相关性在公平排序中的作用

Aparna Balagopalan 美国麻省理工学院 aparnab@mit.edu

Abigail Z. Jacobs 美国密歇根大学 a zjacobs@umich.edu

Asia J. Biega 德国马克斯·普朗克安全与隐私研究所 asia.biega@mpi-sp.org

arXiv:2305.05608v2 [cs.IR] 6 Jun 2023

摘要

在线平台在获取机会的过程中发挥中介作用：基于相关性的排序通过分配曝光，创造并限制人们的选择。在招聘平台上，曝光对象包括职位空缺和求职者；在市场中，则包括卖家。为了负责任地实现这一目标，这些会对社会产生重大影响的系统采用了各种公平性度量和干预措施，其中许多旨在根据应得性分配曝光。由于这些概念通常无法直接观察，平台只能转而使用相关性 等代理分数，并从搜索者点击等行为信号中推断这些概念。然而，相关性是否能在高风险的公平排序中发挥应得性分数的作用，仍是一个尚未解决的问题。

本文结合社会科学、信息检索和机器学习公平性领域的观点与工具，推导出一组相关性分数应满足的期望准则，以便有效指导公平性干预。随后，我们通过一项案例研究，以存在偏差的用户点击数据推断相关性，并以实证方式表明，这些准则并非全部得到满足。我们评估了这些违规情况对系统公平性估计的影响，并分析现有公平性干预是否能够缓解已发现的问题。我们的分析和结果凸显了迫切需求：开发适用于公平排序的相关性收集与生成新方法。

CCS 概念

• **信息系统** → **检索模型与排序**； •**以人为本的计算** →排序，公平性。

关键词

相关性，排序，测量理论，公平曝光，有效性

ACM 参考格式：

Aparna Balagopalan、Abigail Z. Jacobs 和 Asia J. Biega。2023年。相关性在公平排序中的作用。载于 第*46*届 *ACM* 信息检索研究与发展国际会议论文集（*SIGIR '23*），*2023*年7月*23*日至*27*日，台湾台北。ACM，美国纽约州纽约市，13页。https://doi.org/10.1145/3539618.3591933

1 引言

搜索系统生成排序结果，以对待聘候选人、待申请的工作岗位、住房选项以及商业产品或服务进行排序和提供。由于这些排序结果会影响人们获取机会的途径， [10, 62], 人们提出了各种公平性度量和干预措施，以防止有害结果，例如系统性地惩罚边缘化群体或在线市场的新进入者 [8, 15, 28, 29, 48, 70, 78]。这类处理方式将排序视为一种因果干预：它会影响向搜索用户分配排序项目的曝光量 [38, 58]，并直接影响公平性 [10, 62], ，这与针对公平性相关伤害的社会和法律观点一致 [70, 78]。

许多现有的公平性定义都提出，应根据应得性或应得性（例如，某位候选人是否具备某份工作的任职资格）来明确分配排序结果中的曝光 [10, 13, 17, 24, 38, 47, 62]。然而，迄今为止，在公平排序文献中，应得性的概念仍未得到充分界定，取而代之的是相关性：诸如注意力公平 [10]或曝光公平性 [62], 等公平性定义，都要求排序中的项目曝光分配与项目对排序查询的相关性成正比。因此，在具有社会影响的搜索系统中，相关性如今在分配影响生计的关键资源时发挥着双重重要作用——对相关性的估计不公平，继而会导致曝光分配不公平。然而，信息检索中的相关性概念最初是为了指导项目排序而提出的，旨在为搜索用户提供高效用的结果。能否将相关性分数重新用作公平排序中应得性的衡量标准，仍是一个悬而未决的问题，也是本文的关注重点。

我们融合社会科学中的测量理论、 [34], 信息检索以及机器学习公平性等视角，批判性地考察相关性在公平的信息获取中的作用。我们首先对相关性进行概念分析，从历史和定义的角度加以审视，并探究它能否合理地作为衡量价值的代理指标来指导公平性干预。通过分析，我们得以推导出公平排序中相关性应满足的必要条件：分数若要成为基于曝光的公平性度量中的可靠组成部分，就应满足一组条件（可信度、一致性、稳定性、可比性和可用性）。

随后，我们通过一项案例研究，对从有偏差的用户点击数据中推断出的相关性进行考察，实证检验推导出的准则在实践中是否得到满足。我们将这些问题视为复杂系统中的长期动态问题，该系统由算法和人类主体共同构成；因此，我们采用合成数据集和真实世界数据集开展模拟研究，方法与以往研究类似 [11, 20, 25]。

基于点击的方法通过用户浏览模型（即，给定用户点击行为模型时相关性的条件概率 [22, 69]）来测量相关性，确实是大规模收集相关性数据的常用方式，因为除此之外相关性数据并不容易获得。以往研究表明，点击行为模型有时无法很好地代表搜索者行为，反映出偏差，或仅捕捉对相关性的狭隘理解。我们的结果还表明，这些方法

本作品采用知识共享署名 4.0 国际许可协议授权。

SIGIR '23，*2023*年7月*23*日至*27*日，台湾台北©

2023年版权所有者/作者所有。ACM ISBN

978-1-4503-9408-6/23/07。
https://doi.org/10.1145/3539618.3591933

可能会违反公平排序中相关性的某些必要条件。具体而言，基于点击的相关性分数违反了可用性和可比性准则——由于反馈仅针对排名靠前的项目收集 [22],准确的相关性数据可能无法覆盖语料库中的大多数项目，而由此产生的分数也可能无法支持对个人与群体之间的相关性进行准确的比例比较。

本文作出以下贡献：(1) 我们探究了相关性作为公平曝光分配中应得性的代理指标所发挥的作用，并推导出相关性要成为有效且可靠的代理指标所需满足的五项准则。(2) 我们将基于点击的相关性排序作为案例进行研究，并观察到一些必要条件在实践中可能无法满足。(3) 我们探讨了类别不平衡或公平性干预等因素，这些因素可能减轻必要条件在实践中未得到满足所产生的影响，并发现这种影响可能因数据集而异。总之，我们的分析和结果表明，需要批判性地审视相关性，才能将其作为公平曝光分配的有效且可靠目标。研究结果还凸显了对相关性收集和生成新方法（例如，将相关性推断为校准概率）的迫切需求，同时为可行的解决方案提供了一组准则。

2 公平排序：相关性的作用

概率排序原则 [55] 指出，最优排序会按相关性对项目进行排序——其中假定相关性是无偏且一致的。然而，尽管项目采用了这种“最优”排序，不同用户群体 [10, 13, 14, 17, 24, 47, 62, 72, 76]仍可能受到不公平对待。例如，当招聘人员的搜索结果中排名靠前的内容主要是男性候选人，尽管候选池中也有符合条件的女性候选人时，我们可能会观察到不公平的排序。许多论文提出了量化和缓解不公平性的方法，通过按照分配的“应得性”来分配排序曝光 [10, 13, 17, 24, 47, 62, 76]。然而，应得性的概念仍未定义，而且项目的相关性常被用作代理指标 [10, 54, 62]。本节将探讨相关性在公平排序中的作用，并审视其作为应得性代理指标的有效性。

2.1 公平曝光分配的目标

所有排序系统以及公平性的操作化定义都体现了一种规范性目标。因此，在公平排序中，对公平性的不同诠释，建立在各框架所依据的不同歧视规范理论之上 [7]。例如，若目标是实现机会均等，并且认为相似项目应获得同等关注度，那么可以按照项目在排序时应得的关注度，或其“应得性”来分配曝光。另一种做法是以人口统计均等为目标，让所有群体都能同等充分地实现其潜力；在实施中，可以通过均衡各群体获得的曝光水平来实现这一点。项目的理想曝光率也可以依据领域知识或法律规定从外部确定，例如鲁尼规则 [18]。我们建议感兴趣的读者参阅 Zehlike 等人近期的研究 [77]，其中阐述了若干公平排序技术背后的规范性维度。

在许多公平性指标中，针对查询返回的某个排序项的相关性 被视为其应得性的代理指标，而搜索用户对某个项目的关注度理想情况下应与其相关性成正比 [10, 54, 62]。

2.2 衡量和缓解（不）公平性

2.2.1 衡量公平性。 本文聚焦于基于曝光或基于关注度的指标系列，这些指标按相关性的比例分配曝光 [10, 13, 17, 24, 47, 62, 76]。

群体 G_k 在排序 P 下获得的曝光，定义为 G_k中所有项目所获得曝光的平均值：

$$Exposure(G_k|P) = \frac{1}{||G_k||} \sum_{d_i \in G_k} Exposure(d_i|P) \tag{1}$$

类似地，群体 G_k 的相关性定义为其中所有项目推断相关性的平均值：

$$Relevance(G_k) = \frac{1}{||G_k||} \sum_{d_i \in G_k} Relevance(d_i|P) \tag{2}$$

