033 | 大型搜索框架宏观视角：发展、特点及趋势

2017-12-18 洪亮劼

AI技术内参

进入课程 >

讲述：初明明

时长 08:29　大小 3.89M

我们在前几周的专栏里讲解了一系列最经典的信息检索（Information Retrieval）技术以及基于机器学习的排序学习算法（Learning to Rank）。然后我们花了一定的时间讨论了两个关键搜索组件的核心技术要点，包括查询关键字理解（Query Understanding）和文档理解（Document Understanding）。除此之外，我们还详细讨论了如何从线上和线下两个层面来评价一个搜索系统。相信你已经对搜索系统的各个基本组成部分有了一个比较基础的把握。

那么，今天我们就第一次从整体上来看看大型搜索系统框架的演变和历史发展，给你一个宏观的认识。相信有了之前的基础知识铺垫，我们今天的分享会让你感觉到水到渠成。

基于文本匹配的信息检索系统

我们在介绍 TF-IDF 和 BM25 这些经典信息检索系统的时候，其实就已经介绍了不少基于文本匹配的基本的信息检索系统的核心概念。

实际上，**从 20 世纪 50 年代有信息检索系统开始一直到 2000 年前后，这种纯粹基于文本匹配的搜索系统一直都是主流搜索系统的基础所在。**甚至当前的很多开源搜索框架也都是基于这种最基本的信息检索系统的。

总结一下，这种信息检索系统有这么几个特点。

首先，文本匹配系统的基础是一个倒排索引（Inverted Index）。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，也并不一定所有包含这个查询关键字的文档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，往往只是存储文档的编号。

从这个基本的索引结构其实衍生出了很多值得研究而且在实际应用中也很有必要考虑的问题。

比如如何进一步优化构建这个索引。特别是当列表中的文档数目过多的时候，或者当查询关键字也很多的时候，采用某种编码的模式来压缩索引就变得很关键。

同时，索引过大也会带来很多性能上的问题。比如，当索引过大的时候，某一部分索引或者很大部分就无法存放在内存中，这个时候，整个搜索系统的性能就受到了很大的威胁。因为在对查询关键字进行处理的时候，就需要反复在内存和硬盘上切换内容。因此，对于索引进行创新，使得索引能够在内存中使用并且快速查询是一个非常重要的课题。

文本匹配系统的另外一个特点就是对传统的检索方法，例如 TF-IDF 或 BM25 以及它们变种的依赖。这些方法在查询关键字和索引之间架起一座桥梁，使得搜索引擎能够针对每一个查询关键字文档对赋予一个数值。然后我们可以利用这个数值进行排序。

然而，这些方法本质上的最大问题就是，他们都不是基于机器学习的方法。也就是说，这些方法本身都是基于一些研究人员的假设和经验，往往无法针对现有的数据进行适应。也正是因为如此，这种方法的研发工作往往让人感到缺乏理论基础。

最后，传统的文本匹配系统还存在一个问题，那就是很难比较自然地处理多模数据。也就是我们之前说过的，如果数据中有文字、图像、图（Graph）信息等综合数据信息，文本匹配的方法在这方面并没有提供什么理论指导。

那么，文本匹配系统有哪些优势呢？其实，即便是在今天，**文本匹配系统的最大劣势也是其最大优势：不依靠机器学习。**也就是说，如果你要构建一个新的搜索系统或者是某个 App 中有搜索功能，最开始的版本最容易依靠文本匹配系统，因为这时候并不需要依靠任何数据，并且文本匹配系统不需要太多调优就能上线。但是，文本匹配系统的这一优势今天往往被很多人忽视。

基于机器学习的信息检索系统

从 2000 年开始，基于机器学习的信息检索系统思潮逐渐变成了构建搜索系统的主流。在这种框架下的信息检索系统主要有以下这些特点。

第一，基于机器学习的系统开始有了一整套的理论支持。比如我们之前讲过的单点法（Pointwise）排序、配对法（Pairwise）排序和列表法（Listwise）排序等方法，都明确地使用通用的机器学习语言来描述搜索问题。

