034 | 多轮打分系统概述

2017-12-20 洪亮劼

AI技术内参

进入课程 >

讲述：初明明
时长 07:27　大小 3.42M

周一我为你介绍了搜索系统的一个宏观分类，包括传统的文本匹配信息检索系统和机器学习信息检索系统。这个分类可以让你非常清晰地了解信息搜索系统的历史进程，并对这两种搜索系统的特点有所了解。

今天我们就来剖析搜索系统的另一个框架体系：**多轮打分（Scoring）系统。**

多轮打分系统综述

什么是多轮打分系统？为什么搜索系统需要多轮打分？

我们拿上次介绍的机器学习搜索系统为例。从整体来说，机器学习搜索系统的目的是利用机器学习模型来预测文档和搜索关键字之间的相关性。那么，在理想状态下，针对每一个查询

关键字，我们需要对数据集中的每一个文档进行打分。

如果是一个类似互联网搜索引擎的应用场景，那么理论上，每一个查询关键字都需要对几亿甚至十几亿的网页进行打分。显然，仅仅从这个数量级上来说，这样做都是不现实的。

从另一个方面来讲，目前比较通用的机器学习模型，特别是在排序问题上有强劲表现的树模型（Tree Model），比如 GBDT（Gradient Boosted Decision Trees）或者神经网络，都有较高的计算时间复杂度。要想在实时响应的反应时间内（例如几百毫秒内）对相对比较多（我们这里说的是几千甚至上万）的文档进行打分是很困难的，我们刚才提到的整个数据集中是有几亿甚至十几亿的文档，那就更加困难了。

于是在这样的情况下，我们就需要有这么一种机制：对于每个查询关键字而言，能够先有一个方法在整个数据集上快速评价出几百到几千不等（视具体应用）的文档，然后在这个几百到几千不等的集合上运用复杂的模型进行计算并且排序。 **这种需要对文档进行两轮打分的流程叫做“两轮打分框架”**（见参考文献 [3]）。

第一轮打分又常常被称作“顶部 K”（Top-K）提取。你可以看到，在这样的机制下，相对比较简单的模型和方法可以用于第一轮打分，因为这一轮是有可能在全部的数据集上进行操作的。这一轮也是被优化得最彻底的一轮，毕竟需要在海量的数据集中快速找到几百或者几千个合适文档。

然后在第二轮，当文档的数目已经降到了几千甚至几百的时候，我们就可以使用比较复杂的模型了。这其实也是整个多轮打分的一个目的，那就是可以在一个比较适量的数据集上应用复杂模型。

实际上我们不仅可以对文档进行两轮打分，甚至可以扩展到多轮打分，比如雅虎搜索引擎的“三轮打分机制”（见参考文献 [1]）。第三轮根据第二轮打分所产生的文档“上下文特征“（Contextual Feature），从而可以进一步精准地提高搜索结果的质量。类似的思想也可以借鉴参考文献 [2]。

一般来说，**多轮打分系统有两个明显的特点。**一个特点是每一轮都比上一轮使用的文档数目要少。也就是说，多轮打分的目的是每经过一轮都筛选出更少的文档。另外一个特点是每一轮使用的特性数目都比上一轮更加复杂，模型也更加复杂。

第一轮“顶部 K 提取”

我刚才说了一下多轮打分系统的机理。现在我们来看一看第一轮打分，也就是俗称的“顶部 K 提取”都有什么技术特点。

“顶部 K 提取”的一个核心问题就是，如何快速地从非常巨大的数据集中返回有价值的几百到几千的文档。这就需要对获取文档的数据结构以及使用的模型有一定的要求。

首先，“**倒排索引**”（Inverted Index）是一个非常重要的机制。是否能够建立有效的索引是第一轮打分能否达到目的的关键。

传统的倒排索引已经可以在很大程度上有效地“削减”没必要的文档。我再简要地讲解一下这个基本的数据结构，我们一起来复习一下倒排索引的内容。索引中的“字段”是某一个查询关键字，而每个字段所对应的则是包含这个查询关键字的文档列表。

这个文档列表大多按照某种重要的顺序排列。比如，某个文档整体和查询关键字的相关度大，那么就会排列到这个列表的前面。当然，也并不是所有包含这个查询关键字的文档一定都会包含到这个列表中。另外，之所以叫做“索引”，也是因为这个列表中并不实际存储整个文档，而往往是只存储文档的编号。

除了最基本的通过索引来提取文档以外，我们还可以通过一些简单的模型来提取文档，比如**线性模型**。一个经典的方法叫做“**WAND 操作符**”（WAND Operator，参见参考资料 [4]）。

当然，严格来讲，WAND 操作符并不是把一个通用的、普遍的线性模型应用到文档索引上，而是说，如果我们能够把模型给简化为只有正系数的线性模型，那么，整个模型其实可以看做是两个向量的点积，而 WAND 则是对点积在索引上的一种优化。

当然，研发人员不仅想把线性模型直接使用到倒排索引上。实际上，这么多年来也有不少的尝试，希望能够把树模型直接应用到倒排搜索上。但是，因为我们之前提到的性能因素，通常情况下树模型都没法直接应用（这里提供一个参考文档 [5] 供你阅读）。应该说，树模型的优化还处在一个研究的阶段。

