021 | 机器学习排序算法：单点法排序学习

2017-11-20 洪亮劼

AI技术内参 进入课程 >

讲述：初明明
时长 09:33 大小 4.38M

在专栏里我们已经讲解过最经典的信息检索技术。这些技术为 2000 年之前的搜索引擎提供了基本的算法支持。不管是 TF-IDF、BM25 还是语言模型（Language Model），这些方法和它们的各类变种在很多领域（不限文本）都还继续发挥着作用。

然而，自从机器学习的思想逐渐渗透到信息检索等领域之后，一个最直观的想法就是如何用机器学习来提升信息检索的性能水平。这个思想引领了 2000 年到 2010 年这个领域的研究，产生了各类基于机器学习的排序算法，也带来了搜索引擎技术的成熟和发展。

我今天就从最简单也是最实用的一类机器学习排序算法讲起，那就是单点法排序学习（Pointwise Learning to Rank）。这类方法在工业界非常实用，得到广泛应用，在实际效果中也表现得很强健（Robust）。同时，理解这类模型可以为后面学习复杂的排序算法打下基础。

单点法排序学习的历史

早在 1992 年，德国学者诺伯特·福尔（Norbert Fuhr）就在一篇论文中最早尝试了用机器学习来做搜索系统参数估计的方法。三年之前，他还发表过一篇利用“多项式回归”（Polynomial Regression）来做类似方法的论文。诺伯特因其在信息检索领域的长期贡献，于 2012 年获得美国计算机协会 ACM 颁发的“杰拉德·索尔顿奖”。

1992 年，加州大学伯克利分校的一批学者在 SIGIR 上发表了一篇论文，文中使用了“对数几率”（Logistic Regression）分类器来对排序算法进行学习。可以说这篇论文是最早利用机器学习思维来解决排序算法学习的尝试。然而，由于机器学习和信息检索研究在当时都处于起步阶段，这些早期结果并不理想。

2000 年之后支持向量机在工业界和学术界逐渐火热，随之，利用机器学习进行排序算法训练重新进入人们的视野。搜索引擎成了第一次互联网泡沫的重要阵地，各类搜索引擎公司都开始投入使用机器学习来提升搜索结果的精度。这股思潮开启了整整十年火热的机器学习排序算法的研究和开发。

单点法排序学习详解

要想理解单点法排序学习，首先要理解一些基本概念。这些基本概念可以帮助我们把一个排序问题转换成一个机器学习的问题设置，特别是监督学习的设置。

我之前介绍的传统搜索排序算法比如 TF-IDF、BM25 以及语言模型，都是无监督学习排序算法的典范，也就是算法本身事先并不知道哪些文档对于哪些关键字是“相关”的。这些算法其实就是“猜测”相关性的一个过程。因此，传统信息检索发展出一系列理论来知道算法对每一个“关键字和文档对”（Query-Document Pair）进行打分，寄希望这样的分数是反映相关性的。

然而，从现代机器学习的角度看，毫无疑问，这样的排序算法不是最优的，特别是当相关信息存在的时候，是可以直接用这些相关信息来帮助算法提升排序精度的。

要想让训练排序算法成为监督学习，首先来看需要什么样的数据集。我们要建模的对象是针对每一个查询关键字，对所有文档的一个配对。也就是说，每一个训练样本至少都要包含查询关键字和某个文档的信息。这样，针对这个训练样本，就可以利用相关度来定义样本的标签。

在极度简化的情况下，如果标签定义为，某个文档针对某个关键字是否相关，也就是二分标签，训练排序算法的问题就转换成了二分分类（Binary Classification）的问题。这样，任何现成的二分分类器，几乎都可以在不加更改的情况下直接用于训练排序算法。比如经典的“对数几率”分类器或者支持向量机都是很好的选择。

我们说这样的方法是“单点法排序学习”（Pointwise Learning to Rank）是因为每一个训练样本都仅仅是某一个查询关键字和某一个文档的配对。它们之间是否相关，完全不取决于其他任何文档，也不取决于其他关键字。也就是说，我们的学习算法是孤立地看待某个文档对于某个关键字是否相关，而不是关联地看待问题。显然，单点法排序学习是对现实的一个极大简化，但是对于训练排序算法来说是一个不错的起点。

知道了如何构建一个训练集以后，我们来看一看测试集，重点来看如何评估排序算法的好坏。测试集里的数据其实和训练集非常类似，也是“查询关键字和文档对”作为一个样本。标签也是这个“配对”的相关度信息。前面说了，如果这是一个二分的相关信息，那么评估排序算法其实也就变成了如何评估二分分类问题。

对二分分类问题来说，有两个主要的评价指标：第一，精度（Precision），也就是说，在所有分类器已经判断是相关的文档中，究竟有多少是真正相关的；**第二，召回（Recall）**，即所有真正相关的文档究竟有多少被提取了出来。

