新闻推荐算法可信评价研究

刘总真[1,2], 张潇丹[1,2*], 郭 涛[1,2], 葛敬国[1,2], 周 熙[2], 王宇航[1,2],
陈家玓[2], 吕红蕾[1,2], 林俊宇[2]

[1] 中国科学院大学网络空间安全学院 北京 中国 100049
[2] 中国科学院信息工程研究所 北京 中国 100093

摘要 随着 AI、5G、AR/VR 等新技术的快速发展, 内容类应用如电子商务、社交网络、短视频等层出不穷, 导致信息过载问题日益严重。人工智能技术的发展推动了智能算法的爆炸式运用, 作为智能算法的一种, 推荐算法在大数据、应用场景和计算力的推动下, 通过信息过滤技术, 为用户提供适应兴趣及行为的个性化及高质量的推荐服务, 逐步提高了用户的使用体验、内容分发效率, 在一定程度上缓解了信息过载的问题。但推荐算法的潜在偏见、黑盒化特性及内容分发方式也逐渐带来了决策结果不公平性、不可解释性, 信息茧房、侵犯用户隐私等安全挑战。如何提高推荐算法的可解释性、公平性、可信程度等越来越受到国内外政府监管部门、产业及学术界的重点关注, 推荐系统和推荐算法也由此从发展期进入管制期。为此, 本文针对新闻推荐领域, 分析推荐算法的稿件画像、用户画像、推荐推送、反馈干预和人工复审等关键要素, 围绕推荐算法生态的参与者, 如内容生产者、受众、算法模型、新闻平台, 从公平性、可解释性和抗抵赖性三个方面提出了一种新闻推荐算法可信评价体系, 并进行定量或定性分析。公平性、可解释性和抗抵赖性是正相关关系, 当公平性和抗抵赖性越强、可解释程度越高, 新闻推荐算法的可信度越高。希望弥补新闻推荐算法领域的可信研究的空白, 建立可信推荐算法生态, 加速安全推荐系统的建立和推广, 同时为智能算法可信研究提供参考, 为智能算法的监管和治理提供思路。

关键词 新闻; 推荐算法; 可信评价体系; 公平性; 可解释性; 抗抵赖性
中图法分类号 TP309.2 **DOI 号** 10.19363/J.cnki.cn10-1380/tn.2021.09.12

Trust Evaluation of News Recommendation Algorithms

LIU Zongzhen[1,2], ZHANG Xiaodan[1,2*], GUO Tao[1,2], GE Jingguo[1,2], ZHOU Xi[2],
WANG Yuhang[1,2], CHEN Jiadi[2], LV Honglei[1,2], LIN Junyu[2]

[1] School of Cyber Security, University of Chinese Academy of Science, Beijing 100049, China
[2] Institute of Information Engineering, Chinese Academy of Science, Beijing 100093, China

Abstract With the rapid development of new technologies such as AI, 5G, and AR/VR, the applications on content, such as e-commerce, social networks, and short videos et al. have emerged one after another, leading the increasingly serious problem of information overload. The development of artificial intelligence technology has promoted the explosive application of intelligent algorithms. As a kind of intelligent algorithm, driven by big data, application scenarios and computing capability, recommendation algorithms provide the users with personalized and high-quality recommendation services that adapt to their interests and behaviors, which has not only gradually improved the user experience and the efficiency of content distribution, but also alleviated the problem of information overload to a certain extent. However, the potential biases, black-box characteristics and the content distribution methods of recommendation algorithms have gradually brought security challenges such as unfairness and inexplicability on the decision-making results, information cocoon and the infringement of user privacy et al. How to improve the interpretability, fairness, and trust of recommendation algorithms has been paid more and more attention from the regulatory agencies of governments, industries and academia at home and abroad. Therefore, the recommendation systems and recommendation algorithms enter the regulatory period from the development period. To this end, in the news field, by analyzing the key elements of the recommendation algorithm, such as manuscript portraits, user portraits, recommendation, feedback and interventions, and manual reviews, focusing on the participants of the recommendation algorithm ecology, such as the content producers, the audiences, the algorithm models and the news platform, this study proposes a trust evaluation system for the news recommendation algorithms based on fairness, interpretability and anti-denying. At last, we carry out the qualitative or quantitative analysis. Fairness, interpretability, and anti-denying are positively correlated. When the fairness and anti-denying are stronger and the interpretability is higher, the trust of the news recommendation algorithm is higher. It is expected to fill the research gaps in the study of the trust of news recommendation algorithms, establish a trust recommendation algorithm ecology,

通讯作者: 张潇丹, 工学博士, 研究员, Email: zhangxiaodan@iie.ac.cn
本课题得到 中国科学院战略性先导科技专项（C 类）项目(No. XDC02060400)。
收稿日期：2021-04-30；修改日期：2021-08-08；定稿日期：2021-08-10

accelerate the establishment and promotion of secure recommendation systems, provide a reference for research on the trust of intelligent algorithms, and provide better ideas for the supervision and governance of smart algorithms.

Key words news; recommendation algorithm; trust evaluation system; fairness; interpretability; anti-denying

1 引言

信息过载的概念于 1964 年首次被提出[1], 但直至 21 世纪进入大数据时代后, 随着 AI、5G、AR/VR 等新技术快速发展, 电子商务、社交网络、短视频等新应用层出不穷, 互联网中每天产生海量的新闻、商品、视频、音乐等信息, 信息过载问题日益严重。据中国互联网络信息中心(CNNIC)发布的《第 47 次中国互联网络发展状况统计报告》[2]统计, 截至 2020 年 12 月, 我国网民规模达 9.89 亿, 其中网络新闻用户规模达 7.43 亿, 网络购物用户规模达 7.82 亿, 网络视频用户规模达 9.27 亿。数据规模呈现爆发式增长, 根据国际数据集团(IDC)2018 年在文献[3]中预测, 2025 年全球数据总量预计增至 175ZB, 将是 2018 年的 5 倍左右。如何从海量信息中获取价值内容是互联网信息服务平台和用户关心, 且迫切需要解决的问题。因此, 作为信息过滤的有效工具, 推荐算法应运而生。推荐算法旨在从过载的信息中, 通过信息过滤筛选技术, 为用户推荐其感兴趣的高质量内容。在大数据、应用场景和计算力的推动下, 作为人工智能分支之一, 推荐算法在电子商务、新闻等领域得到了广泛的应用, 不仅提高了信息分发效率, 还一定程度上缓解了信息过载问题; 通过解读用户个体兴趣, 进行个性化智能推荐的同时, 给互联网信息服务提供商带来极大的商业价值。

