

## 122 | 数据科学家必备套路之一：搜索套路

2018-07-13 洪亮劼

AI技术内参

[进入课程 >](#)



讲述：初明明

时长 09:13 大小 4.23M



到目前为止，我们已经完整地介绍了搜索、推荐和广告的主流技术，为你呈现了这些产品技术方向的整个生态系统。在这些系列的分享里，我们重点介绍了这些技术方向的基本模型，然后花了不少篇幅讲如何评测模型的好坏，包括如何进行线下评测以及线上评测。同时，我们从传统的经典模型讲到最近几年利用深度学习对这些技术方向的提升，帮助你理顺了这些技术发展的脉络。

尽管我们已经在之前的文章中分享了这些技术的方方面面，但是对于很多经验较少的数据科学家或者人工智能工程师来说，依然会感到无法得心应手地把这些模型和知识给应用到真实场景中。

其实，出现这种情况一方面是个人的经验积累的原因，毕竟从初学者到能够熟练应用各种模型工具应对实际产品的需要，是一个长时间磨炼的结果；然而另一方面，也是因为搜索、推荐

和广告这些产品场景其实是有一些套路，在没有接触到这些套路的时候往往会觉得不得要领，而在慢慢熟悉了这些套路之后，进步也就会慢慢加快。

那么，在接下来的三篇文章里，我就有针对性地对来分享在这三个领域里的一些常见套路。今天，我们首先从[搜索产品套路](#)说起。

## 多轮打分套路

我们前面已经介绍过多轮打分的系统架构了。**当我们想要构建任何一个搜索引擎时，都应该立刻想到多轮打分这个架构**，这是一个基本套路。

我们先来回顾一下多轮打分最基本的模式：针对一个搜索关键词，我们首先从索引中找到第一批，也是数目相对较多的相关文档，这就是第一轮打分；然后，我们再根据一个相对比较复杂的模型，对剩余的文档进行打分排序，这是第二轮打分。

很多时候，两轮打分就已经能够满足需求了，可以直接给用户返回结果集。当然了，我们也常常加入第三轮打分，这个时候经常是实现一个商业逻辑层，可以针对最后的搜索结构加上一些商业规则。

多轮打分这个套路之所以关键，是因为它其实影响了一系列的技术决定。

首先，只有第一轮是直接作用在索引上的。这一轮往往可以并行化，并且不需要太多考虑文档之间的顺序。

我来举个例子说明。在一个大型搜索引擎的架构下，假设我们有一亿个文档，每五百万个文档存放在一个数据节点上。如果我们有一个关键词是“人工智能”，那么就要到这 20 个数据节点上分别查找包含这个关键词的文档。当我们把所有的文档汇集起来以后，排序取出前 1000 个文档，这个时候就进入第二轮。最简单的第二轮打分是在一个计算节点上进行的，而且这个时候我们只针对 1000 个文档进行打分，对计算时间的要求就大幅度降低了。那么，在这样的情况下，第二轮打分使用的模型可以是比较复杂的模型。后面每一轮打分所针对的文档往往是越来越少，因此模型的复杂度可以越来越高。

从解决问题的角度上来说，**第一轮在索引上的打分，是要解决“召回”（Recall）的问题**。这一步有可能是从非常多甚至是成千上万的文档中返回几百到几千不等的文档，因此一旦一些文档没有在这一轮中被返回，就无法在后面的轮数中被重新筛选出来。所以，我们要理清这么一个思路，那就是如果你认定系统有“召回”的问题，也就是说，本该搜出来的东西，

完全搜索不出来，那肯定是在第一轮打分就出了问题。**第二轮以后的打分解解决的就是“精度”（Precision）的问题。**

同时，我们可以看到，什么时候解决“召回”问题，什么时候解决“精度”问题，这其实是**取决于具体的业务场景。**

对于“精度”非常看重的搜索场景，比如说网页的信息类关键词，例如“特朗普”、“比尔盖茨”，人们往往只关注前 10 位，甚至是前 3 位的搜索结果。那么很明显，我们可以先有一个比较简单的第一轮架构，比如就是文字匹配，而把功夫都下在第二轮以后的打分上。

而对于“召回”比较看重场景，比如说法律文档搜索，那必须要做好的就是第一轮的打分，这个时候可能需要采用简单的文字直接匹配和语义的模糊匹配。

## 高频和长尾的套路

刚开始接触搜索产品的朋友往往会有一个困惑，那就是不知道该如何提升一个搜索产品，有一种无从下手的感觉。那么，对于搜索产品的提高有没有什么套路可言呢？

一个比较基本的套路，就是**把搜索关键词按照某种频率或者是流量分为“高频关键词”和“长尾关键词”**，从而为这两类不同的关键词设计排序算法。

为什么要把关键词按照频率分开呢？我来介绍一下最主要的思路。

对于很多搜索网站来说，一些高频的关键词往往占据了相对来说比较大的流量，而很多长尾的关键词，也就是仅仅出现过几次的关键词则并没有太多人搜。因此，如果我们先解决了高频的关键词，也就解决了大部分的搜索问题，从而可以把精力留下来慢慢解决低频的长尾关键词。