由于我们的实验考虑的是二元受保护属性的情况，为简化起见，所有指标均基于两个群体定义（但可以扩展）。常用的指标有三种：

人口统计公平性均等。 群体间曝光的比率，不考虑相关性的基准率 [47]），用于衡量分配给不同群体的曝光差异：

$$\left| \frac{Exposure(G_0|P)}{Exposure(G_1|P)} \right| \tag{3}$$

曝光公平性： 群体间曝光与相关性之比的比率，用于衡量各群体按其相关性比例获得的曝光相差多少 [47]:

$$\frac{Exposure(G_0|P)}{Relevance(G_0|P)} / \frac{Exposure(G_1|P)}{Relevance(G_1|P)} \tag{4}$$

个体公平性：对单个项目的曝光和相关性分数的绝对差值求和，并在一系列排序结果中进行摊销（1， M], 其中包含 N个项目 [10]。

$$\sum_{i=1}^{N} | \sum_{j=1}^{M} Exposure(d_i|j) - \sum_{j=1}^{M} Relevance(d_i|j) | \tag{5}$$

为了估算不同位置的曝光，需要一个权重来表示搜索用户平均会给予排序位置多少关注度。我们对前 k 个位置采用对数衰减的注意力模型，类似于归一化折损累积增益中的折扣 [32]。所有相关性和曝光评分都按照 Biega 等人 [10] 的方法进行（0， 1]归一化。虽然本文重点关注上述指标，但我们也强调，近期文献中还提出了其他基于曝光的公平性指标，并推荐感兴趣的读者参阅 Raj 和 Ekstrand [54] 的研究，以了解比较分析。

2.2.2 公平性干预。 人们已经提出了多种方法，用于训练能够根据上述指标最大化公平性的排序系统。这些方法可以在预处理阶段实施 [4, 64]（例如，Sonoda [64] 考虑训练数据的顺序），也可以在处理阶段实施 [47, 73]（在训练学习排序系统时加入公平性约束），或者作为后处理实施 [10, 31, 43, 62]（对项目进行重新排序，以提高系统的公平性）。值得注意的是，上述公平性

公式中的预期相关性分数可能无法事先得知，而在实践中，人们通常会使用某个预测模型的估计值。具体而言，在我们所考察的系统中，这些分数可能由排序模型本身获得。

2.3 探究相关性作为应得性的代理指标

2.3.1 定义价值性。 公平性干预旨在生成更加公平的排序结果，即这些排序结果所体现的某些因素，并不能完全由推断出的相关性所涵盖。近期有几篇论文明确或隐含地将各种"merit"或"应得性"概念作为其干预措施的依据：

- "基于merit（即相关性）进行曝光分配" [71],
- "我们将文档的merit定义为其与查询的相关性的函数[...]，并指出候选集中的每个文档都应获得与其merit成比例的曝光" [63],
- "在特定搜索任务的语境中，相关性可以被视为应得性的代理指标" [10],
- "要实施基于 merit 的公平约束，我们需要明确估算用户相关性，并将其作为衡量 merit 的指标" [47]。

本文将应得性 视为公平排序旨在推断并在其（可能经过调整的）相关性分数中加以实施的底层构念。然而在实践中，这一构念难以轻易定义。Zehlike 等人 [77] 指出，在公平性指标中，有些指标在衡量曝光差异时会考虑 merit 的概念，但大多将"merit 视为效用分数 Y本身。然而，.....对 merit 的理解取决于世界观以及个人对机会均等的认识。"

按照一种解释，某个项目的应得性，例如在线招聘平台上求职者的个人资料，可以理解为向该项目分配关注度所带来的价值。然而，一个关键问题仍然存在：对谁而言有价值？求职者可能会从获得关注度中受益，尤其是当关注他们的搜索者有可能雇用他们时。另一方面，搜索者可能会从看到符合资质、且有望长期留在雇主身边的求职者中受益。因此，基于不同利益相关者所得价值计算出的应得性分数可能会有所不同。

另一种对应得性的解读，不把评分与分配价值挂钩，而是考察特定应用领域中排序项目本身应得的程度。这两种解读之间的差异，类似于结果公平与过程公平之间的差异；例如，参见 Barocas 等人 [7] 关于分配中 merit 与应得性的讨论。

在阐明应得性定义上的一些细微差异后，本节余下部分将聚焦于本文的核心问题：相关性是否是应得性的恰当且充分的代理指标?

2.3.2 相关性的定义视角。相关性是所有搜索系统的核心构念。不过，需要注意的是，单一的定义可能并不适用于所有情境。尽管此前已详细探讨过相关性这一指导信息寻求行为的构念 [12, 40]，但仍存在多种有效定义，包括将其视为一种取决于用户、可衡量的信息需求、效用或有用性构念，也包括所查看和评估的信息对象的主题相关性等 [12, 57]。因此，随着时间推移，信息系统采用了不同的相关性定义 [57]。例如，在搜索系统中，相关性可能指查询与网页之间的主题匹配 [74]，也可能指二元的"恰当性"评分 [41],，这些评分通常通过众包方式收集并取平均值 [6]。

在理想系统中，项目排序顺序应与潜在相关性相对应，并满足用户的信息需求 [57]。毋庸置疑，在系统训练期间采用的单维相关性概念，未必能满足所有规范和用户需求。相关性是一种复杂的关系，涉及特定用户的信息需求以及排序项目的属性——相关性的客观评分可能并不存在，而只能在特定利益相关者的参照框架中界定 [57]。尽管如此，相关性往往仍通过常见的 点击测量模型 推断，而这些模型对相关性作出了简化的定义性假设。

2.3.3 相关性测量：从定义到实践。诸如相关性这样的构念无法直接操作化，而是通过测量被认为与之相关的可观测属性，并借助测量模型推断得出 [34]。行为主义搜索系统将相关性操作化为用户对排序内容的参与度，这隐含地假设内容的（潜在）相关性越高，用户参与度就越高。因此，一种常见策略是利用点击这种形式的行为数据，估计某个项目与给定查询的相关性。点击会通过用户浏览模型转化为相关性。这类模型通常还会考虑各种测量问题，例如搜索用户的交互偏差和认知偏差，或排序基础设施的物理限制。

另一种测量相关性的方法是众包。在这种方法中，人工标注员会针对给定查询的项目列表，提供"真实标准"相关性判断 [57] 。然而，由于相关性这一概念十分复杂，外部标注员很难估计另一位搜索者的相关性 [5]。此外，标注过程成本高昂，且无法可靠地扩展。事实上，从标注员处获取的相关性判断主要用于系统评估，而非训练 [66]。

在相关性测量中，传统上存在若干假设（摘自 Saracevic [57]）：相关性是二元的（排序项目要么相关，要么不相关）；不同项目的相关性判断彼此独立；相关性具有非动态性（即价值观不会随时间变化）；在非个性化系统中，不同用户的相关性判断方差较低；以及其他假设。从理论理解到支撑数据收集的模型，在将相关性操作化时，务必注意其中若干假设——例如相关性的二元性质——至今仍然普遍存在。

2.3.4 将相关性作为应得性评分所面临的问题。我们假设，迄今为止，将相关性分数用作应得性的代理指标，以实现公平的曝光分配，只是出于方便而做出的选择。我们之所以提出这一观点，是因为相关性这一构念主要是在构建准确排序的背景下得到研究的*40, 49, 57],* 而非作为指导公平曝光分配的应得性评分。

此外，相关性分数 [23] 也存在局限（例如，并非所有排序对象都有可用的分数，或者收集到的判断具有二元性质）。

相关性分数的测量过程也存在局限（例如，浏览模型参数 [39] 设定不当），这些局限所带来的后果也主要是在排序效用的背景下进行研究的。在公平性方面，为解决相关性分数不可用的问题，Kırnap 等人 [42] 提出了一种利用不完整相关性判断来估算公平性指标的方法。然而，这些近似方法适用于系统评估，而不适用于训练能够公平分配曝光的算法。同样，公平排序评估基准依赖穷尽式标注，而在真实系统的规模下，这种标注并不可行 [9]。

除了不完整性问题之外，将相关性作为应得性的代理指标，还隐含着一些关于二者关系的假设：例如，按照应得性对项目进行的相对排序，与按照相关性进行的排序一致。此外，相关性分数中的高方差或噪声也可能导致公平性测量不够稳健，或更加武断 [19]。如果要将相关性作为近似应得性的有效且可靠的候选指标，就必须阐明并检验这两个概念之间关系的相关假设。

关注搜索系统的双边特性，有助于我们进一步理解价值观和位置性如何嵌入相关性之中：在公平排序系统中，定义为对搜索者/消费者效用的相关性，可能与定义为对项目/提供者效用的相关性相冲突 [47, 62]。此外，搜索系统本身也可能有其他目标，例如最大限度地延长用户在平台上停留的时间。因此，明确给定的相关性代理指标代表谁对应得性的判断至关重要。

表1总结了几种常见的相关性测量方法，并探讨了它们作为价值代理指标时可能存在的问题。

3 案例研究：基于点击的相关性

本文以从点击推断相关性为案例研究。图1展示了推断过程概览。[1]

3.1 基于点击相关性的排序

无偏学习排序是信息检索领域中一个研究充分的问题，旨在从有偏训练数据中学习准确的排序结果 39, 65, 68, 69]

