论文推荐中的多维公平性

Reem Alsaffar[1], Susan Gauch[1] 和 Hiba Al-Kawaz[21]
美国阿肯色大学，阿肯色州费耶特维尔[2]伊拉克巴格达大学，巴格达[1]{rbalsaff, sgauch}@uark.edu[2]
hiba.m@csw.uobaghdad.edu.iq

摘要：为防止会议和学术期刊的论文评审与遴选过程中出现潜在偏见，大多数会议和期刊都采用双盲评审。尽管如此，研究表明偏见依然存在。用于论文评审的推荐算法也可能带有隐性偏见。为此，我们提出三种公平方法，专门将作者多样性纳入论文推荐的考量。这些方法能同时针对多种受保护变量实现公平结果，不同于通常只使用一种受保护变量的公平算法。我们的多维作者档案纳入了五项人口统计特征——性别、族裔、职业阶段、大学排名和地理位置。总体多样性方法使用总体多样性得分对出版物进行排序。轮询多样性技术依次从属于各受保护群体的作者中选择论文，而多刻面多样性方法则优先选择能够填补最重要人口统计特征的论文。我们比较了基于布尔特征和连续值特征的作者多样性档案的效果。我们从 SIGCHI 2017、DIS 2017 和 IUI 2017 的论文组别中选取论文，为 SIGCHI 2017 推荐论文，并使用用户画像评估这些算法。我们将推荐的论文与会议实际选出的论文进行比较。我们发现，无论使用布尔特征值还是连续特征值的档案，三种技术都能提升多样性，同时仅略微降低效用，甚至不会降低效用。我们的最佳技术——多刻面多样性——通过选择多样性高出 42.50% 的作者，并使效用提升 2.45%，推荐出一组符合人口统计均等的论文。这一策略也可用于资助提案、会议论文、期刊文章以及其他学术工作。

关键词：用户画像、论文推荐、多样性与公平性

1 引言

我们的现代世界极具多样性，个人和机构都在努力促进包容。尽管立法和社会态度有所进步，但人们仍会因种族、肤色、性别、宗教、原国籍、残障或年龄而受到歧视 [1]。这些群体在美国受到法律保护，但有意或无意的歧视仍存在于社会各处 [38, 13]。不幸的是，学术界也不例外：尽管女性获得的博士学位超过 50%，但在终身教职轨道职位中，女性获聘比例仅为 38% [15]。计算机科学在实现多样性方面仍有很长的路要走。[45] 和[10]指出，计算机科学专业毕业生中女性仅占 18%，少数族裔也仅占 18%。这些数据反映在计算机科学会议缺少多元化演讲者这一现象中，也反映在会议参会者的人口统计信息中：少数族裔代表不足 [28]。审稿人、编辑和程序委员会中的种族、性别及其他形式的歧视可能导致论文发表遴选中的偏见 [33] ，这促使影响力最高的 ACM 会议之一 SIGCHI 于 2020 年宣布明确目标：提高其程序委员会的多样性 [40]。仅采用双盲评审流程并不能解决歧视问题 [9, 30]。审稿人往往能够从

先前出版物或容易获取的电子预印本中推断出论文作者，即使采用双盲评审，评审过程实际上也并非真正的双盲[1, 36]。我们的方法是为每篇论文构建档案，反映论文的整体质量，同时对论文作者的多样性建模。最常被认定为偏见来源的人口统计特征包括性别、族裔[7], 职业阶段[31], 大学排名[16] 和地理位置[26], , 因此这些是我们在用户人口统计档案中采用的特征。随后，我们的公平推荐系统利用这一多维档案推荐可纳入会议的论文，在提升入选作品作者多样性的同时，尽量减少对所发表论文质量造成的任何影响。

在我们之前的论文中，我们提出了两种方法，在从组别中挑选作者子集以实现人口统计均等时使用多个属性。我们在选择过程中纳入作者的多样性和质量，以尽量减少效用损失并最大化多样性。我们在用于提升论文推荐公平性的档案中采用了两种不同的特征加权方案：布尔和连续。本文提出了三种公平推荐算法，在向会议推荐待选论文时，平衡论文的两个方面：论文质量和作者的人口统计特征。由于评审过程的信息通常属于机密，我们通过从特定领域内的相关会议中创建具有不同影响因子的论文组别，来模拟评审过程的结果。影响因子最高的会议论文将代表审稿人评价最高的论文，影响因子居中的会议论文将代表评审结果次优的论文，而发表于三个影响因子最低的会议的论文则视为评审结果较低的论文。本研究的主要贡献如下：

- 利用包含多种人口统计特征的档案对作者人口统计特征进行建模。
- 开发并评估公平推荐算法，以选择兼顾质量与多样性的论文。
- 实现录用作者与所有作者组别之间的人口统计均等。

2 相关工作

我们首先讨论学术界偏见的相关方面，然后回顾以往关于构建用户人口统计特征画像的研究。最后，我们总结近期将公平性纳入算法流程的方法。

2.1 偏见

学术界的偏见：学术研究中的偏见，指在测试或抽样阶段，以及研究的任何阶段（即设计、数据收集、分析、测试和发表）中，偏好某种结果而非其他结果 [37]。Bornmann 和 Daniel 讨论了相关证据，指出性别、主要研究领域和机构隶属关系会导致委员会在授予博士和博士后研究奖学金时作出偏颇决定 [5]。Flaherty 研究了美国高校教师中的歧视问题，重点关注族裔。结果显示，教授中只有 6%

是黑人，而白人教授占 76% [17]。最近，斯坦福大学教育研究生院的研究人员于 2021 年发表的一篇文章指出，在美国，女性获得的博士学位比男性更多。尽管如此，女性获得终身教职、发表研究成果或在学术界担任领导职务的可能性仍低于男性。在分析了美国大学的一百万篇博士论文后，他们发现，研究主题与女性相关或采用 涉及女性的方法论的作者，其职业前景不如研究主题与男性相关的作者 [51]。

同行评审中的偏见：多项研究指出，同行评审流程缺乏公平性，会对哪些论文被会议和学术期刊录用产生重大影响[41]。审稿人倾向于接受作者性别与自己相同、且来自同一地区的论文 [33]。双盲评审并未彻底解决这一问题，一些研究者也证明评审过程中仍然存在偏见。例如，Cox 等人得出结论，与单盲评审相比，双盲评审并未显著提高女性的比例 [10]。

