离群偏差对用户点击的影响

Fatemeh Sarvi
AIRLab，阿姆斯特丹大学，阿姆斯特丹，荷兰 f.sarvi@uva.nl

Ali Vardasbi
阿姆斯特丹大学，阿姆斯特丹，荷兰 a.vardasbi@uva.nl

Mohammad Aliannejadi
阿姆斯特丹大学，阿姆斯特丹，荷兰 m.aliannejadi@uva.nl

Sebastian Schelter
阿姆斯特丹大学，阿姆斯特丹，荷兰 s.schelter@uva.nl

Maarten de Rijke
阿姆斯特丹大学，阿姆斯特丹，荷兰 m.derijke@uva.nl

arXiv:2305.00857v1 [cs.IR] 1 May 2023

摘要

用户交互数据是反事实学习排序（CLTR）中的重要监督信号来源。此类数据受到呈现偏差的影响。无偏学习排序（ULTR）的大量研究聚焦于位置偏差，即排名较高的条目更有可能被检查和点击。项目间依赖关系也会影响检查概率，其中排名中的离群项就是一个重要例子。离群项被定义为能够被观察到地偏离其余条目、因此在排序中脱颖而出的条目。在本文中，我们识别并介绍了离群项带来的偏差：用户倾向于更多点击离群项及其邻近项目。

为此，我们首先进行一项受控实验，以研究异常值对用户点击的影响。接下来，为了检验受控实验的发现是否适用于自然情境，我们探索来自电子商务平台的真实点击日志。我们发现，在这两种情境下，在相同排名中，用户点击离群项的次数往往显著多于非离群项。我们还发现，这种倾向适用于所有位置；也就是说，对于任何特定位置，与呈现为非离群项相比，条目呈现为离群项时会获得更多交互。根据分析，我们得出结论：异常值对点击的影响是一种应在无偏学习排序中加以处理的偏差。因此，我们提出一种异常值感知点击模型，同时考虑异常值偏差和位置偏差，称为异常值感知的位置模型（OPBM）。我们基于 OPBM 估计点击倾向，并通过在真实世界电子商务数据和半合成数据上开展大量实验，验证了该异常值感知点击模型的有效性。结果表明，在排序性能和真实相关性估计方面，OPBM 优于基线方法。

CCS 概念

• **信息系统** →学习排序。

关键词

点击偏差；异常值；反事实学习排序

允许免费出于个人或课堂用途制作本作品部分或全部内容的数字或纸质副本，但不得为营利或商业利益而制作或分发副本，且副本必须附有本声明以及首页上的完整引文。本作品中第三方组件的版权必须得到尊重。如需用于其他用途，请联系所有者/作者。*SIGIR '23，2023*年*7*月*23*日至*27*日，台湾台北©2023 年版权归所有者/作者所有。ACM ISBN 978-1-4503-9408-6/23/07。
https://doi.org/10.1145/3539618.3591745

ACM 引用格式：
Fatemeh Sarvi、Ali Vardasbi、Mohammad Aliannejadi、Sebastian Schelter 和 Maarten de Rijke。2023。《离群偏差对用户点击的影响》。载于第*46*届*ACM*国际信息检索研究与开发会议论文集（*SIGIR '23*），*2023*年*7*月*23*日至*27*日，台湾台北。ACM，纽约，NY，美国，10 页。https://doi.org/10.1145/3539618.3591745

1 引言

排序系统通过优化排序决策来提升用户满意度。隐式用户反馈是重要的监督信号来源，反映了真实用户的偏好。然而，用户交互数据（例如点击）受到呈现偏差的影响，因此将其直接作为训练数据使用可能会产生严重误导 [18]。

无偏学习排序（ULTR）的许多研究都聚焦于位置偏差 [3, 17, 19, 43],，即排名靠前的结果比排名靠后的结果更容易被用户查看并点击 [17]。除位置外，还有其他几个因素会影响用户的考察模型和点击行为 [1, 11, 27,34, 44]。以往研究表明，项目间依赖关系会影响用户的相关性判断以及对条目的考察顺序 [12, 34, 44]。离群项的存在是项目间依赖关系的一种具体情形 [34]。Sarvi 等人 [34] 将排序中的离群项定义为：在商品特征方面明显偏离列表中其他条目的条目，因此格外醒目并能吸引用户的注意力。例如，在电子商务搜索场景中，如果页面上只有一个条目带有“畅销商品”标签，就可以将其视为离群项，因为该标签使它有别于排序中的其他条目，从而吸引用户的注意力。

离群偏差。列表中的离群项可能会改变检查概率，使离群项（如果存在）及其相邻条目的检查概率高于位置偏差假设所分配的概率 [34]。

尽管已有研究表明，离群项会影响检查概率 [34],，但其对用户点击行为的影响尚不明确。在本研究中，我们假设离群项的存在会导致点击产生偏差。我们将这一现象称为离群偏差，并旨在理解和应对这一影响。首先，我们开展了一项用户研究，比较特定条目在两种状态下的点击率（CTR）：一次以离群项形式呈现，另一次则作为列表中的非离群项呈现。我们发现，用户对某个条目的行为会因其离群状态而异。当某个条目作为离群项呈现在排序中时，其点击率始终高于作为非离群项呈现时。接下来，我们将

检查这些发现能否推广到自然情境，我们对来自Bol.com（一家总部位于欧洲的知名电子商务平台）的真实世界搜索日志进行了分析。结果证实了用户研究的发现。此外，我们观察到，平均而言，同一列表中的离群项获得的点击显著多于非离群项。而且，用户往往会更多地与至少包含一个离群项的列表进行互动。

离群偏差与上下文偏差。我们发现，离群偏差会影响用户点击：用户更可能与作为离群项呈现的条目及其相邻条目互动。与离群偏差最接近的概念是新闻推送推荐中的上下文偏差 [44]。存在上下文偏差时，与至少一个非常相似的条目相邻的条目，其点击率低于与不相似条目相邻时的点击率。这与离群偏差不同，后者强调离群项与列表中其余条目之间的差异。此外，根据 [34] 的定义，可观测性是检测排序中离群项的关键因素；而上下文偏差并非如此。

考虑异常值。根据用户研究和日志分析的发现，我们得出结论：在为无偏学习排序消除用户点击的偏差时，应考虑异常值的影响。为此，我们提出了一种基于检查假设的点击模型，称为异常值感知的位置模型（OPBM），该模型同时考虑异常值偏差和位置偏差。OPBM 假设点击概率取决于（i）检查，（ii）相关性，以及（iii）异常值的位置（如果存在）。我们使用基于回归的期望最大化算法，根据所提出的点击模型 OPBM 估计点击倾向。我们验证了所提出的离群值感知模型在同时存在位置偏差和离群偏差时估计倾向值的有效性。按照 [6, 19, 26]，我们在实验中采用半合成设置；该设置提供的真实相关性标签可用于评估相关性估计。此外，借助模拟点击，我们能够控制位置偏差和离群偏差的程度。实验结果表明，在排序性能（NDCG@10）和真实相关性估计方面，OPBM 优于基线方法。

主要贡献。本文的主要贡献如下：(i) 我们识别并研究了一种由项目间依赖关系引起的新型点击偏差，称为离群偏差；(ii) 通过对用户研究结果和真实世界搜索日志进行广泛分析，我们证实了关于离群偏差存在的假设；(iii) 为了应对这一影响，我们提出了一种异常值感知点击模型，该模型能够考虑离群项（如果存在）以及位置偏差；(iv) 通过基于真实世界数据和半合成实验的实证分析，我们展示了所提出的异常值感知模型在估计点击倾向方面的有效性；(v) 我们公开提供用户研究数据，以及实现基线方法和 OPBM 的代码。