而实际上，高频关键词因为有足够多的数据量，反而往往比较容易得以解决，而低频关键词，因为数据的匮乏，往往需要更多的精力和时间。所以说，从投资回报的角度来看，我们也需要做区分，首先来解决高频的搜索关键词。

刚才我们提到了高频关键词的一个特点，就是有足够多的用户数据。那么，这里有一种非常简单的思路，或者说是没有较好模型的时候可以首先使用的一种方法，那就是**每天记录下高频关键词文档的用户点击数据**。然后我们可以直接按照**点击率**，或者是文档的**转换率**排

序，并且把这个排序存在某种存储中。当用户在这一天搜索这些高频关键词的时候，我们甚至可以直接从存储中调出事先算好的排序结果。

更加极端的做法，就是**手工对高频词进行更频繁的标注**，这种做法往往也是非常有效的。例如我们刚才说的“特朗普”的例子，我们可以手工标注好前 10 名的结果，然后存下来。只需要每几天更新一下这个标注，我们甚至不需要使用任何模型就可以提供非常高质量的搜索结果。

当然，使用这种方法，显然无法对几百万的搜索关键词都这么一一处理。不过，我们这里针对的主要是高频关键词，所以，即便是针对最高频的 1 千个关键词进行手工标注，也会对整体的搜索效果有非常明显的提升。

相反，长尾的关键词往往需要花比较多的心思。对于长尾来说，我们还可以细分。比如对于有一定数据量的关键词，我们可以尝试**针对这些关键词单独训练一个模型**。之所以要单独训练一个模型，原因也很简单，如果针对所有的关键词只有一个模型的话，高频的关键词因为流量大，往往就会让模型偏重于去解释高频的信息，而忽略了这些中低频的关键词的作用。

因此，先把**高频词**单独处理了，然后就可以针对依然可以训练的**中频关键词**再选取一个单独的模型。而针对非常**低频的关键词**，我们往往需要借助其他的方法来挖掘这些关键词的信息，例如利用同类的其他关键词的数据，或者利用外界的知识库、知识图谱的信息等。

## 三大模型套路

除了分开处理高频和长尾关键词以外，搜索模型的提升还有一个非常简单的“三大模型套路”。

我们构建一个搜索引擎，从最原始的简单系统，慢慢到比较复杂的以至于到后期非常复杂的系统，从模型上来说要跨越三个台阶。在这里我们主要是针对第二轮的打分系统来进行讨论。

**第一个台阶是使用线性模型。**当我们设置好了最基本的第一轮打分系统以后，首先要做好是能够利用线性模型对文档进行排序。这一步其实往往是**搜索系统从“无人工智能”到“有人工智能”的第一步**。这一步对搜索效果性能的提升可能会有 10%~20%。

**第二个台阶是使用配对法线性模型。**一般来说，这一步搜索效果会有 2%~5% 的提升。



**第三个台阶是使用树模型，特别是 GBDT 模型。**这一步搜效果的提升和第二步相似，约有 2%~5% 的提升。然而，要从第二个台阶到达这个步骤，模型的特性可能会发生不小的变化。这一个台阶可以算是一个比较困难的台阶。

**从工程研发的角度来说，可以采用一年一个台阶的做法。**在已经穷尽了当前台阶所有可能用到的特性以后，再进入到下一个台阶，也就是说要尽可能地“榨干”当前台阶模型效果的“养分”。

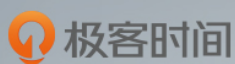
## 总结

今天我为你介绍了做搜索产品的几个套路。

一起来回顾下要点：第一，我们回顾和总结了多轮打分系统的架构套路；第二，我们介绍了区分高频关键词和长尾关键词的套路；第三，我们简单讨论了“三大模型套路”，跨越三个台阶，逐步提升搜索效果。

最后，给你留一个思考题，为什么不鼓励直接采用深度学习模型呢？

欢迎你给我留言，和我一起讨论。

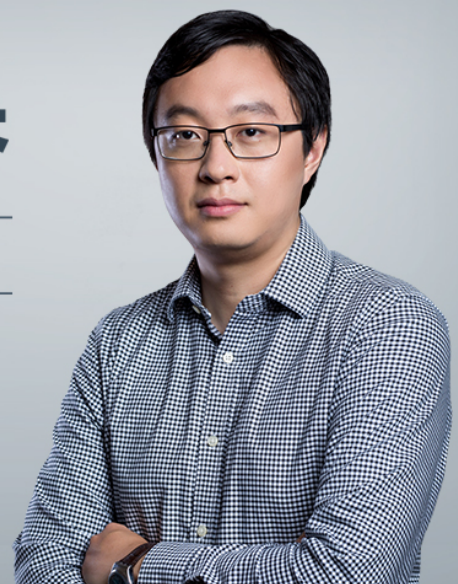



# AI 技术内参

你的360度人工智能信息助理

洪亮劼

Etsy 数据科学主管  
前雅虎研究院资深科学家



新版升级：点击「 请朋友读」，10位好友免费读，邀请订阅更有**现金**奖励。

上一篇 121 | 如何利用机器学习技术来检测广告欺诈?

下一篇 123 | 数据科学家必备套路之二：推荐套路

精选留言 (1)

写留言



Riordon

2018-12-11



写得真好~~~

展开