微软论文概述神经信息检索技术：如何将神经网络用于信息检索？

作者：writer 来源：本站

本文原地址：https://www.iikx.com/news/article/379.html

近日，微软研究人员 Bhaskar Mitra 和 Nick Craswell 在 arXiv 上提交了一篇名为《用于信息检索的神经模型(Neural Models for Information Retrieval)》论文，论文概述了神经信息检索模型背后的基本概念和直观内容，并且将其置于传统检索模型的语境之中。论文的目的在于为神经模型与信息检索之间架起桥梁，互通有无，加快神经信息检索技术的发展。机器对该论文进行了编译，论文链文末。

信息检索(information retrieval，IR)的神经排序模型使用浅层或深层神经网络来根据查询(query)对搜索结果进行排序。传统的学习排序的模型是在手工标注的信息检索特征上使用机器学习技术，与之相反，神经模型可以从原始文本材料(这些材料可以弥合查询与文档词汇之间的差距。)中学习语言的表征。不同于经典的信息检索模型，在可被部署之前，这些新型机器学习系统需要大量的训练数据。该教程介绍了神经信息检索模型背后的基本概念和直观内容，并且该教程也会把它们置于传统检索模型的语境之中。我们以信息检索基本概念介绍和学习文本向量表征的不同神经、非神经进开始。然后，我们回顾一下使用预训练的没有端到端学习信息检索任务的神经项嵌入(term embedding)的浅层神经信息检索方法。之后我们会介绍深度神经网络，讨论热门的深度架构。最后，我们会回顾目前用于信息检索的 DNN 模型，并以讨论的形式对神经信息检索未来可能的发展方向进行总结。

近十年来，计算机视觉、语音识别和机器翻译的性能获得了超乎想象的提升，研究领域和现实世界应用领域了这一切。这些突破大部分由近期在神经网络模型方面的进步所推动，这些神经网络通常有多个隐藏层，我们称之为深度架构。诸如会话代理(agent)和玩游戏达到人类水平的代理这样令人激动的全新应用也相继出现。现在，信息检索社区也开始应用这些神经方法，这将为提升最先进技术或者甚至在其它领域实现突破带来可能。

信息检索的方式有很多。使用者可以文本查询的方式表达其信息需求，这里所谓的文本查询方式可指键盘键入、选择一个查询、声音识别或者图像形式查询，甚至在有些情况下需求不太清楚也可以。检索可以涉及对现存内容的部分进行排序，这些部分可以是文档或简短的文本答案，也可以是通过组合新的答案来具体化检索信息。信息需求和检索结果或许都使用了同样的方式(比如，检索文本文档以响应关键词查询)，亦或也有不同方式(比如，使用文本查询进行图像搜索)。检索系统可能会考虑用户历史、物理定位、信息的时间变化或者排序结果时的其它语境因素。这些因素也可能帮助用户形成其的意图(比如，通过自动完成的查询或者查询)并且/或者可以帮助用户提炼出更易于检查的简练的结果总结(summaries of result)。

神经信息检索指的是将浅层或深层神经网络应用于这些检索任务之上。该教程目的在于介绍神经模型，其回应查询以进行文档排序，这是一项重要的信息检索任务。一条搜索查询通常可能会包含一些词语，然而文档的长度会根据特定的场景而改变，从几个词到成百上千个句子甚至更长。信息检索的神经模型使用文本的向量表征，通常这包含了大量需要调整的参数。带有大型参数集的机器学习模型通常需要大量的训练数据。不同于传统的学习排序的方法(这些方法在一个手工标注的特征集上训练机器学习模型)，信息检索的神经模型通常可以将查询(query)和文档(document)的原始文本(raw text)作为输入。学习文本的恰当表征也需要大量数据训练。因此，不同于经典信息检索模型，这些神经方法非常需要数据，数据越多，性能越好。

文本表征可通过非监督或监督方式习得。监督式方法使用诸如标注的查询文档对(query-document pairs)这样的信息检索数据来习得一个表征，其专为手头任务进行端到端优化。如果没有足够的信息检索标记，那么非监督式方法可仅通过使用查询和/或文档来习得一个表征。在非监督学习方法中，不同的非监督式学习设置可能会导致不同的向量表征，这些表征不同于它们在被表征对象之间所捕获的相似度概念。当应用这些表征时，应该仔细考察非监督学习设置的选择，因此，我们可以产生一个适合于目标任务的文本相似度概念。传统信息检索模型比如潜在语义分析(Latent Semantic Analysis，LSA)可以学习密集的词和文档的向量表征。神经表征学习模型和这些传统方法享有一些共性。几十年来，我们对这些传统方法的大部分理解都可以被扩展成这些现代表征学习模型。

在其它领域，神经网络的进步已经由特定的数据集和应用需求所推动。例如，数据集和成功的架构因视觉对象识别、语音识别和游戏代理而迥然不同。尽管信息检索与自然语言处理领域有一些共同特征，但是它也面临自己的一系列特殊挑战。信息检索系统必须处理可能包含未见过词语的简短查询(short query)，以此来和不同长度的文档进行匹配，找到可能包含了大量不相关文本的相关文档。信息检索系统应该在查询(query)和表明了相关性的文档文本中学习模式，即便查询和文档使用了不同的词汇，甚至即便模式是专用于任务(task-specific)或语境(context-specific)的。

该教程的目标是在传统信息检索研究的语境里介绍神经信息检索的基本内容，用可见的实例展示关键概念和描述关键模型的一致性数学标注(notation)。第二部分会给出一个信息检索的任务、挑战、量度和非神经模型的调查。第三部分会提供简要神经信息检索模型的概览与信息检索的不同神经方法的分类。第四部分介绍学习项嵌入(term embedding)的神经和非神经方法，这些方法不使用来自信息检索标签的监督，而是聚焦在相似度概念上。第五部分调查了合并这些信息检索嵌入的一些特殊方法。第六部分介绍了目前在信息检索中使用的深度模型的基本情况，包括了热门架构和工具包。

第七部分调查了一些在信息检索中实现深度神经网络的特殊方法。第八部分是我们的讨论，包括未来的工作与结论。

由于神经信息检索正在成为一个新兴领域，所以我们撰写了该教程。神经信息检索领域的研究出版物正在逐渐增多，与之同步增长的还有相关线]。由于这种兴趣是最近不久才产生的，所以部分有信息检索专长的研究人员可能对神经模型不太熟悉，而其它熟悉神经模型的研究人员又可能对信息检索不太熟悉。所以该教程的目的即通过描述当下正在使用的相关信息检索概念和神经方法来弥合这条缝隙。