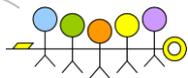


火龙果软件



Interview

# 阿里的大数据算法





专家介绍：孙超，阿里个性化推荐算法工程师

2012年获得浙江大学计算机科学博士学位，专业方向：人工智能，机器学习

2012年就职于百度（上海）网盟广告算法和策略部门，算法工程师

2014年加入阿里巴巴个性化推荐算法部分，任算法工程师

主要方向：计算广告学，推荐系统，自然语言处理，机器学习等

## 访谈关注的话题

- 例如淘宝，用户点击了一个商品，要给客户推荐商品，算法如何设计，
- 数据分析的算法需要数据支持，如果刚开始没那么多数据怎么办
- 阿里算法分析有哪些角色、都做什么
- 大数据算法执行，可能会耗费大量时间，这方面有何优化原则
- 算法也需要需求分析和设计，这方面如何做
- 做好大数据分析工作，需要什么样的知识和技能





问题	解答
例如淘宝，当用户点击了一个商品，要给客户推荐商品，算法如何设计，	一般来讲，推荐算法现在主流使用的方法是两种，基于内容和基于行为的。基于内容的就是根据你购买产品的具体内容（文本，图片等）来判断你的偏好；基于行为的是根据你购买产品的行为，计算有类似行为的用户的其他购买产品，从而扩展给你。典型的是协同过滤的方法。基本上两种方法都会采用，相互补充；并且还有其他的比如用户肖像的分析等不同思路一起实现产品推荐。
数据分析的算法需要数据支持，如果刚开始没那么多数据怎么办	数据量不足实际上在很多场景上都是导致算法失败的重要原因。我个人觉得没有什么太好的解决方案，要么继续积累数据，要么采用一些基于统计或规则的方法。如果非要使用机器学习算法的话，一定要注意数据的置信度和欠拟合问题。
阿里算法分析有哪些角色、都做什么	算法组一般会有：算法产品经理，算法设计和工程实现。产品经理主要是对业务进行了解，能够理解业务需求；算法设计根据业务场景分析数据，设计算法模型；工程同学会对整个模型在实际应用中的实现进行封装，以系统接口或引擎等方式，提供给业务方使用。



问题	解答
大数据算法执行，可能会耗费大量时间，这方面有何优化原则	两个方面：首先在实际运行时间，使用分布式算法，通过Hadoop，Spark等框架来进行数据计算，减少运算时间；另一方面，在算法设计上多做实验，尽量在全量数据运算之前在小数据集上做好预研，保证算法的正确性，避免浪费时间。
算法也需要需求分析和设计，这方面如何做	一般来说，现在主流的算法能够解决一些通用的问题。因此在算法需求分析时，会根据具体的业务需求，确定几个大体的方向：分类问题还是回归问题？线性还是非线性？实时还是离线？...根据这些问题的拆解，能够大体确定一个算法方向，然后就是根据具体的目标和数据分布做分析，如果现有的算法不能满足，那么看在哪些环节上要做算法的改进，改进后会有什么新问题以及改进成本等问题。
做好大数据分析工作，需要什么样的知识和技能	首先需要基本的编程能力，能够实现基本的编程工作；另外是对常见的机器学习算法都有一定了解，原理上要搞明白，需要一定的数学基础；然后就是对于数据问题有敏感性，能够在数据中发现问题。前两个能力可以通过看书和学习来培养，后一个可能需要更多的项目经验。



讲座

2015年4月25日 大型分布式网站架构初探

讲座

2015年5月16日 Qt 在移动端应用开发实践



随时听讲座

每天看新文

追随技术信仰