2.2 公平性

人口统计特征画像：用户画像可用于了解用户意图，并开发个性化服务，以便更好地帮助用户 [18]。近年来，研究人员开始将用户人口统计档案纳入推荐系统，希望借此减少推荐过程中的不公平现象和歧视 [29, 14]。在学术界，年龄、性别、种族和教育等人口统计属性被广泛使用，研究人员通常会根据用户姓名推断这些特征 [8, 39]。

人口统计均等：为实现公平性，许多方法旨在实现人口统计均等，即受保护群体和非受保护群体的成员获得正向结果的可能性相同。然而，这一要求通常会导致效用下降。Yang 和 Stoyanovich 着重开发新指标，以衡量排序结果中人口统计均等的缺失程度 [46]。[47] 和[48] 研究了如何在从大型组别中选取候选者子集时，针对单一二元类型属性改善排序问题的公平性；而我们同时处理多个特征。该方法在满足群体公平标准并确保人口统计均等的前提下最大化效用。我们扩展了这些研究，在从组别中选取作者子集以实现人口统计均等时采用多个属性。我们还在选取过程中考虑作者的多样性和质量，以尽可能降低效用损失并最大化多样性。 [52]的一项研究扩展了这些研究，在从组别中选取作者子集以实现人口统计均等时采用多个属性。他们还使用布尔配置文件，在选取过程中考虑作者的多样性和质量，以尽可能降低效用损失并最大化多样性。后续扩展研究 [53] 比较了用于人口统计特征档案的两种不同特征加权方案——布尔型和连续型，以提高论文推荐的公平性 [52]。最近，一些作者一直在研究广义人口统计均等（GDP），这是一种适用于连续特征和离散特征的群体公平度量，旨在提高公平性度量的准确性。 [54]通过展示联合边际与乘积边际之间的关系，提出了他们的方法

分布距离。他们展示了两种方法，即直方图和核方法，其具有线性计算复杂度。他们的实验表明，GDP 正则化器能够更准确地减少偏见。

机器学习中的公平性：随着越来越多的算法参与财务、学业和职业决策，确保这些算法不会延续针对人口统计学子群体的偏见至关重要。许多研究表明，机器学习方法可能会导致有偏决策 [12]。因此，研究人员正致力于改进分类器，使其在服务于特定目的的分类任务中实现良好效用，同时通过设计能够提供适当数据表示的预测器，减少可能针对受保护群体的歧视 [22，49]。另一些研究人员则试图在不了解受保护群体成员身份的情况下训练机器学习模型，以提高公平性。具体而言， [55] 提出了一种对抗性重加权学习（ARL）方法，旨在模型训练时提升代表性最低的受保护群体的效用。在训练阶段，他们更加依赖非受保护特征和任务标签来识别不公平偏见，并训练模型以提升公平性。在各种数据集上，他们的方法均优于最先进的替代方案。

论文分配公平性：一些研究人员探讨并衡量了在选择合适的审稿人评审论文时的公平性。[32] 和[42] 关注会议同行评审流程中将论文分配给审稿人时的公平性和统计准确性。这些研究大多提出了改进审稿人分配流程质量的方法。我们通过创建包含多种人口统计特征的作者档案，并将其用于新的公平推荐算法，以便在选择纳入会议的论文时实现人口统计均等，从而为这一领域作出贡献。

3 人口统计档案构建

我们首先通过对论文作者的人口统计特征进行建模，为每篇论文构建人口统计特征档案，以便在论文选择过程中使用这些信息。一些人口统计特征属于受保护属性，例如性别、种族，依法应受到防止歧视的特殊保护 [25]。本节将介绍我们如何收集论文组别中每位作者的人口统计特征，以及随后如何构建论文档案。

3.1 数据提取

对于给定的论文，我们的目标是提取其作者的五个人口统计特征，即性别、种族、大学排名、职业阶段和地理位置，并将它们组合起来，为该论文创建档案。每项特征都映射为一个布尔值：根据论文作者是否属于受保护群体，取1（是）或0（否）。随后，我们进一步扩展了现有方法，使用连续值特征对人口统计信息进行建模（每项特征映射为0到1之间的值），以表示计算机科学专业人员中各特征占比的补数。表1概述了每项人口统计特征对应的受保护类别和非受保护类别。

表 **1.** 人口统计特征和类别

特征	受保护/非受保护类别
性别	女性 / 男性
族裔	非白人 / 白人
地理位置	发展中 /发达国家（按国家划分）
	EPSCoR / 非EPSCoR（按美国各州划分）
职业阶段	初级 / 高级
大学排名	小于或等于平均值／大于平均值

性别：为收集作者的性别信息，我们使用 NamSor API v2，这是一种数据挖掘工具，可根据不同语言、字母表、国家和地区的个人名和姓推断其性别。该软件处理了超过40亿个姓名，精确率和召回率分别高达98.41%和99.28%。该工具返回一个介于-1和+1 之间的值；若该值接近+1 ，则表示姓名对应男性，若接近-1，则表示女性。使用该工具进行性别预测的准确率接近99% [3]。收集每位作者的性别后，我们将女性映射为1，因为女性属于受保护群体；将男性映射为0。为计算性别的连续值，我们根据女性和男性在计算机科学领域参与比例的补数进行映射。女性被视为受保护群体，因为她们仅占计算机科学领域专业人员的27% [4]。族裔：为预测族裔，我们再次使用 NamSor 工具，这是一种网络API，可根据名和姓预测族裔，但每月限查500个姓名。它将族裔划分为五类：{White, Black, Hispanic, Asian, other} [3]。非白人被视为受保护群体，因为他们占计算机科学领域专业人员的比例不足40%[4]。族裔的连续值通过将每个类别映射为其在计算机科学专业人员人口中所占比例的补数来计算，数据来源为 [51] （计算机、工程和科学职业，2020年）。白人占计算机科学专业人员的70.46%，因此映射为0.2954。同理，黑人、亚裔、西班牙裔和其他族裔分别映射为0.9295、0.8237、0.9281和0.7400。职业阶段：为提取每位作者的学术职位，我们使用研究人员的 Google Scholar 页面[19] 或其个人主页。主要任职于工业界的研究人员不纳入数据集。随后将结果映射为布尔值：高级研究人员（杰出教授、教授、副教授）映射为0，初级研究人员（助理教授、博士后、学生）映射为1。为计算该特征的连续值，我们按职级递增顺序将其映射为六个均匀分布在[0, ..和1.0] 之间的值，即：杰出教授：0/5 = 0.0；教授：1/5 = 0.2；……；学生：5/5 = 1.0。表2列出了各类别对应的取值。

