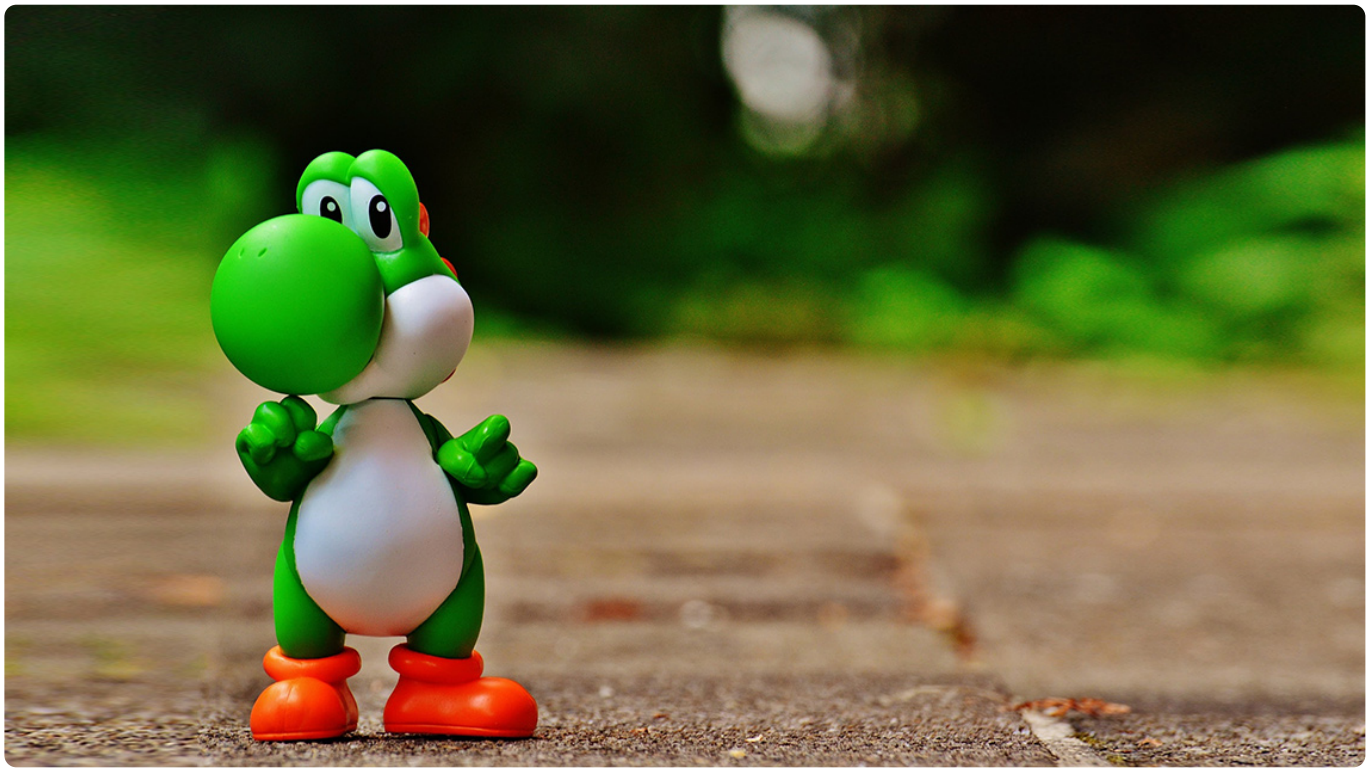
020 | 经典搜索核心算法：语言模型及其变种

2017-11-17 洪亮劼

AI技术内参 进入课程 >



讲述：初明明
时长 10:07　大小 4.64M

在信息检索和文本挖掘领域，我们之前已经讲过了 TF-IDF 算法和 BM25 算法。TF-IDF 因其简单和实用常常成为很多信息检索任务的第一选择，BM25 则以其坚实的经验公式成了很多工业界实际系统的重要基石。

然而，在信息检索研究者的心里，一直都在寻找一种既容易解释，又能自由扩展，并且在实际使用中效果显著的检索模型。这种情况一直到 20 世纪 90 年代末、21 世纪初才得到了突破，一种叫“语言模型”（Language Model）的新模型得到发展。其后 10 多年的时间里，以语言模型为基础的各类变种可谓层出不穷，成了信息检索和搜索领域的重要研究方向。

今天我就来谈谈语言模型的历史，算法细节和语言模型的重要变种，帮助初学者快速掌握这一模型。

语言模型的历史

语言模型在信息检索中的应用开始于 1998 年的 SIGIR 大会（International ACM SIGIR Conference on Research and Development in Information Retrieval，国际信息检索大会）。来自马萨诸塞州大学阿姆赫斯特分校（UMass Amherst）的信息检索学者杰·庞特（Jay M. Ponte）和布鲁斯·夸夫特（W. Bruce Croft）发表了第一篇应用语言模型的论文，从此开启了一个新的时代。