互联网推荐算法和推荐系统的发展历程, 可分为萌芽期、发展期和管制期 3 个阶段。萌芽期起始于 1990 年, 哥伦比亚大学 Jussi Karlgren 首次提出推荐系统[4]的概念, 此后, 明尼苏达大学的 GroupLens 研究组于 1994 年推出名为 GroupLen[5]的新闻推荐系统, 提出协同过滤的思想, 1997 年的基于内容协同过滤算法[6], 2003 年的基于物品协同过滤算法[7]也相继问世。2006 年北美在线视频服务提供商 Netflix 举办的推荐算法竞赛, 极大地推动了推荐系统的发展。此次比赛标志着推荐系统进入发展期。此后, 面向不同应用场景的推荐算法犹如雨后春笋涌现, 效果不断得到提升, 诸如提出矩阵分解方法实现推荐任务的 FunkSVD[8], 首次将深度学习技术与推荐技术结合的 RBCF 算法[9], 首次从概率角度构造 MF 模型的 PMF[10], 结合社交信息[11-12]、基于信任方法[13]、引入注意力机制[14]等以提高推荐算法效果和性能。2016 年深度学习技术被全面应用于推荐系统领域。YouTube 将深度神经网络方法应用于推荐系统中, 打造的工业级推荐系统[15], 实现了大规模内容的高质量推荐,为后续工业级推荐系统的优化开拓了思路。近年来, 随着推荐算法应用领域的日益广泛, 为用户提供基于行为习惯和兴趣偏好的个性化推荐, 极大地提高了用户体验的同时, 逐渐引起的算法偏见、用户隐私问题、信任问题、可解释性、公平性越来越受到用户、相关监管部门等各方的关注, 《网络信息内容生态治理规定》[16]的实施标志着我国进入推荐算法管制期。

互联网新闻是推荐算法应用最为广泛的领域之一, 用户规模大、垂直平台类型繁多, 已成为帮助网民获取信息的主流方式。新闻平台通过推荐算法改变了新闻内容分发方式, 在给网民带来信息筛选便利的同时, 也带来了价值观缺失、信息茧房、算法偏见等问题。为了防止互联网新闻信息运营和服务平台成为传播不良内容传播的帮凶, 警惕算法决定内容、算法偏见, 迫切需要研究新闻推荐算法的公平性、可解释性等, 提高新闻推荐算法的透明性和可信程度。

算法模型机理透明, 推荐内容健康、公平、可解释, 对安全问题抗抵赖是构建可信新闻推荐算法的必备条件。可信新闻推荐算法是建立安全新闻推荐系统及优化推荐结果的核心技术。本文研究具备可解释性、公平性、抗抵赖性的可信新闻推荐算法, 对加速安全推荐系统的建立和推广, 建立可信推荐算法生态, 极具理论意义和应用价值。

本文组织结构如下: 第 2 章扼要介绍了新闻推荐算法的关键要素及分类; 第 3 章梳理分析新闻推荐算法的风险情况及国内外应对现状; 第 4 章提出新闻推荐算法可信评价指标体系; 第 5 章对全文进行了总结, 并提出下一步研究计划。

2 新闻推荐算法关键要素及分类

2.1 新闻推荐算法关键要素

作为推荐系统的核心, 新闻推荐算法的关键要素, 按照根据推荐系统运行过程, 分为稿件、用户和推荐策略, 具体细分为稿件画像、用户画像、推荐推送、反馈干预和人工复审。

新闻稿件的内容形式包括文本、图片、视频。稿件画像是利用文本和多媒体分析技术对稿件的内

容进行挖掘和分析, 生成结构化和分级分类的稿件模型。稿件画像通常分为两个维度, 主题标签和质量标签, 主题标签包括历史、时尚、教育、娱乐等多级标签, 质量标签包括正能量、违法不良、低俗、猎奇/易反感、标题夸张、评论指向等类别。对稿件的内容分析可借助分类器模型、主题模型、实体识别模型、嵌入模型进行内容分类、主题挖掘、角色识别、嵌入语义分析。通常, 推荐算法利用人工标注稿件和用户反馈信息作为训练样本, 训练稿件画像模型。

用户画像, 是将用户信息特征的向量化表示, 用于个性化推荐和精准营销的有效工具。在新闻推荐领域, 用户画像特征体系主要包括人口属性、兴趣属性、行为属性、社交属性和风险控制。其中人口属性主要包含性别、职业、年龄、婚姻状况等, 人口属性相关的标签相对比较稳定, 在较长时间内不需要更新。兴趣属性旨在描述用户兴趣爱好, 具有较强的时效性, 包括长期和短期兴趣。行为属性是另一种刻画用户的常见维度, 可以用以挖掘用户偏好和特征。社交属性被用于了解用户的家庭成员、社交关系、社交偏好、社交活跃程度等。风险控制旨在通过统计账号风险、失信风险、潜在问题用户、无效渠道等信息, 从根源上防止不良内容的产生和传播。

按变化频率, 用户特征分为静态特征和动态特征, 静态特征一般指通常很少发生变化的用户基本属性信息, 如性别、年龄、职业等; 而动态特征通常指与用户兴趣偏好相关, 在时间和空间上是动态变化的特征。按照数据提取和处理维度, 用户特征分为事实特征、模型特征和预测特征。事实特征是指从原始数据中直接提取的用户基本信息, 不需要使用算法模型, 实现简单。模型特征指通过定义规则, 建立模型计算得到的特征实例。预测特征是基于用户的基本信息属性、行为属性、社交属性, 利用机器学习、深度学习等技术预测的特征。