表 **2.** 职业阶段权重分配

职位	权重
杰出教授	0.17
教授	0.33
副教授	0.50
助理教授或讲师	0.67
博士后或研究员	0.83
研究生	1.0

大学排名：我们通过从作者的 Google Scholar 页面 [19] 或其个人主页中提取所在机构的名称，并将其映射到从 [44]获取的世界大学排名，来收集这一特征。我们使用中位数将作者划分为低排名（1）或高排名（0）机构。然后，将大学排名（U_r）除以最低大学排名（L_r），把原始值归一化为连续值：

$$R_C = \frac{U_r}{L_r} \qquad (1)$$

地理位置：我们根据使用提取出的大学名称从其机构主页中提取的信息，确定研究人员的地理位置（国家，以及在美国境内时的州）。如果作者在美国境内工作，我们还会提取州名。我们通过将国家映射到联合国提供的发达经济体和发展中经济体表格中，确定该国所属的类别（发达国家或发展中）。因此，如果研究人员在发达国家（非受保护群体）工作，则将地理位置布尔值设为0；如果在发展中国家（受保护群体）工作，则设为1。对于居住在美国的人员，我们使用EPSCoR（竞争性研究模拟计划）[34] 将地理位置映射为布尔值。获得较少联邦拨款的EPSCoR州属于受保护群体，值为1；非EPSCoR州的值为0。为计算地理位置的连续值，我们使用人类发展指数（HDI）职级的补值[24]。这些值范围为0.957至0.394，表3展示了其中一些值的示例。

表3。 国家人类发展指数样本

国家	HDI
挪威	0.957
爱尔兰	0.955
冰岛	0.949
德国	0.947
瑞典	0.945
澳大利亚	0.944

H指数：我们从每位作者的 Google Scholar 页面提取其 H指数，以便在评估中衡量会议效用。如果作者没有 Google Scholar 页面，我们会使用 Harzing 的 Publish or Perish 工具获取其 H指数。该软件使用一些影响力指标 [23] 来计算学者的 H指数。

总而言之，每位研究者都有人口统计特征档案，其中包含五项特征（性别、族裔、职业阶段、大学排名和地理位置）。每项特征都有一个布尔权重，用于表示候选人是否属于该特征对应的受保护群体；另有一个连续值，用于表示计算机科学专业人员中该特征所占比例的补数。此外，我们通过研究者的 Google Scholar 档案收集其 H指数，或自行计算 H指数，并用它来评估评估中每份录用论文列表的效用。

3.2 论文档案构建

我们将论文所有作者的人口统计特征画像合并，以构建每篇论文的人口统计特征档案。请注意，每位作者都有一个布尔值档案和一个连续值档案。

布尔值：通过对论文作者档案执行按位或运算来创建论文档案。因此，当任何一位作者属于某项人口统计特征对应的受保护群体时，论文在该特征上的档案值为 1。我们曾考虑对作者档案求和，但这样会使作者较多的论文获得优待；而对求和后的档案进行归一化，则会使作者较多的论文处于不利地位。

连续：论文的人口统计特征档案通过从论文作者的档案中为每个特征选取最大值来创建。

3.3 论文质量分析器

衡量论文质量的方法有多种，例如论文的被引次数、出版会议编辑委员会的声誉，或出版会议本身的质量，后者通常以影响因子（IF）来衡量 [6]。尽管对于收录了高质量论文但引用较少的新兴会议而言，影响因子并不准确，但由于我们研究中的

数据集中的会议都已具有良好的发展基础 [50]。我们从 Guide2Research 网站提取每篇论文所属会议的影响因子（IF），该网站于 2019 [21]发布。影响因子是通过 Google Scholar Metrics 计算得出的：利用该指标查找过去 5 年发表的论文中最高的 h 指数 [20]。

3.4 候选池分布

按照下文所述应用我们提出的方法时，我们依靠在人口统计均等的基础上实现目标。这意味着我们选择论文，使录用作者的人口统计信息与候选人池中的人口统计信息相匹配。为此，我们测量候选人池中各项特征对应的参与者比例，并将其存储在一个向量中（PoolParity）。

组别平等=<性别权重,族裔权重,职业生涯权重,大学权重,地理权重 >

其中，每个权重均为来自该受保护群体的作者人数除以组别中的作者总人数。

4 方法

下一个目标是通过应用三种不同的方法，根据各项特征在组别中的分布来遴选论文，并实现人口统计均等，从而最大限度地提升会议的多样性。这样做是为了让高职级会议的论文列表涵盖更多元的人群，同时保持质量水平不变或略有下降。

4.1 整体多样性方法

按照第（3）节所述创建论文人口统计资料后，使用公式（2）根据特征值计算论文多样性分数（PDScore）：

$$PDScore = \sum_{i=1}^{5} f_i \qquad (2)$$

其中，f_i 是每篇论文的人口统计特征值（即每篇论文有五项特征）。我们选择多样化论文列表的第一种方法会考虑两个不同的队列。质量队列（Qquality）包含按照第 3 节所述的影响因子（IF）排名的论文。这会优先选择审稿人排名最高的论文；在我们的案例中，这些论文是在选择性最高的会议上发表的论文。人口统计队列（Qdemog）包含按 PDScore 排名的论文。接下来，我们从（Qdemog）队列的顶部选取论文，直到满足组别中每项特征的人口统计均等要求；然后从质量队列中添加其余论文，以达到会议所需的论文数量。因此，只要组别中有足够的候选论文，我们就能确保满足或超过每个受保护群体的人口统计均等要求。