2 排序中的离群项

排名中的离群项是指在一次呈现给用户的一组条目中，明显与众不同的条目。我们采用 [34] 中的以下定义来引入所谓的离群项：

定义 *2.1*（可观测特征）。可观测的条目特征，F 是列表中条目的一种特征，可以纯粹是呈现形式上的特征（例如图像、标题字体大小和折扣标签）。

定义 *2.2*（异常程度）。设 M 为任意离群检测方法，F_i 为与条目 i 对应的可观测特征，考虑列表 C 中的所有条目。条目 i 的异常程度是 M 针对 F_i 相对于 C 计算出的值，即 $M(F_i|C)$。该值表示相应条目在特征 F 方面与集合中其他元素的差异程度。

定义 *2.3*（排名中的离群项）。对于任意离群项检测方法 M，如果 M 根据某个可观测特征 F 的异常程度，并结合列表的上下文，将排序列表中的条目 i 识别为离群项，我们就称该条目为离群项。

在3.2节中，我们将介绍本文所采用的可观测特征和离群项检测方法。

3 条目异常程度对点击的影响

Sarvi 等人 [34] 发现，离群项会获得用户更多的注意力。然而，尚不清楚条目的异常程度是否也会影响用户的点击。在本节中，我们将回答第一个研究问题：（**研究问题1**）条目排名中是否存在离群偏差？为此，我们首先开展一项用户研究，考察异常程度作为影响点击的唯一变量所产生的效应。接下来，我们需要检验研究发现能否推广到自然情境中。换言之，我们旨在确立其生态效度 [8, 21]。为此，在3.2 节中，我们将分析真实世界的点击日志，以验证研究发现。

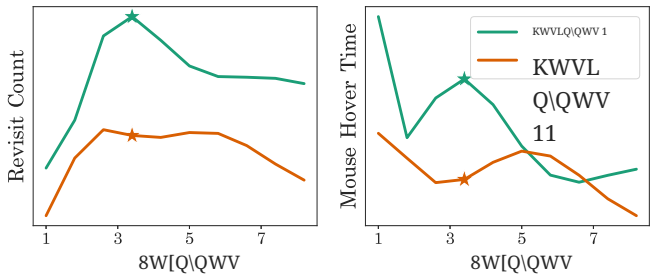
3.1 用户研究

本节将介绍用户研究的结果。我们的主要目标是了解条目的异常程度是否会影响用户点击，而不受条目相关性和位置的影响。

实验设置。我们模拟了Bol.com，这是一个颇受欢迎的欧洲在线市场。我们请参与者像平常一样浏览搜索引擎结果页面，寻找他们偏好且认为相关的条目。我们采用列表视图，每页显示 20 个条目，参与者可以滚动列表查看所有条目。我们设置了两个查询；对于每个查询，我们都将某个特定条目分别作为离群项和普通条目展示一次。我们将该特定条目称为目标条目，并将这两种不同的呈现方式分别称为 条件 *I* 和 条件 *II*。在条件 I 中，目标条目相对于一组可观测特征（例如条目类别、[1] 价格、折扣标签和星级评分）是一个离群项。我们旨在比较用户在每个查询下这两种条件中的行为。我们保持相关性和位置偏差等其他因素在两种条件下不变。为消除位置偏差的影响，我们始终将该条目显示在排名第 4 的位置；为保持其与查询的相关性程度不变，我们只调整周围的条目，以改变目标条目的异常程度。[2]

我们还准备了一份 Qualtrics [31] 调查问卷。问卷包含任务说明、有关说明内容的选择题、查询、示例链接，以及最后的几道人口统计学问题。在说明中，我们介绍了研究的总体目标，并

[1]请注意，此特征也会影响相对于项目图像的异常程度。[2]为检验我们关于项目间依赖的假设，我们在此假设文档的相关性仅取决于查询。



(a) 回访次数 **(b) 鼠标悬停**

图 1：在我们用户研究示例之一的两种条件下，回访次数 (a) 和鼠标悬停时间 (b)。目标项目的位置以星号标记。图中显示，当目标项目作为离群项呈现时（条件 I），用户对其的参与度更高。

请参与者仔细阅读说明。我们描述了如何与结果页面互动，包括浏览结果、滚动列表，以及点击看起来有趣的条目。参与者可以点击条目以打开其详情页面。在说明中，我们鼓励参与者点击自己感兴趣的条目，但点击并非强制要求。我们要求参与者先阅读并理解查询，然后像是自己提交了查询一样浏览结果页面。

参与者。我们通过 Prolific 平台招募了 40 名工作者，他们来自我们市场平台运营的国家 [29]。参与者中，14 人为女性，23 人为男性，3 人填写了其他性别。大多数参与者（27 人）年龄在 25 至 44 岁之间，另有 10 人年龄更小、3 人年龄更大；33 名参与者表示他们至少每月网购一次。

指标。报告中，我们采用三项基于参与者与排序结果交互的度量：（i）回访次数，即参与者平均查看某个条目的次数（由于滚动所致）；（ii）鼠标悬停时间，即参与者平均在某个条目上停留的时间；（iii）每种条件下目标项目的点击率（CTR），这是本研究的主要指标。

研究发现。我们预计，在条件 I 下，用户与目标条目的交互会更多。由于两种条件下其他因素均保持不变，因此用户行为的任何差异都可归因于项目间依赖关系。

图 1 展示了一个示例中不同位置和条件下的回访次数（图 1a）以及鼠标悬停时间（图 1b）。两张图均显示，当目标项目以离群项的形式呈现时，用户参与度更高。第二个示例中也呈现出相同的模式。平均而言，参与者对离群项的回访次数更多，也花费更多时间进行查看。这些发现与 Sarvi 等人开展的眼动追踪实验结果一致 [34],，该实验表明，平均而言，用户会对离群项投入更多注意力。然而，我们的主要目标是研究这种注意力的提升是否会带来更多点击。

表 1 报告了两个示例中目标条目在两种条件下的点击率。在两个示例中，不同条件下报告的点击率都存在显著差异，这表明当目标条目作为离群项展示时，它不仅获得了更多点击，也获得了更多曝光。

表 1：用户研究两个示例中目标项目所在位置的点击率。目标项目以离群项形式呈现时（条件 I）获得的点击更多。

	条件 I	条件 II
查询 1	0.944	0.166
查询 2	0.880	0.091

3.2 真实世界点击日志

第 3.1 节的研究结果通过受控实验设置证实，条目的异常程度会影响用户的点击行为。不过，我们仍需检验这一假设的生态效度。为此，我们展示了对电子商务平台真实世界搜索日志中点击探索情况的观察结果。我们特别关注对数据的探索，以研究排名中是否存在离群项，以及它们对点击数据的影响。请注意，我们仅将这些数据用于点击分析和参数估计（第 5节）。

数据收集。我们收集了连续 20 天的搜索查询日志。数据集的每一行包含七种可观测项目特征，详见表 2，以及用户的交互信号：展示次数和点击。

定义 *3.1*（曝光）。一次曝光表示用户查看搜索引擎呈现的某个条目的次数。如果条目排在低位，可能不会出现在用户可见的窗口中，从而导致曝光次数为零。另一方面，由于滚动，曝光次数可能大于一次。

我们选择了适用于不同类别、用户可观察到，且先前研究已证明会对用户购买决策产生重要影响的商品特征 [4, 20]。由于图像会给本研究带来过度的复杂性，我们在点击行为探索中未纳入商品图片。

大多数搜索引擎将搜索结果页面的多样性视为一项质量指标 [5]。这可能产生副作用，即返回的排名中可能包含离群项。因此，查询日志是研究离群项对用户点击行为影响的重要数据来源。首先，我们根据是否存在离群项，将排名定义为以下两种类型：