第二轮或以后轮数的重排

当我们结束了第一轮之后，就来到了第二个阶段，也是经常所说的**“重排”（Re-rank）阶段**。在这个阶段，文档已经从索引中到达了内存。一般来说，在比较普通的架构下，所有的几百到几千不等的文档在这个时候已经整合到了某一台机器的内存中。

我们在思考第一轮和第二轮的时候，需要先理解这两轮的一个重要区别，才能知道什么样的模型能够比较好地应用在这两个不同的场景中。

首先，第一轮必须能够应用在搜索倒排索引上。现代的索引模式，往往是部署在很多的节点（机器）上的。也就是说，每一个节点都拥有一部分，但不是完整的文档集合。这也就导致了我们之前介绍过的单点法（Pointwise）、配对法（Pairwise）和列表法（Listwise）这些机器学习方法很难在索引的这个级别直接使用，因为每一个节点为了计算效率问题，只能访问到一部分的文档并且进行打分。

因此，**两轮的最大区别就是，第一轮一般都是针对单一文档的打分，而只有第二轮才能利用上配对法或者列表法针对文档打分**。我们之前曾经提过，配对法或者列表法都比单点法的效果要好，因此如何平衡这两者在两轮中的表现差异就变得越来越重要了。

这里我简单提一下第二轮之后的其他轮数。当我们应用了第二轮之后，其实基本上就已经产生了最后的结果集合。为什么还需要其他轮数呢？

我们可能还需要其他轮数至少有两个原因。

第一，很多搜索系统中，相关排序只是搜索系统的一个方面。搜索系统还可能引入“多元化”或者其他的“商业规则”。这些规则或者进一步的重新排序很难完整地在前面的轮数中进行。

第二，当最后文档集合生成之后，有证据表明（参考文献 [1]），我们还可以生成一些更加精细的特性来进一步提高排序的精度。因此，多轮打分是值得探索的。

小结

今天我为你讲了现代搜索技术中一个很重要的思路，多轮打分系统 。 一起来回顾下要点：第一，我们讲了为什么需要多轮打分，多轮打分的核心思路是什么。第二，我们分别讲了第一轮和第二轮以及后面轮数的一些特点。

最后，给你留一个思考题，在多轮打分系统的情况下，如何评测第一轮模型的好坏呢？

欢迎你给我留言，和我一起讨论。

参考文献

1. Dawei Yin, Yuening Hu, Jiliang Tang, Tim Daly, Mianwei Zhou, Hua Ouyang, Jianhui Chen, Changsung Kang, Hongbo Deng, Chikashi Nobata, Jean-Marc Langlois, and Yi Chang. Ranking Relevance in Yahoo Search. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16). ACM, New York, NY, USA, 323-332, 2016.
2. Ruey-Cheng Chen, Luke Gallagher, Roi Blanco, and J. Shane Culpepper. Efficient Cost-Aware Cascade Ranking in Multi-Stage Retrieval. Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '17). ACM, New York, NY, USA, 445-454, 2017.
3. Van Dang, Michael Bendersky, and W. Bruce Croft. Two-Stage learning to rank for information retrieval. Proceedings of the 35th European conference on Advances in Information Retrieval (ECIR'13), Pavel Serdyukov, Pavel Braslavski, Sergei O. Kuznetsov, Jaap Kamps, and Stefan Rüger (Eds.). Springer-Verlag, Berlin, Heidelberg, 423-434, 2013.
4. Andrei Z. Broder, David Carmel, Michael Herscovici, Aya Soffer, and Jason Zien. Efficient query evaluation using a two-level retrieval process. Proceedings of the twelfth international conference on Information and knowledge management (CIKM '03). ACM, New York, NY, USA, 426-434, 2003.
5. N. Asadi, J. Lin and A. P. de Vries. Runtime Optimizations for Tree-Based Machine Learning Models. In IEEE Transactions on Knowledge and Data Engineering, vol. 26, no. 9, 2281-2292, 2014.

上一篇　033 | 大型搜索框架宏观视角：发展、特点及趋势

下一篇　035 | 搜索索引及其相关技术概述

精选留言 (3)

写留言

极客星星
2017-12-20

3

由于第一轮的主要功能是召回 所以我觉得应该以召回率为主要评估指标 不知道理解是否正确

> 作者回复: 是的。但是要衡量召回率需要知道所有可能相关的文档，这几乎是不可能的。具体有什么好办法呢?

jifei
2018-12-08

目前我们的搜索也是做了两轮打分，第一轮搜索引擎排序：结合了文本得分以及物品质量、商业目标定义的得分值；第二轮机器学习算法排序：基于用户的反馈数据，离线训练模型，线上实时预测。下一步打算第一轮扩大召回范围以及个性化召回，让第二轮的数据量提上来在第二轮打分上内部实现分页。老师觉得怎么样呢?

展开

白杨
2018-05-17

从经验上来判断bm25是否在一个范围内，而这个经验的范围可以通过机器学习的方式来拟合出来，这样可行吗?