

044 | 基于深度学习的搜索算法：局部和分布表征下的搜索模型

2018-01-12 洪亮劼

AI技术内参

[进入课程 >](#)



讲述：初明明

时长 07:29 大小 3.43M



周一我们分享了一篇较早利用深度学习技术来进行搜索建模的论文，利用前馈神经网络来对查询关键字和文档进行信息提取，从而学习到更有意义的语义信息。周三我们分享了另外一篇论文，可以说是周一分享文章的一个后续工作，探讨了如何利用卷积神经网络来对搜索表征进行进一步提升。这两篇论文呈现了一个统一的套路，那就是尝试把深度学习的一些经验直接应用到传统的搜索建模上。这些尝试，也都取得了一些初步成绩。

今天我们来聊一篇 2017 年刚刚发表的论文《网页搜索中利用文本的局部和分布表征学习匹配》（Learning to Match Using Local and Distributed Representations of Text for Web Search），这是近期将深度学习模型应用在搜索领域的一个创新。这篇论文发表在世界万维网大会 WWW 2017 上。

论文背景介绍

下面我们来了解一下这篇论文的作者群信息。

第一作者巴斯卡·米特拉（Bhaskar Mitra）是微软研究院在剑桥实验室的一名研究员。他已经发表了多篇利用深度学习技术解决搜索问题的论文。目前，米特拉在伦敦大学学院攻读博士学位。

第二作者是费尔南多·迪亚兹（Fernando Diaz）在文章发表的时候是微软研究院的一名研究员，目前则在 Spotify 工作。迪亚兹长期从事搜索以及信息检索的工作，发表多篇论文，文章总引用数超过三千次。加入微软之前，他曾经在雅虎研究院从事过研究工作。

文章的第三作者尼克·克拉维尔（Nick Craswell）在微软研究院工作，目前是主任级研发经理，长期从事搜索和信息检索的研究，发表多篇论文，文章总引用数达 8 千多次。

局部和分布表征下的搜索模型详解

我们详细讲讲这篇论文的核心思想。要想理解这篇论文提出的思路，我们首先要简单回顾一下这周讲的前两篇文章内容。

本周第一篇介绍的深度结构化语义模型主要是希望利用前馈神经网络来对查询关键字和文档进行信息提取。第二篇文章尝试用卷积神经网络来提取查询关键字和文档的信息。

不论是前馈网络，还是卷积网络，这些尝试都是想从文本中提取高层次的语义信息。那么今天这篇文章说得是，并不是所有的相关信息都是高层次的语义信息。这是什么意思呢？

作者们提出了这样一个观点，那就是在搜索的时候，一个非常关键的需求就是被搜索到的文档应该包含查询关键字；或者反过来说，拥有查询关键字的文档有很大可能是相关的。也就是说，**如果一个模型不能去进行绝对的关键字匹配，那很有可能就无法真正抓住所有的相关信息。**

另一方面，相关信息的提取也需要高层次的语义，比如同义词，或者同一个主题。设想我们需要查找汽车相关的信息，而一个最新品牌的汽车页面也许并不直接包含“汽车”关键字，但很明显是一个相关的页面。因此，**利用同义词或者整个主题的相关性，通常可以提高搜索效果，特别是“召回”（Recall）的效果。**

那么，很显然，一个好的搜索模型应该兼顾这两个方面，也就是说**既能够做到关键字的直接匹配，也能做到在高层次的语义上进行模糊匹配。**

之前讲到的比如利用前馈网络或者卷积网络主要是针对后者，也就是模糊匹配，文章中提到叫做“分布表征”的匹配。那么，这篇文章的新意就是提出一种捕捉直接匹配的方式，文章叫做“**局部表征**”，并且和模糊匹配的分布表征结合在一起，形成一个统一的模型，从而提高搜索的效果。

具体来说，文章提出的模型是这样的。首先，从整体的网络框架来说，整个网络分成两个部分：一部分来学习查询关键字和文档的局部表征，也就是完全匹配；另一部分来学习查询关键字和文档的分布表征，也就是模糊匹配。最后，两个部分分别学习出一个向量，然后两个向量加和就形成了最后的表征。