定义 *3.2*（正常排名）。我们将不包含离群项的排名称为正常排名。正常排名既可以由同质的条目集合构成，也可以由多样化集合构成。

定义 *3.3*（异常排名）。我们将包含至少一个离群项的列表定义为异常排名。

异常值检测。我们根据表 2 中描述的特征，并结合第 2 节所述列表中所有条目的上下文，检查每个条目的异常程度。如果某个条目基于至少一项特征被检测为离群项，则该条目就是离群项。

我们使用四分位距规则检测异常值，并将特征值与上界/下界之间的绝对差值视为相应项目的异常程度（参见第 2节）。我们对特征值进行归一化，使所有可观测特征的异常程度范围均为单位区间。我们将异常程度的阈值设为 0.5，这意味着只有当一个项目的得分与上界/下界之间的绝对差值

表2：用于表示条目的可观测特征描述。

特征名称	缩写	描述
价格	-	商品售价。
促销标签	促销	表示各种促销活动的通用红色标签，例如“价格优惠”和“精选优惠”。
高折扣标签	折扣	商品高折扣红色双片标签（不同于促销）。
库存状态标签	库存数量	表示商品有货或缺货状态的绿色标签。
用户星级评分	评分	商品的用户平均星级评分，以标准的五星模板呈现。
“select”标签	精选	绿色标签，表示该商品为精选商品（类似于 Amazon Prime）。
标题长度	-	项目标题中的词元数量。

表3：用户在所有异常排名中与离群项和非离群项的交互平均值。我们使用 Stu- dent t 检验，并以 $p < 0.001$ 进行统计显著性检验。

	平均点击次数	平均展示次数	平均点击率
异常值	0.202*	1.381*	0.142*
非离群项	0.137	1.346	0.098
总计	0.149	1.352	0.106

大于 0.5 时，才将其标记为离群项。

后处理。我们根据数据中的展示信号，筛除用户未查看的排名部分。这样一来，排名列表大小最小为 3。然而，根据定义，长度少于 4 的列表中的异常程度没有意义，因此我们从数据集中移除了这些排名。我们还移除了包含赞助项目的页面，以避免此类项目对结果产生任何潜在影响。剩余的 10,903 个异常排名平均长度为 10.24，中位数为 8.0。

离群项对点击率的影响。在分析的第一步中，我们旨在了解用户在异常排名中与离群项的交互是否有所不同。为此，我们考察这类排名，并比较离群项平均获得的交互次数与同一排名中非离群项的交互次数。我们将点击作为交互指标。

由于正常排名对我们当前的分析不提供任何信息，我们只保留异常排名。表 3 展示了异常排名中离群项和非离群项的平均点击量、展示次数和点击率。[3] 我们按页面计算点击率（即每个条目的点击次数除以展示次数），并报告所有排名的平均值。我们的研究结果表明，离群项的点击率和平均点击量均显著更高

与同一页面上的非离群项相比，离群项的展示次数更多。此外，我们还发现，离群项的展示次数也显著更高，这与文献 [34]中报告的一项眼动追踪实验的发现一致。

各位置离群项的影响。为确保表3 中报告的较高点击率并非由位置偏差导致，我们考察了各位置的点击率值。图2 展示了结果。总体而言，所有位置上离群项的点击率都较高，这表明这些条目比非离群项获得了更多交互。

接下来，为了进一步研究离群项如何改变用户的点击行为，我们比较整个排序中异常值位置与非离群项所在位置的点击率。为了更好地描绘离群项对不同位置的影响，我们考察包含

[3]请注意，本节报告的数值是根据筛选后的搜索日志子集计算得出的，因此并不能代表数据的真实统计情况。

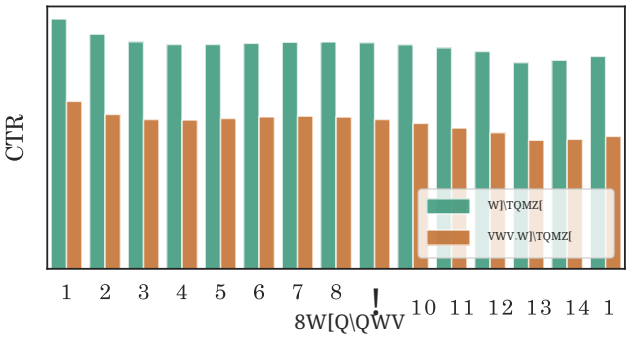


图2：按排名比较离群项与非离群项的点击率。离群项的点击率始终较高。

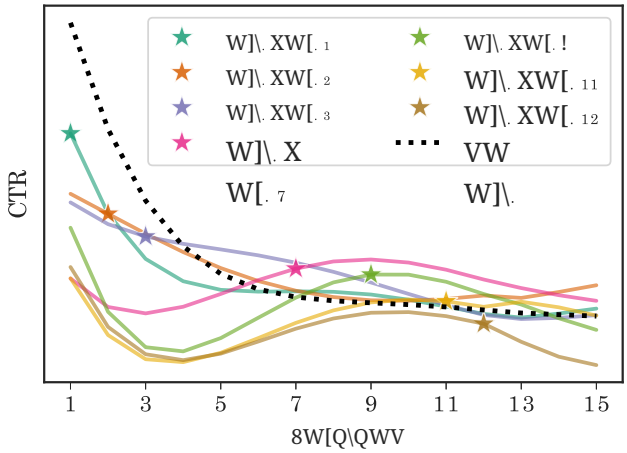


图3：按离群项位置分组的异常排名中各排名的点击率。离群项的位置以星号标记。数值经过 Savitzky-Golay 滤波器平滑处理。彩色查看效果最佳。

恰好一个离群项的排序；我们重点关注前15个位置。值得一提的是，我们的数据中异常排名有不到35% 包含多个离群项。我们根据离群项的位置对异常排名进行分组。图3 展示了结果。黑线表示正常排名的点击率。正如预期，这条线符合位置偏差，即点击某个条目的概率会随着其排名降低而下降。

图 3 中的其他线条显示了在位置 $r \in \{1、\ldots\ldots、15\}$存在一个离群项的各排名组的点击率。为便于查看，我们只展示了部分位置的结果。其他排名也呈现出类似的模式。我们仅展示集合大小至少占整个集合

的 1% 的排名组。在图 3 中，每个星号表示离群项的位置。我们观察到，当排名中存在离群项时，点击率分布不同于基于位置的假设。此外，对于位置 3 之后的各个位置，我们观察到离群项位置及其附近的点击率有所上升。

另一个有趣的观察结果是，与正常页面相比，距离离群项越远的条目，其获得的注意力相对于其排名而言越少。此外，我们发现，对于位置 3 之后的位置，异常值的点击率高于正常页面中相同位置的点击率，这与我们在图 2 中的发现一致。

不同离群项类型的影响。可以认为，不同类型的离群项可能对用户感知产生不同的影响。例如，如果将价格视为可观测特征，那么与便宜的离群项相比，价格极高的离群项被购买的可能性可能更低。我们假设，从统计指标的整体表现来看，这两种类型的影响可能会相互抵消。因此，为了检验这一假设，我们首先尝试根据常识，基于可观测特征采用非正式定义，将离群项划分为正向离群项和负向离群项两组。例如，在前面的例子中，价格昂贵的条目属于负向离群项，而价格便宜的条目则属于正向离群项。根据这一划分方式，对于价格这一特征，我们发现正向离群项和负向离群项的平均点击次数分别为 0.193 和 0.147，二者均显著高于非离群项（0.125）。在所有可观测特征中，无论是曝光还是点击次数，我们都观察到了相同的趋势。基于这些结果，我们否定了上述假设，并继续采用最初的离群项与非离群项划分方式。