算法 1： 总体多样性1 Qq质量， Qd人口统计 ← 初始化两个空优先队列 2 人口统计均等 ← 初始化一个空向量3 Qq ← 插入论文并根据质量分数排序 4 对于每个特征： 5 人口统计均等 [特征] ← 计算人口统计均等 6 对于每篇论文： 7 PDScore ← 计算论文多样性分数8 将论文添加到 Qd人口统计 并使用 PDScore 排序 9 如果有 2 篇或更多论文的 PDScore 相同： 10 使用质量分数对论文排序 11 当人口统计均等未满足时： 12 论文 ← 从顶部选择一篇论文 Qd人口统计 13 从 Qq质量 中删除选中的论文 14 当会议论文数量未满足时： 15 论文 ← 从顶部选择一篇论文 Qq质量

4.2 多面向多样性方法

前一种方法根据每篇论文的多样性总分来选择论文。然而，它无法保证从受保护群体中选出的作者本身具有多样性。例如，最终选出的论文可能多样性得分很高，但作者全是来自发展中国家的女性，完全没有少数群体作者。为避免这种情况，我们在之前的方法基础上扩展，创建五个排名队列（每项特征一个），并一次按一项人口统计特征对论文排序。加上按质量排名的队列，现在共有六个队列。根据组别的人口统计信息，我们优先考虑组别中最稀少的特征，因此会从组别中候选人数最少的特征对应队列里选取论文，直到达到这些特征的人口统计均等目标，从而生成录用论文列表。满足所有受保护群体的人口统计均等要求后，再按质量队列中的顺序添加其余论文。

算法 2：多维多样性

1 FeatureName ← 五个队列名称的列表，每个特征对应一个 2 对 FeatureName 中的每个特征： 3 DivQueue[特征] ← 初始化空优先队列4 QualityQueue ← 初始化一个空优先队列 5 PoolParity ← 初始化一个空向量6 Q ualityQueue ← 插入论文并按质量分数排序 7 对 FeatureName 中的每个特征：

8 PoolParity [特征] ← 计算人口统计均等9 对每篇论文： 10 PDScore ← 计算论文多样性分数11 对 FeatureName 中的每个特征： 12 DivQueue[特征] ← 若论文在此特征上的值为 1，则将其添加13 根据质量分数对论文排序 14 如果两篇或更多论文的质量分数相同：15 使用 PDScore 对论文排序 16 当 PoolParity 不为空时： 17 LowFeature ← min (PoolParity)18 当 LowFeature 尚未达到人口统计均等时 19 论文 ← 选择 DivQueue 中排名最前的论文[LowFeature]20 从 *Q* ualityQueue 中删除选定的论文 21 从 DParity 中删除 LowFeature 22 当会议论文数量尚未达标时： 23 论文 ← 从 *Q*ualityQueue

---

4.3 轮询方法

我们的第三种方法采用与多方面多样性方法相同的人口统计队列，但使用轮询算法访问这些队列。具体而言，该方法针对每个人口统计特征创建一个优先队列，并以轮询方式从每个队列中进行选择，直至形成论文列表。由于所提出的人口统计特征档案对五个特征进行建模，因此我们创建五个优先队列，每个队列都根据PDScore，按照一个特定特征对所有论文进行排序。例如，一个队列根据性别对论文进行排序，而另一个队列则根据学术职位对论文进行排序。所有队列排序完成后，我们通过从每个队列中选取排序最高的人口统计特征档案来应用轮询选择。因此，随后会从所有队列中移除所选档案，以确保不会再次选择相同的档案。该过程会反复进行，直到论文列表达到所需大小。我们将此方法同时用于布尔权重多样性档案和连续权重多样性档案。

算法 3：轮询

---

1 特征名称 ← 每个特征对应一个队列名称，共五个 2 对于特征名称中的每个特征： 3 DivQueue[特征] ← 初始化空优先队列4 *Q*ualityQueue ← 初始化一个空优先队列5 *Q* ualityQueue ← 插入论文并按质量得分排序 6 对于每篇论文： 7 PDScore ← 计算论文多样性得分 8 对于特征名称中的每个特征：

9 将档案添加到 [特征] 使用 PDScore[特征] 作为优先级顺序 10 论文 ← 空列表11 当论文数量 < N：12 特征 ← 特征名称[0]13 重复：14 论文 ← 从 DivQueue 获取并移除档案[特征]15 直到论文不在论文中 16 将论文添加到论文中 17 从特征名称中删除特征。18 现在已选出 N 篇论文。

---

5 实验与结果

现在，我们介绍数据集，并说明评估算法的流程。

5.1 数据集

对于我们的驱动问题，我们重点关注如何从质量和人口统计信息各不相同的论文组别中，为一场具有重大影响力的计算机科学会议筛选论文。为了创建模拟提交给会议的论文的候选论文组别，我们根据以下几项标准选择了三个会议：1）这些会议应发表主题相关的论文；2）这些会议应具有不同的影响力水平（极高、高和中等），以模拟提交的论文经评审后分别被评为高度接受、接受和边缘接受的情况；3）这些会议应拥有数量相对较多的录用论文和作者。根据这些标准，我们选择了SIGCHI（ACM计算机系统中的人为因素会议）、DIS（ACM交互系统设计会议）和IUI（人机交互（HCI）领域与人工智能领域研究人员汇聚的ACM会议）。SIGCHI发表的论文代表被SIGCHI评审评为高度可接受的论文，DIS论文代表被SIGCHI评审评为可接受的论文，而IUI论文则代表被评为边缘可接受的论文。排除来自工业界的作者后，我们为每个会议创建了一个数据集，其中包含录用论文及其作者（见表4）。该数据集包含592篇论文和813位作者，并为这些作者建立了人口统计特征档案。未来，我们将把这项工作扩展到其他会议。

表4. 我们的数据集组成 [56]。

数据集	录用论文 作者 影		因素
SIGCHI17	351	435	87
DIS17	114	231	33
IUI17	64	147	27

图中汇总了各个会议作者的人口统计分布。1。这些数据清楚地表明，在大多数受保护群体中，各会议的作者人数都很少；在影响力最高的会议 SIGCHI 中，参与度最低，性别群体是个例外。例如，SIGCHI 2017 的作者中只有 8.28% 是非白人，DIS 2017 的作者中只有 16.45% 是非白人，而 IUI 2017 中非白人作者占 27.21%。