用户画像构建方法包括两类, 基于统计和基于模型的用户建模[17]。基于统计的用户建模方法, 主要是利用统计方法, 对用户人口属性、历史行为等数据, 将统计结果进行量化和分析。基于统计的构建方法, 简单易实现, 主要应用于结构化信息, 不适用于文本、图片、音视频等非结构化数据。基于模型的构建方法是利用机器学习、深度学习等方法, 针对结构化和非结构化数据, 学习和构建高维稠密向量, 在当前推荐系统中得到广泛应用。

推荐推送技术架构包括召回、排序、重排三个阶段。稿件召回阶段, 考虑用户兴趣偏好、热门内容等多种因素, 通过多路召回进行稿件初筛, 主流召回方法包括基于内容(Content-based)[18]、协同过滤(Collaborative Filtering)[19]、基于知识(Knowledge-based)[20]、混合推荐[6]等传统方法, 基于 FM 模型(Factorization Machines, FM)[21]、基于深度神经网络(Deep Neural Networks, DNN)[22]等深度学习方法。训练模型包括离线模型和实时模型, 分别利用时效性是否敏感的标签进行训练, 以更新推荐模型; 第二阶段是排序, 排序是推荐系统关键环节, 常用模型包括逻辑回归(Logistic Regression, LR)、梯度提升决策树(Gradient Boosting Decision Tree, GBDT)、FM、深度神经网络(Deep Neural Networks, DNN)、Pointwise 等。排序完成后, 进入重排阶段, 根据业务需要和安全策略, 一般需要进行强插过滤、打散, 保证推荐结果的多样性, 常见的重排模型有循环神经网络(Recurrent Neural Network, RNN)、Transformer。

反馈干预主要通过实时收集统计用户阅读、评论、转发、分享等正面反馈, 不喜欢、举报、负评等负面反馈, 更新至推荐模型中实时调整推荐效果。

在工业界, 人工复审环节是重中之重, 对重排结果进行人工二次审核, 审核策略一般按稿件类型和安全等级进行全审和盲审。针对高危、敏感等级稿件进行全审, 其他类型进行盲审。稿件通过人工复审后, 才会形成进入最终推荐稿件列表。

2.2 新闻推荐算法分类

推荐算法是推荐系统中的核心, 在很大程度上决定了推荐系统效果和性能。目前, 对推荐算法的分类并没有统一的标准, 很多学者从不同角度对推荐算法进行分类, 本文从推荐模型角度, 将产业界新闻推荐领域应用较为广泛的推荐算法分成以下几种: 协同过滤方法、矩阵分解方法(Matrix Factorization), 聚类、深度学习方法。

协同过滤是利用集体智慧的一个典型方法, 协同过滤及其扩展方案是最常用的推荐算法之一。当向用户推荐某些新闻内容时, 最合乎逻辑的是找到兴趣相似的人, 分析其行为, 并向用户推荐相同的内容; 或者查看与用户之前的喜好相类似的内容, 并进行推荐。协同过滤两种基本方法: 基于用户的协同过滤(user-based)和基于内容的协同过滤(item-based)。在这两种情况下, 一般推荐的步骤如下: (1)收集用户偏好及行为数据, 如阅读、点赞、评论转发等; (2)对数据进行降噪以及归一化操作得到一个用户偏好的二维矩阵; (3)计算用户间或者内容间相似度, 常见的计算方法有: 欧几里德距离、皮尔逊相关系数、余弦相似度、Tanimoto 系数等。计算得到的两个相似度将作为基于用户、内容的两项协同过

滤的推荐依据。

矩阵分解算法的核心思想是利用用户-内容的评分矩阵, 分解出潜在特征, 然后预测用户对关注或阅读过的内容的评分, 将得分高的内容作为推荐项。在获得用户评分矩阵后, 利用矩阵分解的方法将用户评分矩阵分解为两个低秩矩阵(用户特征矩阵和内容特征矩阵)的乘积, 将用户和内容嵌入到同一个 k 维的向量空间。用户向量和内容向量的内积代表了用户对内容的偏好度。因为 k 维向量空间的每一个维度不具备与现实场景对应的可解释含义, 所以矩阵分解算法的可解释性较差。

协同过滤以及矩阵分解都是有监督的机器学习方法, 在推荐系统中也可以利用无监督的方法-聚类。在推荐中可利用 K-Means、密度聚类(Density-based spatial clustering of applications with noise, DBSCAN)、高斯混合模型(Gaussian Mixed Model, GMM)等聚类算法对用户或者内容的分组, 随后从分组内挑选内容推荐给用户。在实际推荐系统构建中, 聚类方法一般适用于系统初期用户数据量不足的场景, 或者作为协同过滤的补充, 降低计算复杂度。

在过去十年中, 神经网络取得了长足的发展。如今已被广泛应用, 在某些领域正在逐步取代传统的机器学习方法。深度学习模型应用于推荐算法既可以有效获取非线性和重要的用户-内容关系, 还可以在高层中获得更实用的抽象特征, 从大量冗余信息数据中获取复杂的关系, 如上下文、文本、图片等信息。深度学习在推荐系统中既可以作为独立模型使用, 如 Neural Collaborative Filtering (NCF)[23], Cross-domain Content-boosted Collaborative Filtering (CCCFNet)[24], Deep Factorization Machine (DeepFM)[21]等; 也可以结合传统的推荐方法使用, 如利用 MLP 进行用户内容间非线性拟合[25]、利用 CNN 提取局部和全局信息、利用 RNN 提取序列信息[26]、利用 DSSM 进行语义匹配[27-28]等。

3 新闻推荐算法风险分析

3.1 新闻推荐算法风险分析

新闻推荐算法在使用过程中, 存在危害国家和社会安全、用户安全和新闻推荐平台安全等三方面风险。国家和社会方面, 第一主要是推荐算法易被不法人员用于操纵舆论导向, 进行网络意识形态垄断, 窄化人们思想, 威胁国家意识形态安全; 第二是内容质量问题, 如果互联网中充斥着大量的劣质内容, 不利于国家精神文明建设和网民积极向上的价值观的形成。用户方面, 主要是用户数据隐私风险和算法偏见问题、信息茧房问题。新闻推荐平台方面, 主要是用户对新闻推荐平台和推荐结果的信任问题及新闻推荐平台和用户行为的抵赖问题。