进一步说明。我们还考察了异常排名：其中某个特定条目会在固定排名位置上反复出现，至少一次作为离群项，另一次作为正常项。我们汇总了所有此类排名，并观察到，条目作为离群项时平均获得 0.169 次点击，而作为列表中的常规项时平均获得 0.130 次点击。将异常排名与长度分布相似的一组正常排名进行比较（正常排名/异常排名的均值=10.09/10.30，中位数=8.0/8.0，标准差=4.95/5.82），我们发现存在离群项时，每个会话的平均点击次数更高。具体而言，正常排名的平均点击次数为 0.139，异常排名则为 0.149，且具有 $p < 0.001$ 显著性。

总结。总而言之，从第 3.1 节我们得知，用户对条目的行为会因其离群状态而异（即该条目在排序中是否作为离群项呈现）。特定条目作为离群项呈现时，其点击率始终高于作为非离群项呈现时。第 3.2 节证实了用户研究的发现。此外，我们观察到，在相同列表中，离群项平均获得的点击显著多于非离群项。用户也倾向于与至少包含一个离群项的列表进行更多互动。本节证实了离群项对点击的影响。我们将这一效应称为离群偏差。在下一节中，我们提出一种同时考虑离群偏差和位置偏差的点击模型。

4 异常值感知的位置模型

将隐式反馈朴素地用于学习排序可能会产生误导，因为它受到呈现偏差的影响。因此，对检查偏差至关重要 [13, 18]。

基于位置的模型。通常，页面上排名较高的条目更有可能被检查。位置偏差通过检查假设进行形式化建模，该假设指出，用户必须检查某个条目并认为其相关，才会点击它。处理位置偏差的一种广泛使用的点击模型是基于位置的模型（PBM）[19, 43]。尽管 PBM 被认为是一种简单的解决方案，但其效果与更复杂的点击模型相当 [12]。PBM 假设，条目的排名是影响用户对该条目进行检查的唯一参数。检查某个条目意味着在点击等后续交互发生之前先查看并评估它。

给定针对查询 q 的响应中排名为 k 的条目 d，假设采用 PBM，点击 d 的概率为：

$$P(C = 1 \mid q, d, k) = P(E = 1 \mid k) \times P(R = 1 \mid d, q), \quad (1)$$

其中，$P(E = 1 \mid k)$ 表示用户检查排名为 k 的条目的概率，也称为 倾向性；$P(R = 1 \mid d, q)$ 表示条目对 (d, q) 的相关概率。我们分别将这些概率称为 θ_k 和 $\gamma_{q,d}$。

异常值感知的位置模型。PBM 简单地假设，影响倾向性的唯一因素是排名。在第 3节中，我们表明用户更可能点击离群项，因此我们假设倾向性还取决于是否存在离群项。

值得注意的是，即使在离群项之间，我们也观察到异常值间位置偏差——排名较高的离群项会获得更多点击。

因此，为了对这些依赖关系进行建模，我们提出了一种能够感知离群项的基于位置的模型，称为 OPBM。该模型除了考虑位置外，还考虑离群项的影响，具体如下：

$$P(C = 1 \mid q, d, k, o) = P(E = 1 \mid k, o) \times P(R = 1 \mid d, q), \quad (2)$$

其中，o 表示排序中离群项的位置。请注意，PBM 是 OPBM 的一个特例：对于正常排名，OPBM 可简化为 PBM。

我们基于以下假设提出了这个模型（见式（2））：排名 k 处的检查概率除了取决于 o 外，还取决于离群项的位置 k。该模型包含 $K \times O$ 个参数，其中 K 和 O 分别是所有排名和离群项位置的集合，可通过点击数据进行估计。

倾向估计。这里，我们介绍如何根据常规点击估计能够感知离群项的位置偏差。基于基于回归的期望最大化（REM）算法的思想 [43]，我们提出通过回归函数进行估计，同时估计参数 θ_k 、o 和 $\gamma_{q,d}$。

我们采用标准的期望最大化算法（期望最大化算法），旨在找到使整个点击日志的对数似然最大化的参数。生成形式为 $\mathrm{L} = (c, q, d, k, o)$ 的点击日志的对数似然为：

$$\log P(\mathcal{L}) = \sum_{(c,q,d,k,o) \in \mathcal{L}} c \log \theta_{k,o}\gamma_{q,d} + (1-c)\log(1-\theta_{k,o}\gamma_{q,d}). \quad (3)$$

在此，我们旨在基于 L 中的数据点估计参数 $\theta_{k、o}$ 和 $\gamma_{q、d}$。在每次迭代中，EM 在期望步骤和最大化步骤之间交替进行，以计算参数的新估计值。在第 $t+1$ 次迭代的期望步骤中，我们计算

基于第 t次迭代时估计参数、与检查倾向（E）和真实相关性（R）对应的隐变量：

$$
\begin{aligned}
P(E=1,R=1 \mid C=1,q,d,k,o) &= 1,\\
P(E=1,R=0 \mid C=0,q,d,k,o) &= \frac{\theta^t_{k,o}(1-\gamma^t_{q,d})}{1-\theta^t_{k,o}\gamma^t_{q,d}},\\
P(E=0,R=1 \mid C=0,q,d,k,o) &= \frac{(1-\theta^t_{k,o})\gamma^t_{q,d}}{1-\theta^t_{k,o}\gamma^t_{q,d}},\\
P(E=0,R=0 \mid C=0,q,d,k,o) &= \frac{(1-\theta^t_{k,o})(1-\gamma^t_{q,d})}{1-\theta^t_{k,o}\gamma^t_{q,d}}.
\end{aligned} \tag{4}
$$

然后，我们计算 P（$E=1 \mid c$、q、d、k、o）和 P（$R=1 \mid c$、q、d、k）中每个数据点在$\mathcal{L}$上的边缘概率。我们保持γ_q、$_d$ 的估计不变，也就是说，学习排序（LTR）模型在不知道离群项位置的情况下进行训练，只有倾向估计会受到其影响。这得到迭代 $t+1$时的最大化步骤，在该步骤中，我们更新参数，以最大化公式3 中的似然，如下所示：

$$
\begin{aligned}
\theta^{t+1}_{k,o} &= \frac{\sum_{c,q,d,k',o'} \mathbb{I}_{k'=k,o'=o}.(c+(1-c)P(E=1 \mid c,q,d,k,o))}{\sum_{c,q,d,k',o'} \mathbb{I}_{k'=k,o'=o}},\\
\gamma^{t+1}_{q,d} &= \frac{\sum_{c,q',d',k} \mathbb{I}_{q'=q,d'=d}.(c+(1-c)P(R=1 \mid c,q,d,k))}{\sum_{c,q',d',k} \mathbb{I}_{q'=q,d'=d}}.
\end{aligned} \tag{5}
$$

期望最大化算法的最大化步骤需要多次出现对（q，d），其中 d显示在不同的位置。为了解决点击稀疏问题和可能存在的隐私问题，我们修改了迭代 $t+1$时的最大化步骤，通过回归 [43] 估计γ_q、$_d$参数。因此，给定表示对（q，d）的特征向量x_q，我们拟合函数 f（x_q，d）（例如梯度提升决策树（GBDT）），以计算γ_q、$_d$的估计值。因此，我们的最大化步骤是寻找一个回归函数f（x），利用期望步骤中估计的参数，使公式 3 最大化。在REM算法中，通过对二元变量进行采样，将该回归问题转换为分类问题，该二元变量表示来自分布 P（$R=1 \mid c$，q，d，k）的x_q、$_d$的相关性标签。这会得到一个形式为（x_q，$_d$，r_q，d）的训练集，其交叉熵目标如下：

$$
\sum_{x,r} r\log(f(x)) + (1-r)\log(1-f(x)). \tag{6}
$$

说明。 除了单个无偏模型之外，另一种选择是训练多个学习排序模型，作为针对不同异常值位置的无偏专家模型。这种方案主要有两个缺点。第一，采用专家模型意味着每个专家模型仅使用包含特定位置异常值的部分数据进行训练。这不仅可能导致训练效果欠佳，也使得难以将该模型与基于位置偏差模型（PBM）的REM基线进行比较。第二，在现实场景中，将一组 K 专家模型作为排序器并不理想。理想情况下，应有一个无需异常值位置信息即可使用的单一无偏模型。