布鲁斯是信息检索的学术权威。早年他在英国的剑桥大学获得博士学位，之后一直在马萨诸塞州大学阿姆赫斯特分校任教。他于 2003 年获得美国计算机协会 ACM 颁发的“杰拉德·索尔顿奖”，表彰他在信息检索领域所作出的突出贡献。另外，布鲁斯也是 ACM 院士。

从那篇论文发表之后，华人学者翟成祥对于语言模型的贡献也是当仁不让。他的博士论文就是系统性论述语言模型的平滑技术以及各类语言模型的深刻理论内涵。

翟成祥来自中国的南京大学计算机系，并于 1984 年、1987 年和 1990 年分别获得南京大学的学士、硕士和博士学位，2002 年他从美国卡内基梅隆大学计算机系的语言与信息技术研究所获得另外一个博士学位。

翟成祥曾经获得过 2004 年的美国国家科学基金会职业生涯奖（NSF CAREER Award）和 2004 年 ACM SIGIR 最佳论文奖。另外，2004 年翟成祥还获得了著名的美国总统奖（PECASE，Presidential Early Career Award for Scientists and Engineers）。

语言模型详解

语言模型的核心思想是希望用概率模型（Probabilistic Model）来描述查询关键字和目标文档之间的关系。语言模型有很多的类型，最简单的、也是最基础的叫做“**查询关键字似然检索模型**”（Query Likelihood Retrieval Model）。下面我就来聊一聊这个模型的一些细节。

首先，我来描述什么是语言模型。简单来说，**一个语言模型就是一个针对词汇表的概率分布。**比如，词汇表总共有一万个英语单词，那么一个语言模型就是定义在这一万个单词上的离散概率分布。拿骰子来做类比，这里的骰子就有一万种可能性。

一旦语言模型的概念形成。“查询关键字似然检索模型”的下一步，就是认为查询关键字是从一个语言模型中“抽样”（Sample）得到的一个样本。什么意思呢？ 就是说，和我们通

常情况下从一个概率分布中抽样相同，“查询关键字似然检索模型”认为查询关键字是从这个语言模型的概率分布中进行采样，从而产生的一个随机过程。这一观点不仅是这个简单语言模型的假设，也是很多语言模型的核心假设。

我们假设这个语言模型，也就是这个概率分布的参数已知，那么，如何来对一个查询关键字打分（Scoring）就变成了计算在这个概率分布的情况下，一组事件，也就是这组词出现的**联合概率**。现实中，因为联合概率可能会很小，因此很多时候都通过一个**对数变换**来把概率的乘积变成概率对数的加和。

然而，现实情况是，我们事先并不知道这个语言模型的参数，这个信息一般来说是未知的。

要想确定这个语言模型的参数，我们**首先要确定语言模型的形态**。我刚才说过，语言模型本质上就是定义在词汇表上的离散概率分布。那么，这里就有几种经典的选择。首先，**我们可以选择“类别分布”（Categorical Distribution）函数**，也就是多项分布（Multinomial Distribution）去除排列组合信息。这也是最常见的语言模型的实现形式。

在类别分布的假设下，我们认为每一个单词都是从类别分布中采样得到的结果，而单词之间互相独立。那么，定义在一万个单词上的类别分布就有一万个参数。每个参数代表所对应的单词出现的概率，或者说可能性。当然，这个参数是未知的。

除了利用类别分布或者多项分布来对语言模型建模以外，其他的离散概率分布也都曾被提出来用作语言模型。比如，伯努利分布（Bernoulli Distribution）或者泊松分布（Poisson Distribution）。这些不同的假设我今天就不展开讲了。但是在实际应用中，其他概率分布假设的语言模型基本上都还属于纯研究状态。

还是回到刚才说的基于类别分布的语言模型。由于参数是未知的，那么问题的核心就变成了**如何估计这样的参数**，这里就回归到基本的统计参数估计的范畴。

因为类别分布是概率分布，在有观测数据的情况下（这个的观测数据就是现实中的文档和查询关键字），最直接的参数估计算法叫“**最大似然估计**”（Maximum Likelihood Estimation）。在这里我不展开这个算法的细节。

最大似然估计的核心思路就是把参数估计问题变换成一个最大化的优化问题，从而通过求解这个优化问题来达到参数估计的目的。在类别分布的假设下，最大似然估计的最优参数解，恰好有解析形式。

也就是说，在有数据的情况下，我们能够得到一个唯一的最优的参数估计。而且这个解非常直观，也就是每个单词出现的可能性，正好等于这个单词在目标文档中出现的次数，除以所有单词在目标文档中出现的次数。换句话说，每个单词的参数正好等于单词出现的频率。