因为是排序问题，和普通二分分类问题不太一样的是，这里就有一个**Top-K 问题**。什么意思呢？就是说，针对某一个查询关键字，我们不是对所有的文档进行评估，而只针对排序之后的最顶部的 K 个文档进行评估。

在这样的语境下，精度和召回都是定义在这个 K 的基础上的。要是没有这个 K 的限制，在全部数据情况下，精度和召回都退回到了“准确度”，这个最基本的分类问题的评估测量情形。

在实际的应用中，K 的取值往往是很小的，比如 3、5、10 或者 25，而可能被评分的文档的数量是巨大的，理论上来说，任何一个文档对于任何一个查询关键字来说都有可能是潜在相关对象。所以，在评价排序算法的时候，这个 K 是至关重要的简化问题的方法。

除了精度和召回以外，信息检索界还习惯用 F1 值对排序算法进行评估。简单来说，F1 值就是精度和召回“和谐平均”（Harmonic Mean）的取值。也就是说，F1 结合了精度和召回，并且给出了一个唯一的数值来平衡这两个指标。需要指出的是，在很多实际情况中，

精度和召回是类似于“鱼与熊掌不可兼得”的一组指标。所以，F1 值的出现让平衡这两个有可能产生冲突的指标变得更加方便。

刚才我说的评估主要是基于二分的相关信息来说的。而相关的标签信息其实可以定义为更加丰富的多元相关信息。比如，针对某一个查询关键字，我们不再只关心某个文档是否相关，而是给出一个相关程度的打分，从“最相关”、“相关”、“不能确定”到“不相关”、“最不相关”，一共五级定义。在这种定义下，至少衍生出了另外两个评价排序算法的方法。

我们可以使用多类分类（Multi-Class Classification）的评价方法，也就是把五级相关度当做五种不同的标签，来看分类器的分类准确度。当然，这样的评价方式对于排序来说是有问题的。因为，对于一个实际的数据集来说，五种相关类型所对应的数据量是不同的。

一般来说，“最相关”和“相关”的文档数量，不管是针对某个查询关键字还是从总体上来看，都是比较少的，而“不相关”和“最不相关”的文档是大量的。因此，单单看分类准确度，很可能会得出不恰当的结果。

比如说，某个排序算法能够通过分类的手段把大量的“最不相关”和“不相关”的文档分类正确，而可能错失了所有的“最相关”文档。即便从总的分类准确度来说，这样的算法可能还“看得过去”，但实际上这样的算法并没有任何价值。所以，从多类分类的角度来评价排序算法是不完整的。

针对这样的情况，研究者们设计出了**基于五级定义的排序评价方法：NDCG（Normalized Discounted Cumulative Gain）**。在这里针对 NDCG 我就不展开讨论了，你只需要知道 NDCG 不是一个分类的准确度评价指标，而是一个排序的精度指标。

NDCG 这个指标的假设是，在一个排序结果里，相关信息要比不相关信息排得更高，而最相关信息需要排在最上面，最不相关信息排在最下面。任何排序结果一旦偏离了这样的假设，就会受到“扣分”或者说是“惩罚”。

需要特别指出的是，我们这里讨论的 NDCG 仅仅是针对测试集的一个排序评价指标。我们的排序算法依然可以在训练集上从五级相关度上训练多类分类器。仅仅是在测试集上，采用了不同的方法来评价我们的多类分类器结果，而不是采用传统的分类准确度。从某种意义上来说，这里的 NDCG 其实就起到了“**模型选择**”（Model Selection）的作用。

小结

今天我为你讲了单点法排序学习。可以看到，整个问题的设置已经与传统的文字搜索技术有了本质的区别 。

一起来回顾下要点：第一，单点法排序学习起步于 20 世纪 90 年代，直到 2000 年后才出现了更多有显著效果的研究。第二，详细介绍了单点法排序学习的问题设置，包括训练集、测试集以及测试环境。

最后，给你留一个思考题，有没有什么方法可以把我们之前讨论的 TF-IDF、BM25 和语言模型，这些传统的排序算法和单点法排序学习结合起来？

欢迎你给我留言，和我一起讨论。

上一篇　020 | 经典搜索核心算法：语言模型及其变种

下一篇　022 | 机器学习排序算法：配对法排序学习

精选留言 (4)

写留言

黑翼天佑
2017-12-06

6

将TF-IDF、BM25 和语言模型产出的得分作为特征，放到单点法排序中的二分类模型来进行训练?

作者回复: 是的。

黄德平
2018-12-16

用传统方法得到的相关度作为标签，训练一个回归模型，而非分类

展开

梁中华
2018-11-13

可惜了，这个课程听晚了，理论介绍方面做的很好，但对于还没入门的同学能多点例子就更好了。另外最后总结部分只列标题意义不大，而应该真正的提取出要点，而不是提纲

展开

yaolixu
2018-11-07

若建模为二分类问题，那么标签应该是0，,1，那么无法用相关度来定义样本的标签，该怎么处理?
解决方法是直接找个阈值,把训练样本的标签，转换为1或0吗?

展开