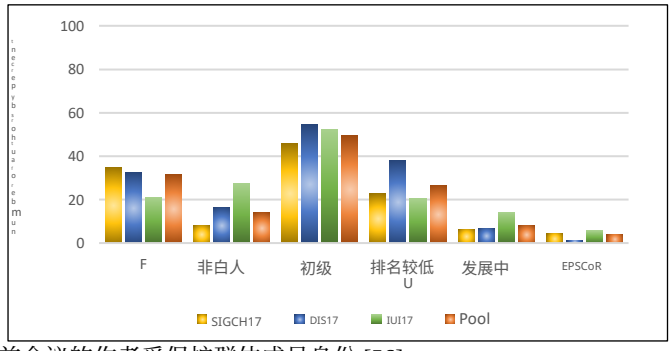

图 1. 三个当前会议的作者受保护群体成员身份 [56]。

我们将人口统计均等定义为：在由三个会议的作者合并而成的组别中，每种人口统计特征的参与率。根据数据集中 813 位作者的数据，表 5 展示了组别中每种特征的平均参与率，这也就是我们的目标人口统计均等。

表 5. 人口统　计参与：三个当　　rre前会议中受保护群体 [56]。

	性别	族裔	CStage	URank	地理位置
SIGCHI	45.01%	7.69%	52.14%	25.64%	8.26%
DIS	57.89%	31.58%	72.81%	55.26%	11.40%
IUI	39.06%	56.25%	76.56%	28.13%	26.56%
平均	47.07%	18.71%	59.55%	32.33%	11.15%

5.2 基线与指标

基线。我们的基线是程序委员会为 SIGCHI 2017 选出的论文原始列表，这些论文均在会议中有所体现。如表2所示，我们基线中受保护群体的分布为：女性占 45.01%，非白人占 7.69%，初级教授占 52.14%，来自排名较低的大学的作者占 25.64%，来自发展中国家的作者占 8.26%。

指标。我们通过计算所提出论文集相较于基线的多样性增益（D_G），来评估算法的有效性：

$$D_G = \frac{\sum_{i=1}^{n} MIN\,(100, \rho_{G_i})}{n} \qquad (3)$$

其中 ρ_{G_i} 表示各特征相对于基线的百分比增益，再除以特征总数 n。每个特征的多样性增益最高封顶为 100，以防单个特征的大幅增益主导整体数值。

通过选择最大化多样性，最终论文的质量可能会略微降低。为了衡量这种质量下降，我们使用论文作者的平均H指数，并通过以下公式计算每个推荐论文列表的效用损失（UL_i）：

$$UL_i = \frac{U_b - U_{P_j}}{U_b} * 100 \quad (4)$$

其中，U_{P_i} 是会议 i 所推荐论文的效用，U_b 是基线的效用。然后，我们按照以下方式计算会议 i 的论文相对于基线的效用节省（Y_i）：

$$Y_i = 100 - UL_i \quad (5)$$

我们计算 F 度量 [27]，以考察我们的算法在平衡多样性增益与效用节省方面的能力：

$$F = 2 * \frac{D_G * Y_i}{D_G + Y_i} \quad (6)$$

为了衡量我们的结果与人口统计均等相差多远，我们计算选定论文与组别之间的欧几里得距离 [11]：

$$\text{DemographicDistance} = \sqrt{\sum_{i=1}^{5}(F1_i - F2_i)^2} \quad (7)$$

其中，F1 表示每个特征在所提出的待选论文列表中的占比，F2 表示该特征在组别中的占比。最后，我们对距离值进行了归一化处理，从而得到我们的结果与组别之间的相似度百分比，如下式所示：

$$\text{DemographicSimilarity} = 1 - \frac{\text{DemographicDistance}}{\text{MaxD}} \quad (8)$$

其中，MaxD 是我们的特征空间中两个向量之间可能的最大距离。

为总结这些方法在提升人口统计均等与保留效用这两项相互竞争的要求之间取得平衡的能力，我们再次使用公式 6 计算 F 度量，该公式使用人口统计相似度和 Y_i。

5.3 结果

我们的推荐系统会生成排名列表，我们从中选出论文，组成录用论文列表，整体目标是提高论文的多样性。本文介绍的两种方法都会从按质量排序的队列以及一个或多个人口统计队列中选取论文。人口统计队列中若出现并列，这些论文会按质量分数排序。

5.4 与基线的比较

我们报告了 SIGCHI 2017 的录用论文与第 4 节所述推荐系统使用布尔配置文件和连续配置文件生成的录用论文之间的差异。如图 2 所示，使用布尔配置文件时，所有算法都成功提高了各人口统计群体中推荐录用论文的多样性。不过，显然，除性别外，轮询方法在所有受保护群体中都实现了最高的多样性。

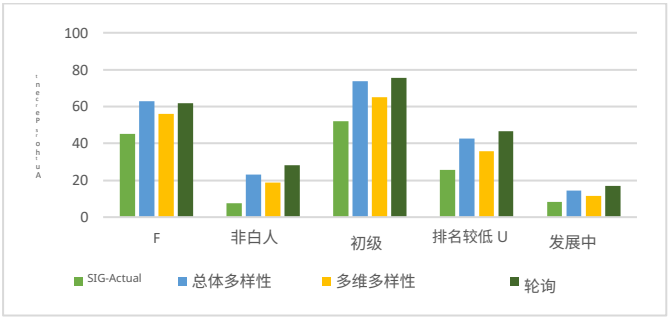

图 2. 使用布尔配置文件时，SIGCHI2017 与我们的论文推荐算法在提升受保护群体参与度方面的比较。

图 3 展示了应用我们提出的推荐算法时，连续配置文件下受保护群体的参与情况。可以看到，所有方法都成功提升了各人口统计群体所推荐论文的接收多样性。

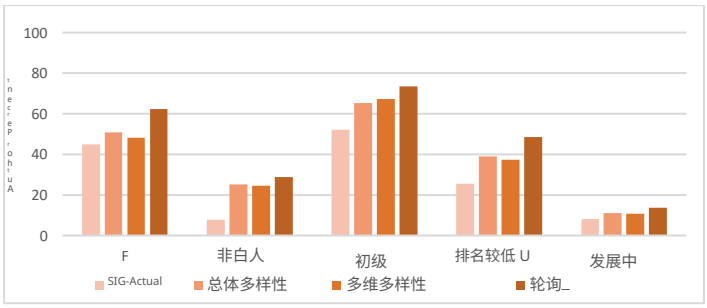

图 3. 使用连续配置文件时，SIGCHI2017 与我们的论文推荐算法在提升受保护群体参与度方面的比较。

表 6 比较了 SIGCHI 实际录用论文中受保护群体的参与情况、我们三种方法推荐的录用论文中受保护群体的参与情况，以及根据数据集中作者组别内受保护群体的参与情况计算的人口统计均等。可以看到，所有算法都提高了各受保护群体作者的多样性。除布尔档案中的女性研究人员外，在所有人口统计群体中，轮询算法提高受保护群体参与度的幅度都超过了多刻面多样性算法。正如预期的那样，