这样的话，每个文档都对应一个类别分布。有多少个文档就有多少个类别分布，而且每个类别分布都可以从自己这个文档中求得所有的参数。

最大似然估计有一个很大的问题，那就是如果某一个单词没有在训练数据中出现过，那么这个单词的参数，根据上面的最优解，就是零。

什么意思呢？也就是说，在最大似然估计的情况下，没有出现过的单词的参数是零，然后模型认为这个词出现的可能性、或者概率就是零。这显然是一个非常悲观的估计。因为你可以认为，不管在任何情况下，就算一个单词没有出现过，但是出现的概率也不应该绝对是零。

那么，如何针对这些为“零”的概率估计，就成了语言模型研究和实践中的一个重要问题。一个通常的技术叫“**平滑**”（Smoothing）。这个技术的基本思想就是，给这些因为最大似然估计所产生的零值一些非零的估计值。**最简单的一个做法，其实也是很有效的一个做法，就是通过整个数据集的频率来做平滑。**

具体来说，就是对于每一个词，我们计算一个目标文档的频率，同时也计算一个全数据集的平率。然后这个单词的最终估计值，是这两个频率的一个加权平均。这个权重就成了另外一组超参数，可以动态调整。

另外一个常见的平滑策略是借助贝叶斯统计推断（Bayesian Inference）的方法。也就是说，为类别概率分布加上一个先验分布，通常是狄利克雷分布（Dirichlet Distribution），并且计算出某个单词在先验分布和数据都存在情况下的后验概率，我这里就不展开这个思路了。

在这里需要注意的是，经过研究人员发现，语言模型的平滑其实是不可或缺的。一方面是为了解决我们刚才提到的零概率估计问题；另一方面，经过一个代数变形，语言模型的平滑其实可以写成一个类似 TF-IDF 的形式。

于是，研究人员指出，这个平滑其实削减了过分流行词汇的概率，使最后的估计值并不完全只是由单词的多少而决定。我在之前介绍 TF-IDF 算法和 BM25 算法的时候，都分别提到

了这个观点，那就是单词出现的多少和相关性的关系问题。从经验上看，这个关系一定是有一个阈值的。

语言模型变种

语言模型有很多类型的变种，我这里简单地提两个比较有代表的方向。

一个方向就是我刚才说的不同类型的平滑策略，比如，结合全数据集平滑和狄利克雷平滑。或者是先把文档分成一些聚类或者不同的类别（例如不同的话题），然后根据不同的类别或者话题进行平滑等等。

另外一个方向其实就是在语言模型本身的的定义上做文章。比如，在查询关键字似然检索模型里，我们假定有一个语言模型，查询关键字是这个模型的一个抽样。乍一看这很有道理，但是仔细一想，这个模型并没有明说目标文档和查询关键字之间的关系。目标文档进入视野完全是为了估计这个语言模型的参数，“相关性”这个概念并没有明确定义。

那么，另外一个主流的语言模型，就是认为有两个模型（分布）。查询关键字从一个分布中产生，目标文档从另外一个分布中产生，而这两个分布的距离，成为了相关性的定义。在这样的结构下，文档和查询关键字形成了一种对称的局面，而相关性也根据距离直接得到定义。

小结

今天我为你讲了文档检索领域或者说搜索技术里一个很有理论深度的技术：语言模型。我们可以看到，语言模型相对于 TF-IDF 以及 BM25 而言，其实更加直观，更好理解。语言模型也是一个强有力的非监督学习方法的文本排序算法。

一起来回顾下要点：第一，简要介绍了语言模型的历史。第二，详细介绍了简单语言模型，即“查询关键字似然检索模型”的主要组成部分。第三，简要地介绍了语言模型的两个变种方向。

最后，给你留一个思考题，如果根据语言模型，也就是概率分布函数的估计，无法得到我们之前提到的最优解析解的话，我们应该怎么求解语言模型的参数呢？

欢迎你给我留言，和我一起讨论。

上一篇　019 | 经典搜索核心算法：BM25及其变种（内附全年目录）

下一篇　021 | 机器学习排序算法：单点法排序学习

精选留言 (7)

写留言

梁中华
2018-11-13

1

感觉有点抽象，对没有基础和背景知识的同学理解起来很累，每句话都懂，但串起来还是把握不了这个知识点

rookie
2019-05-27

EM算法、变分推断、MCMC等

展开

Qi
2018-12-28

这样全用文字讲算法有点似懂非懂，建议举点例子。

展开

黄德平
2018-12-16

可以用神经网络代替解析形式，通过神经网络的参数拟合这些"参数"，梯度下降方法就可以派上用场了

淳韵
2018-10-11

没有解析解，就得用梯度下降等优化方法逼近吧?

展开

淳韵
2018-10-11

没有解析解，就得用梯度下降等优化方法逼近吧?

展开

沉
2017-11-25

Topic modelling 常用的 LDA 是不是就是基是于语言模型发展而来的呢?

展开

作者回复: 不完全是。但我们可以把LDA看作是多个语言模型的某种复杂混合。