内容质量问题。个性化推荐在新闻推荐系统中广泛应用和自媒体的兴起, 低俗内容泛滥, 内容质量无法得到保障, 失去价值引领的属性。内容质量问题由新闻推荐平台管理和推荐算法两方面引起, 新闻推荐平台侧对内容源及质量的分级分类管理体系不够完善, 存在漏检隐患。推荐算法依赖用户画像、行为特征、兴趣特征等推荐主题、关键词相关性较高的内容, 并不对内容来源、质量进行核验。

信息茧房。在个性化推荐领域, 推荐算法向用户推荐的大多是其感兴趣的信息。随着时间的推移, 这将导致推荐内容逐渐同质化、信息阈逐渐收窄, 甚至加重用户群体阶层极化现象。

算法偏见问题。新闻推荐平台为了提高推荐算法推荐的精准性, 训练数据中会引入诸如性别、年龄、职业, 甚至种族等敏感特征, 这一做法违背了算法中立性的原则, 间接造成算法偏见问题。随着用户不断循环反馈, 推荐算法模型迭代调整, 偏见问题将被逐渐加强。

用户数据隐私问题。推荐算法效果的优劣, 关键因素在于用户特征的质量, 而用户特征质量主要由用户数据资源决定。新闻推荐平台在对用户数据采集、分析和挖掘过程中, 存在用户不知情情况下, 过度采集和滥用, 造成用户隐私数据泄露的风险。当前, 用户数据的采集范围、跨平台使用方式、用户对隐私数据的可控程度, 是相关监管部门、用户迫切关心的问题。

信任问题包括用户对个人数据采集和使用的信任、对推荐结果信任。通常用户无法得知推荐平台采集了哪些数据及如何使用, 因此存在对新闻推荐平台关于个人数据信任问题。用户在新闻平台上所见的内容, 主要依赖机器和推荐算法完成。推荐算法是大多使用黑盒化模型, 透明度低, 甚至研发人员都很难解释推荐算法底层机理和推荐结果, 用户更是被动接受推荐结果。如何让用户更大程度上信任新闻推荐平台、推荐算法的决策结果, 引起了学术界、产业界的广泛关注和研究。

抗抵赖问题。新闻推荐平台上常存在一些恶意用户在发布低质内容或者产生一些恶意的行为, 这些内容或者行为会对平台、对其他用户产生一些不利的影响, 事后这些恶意用户可能会尽力去删除或者损毁这些行为证据以逃避、抵赖社会追责。因新

闻推荐平台的封闭性, 新闻平台侧在对用户数据的采集和使用、推荐结果的展示等做出不当行为时, 可能存在删除或损毁操作, 以抵赖相关监管部门的查证。

以上问题可统一归为公平性、可解释性和抗抵赖性三类问题。随着新闻推荐算法应用的广泛性, 相关监管部门和研究学者大多从公平性和可解释性研究内容质量、信息茧房、算法偏见、用户数据隐私等问题, 本文在公平性和可解释性基础上, 首次将抗抵赖性引入作为新闻推荐算法安全问题之一。

3.2 推荐算法安全风险应对现状分析

在推荐算法早期发展和应用的进程中, 产业和学术界通常倾向于追求算法模型的性能指标, 如准确度、精确度和召回率等。近几年, 随着推荐算法应用, 人们逐渐意识到算法安全、公平问题的重要性。因推荐算法属于智能算法的一种, 本文从智能算法安全角度, 梳理国内外政府、学者从政策、标准规范[29]和学术方面对算法安全问题进行的前瞻性研究和探索进展。

(1) 政策方面, 从总体政策举措看, 美国注重在公共数据资源和人工智能安全设计方面要求。2016 年, 美国国防部先进研究项目局(DARPA)资助并启动可解释性人工智能项目 XAI(Explainable AI), 旨在研究实现包含可解释性技术和模型的通用新型机器学习技术, 一方面使得用户理解、信任算法决策结果, 一方面便于算法平台和监管部门有效管理人工智能系统。2017 年发布的《算法透明和可责性声明》[30]中提出了可解释、数据来源保护、可审查性、验证和测试等准则。2019 年在《国家人工智能研究与发展战略计划》中将人工智能系统安全、开发可共享的公共数据集和环境作为战略重点之一。此外, 美国立法者要求 Twitter、 YouTube 和 Facebook 等互联网企业提高算法透明度, 并评估算法是否存在不公平性。2021 年 2 月美国布鲁金斯学会呼吁重启美国会技术评估办公室, 针对人工智能发展可能带来的算法嵌入、算法公平性、算法透明度等问题, 提出缓解建议。

欧盟在隐私数据保护方面较为重视, 已经出台的《通用数据保护条例》(GDPR)中明确赋予个人决定隐私数据使用范围的权利。英国在《人工智能在英国: 准备、志向与能力?》报告中, 提出人工智能应有可理解性和公平性原则, 以及保护个人数据权利或隐私原则, 鼓励在重要领域研制可解释性的人工智能系统, 研究训练数据和算法的审查和测试机制, 探索数据访问和共享的有效措施。2019 年欧盟委员会发布的《可信赖人工智能伦理指南》(Ethics Guidelines for Trustworthy AI)[31]中的公平准则要求人工智能系统的开发、部署和应用要坚持实质公平和程序公平, 确保利益和成本的平等分配、个人及群体免受歧视和偏见。

日本人工智能学会(JSAI)发布的《日本人工智能学会伦理准则》中要求遵循和实践尊重隐私、公正和安全原则。加拿大发布的《可靠的人工智能草案蒙特利尔宣言》中提出隐私是人工智能发展过程中应当遵守的道德原则之一。

我国也已经开展智能算法在相关领域中的规制方法。国务院在 2017 年的《新一代人工智能发展规划》[32]中提出了实现具备高可解释性、强泛化能力的人工智能的目标。此外, 我国已经将算法纳入监管, 2019 年出台的《网络信息内容生态治理规定》[16], 针对算法推荐引发的负面影响, 明确了推荐算法的分发方式, 要求企业持续优化算法模型, 在利用算法决策时, 确保算法的准确性、公平性等。

