核心思想就是：爱动手，爱思考，爱阅读，爱总结。

第二，是理论基础。对于一个从事推荐系统的工程师来说，一定需要有数理基础。高等数学、概率统计、线性代数这些大学基础课一定还在自己心中，没有还给老师。

如果不幸还给老师，你需要重新捡起来，因为整个机器学习都是建立在高等数学基础上的。另外，有一个学科我个人认为很重要，甚至成为人生的指南，那就是信息论。信息论用量化方式确定了什么是信息，很多算法问题也因此可以从通信角度考虑。

第三，是数据可视化思维，在做数据分析和清洗工作时，需要想办法直观地呈现出来，在工具层面，掌握一些常用的绘图工具就很有必要了。Python 中的 Matplotlib，R 语言中的 ggplot2，Linux 命令里面的 Gnuplot，Windows 里的 Excel 等等都是非常常用的绘图工具。

掌握工具并不难，还需要有 show 的冲动，直观方式呈现出数据规律不但对自己优化算法和系统有非常大的作用，还可以与合作伙伴快速达成任务共识，节省沟通成本。

这三个能力，建立起来的难度逐渐增加，需要持之以恒，与《卖油翁》那句著名的“无他，但手熟尔”，规律相同。

除此之外，还有一些非典型工程师的加分项。

1. 学习能力：虽然缓慢，但是科学一直在突破边界，技术更是日新月异地升级了一代又一代，而文化的进化则远远快于人类基因的进化，这些变化，都要求你我要有不断学习的意识，还要有会学习的能力。
2. 沟通能力：在一些中大型厂，一些数据资源分散在不同部门，在技术之外需要去整合这些资源，这需要沟通能力。
3. 表达能力：能把一件事情讲清楚，最直接的好处是在团队内部减少无效的沟通，提高工作效率。

总结

今天我主要谈的是推荐系统中人的因素，包括了团队和个人，这部分内容本来和技术干货内容相比，就有点形而上，但是事实上却又绕不开这部分内容。

因此，我先用一个例子呈现了一个“有下限团队”应该有多少人。这里没有考虑人的个人能力差别，这里就假设大家智商都有一样，没有天才和白痴。

很多时候，其实单机就能搞定很多看上去很复杂的事情，这是我不太推崇分布式的原因，因为多数时候没必要。

最后，我谈了我对推荐系统工程师的能力看法，一共有三个层次，建设起来由易到难，需要不断刻意练习，才可能有较大的能力进步，这一点我和你共勉。

你对工程师的学习路径又有哪些体会呢，可以跟我留言，我们一起分享。

感谢你的收听，我们下次再见。

上一篇 【产品篇】说说信息流的前世今生

下一篇 推荐系统的参考阅读

精选留言 (8)

写留言

shangqiu86
2019-05-09

老师，期待你的书，出来了吗？想买来再学习深入一遍，感觉对推荐系统有了更深的认识

作者回复: 已经在排班中了。

soar

2019-03-21

rerank与rank没有讲解

展开

loirsun

2019-02-21

看到您推荐小团队可以上手qmf，于是尝试qmf，有些困惑，这个包只提供了train model的binary文件，模型预测相关的部分没有给出，是不是需要二次开发？6，7年前从算法方向转工程了，最新的东西都不了解，您方便的话再讲两句qmf可好，或简单回复一下。非常感谢

作者回复: 得到隐因子向量之后，可以用faiss等稠密向量相似计算库来搜索近邻，类似的库还有annoy，kgraph，nmslib。

晴天小雨

2019-01-31

老师，累计产生的日志数量公式是怎么来的，请老师答疑，谢谢!

展开

作者回复: 高中数学。数列求和。

奶爸□ 奶...

2018-12-25

内容消耗的加速度乘以内容的消耗难度。当注意力为正数时，是上马推荐系统的好时机怎么理解？难道一个日活跃平稳的产品就不需要上推荐系统了？

庄

2018-06-11

老师，其他非技术角色指的是哪些哈，举些例子

展开

---

chmit

2018-05-31

邢老师，看到推荐相关相关岗位，多是既要求良好的机器学习算法能力又要求java大数据相关技能您怎么看？这是给双倍工资的嘛。

---

sslrec

2018-05-25

一直有个问题没了解透，向老师请教一下。大多数推荐系统都是预先生产推荐列表等待用户侧来请求，有没有采用交互式的方式实时生成推荐列表的情况呢？有没有这方面的案例

ps：很多不是做推荐业务的同事都会以为我们是交互式的

展开