

## 042 | 基于深度学习的搜索算法：深度结构化语义模型

2018-01-08 洪亮劼

AI技术内参

[进入课程 >](#)



讲述：初明明

时长 08:07 大小 3.72M



近两个月，我们集中系统地分享了搜索核心技术模块。做一个简单的内容梳理，我们讲解了搜索引擎方方面面的话题，从经典的信息检索技术、查询关键字理解、文档理解到现代搜索引擎的架构和索引的核心技术；还从机器学习角度出发分享了搜索引擎的最核心部分，也就是排序算法，深入排序算法的细节讲解了排序支持向量机（RankSVM）、梯度增强决策树（GBDT）以及经典模型 LambdaMART。至此，整个人工智能领域关于搜索的经典话题也就告一段落了。

那么，这个星期，我们来看一些关于搜索算法的前沿思考。火热的深度学习不仅对图像、视频和音频这些领域产生了巨大的冲击，也对自然语言处理、甚至搜索领域有不小的影响。**深度学习带给传统的模型和算法以新的建模能力和新的视角，为以前所不能完成的应用打下了基础。**

今天，我们来看一篇较早利用深度学习技术来进行搜索建模的论文：《使用点击数据学习深度结构化的网络搜索语义模型》（Learning deep structured semantic models for web search using clickthrough data）。这篇论文阐述了一个**深度结构化语义模型**，发表在第 22 届世界信息和知识管理大会 CIKM 2013 上。

## 论文背景介绍

发表于 2013 年的这篇论文应该算是比较早的直接使用深度学习中经验的论文。其主要目的是探索一些经典的深度学习方法能否在搜索的应用中得到合适的效果。

下面我们来了解一下这篇论文的作者群信息。

第一作者黄博森（Po-Sen Huang）是一名来自台湾的学者。在发表论文的时候，他在伊利诺伊大学香槟分校攻读电子工程和计算机博士学位，师从马克·约翰森（Mark Hasegawa-Johnson）。论文是黄博森在微软实习时的工作总结。2015 年黄博森博士毕业，然后于 2016 年加入了微软研究院。到目前为止，他发表了 30 多篇人工智能相关的论文，论文引用次数已经超过 1 千多次。

其他作者均来自当时在微软研究院工作的学者。其中不乏著名学者，比如何晓冬（Xiaodong He）、邓力（Li Deng）、亚历克斯·阿西罗（Alex Acero）和拉里·赫克（Larry Heck）等。下面聊聊比较少被提及的阿西罗和赫克。阿西罗曾长期在微软研究院担任语音相关研究组的经理职位，2013 年之后，他到苹果公司担任 Siri 的资深总监。赫克曾经在雅虎担任搜索和广告业务副总裁，然后到微软研究院担任语音组的首席科学家。文章发表之后，赫克到了谷歌，在一个人工智能组担任总监，并于最近加入三星北美研究院担任资深副总裁。这些学者主要是为这个工作提供支持和指导工作。

这篇论文自 2013 年发表后已经有超过 390 多次的引用，是深度学习在搜索领域应用中被引用次数最多的论文之一。

## 深度结构化语义模型详解

下面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经典的搜索模型构建。

在经典的搜索模型里，不管是 TF-IDF、BM25、语言模型，还是基于机器学习的排序算法模型，整体来说，一个共通的想法就是争取用某种表示（Representation）来表达查询关

键字，然后用相同的、或者类似的表示来表达文档，再通过某种程度的匹配函数来计算查询关键字表示和文档表示之间的距离，然后进行排序。

那么，从深度学习的角度来说，要想针对这个传统的模式进行革新，当然就可以从最主要的三个方面入手：**查询关键字的表达、文档的表达和匹配函数**。

这篇文章也正是沿着这个思路，提出了深度结构化语义模型。

首先，深度结构化语义模型对查询关键字和文档进行了相似的处理。具体来说，就是先把查询关键字或者文档转换为**词向量**（Term Vector），这个词向量可以是最简单的“**词袋**”的表达方式，这也就是最基本的模型的输入。从词向量出发，模型首先学习一个“**词哈希**”（Word Hashing），也就是把 0 或 1 的稀疏词向量转换成为一个稠密（Dense）的向量表达。这一步是**把深度学习方法应用在自然语言处理中所通用的办法，目的就是把稀疏的输入转换为稠密的输入，降低输入的数据维度**。