[1, 58,

。

有些是涉及.....的在线排序算法

预干预并收集实时用户反馈 [50, 60, 75]。其中许多系统还依赖反事实估计模型来消除排序损失中的偏差。一种常见方法是使用逆倾向加权或评分（IPS） [39] ，以应对标准排序损失中用户的各种认知偏差，例如位置偏差 [69], 、信任偏差 [1],或选择偏差 [46]。其他研究提出联合训练排序模型和检查倾向模型 [3, 33]。

在这里，我们介绍一种典型的基于点击的相关性学习排序框架，其中包括学习算法、推断方法和评估指标。

3.1.1 通过浏览模型从点击推断相关性。点击或浏览模型——通过对用户信息检索/搜索行为开展实证研究得出——使我们能够模拟排序系统、近似项目相关性并评估搜索系统 [16]。相关意图假设认为，当且

[1]代码：https://github.com/Aparna-B/FairRankingRelevance

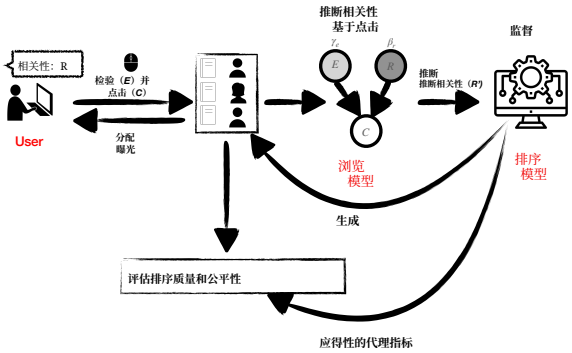


图1：根据搜索者点击推断相关性。搜索者输入一个查询，并对结果的相关性抱有潜在认知（R）。随后，他们检查（E）并点击（C）排序结果，从而为项目分配曝光。然而，他们可能表现出认知偏差（例如，以更高概率检查和点击排名靠前的项目）。为了从这些点击日志中推断项目的相关性，数据必须经过去偏处理。浏览模型通过参数化排序项目的检查概率和相关性来执行这一操作，然后定义一个模型，将查询和项目的点击概率与这些因素联系起来。由此可以估计项目具有相关性的概率，或推断其相关性。随后，为了预测未见过的查询/项目的相关性，会训练一个机器学习模型，其损失项受到浏览模型参数的影响（例如采用倾向加权）。最后，对曝光分配的质量和公平性进行衡量。

只有当项目具有相关性且被检查时，才会产生点击；随后可使用逆倾向得分等算法推断潜在相关性。点击模型或用户浏览模型在不同假设下进行建模的主要变量，是排序列表中项目被检查的概率：o_r@rank r是一个二元变量，表示位置-r上的项目是否被检查，以及项目被排序时的相关性。大多数浏览模型会利用领域知识确定的已知分布，对检查概率进行参数化。当浏览模型发生变化时，特定排名 p处的检查概率$o_r = 1$以及用户行为模型也会发生变化。概括而言，浏览模型是一个包含两个不可见节点——检查和相关性——以及一个可观测节点——点击——的图模型。（有关点击模型的全面概述，请参见 [16]。）

在本文中，我们使用用户行为的位置偏差浏览模型。该模型假设，在给定特定排名 r的情况下，检查概率保持不变。（我们还指出，浏览模型也用于其他应用，例如用户日志去偏。）

3.1.2 倾向加权的无偏排序学习。倾向加权排序算法会推断倾向权重，以最小化特定的预定义损失函数 $l(f)$，并训练排序模型 f。在大多数排序系统中，期望损失可以表示为列表中相关项（或一对项目）的损失项之和。Joachims 等人 [39] 提出的排序算法使用逆倾向得分 (IPS) 加权来处理点击的偏差性质，

价值代理指标	价值的立场性	测量中的潜在问题
众包标签 [6, 23, 41]	取决于标注指引和标注者的视角	许多项目可能没有该值，或者该值为二元值
查询与项目之间的主题匹配 [74]	取决于所使用的属性	可能存在噪声，取决于相似度指标
基于机器学习预测的相关性 [47]	搜索者（根据交互数据推断）	可能存在噪声，也可能无法推广到新的查询
基于点击的相关性 [1, 39]	搜索者	二元取值，容易受到偏差影响，且可能取决于点击日志的规模

表1：将相关性分数用作应得性的代理指标可能存在的局限性。此处，“价值的立场性”指判断理想“应得性”时所处的立场（即，排序所体现的价值取决于谁；详情请参见第2.3.1节）。

例如第 3.1.1 节所述的位置偏差），并相应地为每个项目的损失加权。在逆倾向得分加权中，如果某个项目被点击，其损失将根据该项目被检查的概率进行反向加权。请注意，只有被点击的项目才会对这种逆倾向得分加权损失作出贡献。在下一节中，我们将介绍如何训练和优化排序模型 f。

3.1.3 学习排序模型。这是用于预测相关性并生成排序的机器学习模型 f。我们按照 Joachims 等人 [39] 的方法，考虑使用深度神经网络模型的输出，通过倾向性加权的列表式损失优化的模型。训练排序模型时，我们优化基于列表式Softmax的交叉熵损失 [2, 3]。该损失还会根据倾向性加权，并在模型训练过程中进行优化；最佳参数则根据验证集的表现选取。下一节将列出用于评估排序系统效用的评估指标。

3.1.4 性能评估。最后，对排序算法的质量进行评估。通常，会将模型生成的排序与真实排序（假定其为无偏的黄金标准）进行比较。本文重点关注归一化折损累计增益（NDCG），这是一种经过归一化的排序质量度量，通过对数折损 [35]，使相关性更高的文档出现在较高排序位置时获得更高评价。

4 相关性在公平排序中的必要条件

本节介绍一组必要条件，相关性分数应满足这些条件，才能具有意义并有助于公平分配曝光。借鉴基于相关性的排序系统中的领域知识及其内在假设，这些必要条件为制定有效的 和可靠的 相关性度量提供了一条途径；这些度量中代理分数的属性应在一系列定性维度上符合其理论理想 [34]。

在下列定义中，令 w 表示真实相关性（应得性）分数，r 表示推断得出的相关性分数，即用于分配曝光的 w 测量值。K 表示待排序项目的数量。

可信度。要使推断出的相关性分数发挥作用，它们的表现就必须符合预期。在排序场景中，我们可以提出这样的问题：在多次运行中，真实相关性（应得性）较高的项目，其推断出的相关性评分是否也更高？（较低的是否更低？）如果缺乏可信度，就意味着推断出的相关性分数未能按预期反映“真实”的相关性分数，因此缺乏表面效度和内容效度，也可能无法在输入模型 [34]的相似内容之间保持信度。因此，在“机会均等”等公平性观点下，它可能并非合适的代理指标；按照这些观点，曝光理想情况下应按应得性比例分配。

定义 4.1（可信度）。 如果 i 和 j 是要排序的两个项目，则可信度要求其期望满足：

$$w_i \geq w_j \equiv r_i \geq r_j \tag{6}$$

一致性。 无偏学习排序的一个核心假设是一致性：当数据量足够大时，估计得到的期望相关性分数会收敛。因此，要满足这一统计属性，就必须符合以下标准：推断得出的相关性分数是否会收敛？如果这一标准未得到满足，那么依赖相关性的公平排序指标也可能无法收敛。

定义 4.2 (一致性)。 当数据充足时，预测分数会收敛至可接受的误差范围内。设 N 表示预期训练步数（例如，训练集的大小），$\hat{r}_i$ 表示物品 i 在第N个训练步数时的预测相关性分数。如果 $S_n = \frac{1}{K}\sum_{i=1}^{K}((r_i^n - \hat{r}_i)^2]$ 衡量的是训练相关性预测模型 n 步后得到的相关性分数与$\{\hat{r}_i\}_{i=1}^{K}$之间的均方偏差，且 ϵ 为可接受的误差值（接近 0），则一致性要求：

$$\exists N_0 \text{ such that } S_n \leq \epsilon \forall n \geq N_0 \tag{7}$$

稳定性。 相关性应具有重测信度，即同一模型对同一输入进行测量时，测量结果的变化不应超过预期范围（例如，由不可避免的随机性导致的变化）。在排序设置中，接下来要检验的标准是：在随机参数或初始化略有变化的多次运行中，推断出的相关性分数的变化是否超过既定范围？这与鲁棒性以及公平性估计的噪声程度有关。

定义 4.3（稳定性）。 若 n 表示不重要参数发生变化的实验运行次数，且 ϵ 表示各次运行之间可接受的变化幅度，则：

$$\frac{1}{K}\sum_{j=1}^{K}[\frac{1}{n}\sum_{i=1}^{n}[(r_{ij} - \mu_j)^2]] \leq \epsilon \tag{8}$$

其中 $\mu_j = \frac{1}{n}\sum_{i=1}^{n} r_{ij}$，且 r_{ij}表示第 i次运行中第 j个项目的相关性。

可比性。 若干公平性干预措施认为，应按照推断出的相关性来分配曝光度 [10, 62]，其潜在假设是，这种比例关系也适用于项目/群体的真实相关性（应得性）。因此，我们必须验证相关性分数的相对比率是否与相应的真实应得性比率相匹配，即我们的模型是否具有结构效度——这些相对属性是否按预期表现 [34]。对于群体公平性，平均相关性取属于某一群体的所有项目的汇总值。对于个体公平性，则在每个项目的层面评估这一标准。