(2) 标准规范方面, 2017 年国际标准化组织(ISO/IEC JTC1)成立人工智能的分委员会, 开展的标准研制工作中涉及人工智能可信度、鲁棒性评估、算法偏见等主题。ITU-T 于 2017—2018 年组织的“AI for Good Global”峰会中, 重点关注了人工智能技术可信的战略问题。电气与电子工程师协会(IEEE)正在研制 IEEE P7000 系列标准 IEEE P7002《数据隐私处理》、IEEE P7003《算法偏差注意事项》、IEEE P7011《新闻信源识别和评级过程标准》等。

我国对算法安全标准方面的工作, 集中在算法模型、数据、基础设施、产品和应用相关的安全标准。2018 年我国首个人工智能深度学习算法标准《人工智能深度学习算法评估规范》(T/CESA 1026-2018)发布, 目标旨在发现深度学习算法中影响算法可靠性的因素及如何提高算法可靠性。规范中提出了深度学习算法的评估指标体系、评估流程等内容, 指导深度学习算法相关方对深度学习算法的可靠性开展评估工作。在数据安全领域, 国家标准《信息安全技术 个人信息安全规范》(GB/T 35273-2020)和《信息安全技术 个人信息去标识化指南》(GB/T 37964-2019)等已经发布。国家标准化管理委员会等五部门联合印发《国家新一代人工智能标准体系建设指南》[33]中, 提出人工智能数据、算法和模型安全标准, 包括数据安全、隐私保护、算法模型可信赖等。

(3) 学术研究方面, 算法可解释性和公平性是当前的突出问题和研究重点。

Miller[34]从非数学层面定义可解释性是人们能

够理解决策原因的程度。如果一个推荐算法的决策比另一个推荐算法的决策能让人更容易理解, 则认为前者具有更高的可解释性。算法可解释性的概念起源于 2014 年[35], 近年来算法可解释性问题受到了政府、产业界和学术界的广泛关注和深入研究[36]。

推荐算法的解释目标是以用户为导向的推荐结果解释和以模型为导向的模型机制的解释, 建立用户与推荐平台间的信任关系的同时, 指导算法工程师进行特征工程和调试算法模型。

当前, 算法可解释性方法包含按建模周期流程划分及按解释范围划分。按照建模周期流程划分, 即分为建模前、建模中、建模后三个阶段, (1)建模前的可解释性重点关注数据的可解释性, 针对大规模或高维数据, 通过统计分析及交互式可视化等方法, 多层次角度理解数据的特征分布, 进而支持人类决策; (2)建模中的可解释性是模型有关可解释性, 即深度可解释, 实现对算法模型的机理及执行过程的可解释, 如简化成回归模型、树模型、图模型等进行解释; (3)建模后的可解释性是模型无关可解释, 是当前研究尝试最多的方向。主要通过不同的手段来解释算法模型的决策依据, 测试决策依据对推荐结果的影响程度, 经典方法包括敏感性分析(Sensitivity Analysis)[37]、基于梯度的方法(Gradient-based Methods)[38]、全局或局部代理模型(Surrogate Models)[39]、知识蒸馏(Knowledge Distillation)[40]、隐藏层可视化等。按照解释对象角度, 近年来, 面向用户的推荐可解释性方式, 主要包括异构信息建模[41-42]、知识增强[43-44]和反事实解释[45]等, 在解释推荐结果的同时不断优化推荐质量。按照解释范围分为全局可解释和局部可解释, 全局可解释是从数据及特征、输入参数、模型结构等方面对整个算法模型的决策进行解释, 比如影响决策的关键特征的分布、特征之间如何相互作用等。局部可解释是指在不考虑算法模型内在结构的前提下, 对特定一条样本或一组样本的预测结果进行解释。局部可解释的预测结果可能只依赖于某些线性或单调性的特征, 相对全局可解释, 具有更高的准确性。

算法模型可解释性的工程实现方面, 包含演进式可解释算法模型和全新式可解释算法模型两种思路。演进式可解释算法模型是在不改变现有算法模型的前提下, 将解释模块集成至推荐系统中, 实现算法模型的可解释性; 全新式可解释算法模型是重新设计和实现算法模型, 在设计理念中, 融入可解释性功能。

算法的公平性旨在研究实现推荐算法的决策结果对受众和内容生产者的个人或群体不存在因其固有或后天属性所引起的算法偏见。造成算法不公平性的主要原因包括 4 种: (1)多样性不足, 新闻推荐平台可能为追求利益最大化, 将流量大、热门或存在利益相关的内容排名靠前, 导致曝光内容多样性不足, 进而造成对受众和内容生产者两方的不公平性; (2)算法偏见, 个性化推荐是“千人千面”的差异化推荐, 新闻推荐平台将敏感属性作为训练特征, 优化对不同人群的推荐内容及内容结构。此类算法偏见严重破坏了受众的公平性; (3)信息茧房问题, 在降低用户公平性的同时, 将加剧社会价值分层; (4)优质但冷门的内容得不到曝光机会, 也是对内容生产者的不公平。

针对以上推荐算法中的公平性问题, 近几年, 国内外相关监管部门和研究学者开始重点关注, 但公平性相关解决方案仍处于初期探索阶段。

国外互联网企业如 Facebook、YouTube 等开始尝试探索推荐算法公平性问题, 以便给予用户更大的控制和选择权限。例如, Twitter 曾表达了研究用户对算法选择, 实现用户控制自己使用算法的愿景, 同时宣布启动研究算法公平性的计划, 评估其使用的算法是否存在潜在危害。Facebook 在 2020 年已经成立算法偏见相关问题研究团队。微软在 2018 年表示开发了一套新工具, 用来判断人工智能算法是否存在偏见, 帮助互联网安全使用人工智能算法, 并及时捕获安全风险。YouTube 对其推荐算法模型作出一系列调整, 如拒绝某类内容推送, 以便用户可以更容易地探索主题和内容。