当查询关键字和文档都转换成稠密数组以后，深度结构化语义模型利用了深度学习中的重要经验，那就是通过“**非线性转换**”（Non-Linear Projection）来获取数据深层次的语义信息，而不仅仅只是传统方法中字面上的匹配。这里，查询关键字和文档都使用了简单的“**前馈神经网络**”（Feedforward Neural Network）的方法，对输入向量进行了多层的非线性转换。非线性转换本身通过“**双曲正切函数**”（tanh 函数）实现，这应该算是最传统的深度学习模型的实现方法了。

经过多层转换之后，查询关键字和文档都变成了新的某种表达之后，如何来计算两者间的距离（或者远近）呢？这篇文章采用了非常直接的形式，那就是利用“**余弦函数**”（Cosine）来作为距离函数，描述两个向量之间的距离。在传统信息检索的语境中，也经常用余弦函数来计算向量的距离，所以在这里应该说并没有太多创新的地方。

总体来说，**深度学习在这里的主要应用，就是成为查询关键字和文档的表达的提取器**。和传统方法中人工提取各种类型的文字特性相比，在深度结构化语义模型中，基于前馈神经网络的特征提取器自动提取了文字的深层语义信息。

提出了模型之后，我们来看这个模型是如何被训练出来的。作者们首先利用了用户的点击信息，也就是针对某一个查询关键字，有哪些文档被点击过，作为**正例数据**，其他文档作为**负例数据**，然后把整个建模问题看作一个**多类分类问题**。这样就可以利用标签信息对整个模型进行学习。

整体来说，这个深度学习模型是可以利用“端到端”（End-to-End）的方式进行训练的，并且采用了随机梯度下降（SGD）这样的优化算法，这里就不复述了。

## 深度结构化语义模型的实验效果

因为深度结构化语义模型仅仅使用了查询关键字和文档之间的文字信息，因此提出的模型就无法和完整的、利用很多特性的机器学习排序算法进行比较，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，这也是文章并没有采用一些更为通用的数据集的原因。最终文章在数据集上采用了 Bing 的搜索数据，有 1 万 6 千多的查询关键字以及每个查询关键字所对应的 15 个文档，每个文档又有 4 级相关标签，这样可以用来计算诸如 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如 LSA 和 PLSA。简单来说，深度结构化语义模型在最后的比较中取得了不错的结果，NDCG 在第 10 位的表现是接近 0.5。不过，TF-IDF 的表现也有 0.46，而传统的 PLSA 和 LSA 也有 0.45 左右的表现。所以，可以说深度结构化语义模型的效果虽然很明显但并不是特别惊人。

## 小结

今天我为你讲了深度结构化语义模型的一些基本原理，这是利用深度学习技术对搜索算法进行改进的一个经典尝试。我们在上面的实验结果总结中已经说到，虽然文章仅仅谈到了文本信息的匹配，并没有作为完整的排序算法进行比较，但是也揭开了用深度模型来表征查询关键字和文档的研发序幕。

一起来回顾下要点：第一，我们简要介绍了提出深度结构化语义模型的历史。第二，我们详细介绍了深度结构化语义模型的核心思路以及实验结果。

给你留一个思考题，除了文章中提到的余弦函数可以作为一个距离函数，还有没有其他的函数选择来表达两个向量之间的距离？

欢迎你给我留言，和我一起讨论。

最后，预告一个小活动，本周六（1 月 13 日）晚 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 041 | 机器学习排序算法经典模型：LambdaMART

下一篇 043 | 基于深度学习的搜索算法：卷积结构下的隐含语义模型

## 精选留言 (2)

写留言



麻离弦

2019-05-03



既然是分类问题，那么对doc进行分类后要如何排序呢？

展开 ∨



hello\_word

2018-01-09



可以用 KL divergence，不过不知道是否容易优化。请洪教主指点 😊

展开 ∨