5 实验设置

遵循反事实学习排序（CLTR）领域的大量既有研究 [6, 15, 19, 26, 38], ，我们采用半合成设置来进行我们的实验，即我们从现有的学习排序数据集中采样查询、文档和相关性标签，但根据在专有数据上估计的概率点击模型模拟用户点击。

包含真实相关性标签的学习排序数据集使我们能够评估OPBM及其他基线方法的相关性估计，以及它们对排序性能的影响。此外，半合成设置使我们能够控制模拟点击的位置偏差和离群偏差。

5.1 数据

公开的学习排序数据。 遵循以往关于点击学习排序 [19, 37, 38], 的研究，我们使用 Yahoo! Webscope [10] 和 MSLR-WEB30k [30] 数据集。这两个数据集总共各包含约 30k 个查询，每个查询都对应一个文档列表。Yahoo!和MSLR数据集的查询-文档特征向量维度分别为 501 和 131。这两个数据集都包含 5 个等级的分级相关性标签。按照以往研究的做法，我们将等级 {3,4} 视为相关，将等级 {0, 1, 2} 视为不相关。Yahoo!和MSLR数据集的训练集分别包含 20k 和 19k 个查询，以及 473k 和 2.2M 篇文档。Yahoo!和MSLR数据集的测试集分别包含 6.7k 和 6k 个查询，以及 163k 和 749k 篇文档。

专有数据。我们使用第 3.2 节所述的真实点击日志数据进行实验，并将其称为专有数据。我们使用一个包含 24 个特征的特征向量来表示每个查询-文档对，其中包含相关性特征和产品的可观测特征。我们将这些特征用于学习排序模型。此外，我们还采用第 3.2 节所述的设置，使用四分位距规则，根据可观测特征检测离群项（见表 2）。由于该数据集中排名的平均长度为 10.24，中位数为 8.0，因此实验中使用前 10 个条目。

5.2 点击模拟

我们遵循先前的研究 [6, 19, 26, 37, 38]，从每个公开数据集中均匀随机抽取 1% 的查询，用于训练人工生产排序器。我们在该生产排序器生成的排名上应用概率点击模型，为半合成实验模拟点击。我们采用异常值感知的位置模型，并使用不同的检查概率近似方法。相关性γ_q，$_d$基于数据集中记录的相关性标签。遵循先前的研究 [19, 38]，我们使用二元相关性：

$$
P(R=1 \mid q,d) = \gamma_{q,d} = \begin{cases} 1 & \text{if relevance_label}(q,d) > 2 \\ 0 & \text{otherwise} \end{cases}. \tag{7}
$$

为了模拟离群偏差，我们采用以下两种策略：

OPBM$_{Real}$。我们使用 OPBM 估计的倾向值（见第 4 节）处理我们的专有数据集。在数据集中的所有异常排名中，64% 仅包含一个离群项。由于改善超过一半查询的排序能够在现实场景中带来显著提升，我们首先处理这一多数情况。因此，在该模型中，我们关注包含一个离群项的排名。于是，OPBM 的输出至多是一个 $K \times K$ 矩阵，对应于排名和异常值位置的所有组合，其中 $K=10$ 在我们的实验中。我们使用该矩阵来近似 P（$E=1 \mid k$，o）。

OPBM$_{\mathcal{G}}$。这里，我们假设离群项对用户点击的影响服从以离群项位置为中心的高斯分布。因此，对于每个 k，我们计算离群偏差的线性插值

分布和位置偏差分布，如下所示：

$$OPBM_{\mathcal{G}}(q,d,k,o) = \gamma_{q,d}((1-\alpha)\theta_k + \alpha\mathcal{G}(\mu = o, \sigma^2)), \quad (8)$$

其中，G 是一个高斯分布，参数为 $\mu = o$，用于模拟异常值效应。我们设置 $\sigma = 1$，并通过改变 α 的取值进行实验。为了模拟包含多个异常值的排名中的点击，我们计算所有异常值位置（$O_\prime$）对应的 $\text{OPBM}_{\mathcal{G}}$ 的平均值，如下所示：

$$OPBM_{\mathcal{MG}}(q,d,k,O') = \frac{1}{|O'|}\sum_{i\in O'} OPBM_{\mathcal{G}}(q,d,k,i). \quad (9)$$

根据我们的专有数据，91% 的异常排名最多包含两个异常值。因此，在实验中，我们重点研究最多包含两个异常值的排名。

我们遵循先前的工作 [15, 19, 26, 38]，将位置偏差定义为与项目排名成反比：

$$\theta_k = \frac{1}{k}. \quad (10)$$

我们使用 100 万次模拟点击训练学习排序模型[4]。

5.3 用于比较的方法

我们的主要目标是引入一种新类型的偏差，并研究其对点击倾向的影响。因此，只需将我们的离群项感知点击模型与仅校正位置偏差的基线方法进行比较。为此，我们将 OPBM 与以下估计器进行比较：

- **朴素方法** 是一种不进行校正的模型，其中每次点击都被视为无偏相关性信号。
- **PBM** 是原始的逆倾向评分（IPS）估计器 [19, 43]，仅校正位置偏差。

5.4 评估指标

为了衡量不同方法所达到的排序性能，我们采用归一化折损累计增益（NDCG）。我们还考虑交叉熵（CE），它衡量真实相关性与估计器计算出的无偏相关性之间的差异；该指标反映了模型估计相关性的准确程度，且与学习排序模型无关。由于我们使用二元相关性，因此分别将 PBM 和 OPBM 的校正后的点击（即 c/θ_k 和 $c/\theta_{k,\,o}$）作为预测值，并将真实相关性值作为标签，计算二元交叉熵。为提高可读性，我们报告 CE 的平均值，而不是其总和。

6 结果

在第 3 节中，我们已经回答了 **(研究问题1)** 关于排序列表中是否存在离群偏差的问题。本节将回答以下研究问题：（**研究问题2**）与基线方法相比，我们的离群值感知模型 OPBM 表现如何？（**研究问题3**）在不同离群偏差严重程度条件下，OPBM 的表现如何？（**研究问题4**）对于排名中存在多个异常值的情况，OPBM 的泛化能力如何？

6.1 使用 OPBM 进行倾向估计

我们通过比较 OPBM 在倾向估计方面的整体表现来回答（**研究问题2**）。图4展示了 OPBM 估计的倾向值$_{Real}$（见第5.2节），这些倾向值对应于我们的专有数据集中存在足够数量异常值的前8 个排名。我们发现，

[4]我们使用 allRank 实现学习排序 [28]。

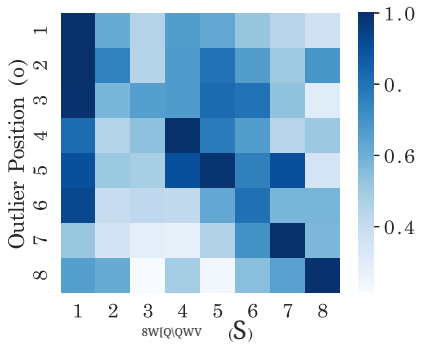


图4：在专有数据上，OPBM 针对每个异常值位置计算的前 8 个排名的点击倾向。异常值及其附近位置的点击倾向更高，这与位置偏差假设相矛盾。

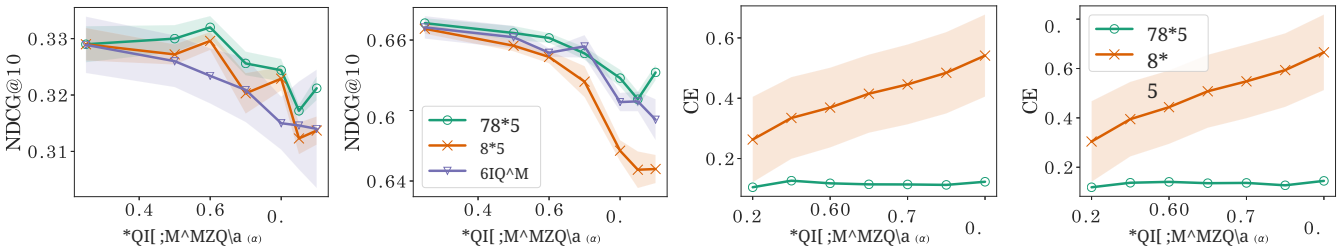
表4：在 Yahoo! 和 MSLR 数据集上，根据 NDCG@10 和交叉熵对 OPBM 与 PBM 进行比较。上标 $*$ 表示与使用 $p < 0.001$的第二佳方法相比存在显著差异。