定义 4.4（可比性）。 如果 i 和 j 是待排序的两个项目或项目群体（例如，根据敏感属性定义），则可比性要求：

$$\frac{w_i}{w_j} \approx t \equiv \frac{r_i}{r_j} \approx t \tag{9}$$

可用性。 信息检索（IR）领域的既有研究已确立，搜索者在查看和检查排序结果时会表现出位置偏差 [22]。因此，在为某个项目或一组排序项目分配曝光时，会对不同位置之间关注度的方差进行建模。随后，曝光会按照未经校正（即未经任何偏差校正）的相关性分数的比例进行分配。因此，这一做法的基本假设是，所有项目的相关性分数均可获取，并且是对真实相关性（或应得性）的无偏 估计。可用性准则检验的正是这一假设：所有项目都能获得无偏的推断相关性分数——若无法做到这一点，就很难声称相关性分数具有有效性或信度。

定义 4.5（可用性）。 如果 f 是一个函数，当 $f(x)$ 有定义时输出 1，否则输出 0，那么，

$$f(r_i) = 1 \forall i \in \{1, 2 \dots K\} \tag{10}$$

此外，我们定义：要满足可用性属性，w 在各项目中的分布应与 r 的分布在统计上无法区分。

请注意，其中一些属性相互关联。例如，可用性可能是检验可比性的必要条件。因此，可以按照固定顺序来检验这些属性。

5 实验设置

我们介绍用于检验第 4 节所述必要条件是否在实际中得到满足的实验设置，其中相关性是根据搜索者点击推断的。

5.1 数据

我们实验中的数据集由一系列项目组成，这些项目需要根据查询进行排序。每个项目都与一些特征和相关性标签相关联。我们假设，效用最高的排序是将项目按照这些真实相关性标签的降序排列。项目还带有表示（受保护）群体成员身份的标签。我们将每个数据集按70-10-20的比例划分为训练集、验证集和测试集。每个数据集的详细信息如下，并汇总于表2中。

数据集	d	相关性分布，%（4=最高）				
		0	1	2	3	4
synth-normal	2	0.13	11.00	61.51	26.65	0.71
synth-pareto	2	86.78	11.39	1.58	0.21	0.04
fairtrec	3	50.00	50.00	-	-	-

表2：数据集统计信息。d 表示特征数量。每列 i 表示给定数据集中相关性等级 i 的相对频率（4：最相关，0：最不相关）。每个数据集包含50000个样本

5.1.1 合成数据。为了在受控条件下进行实验，我们使用 Yang 等人提出的合成数据生成方法，从图模型中合成两个数据集 [72]。在每种情况下，相关性有向无环图（DAG）由四个属性构成：表示 2 个特征和一个相关性标签的连续属性，以及一个二元群体成员属性。所有三个连续属性均服从帕累托分布或正态分布（详情见附录）。采样后，我们根据数值将所有相关性分数均匀离散为 5 个等级，这与标准学习排序设置类似。在所有公平性分析中，我们都考虑二元受保护群体（例如，男性和女性）。我们从两个群体中采样相同数量的项目。从这两种合成分布中，我们各采样 N=50,000 个数据点。

帕累托分布： 我们从参数为 P（2.0,1.0）的帕累托分布中采样相关性。该数据集称为 synth-pareto。

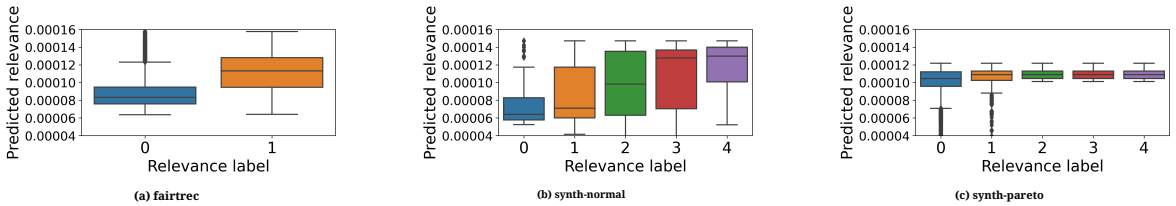
正态分布： 相关性从正态分布中采样，其均值和标准差分别设为：$\mu = 2$，$\sigma = 1$。此数据集称为 synth-normal。

5.1.2 FairTREC 2021。FairTREC 2021 [27] 数据集包含针对文本查询进行排序的维基媒体文章。总共有 50 个训练查询和 50 个测试查询。为了使该数据集的设置与我们的研究保持一致，我们选择一个查询进行分析。我们选取训练集中至少有 25,000 条相关性标注可用的查询（训练查询 ID 6）。此外，FairTREC 数据集中仅提供正向相关性标注。我们假设语料库中的所有其他维基媒体文章都与该查询不相关。最后，我们对数据集进行下采样，使其规模与合成数据集相近，得到一个包含 50,000 个项目的数据集。因此，该数据集中的每个项目都由查询、文章文本和二元相关性判断构成。为了获得用于排序的表示，我们使用一个在 ms-marco [21][2] 任务上训练的交叉编码器模型的预训练检查点。最后，我们使用主成分分析（PCA）将高维表示嵌入到三维空间中。这就是最终的"项目表示"。此外，我们将地理位置元数据用作公平排序的敏感属性。我们将这些类别二值化为"多数"位置（仅欧洲）和其他所有位置（包括没有可用元数据的情况）。该数据集在群体成员身份方面严重失衡，群体分布中多数群体占 90%。

5.2 仿真协议

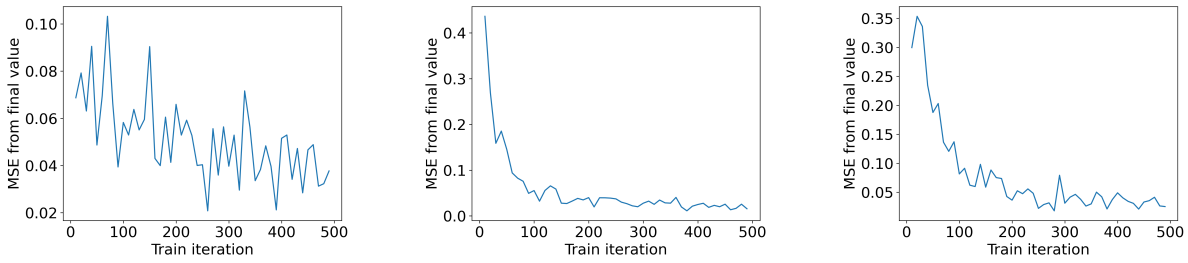
我们在不进行个性化的单查询设置下开展实验。在每一轮中，我们使用位置偏置点击模型（如第 3.1.1 节所述；正点击概率为 1，负点击概率为 0。1，最大相关性等级为 4），根据每个待排序项目的位置偏置检查概率和相关性，生成合成点击，以模拟一组用户点击项目。随后，累积每个项目的平均点击率，以计算其倾向，并使用逆倾向加权优化（第 3.1.2 节）对排序模型进行优化。在实验中，我们使用通过结果随机化预先计算的倾向权重（即预训练的倾向模型），并在训练排序模型时将其用于位置偏置点击模型[3]。每一步优化的指标是带有

[2]https://huggingface.co/cross-encoder/ms-marco-TinyBERT-L-2-v2[3]来自 ULTRA-pytorch 仿真代码库。



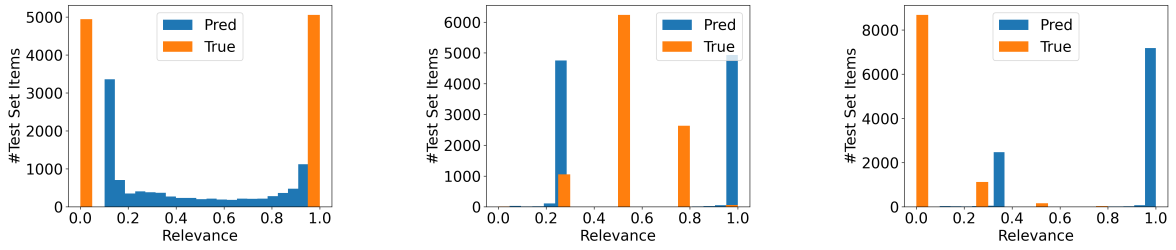
(a) fairtrec (b) synth-normal (c) synth-pareto

图2：推断的分级相关性分数与真实分级相关性分数的分布。真实相关性判断较高的项目必须具有较高的推断相关性，才能满足可信度**标准。我们观察到，平均而言，这一条件确实成立（即中位数的排序是单调的，但不同相关性等级的推断分数之间存在大量重叠）。**



(a) fairtrec **(b) synth-normal** **(c) synth-pareto**

图 3：预测的一致性。相关性持续收敛，评估指标为均方误差（MSE）。



(a) fairtrec **(b) synth-normal** **(c) synth-pareto**

图 4：$[0,1]$**归一化后的推断相关性分数与真实相关性分数的分布：推断过程无法保留分布。**

选择偏差截断值（设为 10）的列表级排序损失。最终排序系统输出的是项目列表，这些项目按照对用户的预期效用降序排列。在训练期间，该过程总共重复 500 次迭代，并对性能进行评估，每 50 次迭代保存一次模型检查点。模型损失通过随机梯度下降进行优化，学习率为 0。01。[4] 在测试集上的性能使用最佳模型检查点进行评估（根据验证性能确定）。该过程使用 10 个随机种子重复执行，并对所有指标取平均值。需要注意的是，此设置中存在多个随机性来源，包括被点击的项目、排序模型中参数的初始化，以及训练期间数据批次的采样。

6 实验与结果

在本节中，我们试图回答以下问题：基于点击的相关性在多大程度上可用于指导公平曝光分配?

