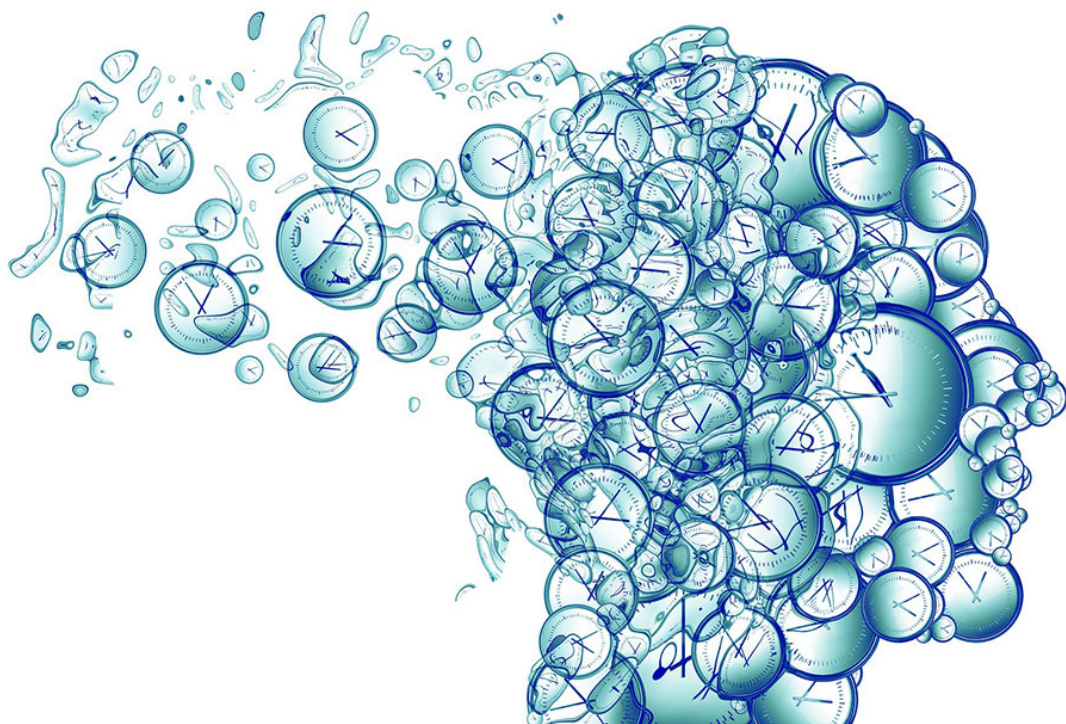


060 | WSDM 2018论文精读：看谷歌团队如何做位置偏差估计

2018-02-19 洪亮劼

AI技术内参

[进入课程 >](#)



讲述：初明明

时长 08:48 大小 4.03M



WSDM (International Conference on Web Search and Data Mining, 国际搜索和数据挖掘大会) 是每年举办一次的搜索、数据挖掘以及机器学习的顶级会议，其从 2008 年开始举办，已经有 11 届的历史。

尽管 WSDM 仅仅举办了 11 届，在计算机科学领域算是一个非常年轻的会议。但是，WSDM 快速积累的影响力已经使其成为了数据挖掘领域的一个顶级会议。根据谷歌学术搜索公布的数据，目前 WSDM 已经是数据挖掘领域仅次于 KDD 的学术会议，而 KDD 已经举办了 20 多年。

WSDM 的一大特点就是有大量工业界的学者参与，不管是投稿和发表论文还是评审委员会或者大会组织委员会的成员，都有很多工业界背景的人员参加。这可能也是 WSDM 备受关

注的一个原因，那就是大家对于工业界研究成果的重视，同时也希望能够从中学习到最新的经验。

2018 年的 WSDM 大会于 2 月 5 日到 9 日在美国的洛杉矶举行。今天，我们就来分享 WSDM 2018 上来自谷歌的一篇文章《无偏排序学习在个人搜索中的位置偏差估计》([Position Bias Estimation for Unbiased Learning to Rank in Personal Search](#))。这篇文章的核心内容是如何结合“因果推断”(Causal Inference)和排序学习(Learning to Rank)来对用户数据进行进一步无偏差的估计。