	MSLR		Yahoo!	
	CE↓	NDCG@10↑	CE↓	NDCG@10↑
预言机	-	0.3451	-	0.6713
朴素方法	0.8205	0.3065	0.9786	0.6489
PBM	0.5474	0.3165	0.6807	0.6406
OPBM	**0.1732***	**0.3233***	**0.1916***	**0.6470***

异常值位置及其附近的倾向值最高，这与我们在第3节中的发现一致。然而，这种效应在前3 个排名中不太明显。这是预料之中的，因为我们观察到，在前3 个排名中，位置偏差占主导地位（见第3.2节），从而削弱了异常值的影响。尽管如此，随着异常值在排序中的位置前移，位置偏差的影响会减弱。例如，当异常值出现在位置1时，前两个排名的倾向值分别为0.99和0.62；而当异常值出现在位置7时，这些数值则降至0.52和0.35。

接下来，我们报告半合成实验的结果。我们使用带有模拟点击的 MSLR 和 Yahoo! 公开学习排序数据集。我们使用由 OPBM 计算的倾向值$_{Real}$，该 OPBM 在我们的专有数据上进行训练。我们从相关性估计（交叉熵）和排序性能（NDCG@10）两个方面比较 OPBM 和 PBM。表 4 总结了结果；在两个数据集上，就交叉熵而言，OPBM 的表现都显著优于 PBM（$p < 0.001$），这表明 OPBM 能够更有效地近似点击倾向——它能更准确地估计一个 $(q,\ d)$ 对的真实相关性。在基于曝光的公平排名等领域，提供真实相关性的准确估计至关重要 [9, 24, 36],，因为相关性被用作衡量项目价值的指标 [9, 14, 24, 34, 36, 39],，并且会对公平性估计产生重大影响。表 4 还显示，在两个数据集上，OPBM 的排序得分（NDCG@10）均显著优于 PBM基线。

总之，在受离群项影响的排序中，使用 OPBM 能够更准确地估计倾向性，并更准确地近似真实相关性。我们还观察到，在 Yahoo!和MSLR数据集上，OPBM 的排序性能显著优于 PBM。



(a) MSLR (b) Yahoo! (c) MSLR (d) Yahoo!

图5：在不同离群偏差水平下，各种估计器在 NDCG@10（(a) MSLR 和 (b) Yahoo!）及交叉熵（(c) MSLR 和 (d) Yahoo!）方面的比较。结果为 8 次运行的平均值；阴影区域表示标准差。

表 5：在 Yahoo! 和 MSLR 数据集上比较 OPBM、OPBM^{lazy} 和 OPM$_{\alpha,=}$ 离群偏差严重程度为 0.75 10 * ，并以 NDCG@ 和交叉熵为指标。上标表示与带有 $p < 0.001$的 PBM 相比存在显著差异。

	MSLR		Yahoo!	
	CE↓	NDCG@10↑	CE↓	NDCG@10↑
朴素方法	0.5704	0.3159	0.6776	0.6564
PBM	0.3126	0.3219	0.3958	0.6497
OPBM^{lazy}	0.1374*	0.3223	0.1548*	0.6566*
OPBM	0.1283*	0.3229	0.1407*	0.6572*

6.2 离群偏差严重程度的影响

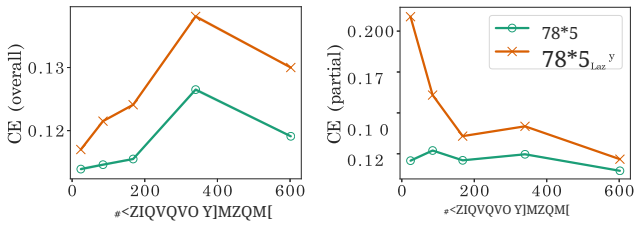
接下来，我们通过考察离群偏差严重程度对 OPBM 性能的影响来回答 **(研究问题3)**。为简单起见，我们假设离群项对倾向分布的影响相同，且与其位置无关；我们使用 OPBM$_{\mathcal{G}}$ (见第 5.2 节)进行点击模拟。OPBM α中的参数 $\mathcal{G}$ 使我们能够控制离群偏差的严重程度。图 5 展示了实验结果。在排序性能方面，OPBM 始终优于 PBM。Yahoo! 数据集上的结果（图 5b）清楚地表明，随着离群偏差严重程度的增加，两种模型在排序性能方面的差距也随之扩大。在 MSLR 的情况下（图 5a），我们观察到 OPBM 的性能波动更大。这一点也体现在朴素方法在不同次运行中的性能具有较高方差；与朴素方法和 PBM 相比，OPBM 的表现更加稳健。此外，结果表明，即使在最差情况下，OPBM 的性能也与 PBM 相近，因此，无论离群偏差的严重程度如何，将其作为用户检查模型都是更可靠的选择。这与我们的理论相符，该理论指出 PBM 是 OPBM 的一个特例（见第 4 节）。

就交叉熵而言（图5c和5d），OPBM的表现始终显著优于PBM。此外，PBM表现出的高方差进一步凸显了OPBM在相关性估计方面相比PBM具有更强的稳健性。

总之，与PBM相比，使用OPBM估计器能够得到更优的排序模型，尤其是在存在严重离群偏差时。这一发现同样适用于准确估计真实相关性分数（即交叉熵）。在离群偏差较轻的情况下，OPBM的表现与PBM相近，使其
这是一个自然的选择，因为它被证明是可靠的。

6.3 推广到多个异常值

我们通过考察 OPBM 如何推广到排序中的多个离群项，来探讨**（研究问题4）**。对于点击模拟，我们使用公式 9



(a)所有排序 (b) 两种异常值排名

图 6：在包含多个异常值的不同规模查询上，对比 OPBM 与 OPBM^{lazy}。

存在严重离群偏差的情况（$\alpha = 0.75$）。如前所述，我们的专有数据表明，91% 的异常排名最多包含两个离群项。因此，我们报告 $|O| = 2$ 的结果。这里，除了前面实验中的单异常值排名外，我们的半合成数据还包含在位置 4 和 9 上有两个离群项的排名。如前所述，排名前 3 的位置偏差很严重，因此我们将第一个离群项放在列表的第 4 个位置。然后，为了分别观察离群项的影响，我们选择了间隔一定距离的第二个位置（排名 9）。