6.1 研究相关性的性质

我们在三个数据集上通过实证检验，实践中是否满足相关性的预期准则。我们假设真实相关性标签与真实应得性成正比，并将真实标注的相关性判断标签称为真实相关性，将推断出的相关性标签称为预测相关性。我们指出，用于操作化形式准则的检验（见第 4 节）以及满足这些准则的阈值，本身也是人为选择的。例如，对于点估计，可信度可以通过推断出的相关性分数与真实相关性分数之间的平均偏差来衡量，也可以通过衡量二者整体的 Spearman 相关系数来衡量（后者可能更接近排序定义）。此外，要判定某项准则是否满足，可以设定接受阈值，例如，只有当特定相关系数大于 0.5 时，才认为满足某项准则。因此，我们强调，满足这些准则可能是一个连续过程，而非二元结果。研究结果概要见表 3。

[4]我们验证了 Adagrad 优化器能够产生相似的训练损失。

6.1.1 可信度。我们研究在这一典型设置下，根据点击推断出的相关性是否可信。具体而言，我们检验学习排序系统在多次运行中（即使用不同随机种子时）排名靠前的项目，是否仍然是排名靠前的项目（即排序是否保持不变）。我们绘制箱线图，展示每个相关性判断标签等级对应的相关性分数分布（为便于解读，完整列表长度设为测试集大小；每个排序中的所有 softmax 分数均纳入其中）。理想情况下，我们预期相关性等级较高的项目会因推断出的相关性分数较高而排在较低的位置。如果不同相关性等级对应的总体中位数在 0.05 的显著性水平下没有显著差异，我们便认为这些相关性分数不可信。

结果。 从图2可以看出，随着真实相关性判断值的增加，所有数据集中项目的推断相关性分数中位数均单调递增。通过 Kruskal-Wallis H 检验，我们发现这些中位数之间存在显著差异($p < 0.001$)。因此可以得出结论：对于所有三个数据集，这些相关性分数在不同运行中确实是可靠的。不过，可能需要进行事后检验，以检验特定群体的中位数（例如，相关性等级为2与3的群体）是否存在显著差异。

6.1.2 一致性。 我们计算第 i次迭代的预测相关性分数与最终迭代（即 $i = 500$）之间的均方误差，取值为 $i = 10,\ 20,\ 30 \ldots 490$。请注意，我们使用的是 softmax 运算之前的 logit 分数预测。这些分数是在验证集而非测试集上计算的（因为测试集在整个训练过程中始终不可见）。随后，我们评估该均方误差是否收敛。根据一致性的定义（见第4节），我们需要选择一个 ϵ 值来评估一致性。这里，我们设定$\epsilon = 0.1$

结果。 我们观察到，所有数据集的均方误差都收敛至最终值，不过 fairtrec 中存在一些噪声。因此，在验证集上，所有数据集均满足一致性标准。（训练集上的数据集也满足一致性。）*6.1.3* 稳定性。 我们测量多次运行中推断出的相关性值的偏差。我们计算特定项目在十个随机种子下相关性测量值的平均标准差，并对所有项目的该值取平均。为确保不同数据集间的所有值都可比较，我们对给定列表中每个项目的预测分数进行缩放：减去均值，再除以标准差；均值和标准差均根据该列表中所有项目的预测值计算得出。

结果。 在合成数据集中，我们发现平均变化介于 $0.17 - 0.18$ 之间。由于所有相关性分数都已归一化为单位方差，这相当于完整相关性分布的一个标准差不到。fairtrec 数据集的结果呈现出类似趋势（方差分数为 ~0.41）。因此，我们认为所有数据集的相关性分数都相当稳定。

6.1.4 可比性。 我们计算两个指标来评估可比性标准：(1) 所有项目的预测相关性分数与真实相关性分数之间的斯皮尔曼相关系数；(2) 按真实相关性分数和预测相关性分数计算的各群体平均相关性之比。若 (1) 中的相关系数为 ≥ 0.3，且 (2) 中两种比值的差异不超过 0.05，则认为该标准已满足。请注意，(1) 与个体公平性密切相关，因为它考虑了所有项目的预测结果。另一方面，(2) 直接影响群组级别的曝光公平性：具体而言，根据若干公平排序方案 [10, 37, 54]，曝光会按预期群体相关性（预测值或真实值）成比例分配。

结果。 synth-normal 中，真实值与预测值相关性评分之间的 Spearman 相关系数为 0.09（各子群体均为 0.09）。对于 synth-pareto，真实值与预测值相关性评分之间的 Spearman 相关系数为 0.23（女性和男性子群体分别为 0.20 和 0.26）。这表明，在这里，相关性评分的可比性准则对一个子群体更有利。相比之下，在 fairtrec 数据集上观察到的趋势有所不同：所有项目的相关系数为 0.27（子群体相关系数分别为 0.27 和 0.29）。

使用真实相关性分数时，两个合成数据集的子群体归一化平均相关性之比都接近 1（这是通过构造实现的）。对于 synth-normal 和 synth-pareto，推断出的群体平均相关性分数之比分别也接近 1.00。fairtrec 的差异相似，真实相关性和推断相关性对应的两个比值分别为 1.04 和 1.02。因此，使用真实相关性分数还是预测相关性分数，曝光比率的目标可能会略有变化。在本例中，根据设定的准则，并非所有数据集都满足可比性准则。所有数据集均满足与群体公平性相关的可比性准则。*6.1.5* 可用性。我们分析推断相关性分数的分布，并将其与真实相关性标签的分布进行比较。**结果**。根据设计，由于机器学习模型可以对所有项目进行预测，因此所有项目都具有连续相关性分数预测值。不过，在查看预测相关性和真实相关性的直方图时（图 4 中的图表），我们发现分布的特征并未保留。我们计算了双样本 Kolmogorov-Smirnov 检验 [44]以检验拟合优度，其零假设为两个样本——真实相关性分数和推断相关性分数——来自同一分布。我们发现，在所有情况下均可拒绝零假设，且 $p \ll 0.001$。因此，我们得出结论：至少对于排序项目的一个子集，可用性准则并未满足。这是合理的，因为这里可用的"应得性"判断是分级的，而预测值是连续的。

数据集	可信度	一致性	稳定性	可比性	可用性
fairtrec	✓	✓	✓	✗	✗
synth-normal	✓	✓	✓	✗	✗
synth-pareto	✓	✓	✓	✗	✗

表3：在三个数据集上对相关性评分必要条件进行的实证测试。在本表中，我们总结了五项期望准则是否在数据集中得到了实证满足。

6.2 在实践中违反准则的影响

虽然相关性是排序系统衡量的构念，但面向用户的输出是项目的排序列表（它是推断出的相关性的函数）。因此，系统评估是优化排序系统的重要组成部分。我们从两个维度评估排序系统：（1）排序的质量或效用；（2）排序的公平性。对于所有数据集，我们使用 NDCG@10 分数衡量排序结果的质量，fairtrec、synth-pareto 和 synth-normal 的得分分别为 0.90（接近满分）[5]、0.29 和 0.26。

[5]我们使用预训练排序模型生成嵌入；对此进行修改得出了类似的结果。

	个体公平性		曝光公平性		曝光公平性 + 干预	
数据集	True	预测值	True	预测值	DetConstSort	DetCons
fairtrec	19.703 ± 0.01	$19.776 \pm 0.01^*$	0.942 ± 1.03	$0.953 \pm 1.03^*$	$1.023 \pm 1.11^*$	1.020 ± 0.93
synth-normal	19.971 ± 0.03	$19.966 \pm 0.03^*$	3.420 ± 0.31	$3.407 \pm 0.31^*$	$3.131 \pm 0.34^*$	$1.420 \pm 0.18^*$
synth-pareto	19.523 ± 0.03	$19.949 \pm 0.03^*$	2.180 ± 1.01	$2.210 \pm 1.01^*$	2.154 ± 0.91	1.901 ± 0.69
	(a)		(b)		(c)	

表 4：（a）个体公平性；（b）在测试集上测得的群体曝光公平性，对同一组排序结果使用真实相关性分数与预测相关性分数，取 10 次运行的平均值；（c）群体层面公平性干预的影响。我们观察到，使用推断出的相关性分数时，公平性指标值存在显著差异（配对 Wilcoxon 检验；$p < 0.05$；非配对检验中差异不显著），但差异幅度通常较小。个体公平性测量所报告的偏差对应于项目间的方差。（c）展示了不公平性缓解算法进行干预后，使用预测相关性分数测得的公平性。