这些基于多样性的推荐方法矫枉过正，纳入的受保护群体作者比例高于整个组别中的比例。

表 6. 使用布尔和连续档案的推荐算法中受保护群体的参与情况。

特征	SIGCHI	总体 多样性 (B)	总体 多样性 (C)	多 刻面的 (B)	多 刻面的 (C)	轮 询 (B)	轮次 罗宾 (C)	Pool
女性	45.01%	62.96%	50.71%	56.13%	48.15%	61.82%	62.39%	47.07%
非白人 7.69%		23.08%	25.36%	18.80%	24.50%	28.21%	28.77%	18.71%
初级	52.14%	73.79%	65.24%	64.96%	67.24%	75.50%	73.50%	59.55%
Low 排名 大学	25.64%	42.45%	39.03%	35.90%	37.32%	46.44%	48.43%	32.33%
发展 国家	8.26%	14.53%	11.11%	11.68%	10.83%	16.81%	13.68%	11.15%

推荐论文由我们数据集中的三场会议论文按不同比例构成，如表7所述。多面向多样性方法从实际 SIGCHI 论文中选取了比例最高的推荐论文：85.8%（Bool）和 78.06%（Cont.）；而总体多样性和轮询也分别从原始 SIGCHI 选定论文中选取了大多数论文，比例分别为 75.5%（Bool）、62.11%（Cont.）和 61.53 %（Bool）、60.11%（Cont.）。我们还发现，总体多样性和多方面算法选取的大多数论文来自人口统计队列，只有少量来自按质量排序的队列。总体多样性方法从人口统计队列中选取了 67.24%（Bool）和 66.67%（Cont.）的录用论文，而仅有 32.76%（Bool）和 33.33%（Cont.）来自质量队列。相比之下，多面向多样性方法选取的录用论文几乎全部来自五个人口统计队列之一，占 92.88%；仅有 7.12% 来自质量队列。

表7。来自各会议的推荐论文比例。

	总体多样性 (B)	总体多样性 (C)	多重- 刻面的（B）	多重- 刻面的 (C)	轮询 (B)	轮询 (C)
SIGCHI	265 (75.5%)	218 (62.11%)	301 (85.8%)	274 (78.06%)	216 (61.53%)	211 (60.11%)
DIS	59 (16.8%)	87 (24.79%)	47 (13.4%)	61 (17.38%)	84 (23.93%)	90 (25.64%)
IUI	27 (7.7%)	46 (13.11%)	3 (0.9%)	16 (4.56%)	51 (14.52%)	50 (14.25 %)
论文 #	351	351	351	351	351	351

我们还比较了各算法在所选录用论文质量方面的表现。表8汇总了使用布尔和连续档案时，各算法提出的录用论文的多样性增益（D_G）、效用节省（Y_i ）和F分数。所有方法为所提出的录用论文集合带来的多样性增益均超过40%，其中最大增益出现在轮

罗宾算法中。轮询算法带来多样性增益的同时，效用节省为95.14%（B）和94.06（C）；相比之下，总体多样性算法和多维多样性算法的效用节省分别为93.47%（B）、97.52%（B）以及97.52%（C）、102.45%（C）。根据这些结果，我们得出结论：轮询算法优于另外两种算法；在考虑作者人口统计特征并以实现人口统计均等为目标时，选定论文的质量实际上有所提高。

表8. 我们的算法相对于布尔和连续档案的基线所带来的多样性增益和效用节省。

	总体多样性	多维多样性轮询	
D_G （布尔）	64.58%	46.00%	72.65%
Y_i (Bool)	93.47%	97.52%	95.14%
F分数（Bool）	76.39	62.51	82.39
D_G （续）	44.90%	42.50%	66.80%
Y_i （续）	102.49%	102.45%	94.06%
F分数（连续）	62.44	60.08	78.12

基于多样性的算法可能会过度纠正，从而导致反向歧视；或者，多样性增益可能全部集中在某一个组别，而忽视其他代表性不足的人口。表 9 和表 10 分别展示了使用布尔和连续特征评估我们的算法实现人口统计均等能力时的结果。我们观察到，根据这一标准，多方面多样性算法产生的结果最接近人口统计均等：在使用布尔配置文件时，其与组别的相似度为 95.01%，效用损失仅为 2.48%。

我们进一步发现，使用连续值特征时，多方面方法能实现更高的人口统计均等，达到 95.12%，且实际上使效用提升了 2.45%。这意味着，在选择论文时考虑作者多样性并以实现人口统计均等为目标，实际上可以提高会议接收论文的质量。

表 9. 我们的算法与基线在人口统计均等相似度和效用节省方面的比较（布尔）。

方法	人口统计 相似度	Y_i	F分数
总体多样性	89.15%	93.47%	91.26
多方面	95.01%	97.52%	96.24
轮询	87.40%	95.14%	91.11

表 **10.** 我们的算法与基线（连续）的 人口统计均等相似度和效用节省。

方法	人口统计 相似度	Y_i	F分数
总体多样性	94.80%	102.49%	98.27
多方面	95.12%	102.45%	98.44
轮询	87.38%	94.06%	90.60

5.5 讨论

在前文各节中，我们通过将所提出的方法与基线进行比较，使用多样性增益（D_G）、效用节省（Y_i）、F分数指标和人口统计相似度对这些方法进行了评估。我们使用人口统计特征档案中的布尔权重特征和连续权重特征，对推荐系统的三种算法进行了比较。

5.5.1 多样性增益对比

表 11 总结了使用两种档案权重、以多样性增益、效用损失和 F度量评估的各算法实验结果。从中可以看出，采用布尔配置文件的总体多样性方法使多样性增益最大，而采用连续权重的总体多样性方法使效用损失最小。不过，采用布尔权重的总体多样性方法取得了最高的 F分数，在两者之间实现了平衡。