作者群信息介绍

这篇论文的所有作者都来自谷歌，我们这里对作者群做一个简单的介绍。

第一作者王选挥 (Xuanhui Wang) 2015 年起在谷歌工作。他之前在 Facebook 工作了三年，一直从事广告系统的开发；再往前，是在雅虎担任了两年的科学家。王选挥于 2009 年毕业于伊利诺伊大学香槟分校，获得计算机博士学位，他的博士生导师是信息检索界著名的华人学者翟成祥 (Chengxiang Zhai)。

第二作者纳达夫·古尔班迪 (Nadav Golbandi) 于 2016 年加入谷歌，之前在雅虎研究院担任了 8 年的主任级研究工程师 (Principal Research Engineer)，一直从事搜索方面的研发工作。在雅虎研究院之前，古尔班迪在以色列的 IBM 研究院工作了 6 年。他拥有以色列理工大学的计算机硕士学位。

第三作者迈克尔·本德斯基 (Michael Bendersky) 于 2012 年加入谷歌，一直从事个人以及企业信息系统 (Google Drive) 的研发工作。本德斯基于 2011 年从马萨诸塞州阿姆赫斯特分校 (University of Massachusetts Amherst) 毕业，获得计算机博士学位，他的导师是信息检索界的学术权威布鲁斯·夸夫特 (Bruce Croft)。

第四作者唐纳德·梅泽尔 (Donald Metzler) 也是 2012 年加入谷歌的，一直负责个人以及企业信息系统 (Google Drive) 搜索质量的研发工作。梅泽尔曾在雅虎研究院工作过两年多，然后还在南加州大学 (University of South California) 担任过教职。梅泽尔是 2007 年从马萨诸塞州阿姆赫斯特分校计算机博士毕业，导师也是信息检索界的学术权威布鲁斯·夸夫特。

文章的最后一个作者是马克·诺瓦克 (Marc Najork) 于 2014 年加入谷歌，目前担任研发总监 (Research Engineering Director) 的职位。诺瓦克之前在微软研究院硅谷分部工作

了 13 年，再之前在 DEC 研究院工作了 8 年。诺瓦克是信息检索和互联网数据挖掘领域的学术权威，之前担任过 ACM 顶级学术期刊 ACM Transactions on the Web 的主编。他发表过很多学术文章，引用数在七千以上。

论文的主要贡献

按照我们阅读论文的方法，首先来看这篇文章的主要贡献，梳理清楚这篇文章主要解决了什么场景下的问题。

众所周知，所有的搜索系统都会有各种各样的“**偏差**”（Bias），如何能够更好地对这些偏差进行建模就成为了对搜索系统进行机器学习的一个重要的挑战。

一种方式就是像传统的信息检索系统一样，利用人工来获得“**相关度**”（Relevance）的标签，不需要通过通过人机交互来获取相关度的信息。所以，也就更谈不上估计偏差的问题。

第二种，文章中也有谈到的，那就是利用传统的“**点击模型**”（Click Model）。点击模型是一种专门用来同时估计相关度和偏差的概率图模型，在过去 10 年左右的时间内已经发展得相对比较成熟。文章中也提到，大多数点击模型的应用主要是提取相关度信息，而并不在乎对偏差的估计是否准确。

第三种，也是最近几年兴起的一个新的方向，那就是利用“**因果推断**”（Causal Inference）和排序学习的结合直接对偏差进行建模。在 WSDM 2017 的最佳论文 [1] 中，已经让我们见识了这个思路。然而，在去年的那篇文章里，并没有详细探讨这个偏差的估计和点击模型的关系。

简言之，**这篇论文主要是希望利用点击模型中的一些思路来更加准确地估计偏差，从而能够学习到更好的排序结果**。同时，这篇文章还探讨了如何能够在较少使用随机数据上来对偏差进行更好的估计。这里，作者们提出了一种叫作“**基于回归的期望最大化**”（Regression-based EM）算法。

论文的核心方法

文章首先讨论了如果已知“**偏差值**”（Propensity Score），也就是用户看到每一个文档或者物品时的概率，我们就可以构造“**无偏差**”的指标，比如“**无偏差的精度**”（Unbiased Precision）来衡量系统的好坏。