为了测试违反准则在实践中是否重要，我们使用第 2 节定义的指标来衡量曝光公平性；所有指标均基于测试集上的真实值、分级或推断得出的连续相关性分数计算。根据表 4b 和表 4a，我们观察到，公平性评估可能因使用真实值还是推断得出的相关性分数而异，且差异大小通常取决于数据集。具体而言，对于合成数据集，这种差异似乎并不特别明显。然而，按随机种子配对进行 Wilcoxon 符号秩检验（随机种子会影响预测模型的随机初始化）后发现，所有数据集的曝光公平性指标均存在显著差异。个体公平性也呈现出类似趋势：在项目级配对（即 10000 个项目）时，差异显著。我们观察到，这些结果对相关性分数的归一化方式较为敏感。

因此，公平性指标会因使用真实值还是预测值相关性分数而有所不同。值得注意的是，在全部三个数据集上，可比性准则也遭到了违反（尤其是个体公平性）（第 6.1 节）。分析某项准则被违反的程度可能会带来有趣的见解。例如，在可比性评估过程中评估的群体级相关性比例差异，对于合成数据集而言通常较小（第 6.1.4 节），群组级曝光公平性的差异也是如此（表 4b）。因此，使用真实值相关性和预测值相关性计算得出的公平性指标之间的差异，可能与预测值分数的可比性准则密切相关。

6.3 缓解测量问题

本节将举例说明可能缓解相关性测量问题的条件与干预措施。

6.3.1 机制：公平性干预。我们研究 Geyik 等人 [31] 提出的两种算法所构成的事后公平性处理方案。这两种公平重排序算法（DetCons、DetConstSort）都会以贪心方式重新排列排序列表中的前 k 个项目，使群体分布与用户指定的分布相匹配（详见 Geyik 等人 [31]）。从群体公平性视角出发，我们将某个群体在前 k 个项目中所占的目标比例设为该群体中项目的平均相关性除以所有群体项目平均相关性的总和。

我们研究这些干预措施对估算的群体层面曝光公平性的影响。这些干预措施能否缓解公平性评估中的问题？从表 4c 可以看出，公平性干预成功降低了系统的不公平程度（与表 4b 中干预前的分数相比，比率更接近 1）。在六种情况（两种算法、三个数据集）中的部分情况里，重新排序算法使系统公平性变得与使用真实相关性分数时的干预前数值无法区分（例如，fairtrec+DetCons 实验和 synth-pareto+DetConstSort 实验）。我们观察到，结果对重新排序所选的前“k”个项目十分敏感。因此，公平性干预可能是减少公平性评估误差的一种机制，但其影响可能因算法而异。

6.3.2 数据：群体不平衡。 在合成数据生成过程中（第 5.1 节），我们确保了群体样本量相等。尽管类别不平衡并非相关性测量模型的明确组成部分，但它往往是排序系统设计的产物，因此可能影响系统公平性评估。在此，我们考察数据集中不同程度的数据不平衡，并研究其对曝光公平性的影响——也就是说，训练数据不平衡会缓解或掩盖相关性期望条件违反的影响吗？我们通过对合成数据集进行下采样，将不平衡程度从 50-50%（即没有不平衡）模拟到 90-10%（即严重不平衡），共取 25,000 个数据点。

对于这两个数据集，我们观察到不平衡程度会影响使用真实相关性与预测相关性计算得出的曝光公平性指标之间的差异：对于 synth-pareto，差异约为 -0.10 至 0。12；对于 synth-normal，差异约为 -0.03 至 0。00。此处的负值意味着，使用预测相关性分数评估的曝光公平性低于使用真实相关性评估的结果。当此处的公平性测量值大于 1 时（在这里的大多数不平衡比例下确实如此），这意味着使用预测相关性时，系统看起来更加公平。因此，在某些特定的不平衡比例下，系统看起来可能更加公平，但实际上并非如此。总之，群体样本量不平衡可能会影响推断出的相关性分数，进而影响系统公平性。

7 信息检索的未来展望

对排序中公平性 [14] 和公平待遇 [10] 的研究转向，推动了信息检索领域的新研究议程。同时，关注排序平台的双向特性拓宽了公平排序系统的必要条件 [26, 51, 53]。既有研究提出

通过干预措施公平分配机会，并避免基于某种潜在且复杂的价值判断标准加剧社会偏差 [10, 30, 45, 56, 61, 67]。归根结底，这些干预措施依赖于将相关性 作为一种便捷的代理指标来分配曝光。我们的研究指出，信息检索亟需重新关注公平排序背景下的相关性研究。

批判性地评估相关性作为应得性的代理指标的局限性。相关性可以被合理地视为应得性的代理指标；然而，必须通过在特定情境下切实确立相关性的效度与信度，才能为这一观点提供充分依据。对于公平排序这一目标而言，这要求证明相关性分数的属性及其产生的排序结果与预期的公平性目标相一致。在本文中，我们展示了验证若干此类属性的示例。具体使用场景（例如招聘、市场平台）以及预期的规范性目标，应当指导相关性可接受的局限性。

本文提出的属性，是作者基于机器学习、社会技术系统、测量理论和信息检索等跨学科视角，经过广泛讨论后形成的概念。根据具体应用，有些属性可能比其他属性更为重要。我们还强调，在各种使用场景中，可能仍需识别并测试其他属性。尽管如此，我们相信，先阐明属性、再检验属性的框架，为评估某个特定代理指标（如相关性）能否可靠地作为公平排序的目标迈出了一步。在一项采用不同观点的同期研究中，Schumacher 等人 [59] 推导出了群体公平性指标所需的属性，并指出了依赖相关性的指标所具有的类似细微差异，从而进一步验证了我们的研究发现。相比之下，我们推导出了相关性的必要条件，以指导公平曝光分配。

为不同排序应用领域界定"应得性"。对应得性的不同理论概念化方式，将会对应不同的规范性目标与正义理论。例如，我们可以区分程序公平（应得性评分可能编码排序对象的 merit）与结果公平（应得性评分可能对应于某个对象排名第一时对特定利益相关者的效用）。系统的设计目标也会体现其应用领域：在线卖家市场对应得性的看法，可能不同于招聘平台。将应得性的不同概念化方式与其在公平排序系统中的具体实现方式准确关联起来，将有助于研究人员、开发者和审计人员提升系统公平性。

获取应得性分数的新方法。除了需要新的定义之外，本文还指出，有必要考虑在公平排序的语境下如何对"应得性"进行操作化，同时揭示了采用新操作化方法的必要性。潜在的方法可能包括寻找更好的代理指标、开发直接开展应得性众包的新方法、寻找新的方法来应对标注者偏差、校准浏览模型对相关性的预测，或提出超越现有信息检索相关性方法的新途径。

8 结论

排序系统在各种具有重大影响的场景中调节机会 [36]。在本文中，我们借助跨学科视角，展示了公平排序系统可能无法实现公平排序的预期目标，转而依赖"相关性即应得性"这一并不稳固的假设。基于关于基于相关性的公平排序的领域知识，我们推导出一组准则：相关性必须满足这些准则，才能作为指导公平曝光分配的有效且可靠目标：可信度、一致性、稳定性、可比性和可用性。我们以基于点击的相关性为案例研究，在三个数据集上检验了这些准则是否能够在实证上得到满足。我们观察到，实践中可能无法满足其中一部分准则。除基于点击的相关性之外，还需要针对其他形式的相关性估计（例如，众包相关性判断）检验类似准则。我们的工作贡献在于建立了一个以测量理论为基础的框架，用于思考相关性在公平曝光分配场景中的作用。更广泛地说，我们的工作强调，在公平排序场景中，需要采用新颖的方法，以有效且可靠的方式生成和收集相关性分数。总体而言，这揭示了公平曝光分配中大规模相关性测量仍存在若干未解决的问题。