中国信息通信研究院发布的《人工智能安全框架(2020)》[46]中提出算法公平性保障是算法安全技术之一, 可从算法公平性约束和偏见后处理两方面保障算法公平性。推荐算法公平性的研究主要从数据公平性、内容公平性、用户公平性、算法模型公平性等角度进行研究。推荐算法的公平性是涵盖受众、内容生产者、推荐平台三方的多目标公平性[47-49], 一般采用多目标优化方法, 既保障对受众的公平性, 也保障内容的多样性。当前研究方向主要集中在数据角度、受众角度、内容角度、多目标角度等。

数据角度, 如果算法模型输入数据未使用诸如性别、年龄、受教育程度、种族等敏感属性, 则视为是公平的。一般通过机器或人工干预机制对推荐结果核查来解决数据公平性问题。

用户角度, Hongyu Lu 等人[50]从受众满意度角度, 研究受众在阅读前、阅读后、后任务三阶段的动态偏好, 提升受众偏好的捕获准确度, 并同步提高推

荐质量。组推荐旨在向兴趣相同、社会关系粘性强的群组推荐内容, 例如谷歌 Beutel Alex 等人[51]通过提出成对公平性、组内成对公平性和组间成对公平性指标, 实现对推荐系统排名公平性的无偏度量。文章[52]认为同一群组内的受众, 感兴趣的内容有相似之处, 因此将组分为长期型组, 如一家人、长期好友; 另一种是短期型组, 如兴趣爱好暂时趋于相同的一群人。文章[53]以排名敏感的方式平衡被推荐内容在组成员间的相关性, 并利用贪心算法 GFAR 寻找 top-N。文章[54]设计了一种重新排序的方法, 通过在评估指标上添加约束来缓解优势组和劣势组的推荐质量的不公平性问题。

内容角度, 研究人员主要从流行度偏差[55]、位置偏差[56]、曝光偏差[57]等方面研究如何提高内容在推荐选择、排名等方面的公平性[58]。流行度偏差主要问题是热度低或不流行的内容得不到有效推荐, 一般解决方法是对内容赋予热度权重, 通过升权和降权, 调整内容展示的机会、位置等。位置偏差中排名靠前的内容更容易被用户注意且产生互动, 但这不足以代表用户的真实偏好。算法模型在获取用户偏好时出现偏差, 一般缓解方法是将位置特征作为输入参数, 或者构建用户行为模型并应用于推荐模型。曝光偏差是对没有机会展示的内容的不公平, 进而产生马太效应问题, 简单解决办法是通过探索机制, 对于新内容和历史曝光机会比较少的内容, 给予一定的探索机会, 提升用户对内容的可见度, 如使用汤普森采样的方法将排序较后的内容, 设置一定的概率呈现在较前的位置、设置用户行为无关的内容特征、利用贪心方法进行推荐结果校准等。

理论上, 在保障多目标公平性时, 提升一方的公正性, 另一方的公正性则会降低, 同时降低整体推荐质量。近年来, 研究学者关注于多目标公平均衡问题, 寻求一种解决方案, 平衡内容提供者和受众的多方公平性, 如文章[59]和[60]。前者通过分析推荐质量、受众公平性和内容提供者公平性之间的关系, 提出一种面向受众和内容提供者的双方公平性的推荐模型 TFROM, 以保障双方的公平性。后者将推荐公平性问题映射为不可分割物品的公平分配问题, 以此提出 FairRec 算法, 保证大多数内容提供者中至少一个能够获得最大份额的曝光率, 而且每个受众拥有相对较好的公平性。Robin Burke 等人[49]证明了一种改进的稀疏线性方法 SLIM, 可以改善受众和推荐内容邻域之间的平衡, 在提高推荐公平性的同时, 最大程度降低排序性能损失。

4 可信评价指标体系

当前, 国内外对推荐算法的可信评价研究仍是空白。本文将从公平性、可解释性、抗抵赖性三方面建立评价新闻推荐算法的可信指标体系, 共划分成三级指标, 如表 1 所示。一级指标包括公平性、可解释性和抗抵赖性。公平性包括生产者侧、内容侧、受众侧和算法模型侧; 可解释性包括数据可解释性、模型可解释性和推荐结果可解释性; 抗抵赖性包括受众侧和平台侧。公平性、可解释性和抗抵赖性是正相关关系, 当公平性和抗抵赖性越强、可解释程度越高, 新闻推荐算法的可信度越高。

表 1 推荐算法可信评价指标

Table 1 The trust evaluating indicators of recommendation algorithms

一级指标	二级指标	三级指标
公平性	生产者侧	生产者被推荐率
		稿源可信度
	内容侧	稿件池主题覆盖度
		新闻实时性
		新闻真实性
		内容规范性
	受众侧	相似个体间推荐内容偏差
		是否进行群组划分
		是否进行兴趣探索
	算法模型侧	训练数据是否使用敏感属性
		推荐内容分布与受众兴趣分布偏差
		推荐内容覆盖率
		最近 N 小时新闻推荐率
		同质内容推荐率
		受众覆盖率
		冷门内容推荐率
可解释性	数据可解释性	受众关键特征
		内容关键特征
		特征及关系可视化
	模型可解释性	参数可解释
		模型可解释程度
		是否具备解释功能或模块
	推荐结果可解释性	解释准确率
		解释有效性
抗抵赖性	受众侧	发布内容存证
		交互行为存证
	平台侧	采集内容存证
		历史推荐内容存证
		训练数据有效特征存证

(1) 公平性

公平性应从内容生产者、内容本身、受众、算

法模型等方面实现对用户的公平。

生产者侧指标包括生产者被推荐率(recommendation rate of producer, RRP)和稿源可信度(credibility of sources, CS)。

生产者被推荐率是指新闻推荐算法对平台上内容生产者的推荐率。计算公式如下:

$$RRP = \frac{User_R}{User_{All}}$$

其中, $User_R$ 是被推荐的受众数, $User_{All}$ 是新闻平台所有受众数。生产者被推荐率越高, 对生产者越公平。

稿源可信度是指新闻稿件来源的可信, 新闻稿件来源于国家互联网信息办公室发布的《互联网新闻信息稿源单位名单》[61], 政务机构、新闻媒体机构和其他被授权发布时政信息的单位的, 可信度高。此外, 依据包含除此之外的来源比例, 逐步降低可信度。