为对多个异常值的影响进行建模，我们提出两种策略：(i) 根据 OPBM 的原始描述（见第 4 节），我们将存在多个异常值的情况视为 o的一个独立取值，即当排序中位置 4 和 9 处存在两个异常值时，我们分别计算 k 个排名的点击倾向。(ii) 我们简化问题，仅考虑第一个异常值位置，并将其称为 OPBM ^{lazy}。我们比较了这两种策略下 OPBM 的性能，并将其与 PBM 进行比较。表 5 总结了结果。总体而言，两种 OPBM 变体在 NDCG@10 和交叉熵方面均优于 PBM。正如预期，OPBM 在这两项指标上都优于其简化版本 OPBM ^{lazy}；交叉熵有显著改善，而排序性能的提升则较为有限。我们可以得出结论：对于存在多个异常值的情况，作为精确解的原始 OPBM 版本表现更好。然而，在数据稀疏的情况下，我们可以将问题简化为单异常值设置，其结果仍优于 PBM。

最后，我们探讨训练数据规模如何影响 OPBM 相较于 OPBM^{lazy} 的性能。我们逐步增加包含两个离群项（位置 4 和 9 ）的排名中的训练查询数量，同时保持训练集的其余部分不变。我们比较所有排名（整体）以及仅包含两个异常值排名上的交叉熵（CE）性能

（部分结果）。见图6。可以看到，即使只有24 个包含两个离群项的排名，OPBM 也能比 OPBM 更好地学习倾向值^{lazy}。然而，随着两个异常值排名对应的训练样本数量增加，两种模型的整体性能差异（图6a）也随之扩大，这表明在训练数据量充足时，OPBM 是一种自然的选择。

总之，当训练数据中有足够多对应多个异常值位置的样本时，最好使用带有特定 o 的 OPBM，以表示当前情况。否则，将这些样本简化为单一离群项设置，仅考虑第一个异常值位置，其性能仍优于 PBM。

7 相关工作

离群项。 离群项是偏离一般数据分布的异常对象 [40]。无论是有趣的观测结果还是可疑的异常，离群项都可能影响统计分析。识别这些离群样本对于许多研究领域都至关重要 [22, 40]。人们已提出许多用于检测离群项的方法 [16, 22, 32, 33, 35, 47]。离群项的定义和处理取决于应用领域 [40]。我们沿用 [34] 中对排名中的离群项的定义：离群项是指在排序中相对于可观测项目特征而言格外突出的条目。他们通过眼动追踪实验研究此类条目对曝光分布的影响，并进一步探讨离群项对基于曝光的公平性的影响。相比之下，本文关注的是由这一现象导致的点击偏差。我们首次研究了现实世界搜索点击日志中离群偏差的存在，并提出一种无偏学习排序模型来校正离群偏差和位置偏差。

隐式反馈中的偏差。 用户的隐式反馈（例如点击）可以成为点击学习排序的宝贵监督信号 [2]。然而，点击数据中的偏差可能导致点击概率与相关概率不同，从而产生误导。近年来，研究人员对不同类型的偏差进行了研究，例如位置偏差 [17, 19]、、展示偏差 [46]、、选择偏差 [27]、、信任偏差 [2, 38]、、流行度偏差 [1], 和新近性偏差 [11]。影响用户感知条目相关性的另一个因素是项目间依赖 [12, 34]。我们引入了离群偏差，它是项目间依赖的一种类型。由于离群偏差考虑了项目间关系，因此不同于前文提到的偏差类型。我们的研究表明，用户倾向于更多地与离群项互动，因此，当排序中存在离群项时，位置偏差所假设的检查概率会发生变化。

呈现偏差 [46] 考虑了一种相关现象：标题中含有粗体关键词的条目看起来更具吸引力。这与离群偏差不同，后者将条目的吸引力定义为独立于其周围条目的属性。此外，在搜索结果页的顶部位置增加更多图像会影响点击率 [23]。然而，此类操纵对点击偏差的影响尚未得到研究。与我们的工作最接近的概念是新闻推送推荐中的上下文偏差 [44]；与周围是相似产品的情况相比，当产品周围至少有一个非常相似的产品时，其点击率更低。这与离群偏差不同，后者强调离群项与其他条目之间的差异。此外，可观测性是检测离群项的关键因素 [34], ，但上下文偏差并未考虑这一因素。与以往工作不同，我们关注离群项对点击的影响；这种影响可被用户观察到，并源于项目间依赖关系。

无偏学习排序。 无偏学习排序方法利用存在偏差的用户反馈直接训练无偏排序模型 [7]。这些方法可分为反事实学习排序算法 [6, 19, 42] 和赌博机学习算法 [25, 41, 45]。本文关注的是点击学习排序。点击学习排序算法的关键在于先估计检查概率 [6, 43]，再使用 IPS [19, 42] 消除点击偏差。估计值可以通过在线结果随机化 [42]、、在线交错 [19], ，或从多个排序器收集的干预数据 [3]得出。然而，干预可能损害用户体验；Ai 等人 [6] 提出了一种双重学习算法，可利用离线数据自动学习排序模型和倾向值。同样，Wang 等人 [43] 使用基于回归的期望最大化算法，计算每个查询中观测到点击的似然。我们以 [43] 为基础，提出一种无偏排序模型，通过添加一个用于考虑离群项位置的参数，同时校正位置偏差和离群偏差。

8 结论

我们引入并研究了一种新型点击偏差，即离群偏差。我们开展了一项用户研究，比较特定条目在两种条件下的点击率：一次以离群项的形式展示，另一次则作为列表中的非离群项展示。我们发现，与作为非离群项展示时相比，当条目以离群项形式展示时，其点击率始终更高。此外，我们对真实世界搜索日志的分析也证实了用户研究的发现。平均而言，在相同列表中，离群项获得的点击显著多于非离群项。

为了解释这一影响，我们提出了 OPBM，这是一种基于检查假设的点击模型，能够同时考虑离群偏差和位置偏差。我们使用基于回归的期望最大化算法，根据所提出的点击模型 OPBM 估计点击倾向。实验结果表明：（i）在排序性能方面，OPBM 优于对比模型；（ii）真实相关性估计中存在离群偏差。与 PBM 相比，我们还表明，OPBM 在异常值偏差严重程度的各个级别上表现得更加稳健。此外，结果显示，在最坏情况下，OPBM 的表现与 PBM 相似，因此是更可靠的选择。

我们工作的一个局限是，对于包含多个异常值的排名，我们假设每个异常值的影响都独立于其位置及其他异常值。我们计划研究同一排名中的多个异常值如何相互影响，以及它们如何影响周围的条目。最后，我们工作的一个自然延伸是研究离群偏差如何补偿前 k 个排名位置中的位置偏差，并探索其在不同领域中的应用，例如曝光公平性。

数据与代码。 为便于复现我们的工作，所有代码和参数均已共享于 https://github.com/arezooSarvi/outlierbias/。

致谢

本研究获得了 Ahold Delhaize 和 Hybrid Intelligence Center 的支持；该中心是一个为期10年的项目，由荷兰教育、文化与科学部通过荷兰科学研究组织资助，https://hybrid-intelligence-centre.nl。所有内容均代表作者的观点，不一定得到其各自雇主和/或赞助方的认同或认可。

