124 | 数据科学家必备套路之三：广告套路

2018-07-18 洪亮劼

讲述：初明明
时长 06:52 大小 3.15M

讲完了搜索产品和推荐系统的套路，今天我们继续来看数据科学家应该掌握的广告产品的一些套路。

利用搜索和推荐的套路

前面我们讲过两种普遍使用的互联网广告模式，搜索广告和展示广告。对于搜索广告而言，一个基本套路就是尽量利用现有的搜索系统来推送广告。而对于展示广告而言，一个基本套路就是尽量利用现有的推荐系统来推送广告。

我们在介绍推荐套路时提过，推荐系统和搜索系统的很多方面其实都有重叠，所以做好一套搜索系统是非常有必要的，几乎所有的广告应用其实最终也可以在搜索系统的架构上搭建。因此，我们可以说，**搜索系统是很多现代人工智能系统应用的一个核心技术组件。**

具体来说，广告其实也和一般的文档一样，首先利用搜索引擎的索引把这些广告都存储起来。对于搜索广告来说，利用关键词的倒排索引，可以轻松地找到相关的广告，这和找到相关文档的原理其实是一样的。

当然，我们前面也提到过，广告的排序和普通文档有一个不一样的地方，那就是竞价。因此，在从索引中提取广告的时候，我们必须要去思考一个问题，如何让广告竞价的赢家能够从索引中被提取出来？

我们知道，广告的竞价常常是以点击率和出价的乘积来作为排序的依据。这就会有一个问题，如果我们从索引中提取广告的时候，仅仅看哪些广告从关键词的角度是相关的，而忽略了点击率和出价，那么，最后提取出来的广告很有可能不是真正能够赢得竞价的广告。

如何来对这个问题进行修正呢？一种做法是在索引里面增加点击率信息。也就是针对每一个关键词，我们不是按照文本的相关度去索引最相关的文档，而是按照点击率去索引点击率最高的一系列文档。

那么，当需要针对某一个关键词提取广告的时候，我们就直接从这个关键词所对应的索引中提取点击率最高的几个广告。这个时候，我们再从某一个存储出价的数据库中读取这些广告的出价，并且进行竞价排序。

从这个流程我们可以看出，最终的竞价排名很可能并不是完全依赖点击率和出价的乘积，而是在点击率先有了一定的保证下的这个乘积的排序。这种有保障的点击率常常被叫做“**质量值**”（Quality Score），用来描述这些广告的点击率高于一个设定的阈值。

接下来，我们来看广告提取的另外一个重要的要求，就是需要满足广告投放的业务逻辑。比如，有一个广告的投放要求是针对男性，现在有一个女性用户，那么，我们就不应该针对这个用户显示这个广告，而不管这个广告的点击率和出价信息是怎样的。

如何实现这样的效果呢？我们依然可以利用索引。在索引中，我们插入广告的各种投放条件作为被索引的对象，然后把在这个投放条件下的各种广告作为文本。这样，我们就可以提取满足任意投放条件的广告了。

针对这些投放条件的组合，例如投放条件是“女性、在北京”，我们可以认为是**在索引上进行“且”操作**，也就是提取出同时满足两个关键词的操作。事实上，针对任意一个关键词的广告，我们都是进行了多个“且”操作。例如，针对“可乐”这个关键词，我们可能是需要

提取这个关键词点击率最高的 100 个广告（如果有那么多的话），并且这些广告的投放条件都满足“女性、在北京”。当提取出了这些广告之后再进行竞价排名。

当然，在这样的架构下，我们就需要**对索引有快速更新的能力**，例如某一个广告的点击率或者投放条件都有可能发生变化。

层次建模套路

对于广告系统的建模有一个基本的套路，那就是**层次建模**（Hierarchical Modeling）。什么是层次建模呢？在广告的生态系统中，至少有广告商、广告推广计划、单一广告这三个层次的实体。提高广告投放精准度的一个核心问题，就是如何能够对这这三种实体进行有效建模。

当我们对当前的广告商一无所知的时候，需要看一看过去有没有其他类似的广告商在平台投放过广告，如果有，那么能否借鉴那些过去的数据。当这个广告商开始投放广告以后，我们就可以积累数据，慢慢就能够增强对这个广告商的建模能力。

类似的，当我们计划推出某一个广告推广计划的时候，我们先看一看同一个广告商有没有类似的推广计划，或者看一看其他类似的广告商有没有相近的推广计划。当某一个广告开始运行的时候，我们看一看同一个推广计划下其他广告的表现，或者是同一个广告商下其他广告的表现。

层次建模的一个重要的特点就是利用可以利用的一切其他信息来进行建模。在计算广告中，经过验证，层级信息往往是最有用的特性。

具体和泛化的套路

这个套路其实并不是完全针对广告的。就像我们之前所说的广告、搜索和推荐之间的关系，这个套路其实也可以应用在搜索中。

前几年，Google 的工程师发现，如果仅仅利用深度学习模型来学习抽象的特性，从而寄希望模型的性能得以提升，这种方法也许可以很好地解决计算机视觉的一些问题，但是对于搜索、广告和推荐的效果则并不好。

下面我们聊聊工程师们发现的这里面的原因 [1]。

一个好的模型必须具备两种能力。第一，能够**对具体的关键词进行匹配。**比如，我需要匹配“可口可乐”，那么任何与“百事可乐”相关的广告其实都是不能显示的。这就要求模型中针对每一个具体的关键词能够进行**字对字的匹配**，而不是模糊匹配。第二，那就是**具有泛化能力。**比如，我们要去对“可口可乐”在 2018 年的广告推广进行建模，模型就能够借鉴“可口可乐”在 2017 年的推广数据，以及借鉴“可口可乐”公司其他推广的数据。这里面的借鉴能力其实就是模型的泛化能力。

由此，我们可以得到一个好模型的重要套路：**一个好的模型既要能够精确记忆某一种关键词，又要能够在广告层次上进行泛化。**

总结

今天我为你介绍了做广告产品的几个套路。

一起来回顾下要点：第一，搜索系统是很多现代人工智能系统应用的一个核心技术组件，广告系统也可以借鉴搜索系统的套路；第二，广告生态中层次建模的套路，就是利用可以利用的一切其他信息来进行建模；第三，一个好模型的套路，关键是模型的具体能力和泛化能力并存。

最后，给你留一个思考题，为什么在计算机视觉中，对于具体匹配的要求没有那么高呢？

欢迎你给我留言，和我一起讨论。

参考文献

1. Heng-Tze Cheng, Levent Koc, Jeremiah Harmsen, Tal Shaked, Tushar Chandra, Hrishi Aradhye, Glen Anderson, Greg Corrado, Wei Chai, Mustafa Ispir, Rohan Anil, Zakaria Haque, Lichan Hong, Vihan Jain, Xiaobing Liu, and Hemal Shah. **Wide & Deep Learning for Recommender Systems**. Proceedings of the 1st Workshop on Deep Learning for Recommender Systems (DLRS 2016). ACM, New York, NY, USA, 7-10, 2016.

上一篇　123 | 数据科学家必备套路之二：推荐套路

下一篇　125 | SIGIR 2018论文精读：偏差和“流行度”之间的关系

精选留言 (1)

写留言

silverhawk
2018-07-26

1

CV中对匹配要求不高是因为图像信息天生即使部分缺失或者不准确也能大概率拟合还原真实，类似模糊老照片，黑白照片缺失信息不影响。但是NLP类的关键词信息极其重要。总之图像是原始信息信息量大，语言类本来就被抽象一次再损失关键信息就难