内容侧指标细分为稿件池主题覆盖度(coverge of news theme, CNT)、新闻实时性(real time, RT)、新闻真实性(authenticity)和内容规范性(normativity)。

稿件池主题覆盖度是新闻平台稿件池中内容主题占新闻领域总主题数的比率。计算方式如下:

$$CNT = \frac{T_{Exist}}{T_{All}}$$

其中, T_{Exist} 是新闻平台稿件主题数, T_{All} 是新闻主题类别总数。稿件池主题覆盖度应接近 100%, 且覆盖度越高, 对受众的公平性越高。

新闻实时性指新闻被推荐给受众的时效。新闻实时性不应晚于新闻本身的时效。新闻实时性越高, 对受众的公平性越高。

真实性是新闻应具备的基本原则之一。新闻内容在呈现给受众前, 应进行真实性检测。当真实性检测模型准确率达 90%以上时认为具备真实性。新闻真实性越高, 对受众的公平性越高。

内容规范性是指新闻内容应满足健康向上、能够弘扬正确价值观, 不含违法不良信息。当一条新闻内容违反内容规范性要求时, 直接违背了受众获取健康内容的公平性权利, 极大降低了推荐算法的可信度。

受众侧指标可细分为相似个体间推荐内容偏差(news bias of similar individual, NBSI)、是否进行群组划分(is groups, IS)和是否进行兴趣探索(explore interest, EI)。

通常相似个体间接收的推荐内容一致或接近一致, 被认为具有高公平性。相似个体间推荐内容偏差是指相似个体间被推荐内容的不同程度。计算方式如下:

$$NBSI = \frac{\left|News_A \cup News_B\right| - \left|News_A \cap News_B\right|}{\left|News_A \cup News_B\right|}$$

其中, $News_A$ 和 $News_B$ 是对个体受众 A 和 B 推荐的新闻。考虑个体相似度计算偏差, 可对 $NBSI$ 设置一个阈值, 当 $NBSI$ 大于该阈值时, 认为是对受众存在不公平性, 而且将随着 $NBSI$ 增大而加剧。

通常, 群组划分是平台为了提高个性化推荐效果。但根据用户偏好、位置、职业等属性将用户划分群组, 在一定程度上降低了受众公平性, 将加剧社会分层。

兴趣探索是为了挖掘用户兴趣点, 拓宽推荐范围和优化推荐效果。若推荐算法中运用兴趣探索机制, 不仅一定程度上增加了内容多样性, 还将增大对受众的公平性。

算法模型侧的三级指标包括训练数据是否使用敏感属性(sensitive attributes, SA)、推荐内容分布与受众兴趣分布偏差(bias between news and user, BBNU)、推荐内容覆盖率(coverge of recommended news, CRN)、最近 N 小时新闻推荐率(recent N-hour recommendation rate of news, RNRR)、同质内容推荐率(recommendation rate of homogeneous content, RRHC)、受众覆盖率(coverge rate of user, CRU)和冷门内容推荐率(recommendation rate of non-popular news, RRNN)。

训练数据中使用敏感属性是间接地对受众实施分级分类, 造成的显式不公平。明确敏感属性范围和分级分类, 根据使用敏感属性的级别和类别, 对模型侧公平性进行分级量化。

推荐内容分布与受众兴趣分布偏差是指因算法模型推荐的精准度问题, 造成推荐的内容分布与受众兴趣分布有一定偏差。该偏差与公平性是负相关, 偏差越小, 公平性越高。

$$BBNU = Q_{user} - F_{news}$$

其中 Q_{user} 是受众兴趣分布, F_{news} 推荐算法为受众 user 推荐的内容分布, 若 $BBNU$ 为常量, 认为无偏差。

推荐内容覆盖率是指推荐算法推荐的内容占稿件池中内容的比率, 计算公式如下:

$$CRN = \frac{C_u}{C_{All}}$$

其中, C_u 是推荐给受众 u 的内容量, C_{All} 是稿件池中内容总量。推荐内容覆盖率越高, 表示推荐算法的内容分发能力越强, 公平性也越强。

最近 N 小时新闻推荐率表示在某一时间点, 推

荐最近N小时内生产的新闻的量与总推荐量的比率。计算公式如下:

$$RNRR = \frac{N_{news}}{R_{All}}$$

其中, N_{news}是推荐的最近N小时内生产的新闻数量, R_{All}是推荐新闻总量。最近 N 小时新闻推荐率越高, 代表推荐算法对新闻推荐的时效性越高, 推荐算法公平性越强。

属于同一分类、同一话题和同一作者的内容称为同质内容。同质内容推荐率是指在向单个受众的一次推荐中, 同质内容量与所有推荐内容的比率。计算公式如下:

$$RRHC = \frac{R_h}{R_{All}}$$

其中, R_h是推荐的同质内容量, R_{All}推荐内容总量。同质内容推荐率越高, 推荐算法在内容多样性方面能力越低, 推荐算法的公平性越低。

受众覆盖率是指被推荐受众占总受众量的比率。计算公式如下:

$$CRU = \frac{R_u}{U_{All}}$$

其中, R_u是被推荐受众量, U_{All}是总受众量。对一条新闻来讲, 受众覆盖率越高, 对受众越公平。但不同类别新闻的受众覆盖率不同, 如时政类新闻应比娱乐类新闻受众覆盖率高。在推荐平台经济效益的策略下, 受众覆盖率应设置一个最低阈值, 每种类别新闻的受众覆盖率不应低于该阈值。

保证冷门内容推荐率是推荐内容多样性的保障措施之一, 冷门内容推荐率是指在一次推荐中, 对非流行的新闻内容占推荐总量的比率。计算公式如下:

$$RRNN = \frac{R_n}{R_{All}}$$

其中, R_n是推荐的冷门内容量, R_{All}是总推荐内容量。冷门内容推荐率的合理性, 是对冷门内容公平性的有效保障。

可解释性是增强推荐算法透明性的有效方法, 可解释性越强, 推荐算法可信程度越高。本文从数据层面、算法模型层面和推荐结果层面分析推荐算法可解释性评价指标, 具体分为数据可解释性、模型可解释性、推荐结果可解释性。