这里，无偏差的效果主要是来自于重新对结果进行权重的调整。意思就是说，并不是每一个点击都被认为是同样的价值。总的来说，如果文档位于比较高的位置上，那权重反而会比较低，反之，如果文档位于比较低的位置上，权重反而较高。**这里的假设是一种“位置偏差”（Position Bias）假设。意思就是不管什么文档，相对来说，放在比较高的位置时都有可能获得更多的点击。因此，在较低位置的文档被点击就显得更加难得。**

这种情况下，一般都无法直接知道“偏差值”。因此，如何去估计偏差值就成了一个核心问题。

这篇文章在进行“偏差值”估计的方法上，首先利用了一个叫“**位置偏差模型**”（Position Bias Model）的经典点击模型，对偏差值和相关度进行了建模。“位置偏差模型”的假设是用户对于每一个查询关键字的某一个位置上的文档点击概率，都可以分解为两个概率的乘积，一个是用户看到这个位置的概率，一个就是文档本身相关度的概率。那么，位置偏差模型的主要工作就是估计这两个概率值。

如果我们能够对每一个查询关键字的结果进行随机化，那么，我们就不需要估计第一个概率，而可以直接利用文档的点击率来估计文档的相关度。但是，作者们展示了，彻底的随机化对于用户体验的影响。

另外一种方法，相对来说比较照顾用户体验，那就是不对所有的结果进行随机化，而仅仅针对不同的“配对”之间进行随机化。比如，排位第一的和第二的文档位置随机互换，然后第二的和第三的随机互换等等。在这样的结果下，作者们依然能够对偏差和相关度进行估计，不过用户的体验就要比第一种完全随机的要好。只不过，在现实中，这种方法依然会对用户体验有所损失。

于是，作者们提出了第三种方法，那就是**直接对位置偏差模型进行参数估计**。也就是说，不希望利用随机化来完全消除其中的位置概率，而是估计位置概率和相关度概率。

这里，因为有两个概率变量需要估计，于是作者利用了传统的“期望最大化”（EM）算法，并且提出了一种叫做“基于回归的期望最大化”的方法。为什么这么做呢？原因是在传统的期望最大化中，作者们必须对每一个关键字和文档的配对进行估计。然而在用户数据中，这样的配对其实可能非常有限，会陷入数据不足的情况。因此，作者们提出了利用一个回归模型来估计文档和查询关键字的相关度。也就是说，**借助期望最大化来估计位置偏差，借助回归模型来估计相关度。**

方法的实验效果

这篇文章使用了谷歌的邮件和文件存储的搜索数据，采用了 2017 年 4 月两个星期的日志。数据大约有四百万个查询关键字，每个关键字大约有五个结果。作者们在这个数据集上验证了提出的方法能够更加有效地捕捉文档的偏差。利用了这种方法训练的排序模型比没有考虑偏差的模型要好出 1% ~ 2%。

小结

今天我为你讲了 WSDM 2018 年的一篇来自谷歌团队的文章，这篇文章介绍了如何估计文档的位置偏差，然后训练出更加有效的排序算法。

一起来回顾下要点：第一，我们简要介绍了这篇文章的作者群信息；第二，我们详细介绍了这篇文章要解决的问题以及贡献；第三，我们简要地介绍了文章提出方法的核心内容。

最后，给你留一个思考题，如果要估计位置偏差，对数据的随机性有没有要求？

欢迎你给我留言，和我一起讨论。

参考文献

1. Thorsten Joachims, Adith Swaminathan, and Tobias Schnabel. [Unbiased Learning-to-Rank with Biased Feedback](#). Proceedings of the Tenth ACM International Conference on Web Search and Data Mining (WSDM '17). ACM, New York, NY, USA, 781-789, 2017.


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 059 | 2017人工智能技术发展盘点

下一篇 061 | WSDM 2018论文精读：看京东团队如何挖掘商品的替代信息和互补信息

精选留言 (1)

 写留言



rkq@geekba...

2018-02-19

从事搜索领域请问有哪些会议和期刊需要关注呢？

展开 ∨