参考文献

[1] Himan Abdollahpouri、Robin Burke 和 Bamshad Mobasher。2017。控制学习排序推荐中的流行度偏差。载于 RecSys。42–46。[2] Aman Agarwal、Xuanhui Wang、Cheng Li、Michael Bendersky 和 Marc Najork。2019。解决无偏学习排序中的信任偏差。载于 WWW。4–14。[3] Aman Agarwal、Ivan Zaitsev、Xuanhui Wang、Cheng Li、Marc Najork 和 Thorsten Joachims。2019。无需侵入式干预的倾向估计。载于 WSDM。474–482。[4] Praveen Aggarwal 和 Rajiv Vaidyanathan。2016。字体大小是大问题吗？关于比较性价格促销的交易—获取效用视角。《消费者营销期刊》（2016）。[5] Rakesh Agrawal、Sreenivas Gollapudi、Alan Halverson 和 Samuel Ieong。2009。搜索结果多样化。载于 WSDM。5–14。[6] Qingyao Ai、Keping Bi、Cheng Luo、Jiafeng Guo 和 W Bruce Croft。2018。通过无偏倾向估计进行无偏学习排序。载于 SIGIR。385–394。[7] Qingyao Ai、Tao Yang、Huazheng Wang 和 Jiaxin Mao。2021。无偏学习排序：在线还是离线？《ACM 信息系统汇刊》（TOIS）39，2（2021），1–29。[8] Chittaranjan Andrade。2018。研究设计、实施与评估中的内部效度、外部效度和生态效度。《印度心理医学杂志》40，5（2018），498–499。[9] Asia J Biega、Krishna P Gummadi 和 Gerhard Weikum。2018。注意力公平性：在排名中摊销个体公平性。载于 SIGIR。405–414。[10] Olivier Chapelle 和 Yi Chang。2011。Yahoo! 学习排序挑战概述。《机器学习研究期刊》14（2011），1–24。[11] Ruey-Cheng Chen、Qingyao Ai、Gaya Jayasinghe 和 W Bruce Croft。2019。纠正职位推荐中的新近度偏差。载于 CIKM。2185–2188。[12] Aleksandr Chuklin、Ilya Markov 和 Maarten de Rijke。2015。《网络搜索点击模型》。Morgan & Claypool Publishers。[13] Zhichong Fang、Aman Agarwal 和 Thorsten Joachims。2019。通过干预收集进行上下文相关的检查偏差估计。载于 SIGIR。825–834。[14] Maria Heuss、Fatemeh Sarvi 和 Maarten de Rijke。2022。基于不完整曝光估计审视曝光公平性。载于 SIGIR。759–769。[15] Rolf Jagerman、Harrie Oosterhuis 和 Maarten de Rijke。2019。建模还是干预：基于用户交互的反事实学习排序与在线学习排序比较。载于 SIGIR。15–24。[16] Wen Jin、Anthony K. H. Tung、Jiawei Han 和 Wei Wang。2006。利用对称邻域关系对离群项进行排名。载于 PAKDD。577–593。[17] Thorsten Joachims、Laura Granka、Bing Pan、Helene Hembrooke 和 Geri Gay。2005。将点击数据准确解读为隐式反馈。载于 SIGIR。154–161。[18] Thorsten Joachims、Laura Granka、Bing Pan、Helene Hembrooke、Filip Radlinski 和 Geri Gay。2007。评估网络搜索中点击和查询改写所产生的隐式反馈的准确性。《ACM 信息系统汇刊》（TOIS）25，2（2007），7–es。[19] Thorsten Joachims、Adith Swaminathan 和 Tobias Schnabel。2017。基于有偏反馈的无偏学习排序。载于 WSDM。781–789。[20] Karen C Kao、Sally Rao Hill 和 Indrit Troshani。2020。在线团购中线索一致性与感知线索真实性的影响。《互联网研究》（2020）。[21] David J Lewkowicz。2001。生态效度的概念：其局限性是什么？无效是否有害？《婴儿期》2，4（2001），437–450。[22] Zheng Li、Yue Zhao、N Botta、C Ionescu 和 X COPOD Hu。2020。COPOD：基于联结函数的异常值检测。载于 ICDM。17–20。[23] Pavel Metrikov、Fernando Diaz、Sebastien Lahaie 和 Justin Rao。2014。整页优化：页面元素如何与位置拍卖互动。载于 EC。583–600。[24] Marco Morik、Ashudeep Singh、Jessica Hong 和 Thorsten Joachims。2020。控制动态学习排序中的公平性与偏差。载于 SIGIR。429–438。[25] Harrie Oosterhuis 和 Maarten de Rijke。2018。可微无偏在线学习排序。载于 CIKM。1293–1302。[26] Harrie Oosterhuis 和 Maarten de Rijke。2020。面向 Top-k 排名的策略感知无偏学习排序。载于 SIGIR。489–498。[27] Zohreh Ovaisi、Ragib Ahsan、Yifan Zhang、Kathryn Vasilaky 和 Elena Zheleva。2020。校正学习排序系统中的选择偏差。载于 WWW。1863–1873。[28] Przemyslaw Pobrotyn、Tomasz Bartczak、Mikolaj Synowiec、Radoslaw Bialobrzeski 和 Jaroslaw Bojar。2020。结合自注意力的上下文感知学习排序。arXiv abs/2005.10084（2020）。[29] Prolific。2023。数据标注。https://www.prolific.co/。[30] Tao Qin 和 Tie-Yan Liu。2013。介绍 LETOR 4.0 数据集。arXiv 预印本 arXiv:1306.2597（2013）。[31] Qualtrics。2023。XM。https://www.qualtrics.com/。[32] Sridhar Ramaswamy、Rajeev Rastogi 和 Kyuseok Shim。2000。从大型数据集中挖掘异常值的高效算法。载于 SIGMOD。427–438。[33] Peter J Rousseeuw 和 Katrien Van Driessen。1999。最小协方差行列式估计量的快速算法。Technometrics 41，3（1999），212–223。[34] Fatemeh Sarvi、Maria Heuss、Mohammad Aliannejadi、Sebastian Schelter 和 Maarten de Rijke。2022。理解并缓解异常值对公平排名的影响。载于 WSDM。861–869。[35] Bernhard Schölkopf、John C Platt、John Shawe-Taylor、Alex J Smola 和 Robert C Williamson。2001。估计高维分布的支持集。Neural Computation 13，7（2001），1443–1471。[36] Ashudeep Singh 和 Thorsten Joachims。2018。排名中的曝光公平性。载于 KDD。2219–2228。[37] Ali Vardasbi、Maarten de Rijke 和 Ilya Markov。2021。用于反事实学习排序的位置偏差和信任偏差的基于混合模型的校正。载于 CIKM。1869–1878。[38] Ali Vardasbi、Harrie Oosterhuis 和 Maarten de Rijke。2020。当逆倾向评分不起作用时：无偏学习排序的仿射校正。载于 CIKM。1475–1484。[39] Ali Vardasbi、Fatemeh Sarvi 和 Maarten de Rijke。2022。概率置换图搜索：用于排序公平性的黑盒优化。载于第45届 ACM国际信息检索研究与开发会议（西班牙马德里）（SIGIR '22）。715–725。[40] Hongzhi Wang、Mohamed Jaward Bah 和 Mohamed Hammad。2019。异常值检测技术进展：综述。IEEE Access 7（2019），107964–108000。[41] Huazheng Wang、Ramsey Langley、Sonwoo Kim、Eric McCord-Snook 和 Hongning Wang。2018。在线学习排序中对梯度空间的高效探索。载于 SIGIR。145–154。[42] Xuanhui Wang、Michael Bendersky、Donald Metzler 和 Marc Najork。2016。个人搜索中考虑选择偏差的学习排序。载于 SIGIR。115–124。[43] Xuanhui Wang、Nadav Golbandi、Michael Bendersky、Donald Metzler 和 Marc Najork。2018。个人搜索中无偏学习排序的位置偏差估计。载于 WSDM。610–618。[44] Xinwei Wu、Hechang Chen、Jiashu Zhao、Li He、Dawei Yin 和 Yi Chang。2021。信息流推荐中的无偏学习排序。载于第14届 ACM国际网络搜索与数据挖掘会议。490–498。[45] Yisong Yue 和 Thorsten Joachims。2009。将信息检索系统作为对抗老虎机问题进行交互式优化。载于 ICML。1201–1208。[46] Yisong Yue、Rajan Patel 和 Hein Roehrig。2010。超越位置偏差：考察结果吸引力这一点击数据中呈现偏差的来源。载于 WWW。1011–1018。[47] Yue Zhao、Zain Nasrullah、Maciej K Hryniewicki 和 Zheng Li。2019。LSCP：并行离群值集成中的局部选择性组合。载于 ICDM。585–593。