127 | SIGIR 2018论文精读：如何对搜索页面上的点击行为进行序列建模?

2018-07-25 洪亮劼

AI技术内参

进入课程 >

讲述：初明明

时长 07:13　大小 3.31M

今天我们将继续来精读 SIGIR 2018 的论文。

我们已经分享了 SIGIR 2018 的最佳论文，介绍了如何对推荐系统中的偏差进行建模，从而能够利用这种对偏差的理解，来更加准确地对待基于流行度的推荐结果。周一我们分享了本次大会的最佳短论文，主要讲了如何利用对抗学习的技术来让学习的排序模型更加“健壮”，可以被应用到多个领域上。

今天我们分享的论文题目是《页面搜索的点击序列模型》（A Click Sequence Model for Web Search）。

文章的第一作者阿列克谢·博里索夫（Alexey Borisov）来自俄罗斯的搜索引擎 Yandex，并且在阿姆斯特丹大学攻读博士学位。之前，他已经发表过了多篇关于“点击模型”（Click Model）和深度学习模型结合的论文。

文章的第二作者马丁·万德纳（Martijn Wardenaar）、第三作者伊雅·马尔科夫（Ilya Markov）和最后的作者马顿·德里克（Maarten de Rijke）也都来自阿姆斯特丹大学。其中，马顿是荷兰的计算机科学家，欧洲的信息检索学术权威，并且还是荷兰皇家科学院成员。

论文的主要贡献

我先对这篇论文的核心思想做一个提炼，就是利用深度学习模型，来对用户在搜索页面上的点击行为进行建模。

传统上，这种对用户在搜索页面上的点击行为进行建模的思路就是“**点击模型**”。从 2000 年开始，对点击模型的研究，就成为了信息检索以及搜索领域中一个非常活跃的课题。在最近 10 年的时间里，研究人员提出了几十种不同的点击模型。总体来说，不同的点击模型主要是对不同的用户行为进行编码，从而能够更加准确地对用户的点击行为进行预测。

在很多传统的点击模型中，为了简化模型，经常使用的一个假设是：针对每一个查询关键词，用户在搜索结果页只进行一次点击。在这种简化了的假设下，研究人员对用户的浏览、点击以及页面的偏差（例如位置偏差）进行建模，就会变得更加容易。然而，在很多场景中，这种假设就显得过于简化了。在同一个查询关键词的搜索结果页面下，很多用户都会点击多个结果。因此，对于多个点击结果的建模就变得重要起来。

这篇论文就是**针对用户在搜索页面上的点击行为进行了序列建模**，使得我们可以轻松地对每一个搜索页面进行预测，比如会有多少点击以及在什么位置点击等。

同时，这篇论文还有一个贡献，就是**利用了深度学习中的循环神经网络（RNN）来对查询关键词的结果进行建模，扩宽了传统的完全基于概率建模的点击模型在深度学习时代下的表现力**。

论文的核心方法

论文提出方法的核心思路是针对每一个查询关键词，模型需要对所有可能的点击序列进行建模。这个任务是通过构建一个神经网络来完成的。

具体来说，文章提出的模型有两个重要的模块，**编码器**（Encoder）和**解码器**（Decoder）。

编码器的作用是利用查询关键词和搜索结果为输入，生成它们的“**嵌入向量**”（Embedding Vector）。近年来，嵌入向量是深度学习建模中的一个重要技术手段，它的目的往往是先把一些离散变量信息转化为连续信息。在这里，查询关键词和搜索结果都可以首先表征为离散的输入信息，然后需要映射到一个共同的语义空间。这可以被认为是一个中间结果，或者在概率模型中，这往往被认为是一个隐含变量。

解码器的作用是根据这个中间的嵌入向量表达下的查询关键词和搜索结果，然后决定在哪一个位置上可能会或者不会发生点击。这其实就是一个**多类的分类问题**。那么，怎么才能让解码器终止在某一个状态呢？作者们引入了一个特殊的符号代表序列的终止。这样，解码器也需要预测是否需要终止。类似的对解码器的操作在深度序列建模中十分常见。

可以说，作者们在设计编码器和解码器的结构上也是费了一番功夫的。

对于编码器而言，作者们认为一个好的嵌入向量必须包含当前的结果信息，以及当前结果周围的结果，或者说是上下文的信息，以及查询关键词的信息。这样，可以把每一个搜索结果都当做是一个独立的单元，有着足够丰富的信息来进行后面的建模。

因此，作者们首先把查询关键词和每一个搜索结果转换成为第一个层次的嵌入向量，组成一个大的第一层次的嵌入向量。然后，作者们利用这个第一层次的嵌入向量，并且引入了循环神经网络，来对当前结果前后的结果进行了两次编码，一次正向，一次逆向，从而形成了第二层次的嵌入向量。这个第二层次的嵌入向量就是最终表征每一个搜索结果的向量。

对于解码器而言，作者们利用了“**关注**”（Attention）机制来对每一个搜索结果施加不同的权重，或者说是关注度。每个时间点，也就是每一次做“是否要点击”的决策之后，都会重新生成一个关注向量，或者说是一组新的关注权重。这里的核心是一个**循环神经网络**，自己更新内部的状态变量，并且根据关注向量以及输入的嵌入向量，来预测下面一个点击的位置。

有了编码器和解码器之后，一个难点是**如何生成最有可能的点击序列**。我们刚才提到了，整个模型其实可以预测多种不同的点击序列。因此，**生成最优可能的 K 个序列**就成为了必要的一个步骤。在这篇文章里，作者们利用了“**集束搜索**”（Beam Search）的方法来近似生成最佳的 K 个序列，在文章中，K 的值是 1024。

模型的训练采用了标准的**SGD**以及**Adam 优化法**，同时作者们还采用了“**梯度裁剪**”（Gradient Clipping）的方式来防止在优化过程中发生“爆炸问题”（Gradient Clipping）。

实验结果

作者们在 Yandex，俄罗斯的搜索引擎数据上进行了实验。因为之前没有类似的模型，因此文章并没有可以直接比较的其他模型。作者们主要进行评估的地方是，看历史数据中已经发生的点击序列，会不会被正确预测出，会不会出现在 K 个模型认为最有可能发生的点击序列中。这也就是作者们为什么选择 K 等于 1024 的原因，因为在这种情况下，接近 97% 的历史序列都在模型的预测序列中。

作者们还评估了模型能否预测出总的点击次数等一系列和点击预测有关的任务，论文中提出的模型都能够以接近 1 的概率预测所有的点击，并击败一些过去的基于概率的点击模型。可以说，提出的模型的确可以对用户在搜索页面的点击行为进行有效的建模。

小结

今天我为你讲了今年 SIGIR 2018 的一个篇精彩论文。

一起来回顾下要点：第一，我们详细介绍了这篇文章要解决的问题以及贡献，主要是对用户在搜索页面上的点击行为进行序列建模；第二，我们简要介绍了文章提出方法的核心内容，主要是编码器和解码器两个模块；第三，我们简单介绍了论文的实验结果。

最后，给你留一个思考题，如果针对多个连续的查询关键词的点击行为进行建模，你能否用这篇论文提出的思路来扩展模型呢？

欢迎你给我留言，和我一起讨论。

上一篇　126 | SIGIR 2018论文精读：如何利用对抗学习来增强排序模型的普适性？

下一篇　128 | CVPR 2018论文精读：如何研究计算机视觉任务之间的关系？

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。