表 11. 布尔权重档案和连续权重档案的多样性增益与效用节省

	档案	D_G	Y_i	F分数
总体 多样性	布尔	**64.58**	93.47	**76.39**
	连续	44.90	**102.49**	62.44
	平均	54.74	97.98	69.41
多方面 多样性	布尔	**46.00**	97.52	**62.51**
	连续	42.50	**102.45**	60.08
	平均	44.25	99.98	61.29
轮询	布尔	**72.65**	**95.14**	**82.39**
	连续	66.80	94.06	78.12
	平均	69.73	94.60	80.26
平均	布尔	61.08	95.38	73.76
	连续	51.40	99.67	66.88

所有 方法比较

对两种特征权重取平均后，轮询指标带来的多样性增益更高，分别为 69.73%、44.25% 和 54.74%。多面向多样性方法平均带来的效用下降较小，为 0.02%，而其他方法为 2.02%。不过，在平衡多样性增益与效用下降时，轮询方法仍表现更好，其平均 F分数为 80.26，而总体多样性方法和多面向方法分别为 69.41 和 61.29。

布尔与连续特征权重

对三种方法取平均后，布尔配置文件产生了更高的多样性增益，达到 61.08%，而连续配置文件为 51.40%。连续配置文件的平均效用下降幅度略小，为 0.33%，而布尔配置文件的效用下降幅度为 4.62%。然而，在综合权衡多样性增益与效用下降时，布尔配置文件仍然表现更好，其平均 F分数为 73.76，而连续配置文件为 66.88。

5.5.2 人口统计均等性比较

表12总结了使用两种档案权重、采用人口统计相似度、效用损失和F度量进行评估的各算法实验结果。从中可以看出，采用连续配置文件的多方面方法使人口统计相似度最大化，而采用连续权重的总体多样性方法将效用损失降至最低。事实上，它甚至带来了效用增益！采用连续权重的多方面方法还取得了最高的F分数，能够平衡这些指标。

表12. 布尔和连续权重配置文件的人口统计相似度

	档案	人口统计相似度	Y_i	F分数
总体多样性	布尔	89.15	93.47	91.26
	连续	**94.80**	**102.49**	**98.27**
	平均	91.97	97.98	94.76
多方面多样性	布尔	95.01	97.52	96.24
	连续	**95.12**	**102.45**	**98.44**
	平均	95.06	99.98	97.34
轮询	布尔	**87.40**	**95.14**	**91.11**
	连续	87.38	94.06	90.60
	平均	87.39	94.60	90.86
平均	布尔	90.53	95.38	92.87
	连续	92.43	99.67	95.77

所有方法比较

对两种特征权重取平均后，多维多样性度量的人口统计相似度更高，为95.06%，而另外两种方法分别为91.97%和87.39%。多维多样性方法平均造成的效用下降较小，为0.02%，而另外两种方法分别为2.02%和5.4%。此外，在平衡多样性增益与效用下降时，多维多样性方法的结果仍然更好，平均F分数为97.34%；总体多样性方法和轮询方法的平均F分数则分别为94.76%和90.86%。

布尔与连续特征权重

对两种方法取平均后，连续配置文件的平均人口统计相似度更高，为 92.43%，而布尔配置文件为 90.53%。连续配置文件的平均效用降幅略低，为 0.33%，而布尔配置文件的效用降幅为 4.62%。然而，在平衡多样性增益与效用降幅时，连续配置文件仍表现更好，平均 F分数为 95.77，而布尔配置文件为 92.87。

6 结论

我们的目标是在为会议推荐论文时，最大限度地提高多样性，同时减少质量损失。因此，我们提出了新的推荐算法，以选择适合在会议上发表的论文。我们的方法在考虑论文质量的基础上，进一步纳入作者的多种人口统计特征，从而促进多样性。早期研究大多集中于基于单一布尔属性（如族裔、性别或残障状况）确保公平性的算法。与此不同，我们根据五项属性对作者建立档案，即性别、族裔、职业阶段、大学排名和地理区域。此外，我们通过比较布尔型与连续型特征权重，拓展了先前的研究。为了在真实任务中比较我们的算法，我们创建了一个作者数据集，其中包含其论文曾被选中在不同影响因子的计算机科学会议上发表的作者；我们借此模拟由审稿人对论文进行不同程度的可接受性评判，以展示我们的方法。

我们评估了三种算法：总体多样性、多方面多样性和轮询。总体多样性方法根据总体多样性得分对论文进行评级；多面向多样性方法首先选择满足最高优先级人口统计属性参与目标的论文；而轮询方法则每次为每个人口统计属性选择一篇论文。随后，我们从多样性增益和效用节省两个方面，将所得的推荐论文与 SIGCHI 2017 实际接收的论文进行了比较，其中效用节省是根据选定论文作者的 H 指数下降幅度（如果发生下降）来衡量的。结果显示，根据 F 度量，采用布尔特征权重的轮询方法是最大化多样性的最有效策略。该方法使多样性提高了 72.6%，而效用仅降低了 4.8%。同样根据 F度量，在人口统计特征档案中使用连续特征权重的多面向多样性算法，是实现人口统计均等的最有效策略。在多样性提升 42.50%、效用提高 2.45%的情况下，该方法成功实现了与所有投稿论文作者组别的人口统计信息达到 95.12% 的相似度。这一发现表明，