致谢

本文是 FINDHR（人类推荐中的公平与交叉性非歧视）项目的一部分。该项目获得了欧盟"地平线欧洲"研究与创新计划的资助，资助协议编号为 101070212。

参考文献

[1] Aman Agarwal、Xuanhui Wang、Cheng Li、Michael Bendersky 和 Marc Najork。2019年。解决无偏学习排序中的信任偏差。载于 WWW。4–14。[2] Qingyao Ai、Keping Bi、Jiafeng Guo 和 W Bruce Croft。2018年。学习用于排序优化的深度列表式上下文模型。载于 ACM SIGIR。135–144。[3] Qingyao Ai、Keping Bi、Cheng Luo、Jiafeng Guo 和 W Bruce Croft。2018年。通过无偏倾向估计实现无偏学习排序。载于 SIGIR。385–394。[4] Abolfazl Asudeh、HV Jagadish、Julia Stoyanovich 和 Gautam Das。2019年。设计公平排序方案。载于 SIGMOD。1259–1276。[5] Peter Bailey、Nick Craswell、Ian Soboroff、Paul Thomas、Arjen P de Vries 和 Emine Yilmaz。2008年。相关性评估：评审员是否可以互换，这是否重要？载于 SIGIR。667–674。[6] David La Barbera、Kevin Roitero、Gianluca Demartini、Stefano Mizzaro 和 Damiano Spina。2020年。众包真实性：判断量表和评估者偏差的影响。载于欧洲信息检索会议。Springer，207–214。[7] Solon Barocas、Moritz Hardt 和 Arvind Narayanan。2019年。《公平性与机器学习》。fairmlbook.org。[8] Alex Beutel、Jilin Chen、Tulsee Doshi、Hai Qian、Li Wei、Yi Wu、Lukasz Heldt、Zhe Zhao、Lichan Hong、Ed H Chi 等。2019年。通过成对比较实现推荐排序公平性。载于第25届 ACM SIGKDD 知识发现与数据挖掘国际会议论文集。2212–2220。[9] Asia J Biega、Fernando Diaz、Michael D Ekstrand、Sergey Feldman 和 Sebastian Kohlmeier。2021年。TREC 2020 公平排序赛道概述。arXiv 预印本 arXiv:2108.05135（2021）。[10] Asia J Biega、Krishna P Gummadi 和 Gerhard Weikum。2018年。注意力公平：在排序结果中摊销个体公平性。载于 SIGIR。405–414。[11] Eszter Bokányi 和 Anikó Hannák。2020年。通过模拟理解网约车服务中的不平等。Scientific Reports 10，1（2020），1–11。[12] Pia Borlund。2003年。信息检索中的相关性概念。《美国信息科学与技术学会期刊》54，10（2003），913–925。[13] Amanda Bower、Hamid Eftekhari、Mikhail Yurochkin 和 Yuekai Sun。2020年。个体公平排序。载于 ICLR。[14] Carlos Castillo。2019年。排序中的公平性与透明度。载于 ACM SIGIR Forum，第52卷。ACM New York, NY, USA，64–71。[15] Mingming Cheng 和 Carmel Foley。2018年。共享经济与数字歧视：以 Airbnb 为例。Int. J. Hosp. Mgmt. 70（2018），95–98。[16] A Chuklin、I Markov 和 M de Rijke。2015年。网页搜索的点击模型。Synthesis Lectures on Information Concepts, Retrieval, and Services 7，3（2015），1–115。[17] Leonid Churilov 和 Andy Flitman。2006年。迈向奥运会成绩的公平排序：悉尼2000年奥运会案例。Comput. & Op. Res. 33，7（2006），2057–82。[18] Brian W Collins。2007年。应对招聘实践中的无意识偏见：鲁尼规则的困境。NYUL Rev. 82（2007），870。[19] A Feder Cooper、Solon Barocas、Christopher De Sa 和 Siddhartha Sen。2023年。公平分类中的方差、自洽性与任意性。arXiv:2301.11562（2023）。[20] Michael D Cooper。1973年。信息检索系统的仿真模型。Information Storage and Retrieval 9，1（1973），13–32。[21] Nick Craswell、Bhaskar Mitra、Emine Yilmaz、Daniel Campos 和 Jimmy Lin。2021年。MS MARCO：大规模数据场景下排序模型的基准测试。载于第44届国际 ACM SIGIR 信息检索研究与发展会议论文集。1566–1576。[22] Nick Craswell、Onno Zoeter、Michael Taylor 和 Bill Ramsey。2008年。点击位置偏差模型的实验比较。载于 WSDM。87–94。[23] Carlos A Cuadra 和 Robert V Katter。1967年。揭开“相关性”黑箱。Journal of Documentation 23，4（1967），291–303。[24] Fernando Diaz、Bhaskar Mitra、Michael D Ekstrand、Asia J Biega 和 Ben Carterette。2020年。使用预期曝光评估随机排序。载于 CIKM。275–284。[25] Michael D Ekstrand、Allison Chaney、Pablo Castells、Robin Burke、David Rohde 和 Manel Slokom。2021年。SimuRec：推荐系统研究中的合成数据与仿真方法研讨会。载于 RecSys。803–805。[26] Michael D Ekstrand、Anubrata Das、Robin Burke、Fernando Diaz 等。2022年。信息访问系统中的公平性。Foundations and Trends in Information Retrieval 16，1-2（2022），1–177。[27] Michael D. Ekstrand、Graham McDonald、Amifa Raj 和 Isaac Johnson。2022年。TREC 2021 公平排序赛道概述。载于第三十届文本检索会议（TREC 2021）论文集。[28] Francesco Fabbri、Francesco Bonchi、Ludovico Boratto 和 Carlos Castillo。2020年。同质性对人物推荐系统中少数群体差异化可见度的影响。载于国际 AAAI 网络与社交媒体会议论文集，第14卷。165–175。[29] Anna Farmaki 和 Stella Kladou。2020年。Airbnb 房东为何歧视？考察房东实践中歧视的来源与表现。《酒店与旅游管理期刊》42（2020），181–189。[30] Andres Ferraro、Xavier Serra 和 Christine Bauer。2021年。何为公平？从艺术家视角探索音乐流媒体平台的公平性。载于 IFIP *Conf*类人机交互会议 *p*交互。S pringer, 562–584。[31] Sahin Cem Geyik、Stuart Ambler 和 Krishnaram Kenthapadi。2019年。搜索与推荐系统中的公平性意识排序及其在 LinkedIn 人才搜索中的应用。载于第25届 ACM SIGKDD 国际知识发现与数据挖掘会议论文集。2221–2231。[32] Avijit Ghosh、Ritam Dutt 和 Christo Wilson。2021年。当公平排序遇上不确定推断。载于 SIGIR。1033–1043。[33] Ziniu Hu、Yang Wang、Qu Peng 和 Hang Li。2019年。无偏 LambdaMART：一种无偏的成对学习排序算法。载于 WWW。2830–2836。[34] Abigail Z Jacobs 和 Hanna Wallach。2021年。测量与公平性。载于《2021年 ACM 公平性、问责制与透明度会议论文集》。375–385。[35] Kalervo Järvelin 和 Jaana Kekäläinen。2017年。用于检索高度相关文档的信息检索评估方法。载于《ACM SIGIR Forum》，第51卷。243–250。[36] T Joachims、L Granka、B Pan、H Hembrooke 和 G Gay。2017年。将点击数据准确解读为隐式反馈。载于《ACM SIGIR Forum》，第51卷。4–11。[37] Thorsten Joachims、Laura Granka、Bing Pan、Helene Hembrooke、Filip Radlinski 和 Geri Gay。2007年。评估网页搜索中点击和查询改写所产生的隐式反馈的准确性。《TOIS》25卷2期（2007年），7–es。[38] Thorsten Joachims、Ben London、Yi Su、Adith Swaminathan 和 Lequn Wang。2021年。作为干预措施的推荐。《AI Magazine》42卷3期（2021年），19–30。[39] Thorsten Joachims、Adith Swaminathan 和 Tobias Schnabel。2017年。基于有偏反馈的无偏学习排序。载于 WSDM。781–789。[40] Yuri Kagolovsky 和 Jochen R Möhr。2001年。信息检索（IR）中“相关性”概念的新方法。载于 MEDINFO 2001。IOS Press，348–352。[41] Jaana Kekäläinen。2005年。信息检索评估中的二元相关性与分级相关性——比较其对信息检索系统排序的影响。《Information Processing & Management》41卷5期（2005年），1019–1033。[42] Ömer Kırnap、Fernando Diaz、Asia Biega、Michael Ekstrand、Ben Carterette 和 Emine Yilmaz。2021年。基于不完整判断估算公平排序指标。载于《2021年 Web Conference 论文集》。1065–1075。[43] Till Kletti、Jean-Michel Renders 和 Patrick Loiseau。2022年。基于 DBN 曝光模型的排序中帕累托最优的公平性—效用折中。载于 SIGIR 2022——第45届国际 ACM SIGIR 信息检索研究与发展会议。[44] Frank J Massey Jr. 1951年。用于拟合优度检验的 Kolmogorov-Smirnov 检验。《Journal of the American Statistical Association》46卷253期（1951年），68–78。[45] Rishabh Mehrotra、James McInerney、Hugues Bouchard、Mounia Lalmas 和 Fernando Diaz。2018年。迈向公平市场：推荐系统中相关性、公平性与满意度之间权衡的反事实评估。载于第27届 ACM 国际信息与知识管理会议论文集。2243–2251。[46] Massimo Melucci。2016年。查询样本选择偏差对信息检索系统排序的影响。载于 DSAA。IEEE，341–350。[47] Marco Morik、Ashudeep Singh、Jessica Hong 和 Thorsten Joachims。2020年。控制动态学习排序中的公平性与偏差。载于 SIGIR。429–438。[48] Mohammadmehdi Naghiaei、Hossein A. Rahmani 和 Yashar Deldjoo。2022年。CPFair：面向推荐系统的个性化消费者与生产者公平性重新排序（SIGIR '22）。770–779。[49] Harrie Oosterhuis。2022年。走到无偏性的尽头：揭示基于点击的学习排序的隐性局限。载于《2022年 ACM SIGIR 信息检索理论国际会议论文集》。264–274。[50] Harrie Oosterhuis 和 Maarten de Rijke。2018年。可微分的无偏在线排序学习。载于 CIKM。1293–1302。[51] Gourab K Patro、Lorenzo Porcaro、Laura Mitchell、Qiuyue Zhang、Meike Zehlike 和 Nikhil Garg。2022年。公平排序：关键综述、挑战与未来方向。载于《2022年 ACM 公平性、问责制与透明度会议》。1929–1942。[52] Fabian Pedregosa、Gaël Varoquaux、Alexandre Gramfort、Vincent Michel、Bertrand Thirion、Olivier Grisel、Mathieu Blondel、Peter Prettenhofer、Ron Weiss、Vincent Dubourg 等。2011年。Scikit-learn：Python 中的机器学习。《The Journal of Machine Learning Research》12卷（2011年），2825–2830。[53] Milo Phillips-Brown。2023年。算法中立性。arXiv 预印本 arXiv:2303.05103（2023年）。[54] Amifa Raj 和 Michael D Ekstrand。2022年。衡量排序结果中的公平性：分析与实证比较。载于第45届国际 ACM SIGIR 信息检索研究与发展会议论文集。726–736。[55] Stephen E Robertson。1977年。信息检索中的概率排序原则。《J Doc》（1977年）。[56] Piotr Sapiezynski、Wesley Zeng、Ronald E Robertson、Alan Mislove 和 Christo Wilson。2019年。量化用户关注度对排序列表中群体公平代表性的影响。载于《2019年万维网大会伴随论文集》。553–562。[57] Tefko Saracevic。2016年。信息科学中的相关性概念：人人都知道相关性是什么。但它究竟是什么？《Synthesis Lectures on Information Concepts, Retrieval, and Services》8卷3期（2016年），i–109。[58] Tobias Schnabel、Adith Swaminathan、Ashudeep Singh、Navin Chandak 和 Thorsten Joachims。2016年。作为干预措施的推荐：消除学习偏差