什么叫做通用的机器学习语言？那就是，有一个明确的目标函数，有明确特性（Feature），有明确的算法来求解在这些框架下的机器学习问题。同时，机器学习的一系列基本的方法论，比如训练数据、测试数据、评测方法等等都可以应用到信息检索的场景中来。这对于搜索系统的性能以及整体搜索系统的研发都有了非常重要的指导意义。

同时，这也开启了一个非常便利的提高搜索系统效果的大门。那就是任何机器学习领域内部的发展，很多都可以被借鉴到搜索系统中。比如，最近几年深度学习的大力发展，就可以在已经铺就的基于机器学习的搜索系统框架下很容易地进行尝试。

第二，基于机器学习的搜索系统能够很容易地利用多模数据。对于机器学习而言，多模数据，或者说是多种类型的数据的融合，可以很自然地通过特性以及不同类型的特性来表达。因此，对于多模数据，机器学习有天然的优势。**通过学习这些特性之间的联系从而预测相关度，是机器学习的强项。**

因此，理解搜索系统各个部分的数据并把这些信息用在排序算法中，这样的方式就如雨后春笋般大量地出现了。比如，我们之前提到过的查询关键字理解中的查询关键字分类和查询关键字解析，以及文档理解中的文档分类所产生的特性，很难想象这些内容在传统的文本匹配

系统中得以应用。但在基于机器学习的搜索系统中，这些信息则往往成为提高相关度建模的重要工具。

同时，我们也在之前的分享中介绍了，针对多模数据，机器学习中专门有相关的研究，思考如何把不同类型的数据能够更好地融合在一起来建模。这类研究在传统的文本匹配搜索系统中根本不存在。

基于机器学习的搜索系统也不是完美无瑕。实际上，如果没有各种保证，机器学习并不一定能在实际中获得满意的效果，因为基于机器学习的搜索系统对整个系统而言有了较高的要求。

机器学习往往需要大量的数据，而在一个现实的软件产品中，如何能够构建可靠并且干净的数据就是一个不简单的任务。如果没有可靠的数据，对于一般的机器学习算法而言，就是“垃圾进入，垃圾出来”，实际效果往往比不使用机器学习还要糟糕。

同时，机器学习系统可能会有特性异常、模型异常、数据异常等等其他软件系统所不具备的各种问题。如果在生产系统中对这些情况没有一个估计和处理，机器学习搜索系统往往也会不尽人意。

更加智能的搜索系统

很明显，搜索系统不会仅仅停留在应用普通的机器学习算法。近几年，搜索系统的发展有两个方面。

一方面，当然就是依靠深度学习发展的春风，不少学者和研究人员都在思考，如何能够利用深度学习技术让搜索系统更上一层楼。在这方面的研发中，不仅仅是针对普通的深度学习算法，而是看如何应用深度学习所特有的一些模式，比如**深度强化学习**等方式来重新思考搜索问题。

另一方面，就是从用户的角度来说，研究更加有意义的评测方式。也就是说，如何能够真正抓住用户对这个系统的偏好，并且能够进一步地去优化这个系统的性能。

小结

今天我为你讲了现代搜索技术框架的发展，并简单提及了搜索系统目前发展的趋势 。 一起来回顾下要点：第一，我们讲了基于文本匹配的经典搜索系统的特点；第二，我们讲了基于

机器学习的搜索系统的特点。

最后，给你留一个思考题，在机器学习和深度学习的思潮中，传统搜索系统的核心，也就是我们说过的索引，能否依靠机器学习来生成呢？

欢迎你给我留言，和我一起讨论。

上一篇　032 | 文档理解的重要特例：多模文档建模

下一篇　034 | 多轮打分系统概述

精选留言 (1)

写留言

范深
2017-12-18

通过机器学习决定索引的排序顺序，是否有助于索引的效率优化和查全率？

展开

作者回复: 这是目前的一个研究方向。