与预期相反，提高多样性实际上提升了获接收出版物的质量 。

未来，我们将研究在每次论文选择后调整推荐目标的动态爬山算法。此外，我们还将研究公平深度学习方法在论文推荐中的有效性。

参考文献

1. Barak, B. (2018). 关于理论会议中的双盲评审。Windows on Theory。2. NamSor 博客。(2018). 了解 NamSor API 在性别推断方面的精确率。《使用个人姓名推断世界的性别和族裔多样性》。3. NamSor 博客。(2019). NamSor 美国“种族”/族裔模型有助于估算芝加哥警察局的多样性。《使用个人姓名推断世界的性别和族裔多样性》。4. Khan, B., Robbins, C., and Okrent, A. (2020). 科学与工程指标。美国国家科学基金会。https://ncses.nsf.gov/pubs/nsb20198/demographic-trends-of-the-s-e-workforce 5. Bolt
事会 靠性、公平性 项 科学计量学》，63(2), 297-320。6. Bolt 式吗？《欧洲分子生物学组织报告》，10(1), 2-6。7. Cannon, S., Reid, D. A., McFarlane, K., King, L., MacKenzie, L., Tadaki, M., and Koppes, M. (2018). 学术会议 S., Lakkaraju, P., & Luong, H. P. (2008, July 推荐。在《国际自适应超媒体与自适应网络系统会议》（第83-92页）。施普林格，柏林，海德堡。9. Code.org. (2020). 女性计算机科学专业毕业生人数终于超过17年前创下的纪录，但占比 M 10. Cox, A. K., & Montgomerie, R. (2019). 支持和反对双盲评审的论据。PeerJ，7, e6702。Data USA，计算机、工程和科学职业。(2020). 11. 距离次 计算研讨会论文集》（SNC 14）中。ACM。（第9-16页）。人知实 理论计算机科学创新会议论文集》（第214-226页）。13. eeoc.gov。(n.d.). 美国平等就业机会委员会。14. Farnadi, G., Kouki, P., Thompson, S. 本 arXi (2016). 教师多样性提升，但未进入终身教职轨道。Inside Higher ED。16. Flaherty, C. (2018). 当学术期刊偏袒某些人时。Inside Higher ED。17. Flaherty, C. (2019). 教授是白人的可能性仍高于学生。Inside Higher Ed。18. Gauch, S., Speretta, M., Chandramouli, A., & Micarelli, A. (2007). 用于个性化信息访问 https://scholar.google.com/ 20. Google Scholar (n.d.). https://scholar.google.com/citations?view_op=top_venues&hl=en&vq=eng 21. Guide2Research。(2020). 顶级计算机科学会议。Guide2Research。http://www.guide2research.com/topconf/

22. Hardt, M., Price, E., & Srebro, N. (2016). 监督学习中的机会平等。arXiv 预印本 arXiv:1610.02413。23. Harzing, A. (2016). Publish or Perish。https://harzing.com/resources/publish-or-perish 24. 《人类发展报告》。(n.d.). 人类发展数据中心。《人类发展报告》。25. Inc. US Legal。(n.d.). 受保护群体成员法及其法律定义。https://definitions.uslegal.com/p/protected-group-member/ 26. Jacob, B. A., & Lefgren, L. (2011). 科研经费资助对科研生产力的影响。《公共经济学杂志》，95(9-10), 1168-1177。27. Jardine, N. & (1971). 层次聚类在信息检索中的应用。《信息存储与检索》，(第7[illegible]240页。). 28. Jones, T. M., Fa[illegible]fear, R., Symonds, M. R., [illegible] Labille, K., Gauch, S., Joseph, A. S., Rogers, T., & Koolen, M. (2015). 面向 CiteSeerx 的基于[illegible]). Ler[illegible] D. [illegible] Lerback, J., & Hanson, B. (2017). 学术期刊邀请担任审稿人的女性太少。《自然新闻》，541(7638), 455。32. Long, C., Wong, R. C. W., Peng, Y., & Ye, L. (2013, December). 关[illegible]第1145-1150页）中。[illegible] 33. Murray, D., Siler, K., Larivière, V., Chan, W. M., Com[illegible], J., & Sugimoto, C. R. (2019). 性别和国际多样性提升[illegible]评审公平[illegible]0515。34. Naional Sciencef Foundation。(2019). 促进竞争性研究建立[illegible]形势与[illegible] 2020。访问于，20。36. Palus, S. (2015). 双盲评审更好吗？美国物理学会。37. Pannucci, C. J., & Wilkins, E. G. (2010). 识别并避免研究中的偏见。《整形与重建外科》，126(2), 619。38. 受保护群体。(2020). 维基百科。https://en.wikipedia.org/wiki/Protected_group 39. Santamaria, L., & Mihaljević, H. [illegible] IGCHI。(2019). CHI 2020 程序委员会的多样性。41. Sikdar, S., Marsili, M., Ganguly, N., & Mukherjee, A. (2016, October). 同行评审系统中的异常：以《高能物理学杂志》为例。在《第[illegible]和识管理国际会议论文集》[illegible]2245-2250页）中。42. Stelmakh, I., Shah, N. B., & Singh, A. (2019, March). PeerReview4All：实现公平且准确的同行评审[illegible] Nation, M., Yuan, N. P., Kuperminc, G. P., Hassoun Ayoub, L., & Hamby, S. (2018). 仇恨与暴力：应对[illegible]、宗教、性取向和性别认同的歧视。《暴力心理学》，8(6), [illegible]。(2020). 世界大学排名。[illegible] Com[illegible] 衡量排序结果中的公平性。在《第29届科学与统计数据库管理国际会议论文集》（第1-6页）中。

47. Zehlike, M., Bonchi, F., Castillo, C., Hajian, S., Megahed, M., & Baeza-Yates, R.（2017 [illegible] 2017 年 ACM 信息与知识管理会议论文集（第 1569-1578 页）。48. Zehlike, M., & Castillo, C.（2020 年 4 月）。减少排名中的差[illegible]排序方法。载于 2020 年万维网会议（第[illegible]。49. Zhong, Z.（2018）。机器学习公平性教程。 *Towards Data Science*。50. Zhuang, Z., [illegible]议质量。载于第七届 ACM/IEEE 数字图书[illegible]集（第 225-234 页）。51. Andrews, L., Edmund.（2021 年 12 月）。斯坦福研究揭示学术界女性面临的一项隐性障碍。斯坦[illegible]
h
tt
[illegible] women-academia/
51. Zweben, S., & Bizot, B.（2018）。2018 年 CRA Taulbee 调查。 *Computing Research* [illegible]21 年[illegible]存。载于 2021 年 ACM/IEEE 数字图书馆联合会议（JCDL）[illegible]多[illegible]8 页）。54. Jiang, Z., Han, X., Fan, C., Yang, F., Mostafavi, A., & Hu, X.（2022 年 9 月）。群体公平的广义人口统计均等。载于国际学习表征会议[illegible] & Chi, E.（20[illegible]学习实现无人口统计信息的公平性。NeurIPS 2020。http://alexbeutel.com/papers/NeurIPS-2020-fairness-without- demographics.pdf 56. [illegible]于 KDIR（第[illegible]