和评估。在*ICML*。1670–1679。[59] Tobias Schumacher、Marlene Lutz、Sandipan Sikdar 和 Markus Strohmaier。2022。排序的群体公平性指标的性质。*arXiv*预印本 *arXiv:2212.14351* (2022)。[60] Anne Schuth、Harrie Oosterhuis、Shimon Whiteson 和 Maarten de Rijke。2016。用于快速在线排序学习的多重交错梯度下降。在*WSDM*。457–466。[61] Dougal Shakespeare、Lorenzo Porcaro、Emilia Gómez 和 Carlos Castillo。2020。探索音乐推荐中的艺术家性别偏见。*arXiv*预印本 *arXiv:2009.01715* (2020)。[62] Ashudeep Singh 和 Thorsten Joachims。2018。排序结果中的曝光公平性。在*KDD*。2219–2228。[63] Ashudeep Singh 和 Thorsten Joachims。2019。排序公平性的策略学习。神经信息处理系统进展 32 (2019)。[64] Ryosuke Sonoda。2021。一种用于排序公平性的预处理方法。*arXiv*预印本 *arXiv:2110.15503* (2021)。[65] Ali Vardasbi、Maarten de Rijke 和 Ilya Markov。2020。用于反事实排序学习的基于级联模型的倾向度估计。在*SIGIR*。2089–2092。[66] Ellen M Voorhees、Donna K Harman 等。2005。TREC：信息检索中的实验与评估，第1卷。[67] Johannes Wachs、Anikó Hannák、András Vörös 和 Bálint Daróczy。2017。为什么男性获得更多关注度？探究在线设计社区中成功背后的因素。在第十一届*AAAI*国际网络与社交媒体会议。[68] Xuanhui Wang、Michael Bendersky、Donald Metzler 和 Marc Najork。2016。个人搜索中存在选择偏差时的学习排序。在 *SIGIR*。115–124。[69] X Wang、N Golbandi、M Bendersky、D Metzler 和 M Najork。2018。个人搜索中无偏排序学习的位置偏差估计。在*WSDM*。[70] Lu Xian 和 Abigail Z Jacobs。2022。作为过程的伤害：美国抵押贷款市场中的机会伤害。预印本(2022)。[71] Himank Yadav、Zhengxiao Du 和 Thorsten Joachims。2020。基于隐式反馈的公平学习排序。在*SIGIR*。[72] Ke Yang、Joshua R Loftus 和 Julia Stoyanovich。2020。公平排序的因果交叉性。*arXiv*预印本 *arXiv:2006.08688* (2020)。[73] Tao Yang 和 Qingyao Ai。2021。最大化动态排序学习的边际公平性。在*WebConf*。137–145。[74] Dawei Yin、Yuening Hu、Jiliang Tang、Tim Daly、Mianwei Zhou、Hua Ouyang、Jianhui Chen、Changsung Kang、Hongbo Deng、Chikashi Nobata 等。2016。雅虎搜索中的排序相关性。在第*22*届*ACM SIGKDD*知识发现与数据挖掘国际会议论文集。323–332。[75] Yisong Yue 和 Thorsten Joachims。2009。将信息检索系统的交互式优化视为对抗老虎机问题。在*ICML*。1201–1208。[76] Meike Zehlike、Ke Yang 和 Julia Stoyanovich。2021。排序公平性：综述。*arXiv*预印本 *arXiv:2103.14000* (2021)。[77] Meike Zehlike、Ke Yang 和 Julia Stoyanovich。2022。排序公平性，第二部分：学习排序与推荐系统。计算机调查55, 6 (2022), 1–41。[78] Elana Zeide。2022。硅天花板：算法评估如何构筑无形的机会壁垒。*UMKC Law Rev.* (2022)。

合成数据生成图

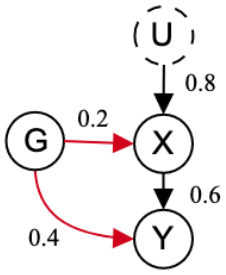


图5：合成数据：Yang 等人合成数据的生成图。节点 U、X、Y均服从帕累托分布或正态分布。“U”表示特征“X”和“Y”对应数据点的效用或相关性。“G”表示群体信息（例如男性或女性）

数据生成过程如下：在每种情况下，因果模型都由以下节点组成：表示数据点（在我们的情形中为待排序项目）特征或属性的节点“X”和“Y”，“U”表示相关性或效用，“G”表示受保护群体标签。生成过程如下：我们从两种相关性分布中采样（所有参数见附录图5），以创建一组待排序的候选人或对象。在每种情况下，我们模拟一个包含4个变量的数据集：群体变量“G”的取值为“M”和“F”，且每个取值出现的概率均为0.5，即“M”和“F”各占50%。两个特征“X”和“Y”之间的相关性由“G”和“U”决定，权重预先设定（此处，“X”由“G”和“U”决定，权重分别为0.2和0.8；“Y”由“G”和“X”决定，权重分别为0.4和0.6）。分箱后的取值见正文表2。我们在所有数据生成过程中都使用了支持用户指定因果关系的开源实现。

B 可复现性详情

所有模拟均使用 ULTRA-pytorch 工具箱[6]完成

原始数据。 所有数据均以 *libsvm* 格式存储。由于我们研究的是全局排序，因此采用单查询设置。在我们的数据集背景下，这相当于使用特定字符串进行查询。对于每种情况，我们都使用训练集得出的中位数和四分位距对所有特征进行标准化 [52]。

学习算法。 我们使用 IPS 学习算法和神经网络排序模型。在该模型中，我们使用预训练的点击模型，以及采用 Wang 等人提出的随机化方法推断出的倾向权重 [68]： 该方法会随机打乱前 k 个结果，并使用各位置的平均点击率来估计倾向。我们使用与点击模型对应的预训练倾向权重，而不是将整个流程重新训练两次。

[6]https://github.com/ULTR-Community/ULTRA_pytorch

排序模型。 在所有情况下，我们都使用具有 2 个隐藏层的神经网络，每层分别包含 256 个和 128 个单元。我们使用随机梯度下降 (SGD) 优化器最小化列表式损失，初始学习率为 0.01。批量大小设为 256。每隔 10 次迭代，都会根据验证集为排序模型的权重保存一个检查点。最后，使用验证集上 NDCG@10 得分最高的检查点生成排序结果。请注意，在测试集上，我们使用预测 logits 经 softmax 变换得到的预测值来生成并排序项目；训练期间，则在列表式损失中使用 softmax 激活。由于 softmax 是单调函数，因此使用 logits 进行排序——并将其作为“相关性”——是合理的。不过，与准则检验相关的结果可能会因此有所变化，但在我们进行基准测试的设置中，并未观察到明显变化。

系统评估。 我们使用多种指标（如归一化折损累积增益）评估所有系统，并对 10 个随机种子的测试集结果取平均值。在所有指标中，我们都研究了前 k 个位置 $k = 10$，除非另有说明。

公平性指标。 除非另有说明，否则我们假设对数分布在 k 处截断=10 。

C 选择公平性干预的理由

我们使用群组级别的项目相关性分数，通过 Geyik 等人提出的公平重新排序算法计算群体的期望分布 [31]。尽管缓解不公平的算法在重新排序期间并不考虑关注度分布（只能对项目重新排序，以满足群体的 比例要求），Geyik 等人 [31] 表明，这些算法也能改善归一化折损累积 KL 散度（NDKL）；该指标计算已获得的群体分布与期望群体分布之间经过对数折损的 KL 散度。我们假设采用对数衰减注意力分布来衡量曝光度，因此我们认为该算法非常适合我们的设置。