完全匹配的局部表征技巧来自于数据的输入。和之前介绍的模型不同，因为我们需要学习查询关键字和文档之间的匹配信息，因此，网络的输入信息就不单单是查询关键字和文档本身，而是两者的一个“**点积**” (Dot-Product)，也就是说，网络的输入信息就是两者是否有匹配。把这个信息作为输入向量之后，这篇文章采用了我们分享过的卷积神经网络的结构，来进一步提取点积过后的输入向量。

在模糊匹配的分布表征部分，整体的框架和上次分享的模型很类似，也就是对查询关键字和文档分别进行建模，分别利用卷积神经网络提取高层次的语义信息。然后在高层次的语义信息上再进行查询关键字和文档表征的乘积（这里是矩阵相对应元素相乘）。最后，在经过基层的隐含转换（其实就是前馈网络），形成分布表征的最后唯一结果。

从整个模型来看，局部表征和分布表征的主要区别在于如何处理查询关键字和文档的匹配信息。如果是在原始数据上直接匹配，然后学习匹配后的高层语义，这就是局部表征。如果是先学习高层语义然后再匹配，这就是分布表征。

整个模型利用相关标签，进行的是监督学习流程，并且采用了 SGD 来优化。

局部和分布表征的搜索模型实验效果

这篇论文提出的模型还是仅仅使用了查询关键字和文档之间的文字信息，因此和上两篇分享一样，提出的模型就只能和文字型的排序算法例如 TF-IDF、BM25 和语言模型进行比较。文章在数据集上采用了 Bing 的搜索数据，有 19 万多的查询关键字，总共有将近百万的文档数。这比之前两个分享里的数据都要大。不过遗憾的是，这三篇文章都是不同的数据集。每个文档又有 4 级的相关标签，可以用来计算诸如 NDCG 这样的指标。

在这篇文章里，作者们比较了一系列的方法，比如 TF-IDF、BM25，以及一些传统的降维方法比如 LSA，然后还比较了之前两个分享中提到的模型。简单来说，本文模型在最后的比较中取得了非常不错的成绩，NDCG 在第 10 位的表现接近 0.53，而之前提出的一系列深度搜索模型，包括我们分享的两个模型达到了差不多 0.45~0.48 左右。看来，既需要完全匹配还需要模糊匹配的确能够带来性能上的提升。在这个数据集上，传统方法其实也不差，比如 BM25 的表现有 0.45 左右，而传统的 LSA 也有 0.44 左右的表现。

小结

今天我为你分享了搜索专题的最后一篇内容，那就是利用深度学习技术对搜索算法进行改进的又一个尝试：一个结合了学习完全匹配的局部表征和模糊匹配的分布表征的统一的搜索模型。

一起来回顾下要点：第一，我们简要介绍了局部和分布表征搜索模型提出的历史。第二，我们详细介绍了局部和分布表征搜索模型的核心思路以及实验结果。

给你留一个思考题，我们这周分享了三个经典的深度学习和搜索相结合的尝试，你觉得目前深度学习在搜索领域取得的成果，有让你感到特别惊讶的结果吗？

欢迎你给我留言，和我一起讨论。

最后，预告一个小活动，明晚（1 月 13 日）8:30 我会在极客时间做一场直播，欢迎你参加。主题是“人工智能 20 问”，如果你有想交流的问题，欢迎给我留言，我们周六直播见！

极客

人工智能20问

1月13日(周六) 20:30直播

洪亮劼

Etsy 数据科学主管



极客时间

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。

© 版权归极客邦科技所有，未经许可不得传播售卖。页面已增加防盗追踪，如有侵权极客邦将依法追究其法律责任。

上一篇 043 | 基于深度学习的搜索算法：卷积结构下的隐含语义模型

下一篇 045 | 职场话题：当数据科学家遇见产品团队

精选留言

写留言

由作者筛选后的优质留言将会公开显示，欢迎踊跃留言。