数据可解释主要从受众和新闻内容的具有影响力特征及关系是否可视化来评价, 评价指标包括受众关键特征、内容关键特征和特征及关系可视化。如果推荐平台具备解释关键特征和可视化展示特征及关系的能力, 说明具备数据可解释性。

算法模型的可解释性主要面向算法开发人员, 有利于增强对模型的调参和优化。通常情况下, 算法模型基本是一个黑盒, 因此可从参数可解释、模型可解释程度及是否具备解释功能或模块三个指标进行解释。

参数是推荐算法模型的重要组成部分, 在模型构建和优化的过程中, 应对参数的初始化值和调参的依据、目标进行明确和记录, 增强模型参数的可解释性。

推荐算法可根据使用的模型类别, 判断模型可解释程度。如逻辑回归、树、图等统计学模型可根据规则进行推理解释, 因此可解释程度较高。而深度学习模型、混合模型等因网络复杂度高、黑盒化程度高, 可解释程度低。

当前算法解释功能或模块的研究仍处于研究初期, 主要路线分为演进式和全新式, 演进式推荐算法模型是独立于当前推荐算法模型运行的解释模块, 即浅层解释方法, 主要解释输入和输出的因果关系。全新式推荐算法模型致力于研究内置的解释功能, 即深层解释方法, 从算法模型原理角度解释每一步执行的过程。若一个推荐算法模型具备可解释功能或模块, 代表具备可解释性, 且深层解释方法比浅层解释方法具有更高的可解释程度。

推荐结果可解释性主要面向受众, 使其更好的理解和信任推荐结果。当前, 推荐结果的解释已经广泛应用于新闻推荐平台, 且大幅度提高了流量。如基于好友关系的解释, 可通过说明多少好友感兴趣、哪位好友已经关注等方法进行解释。推荐结果可解释性主要包括推荐准确率和推荐有效性。解释准确率一方面标识着用户对推荐结果的信任度, 另一方面代表推荐算法的解释能力。解释有效性可通过可解释性推荐结果产生的效益进行量化。解释准确率越高、解释有效性越强, 表示推荐算法可解释性越高。

一个可信的推荐算法应避免和预防受众、内容生产者等用户和新闻平台对各自的违法或不当行为进行抵赖, 因此抗抵赖性也应是推荐算法可信评价中的一个重要指标。

抗抵赖性指标按照对象可分为指标用户侧、新闻平台侧两个维度, 分别对用户、平台两方的行为进行评价。

用户在系统中发布内容, 产生评论、点赞等行为, 新闻平台应有能力记录详尽的产生、传播、销毁等过程, 即分别对发布内容和用户与平台间的交互行

为进行存证。按照存证期限可将系统划分 5 个级别: 日、周、月、季、年; 按照存证粒度的粗细可划分两个级别: 最终版本存证、历史修改存证。

新闻平台侧也应记录自身系统数据流转过程中的采集、训练、干预等行为和推荐结果数据, 供相关部门或者社会进行监管。按照推荐算法数据流程, 一般分为数据内容采集、模型训练、结果干预、结果推荐 4 个主要步骤。针对该 4 个步骤, 新闻平台均应具备记录存证的能力, 即分别需要记录推荐算法采集的数据源、数据范围、数据类型等内容, 记录算法模型训练日志、训练参数, 记录面向不同受众的历史召回、排序的结果。

5 总结与展望

近些年, 互联网信息数据量急增, 信息过载问题日益严重, 随着人工智能技术迅速发展, 推荐算法尤其是个性化推荐得到了前所未有的发展。作为推荐算法一大应用场景, 新闻推荐不仅改变了内容分发方式, 且便利了用户获取自身需要的新闻内容。但依靠算法实现推荐推送, 用户被动接收新闻推荐的形式, 逐渐引起信息茧房、算法偏见等问题。因此国内外监管部门、研究学者越来越关注如何提高推荐算法可解释性、公平性等。但针对新闻推荐算法的可信评价的研究仍是空白。

本文主要研究新闻推荐算法的可信评价。首先深入分析新闻推荐算法的关键要素, 研究产业界当前应用的主流推荐算法。通过分析新闻推荐算法现存的风险, 梳理了国内外相关监管部门、研究学者及推荐算法一线研发人员, 从政策、标准规范、学术研究等方面在算法公平性、可解释性等方向的研究探索的成果。本文最后面向生产者、受众、算法模型、新闻平台等参与角色, 从公平性、可解释性和抗抵赖性三方面建立了一套新闻推荐算法可信评价指标体系, 分析各指标对新闻推荐算法可信的影响, 并定性或定量分析。本文提出的推荐算法可信评价指标体系填补了推荐算法可信评价研究领域的空白, 同时为新闻推荐算法在可信方向的技术演进提供思路, 为智能算法治理相关工作提供有力参考。

未来研究工作重点将在本文中提出的指标评价体系的基础上, 建立更为严谨的评价指标, 结合推荐算法具体应用场景, 研究可量化的评价方法。此外, 研究高效、准确的自动化推荐算法可信测评技术和工具, 探索推荐算法可信性分析, 验证其在新闻等多领域的有效性, 同时为智能算法的可信研究给予参考。

致 谢 本课题得到中国科学院战略性先导科技专项(C 类)项目(No.XDC02060400)资助。

参考文献

[1] Melville D, Gross Bertrand M. The Managing of Organizations: The Administrative Struggle[J]. *American Sociological Review*, 1965, 30(4): 606.

[2] CNNIC. The 47th China Statistical Report on Internet Development. http://www.cnnic.net.cn/hlwfzyj/hlwxzbg/hlwtjbg/202102/t20210203_71361.htm., Feb. 2021.
(中国互联网络信息中心. 第 47 次中国互联网络发展状况统计报告. 2021-02)

[3] Reinsel David, John Gantz, and John Rydning. Data age 2025: The Digitization of The World from Edge to Core[J]. *Seagate Data Age*, 2018.

[4] Karlgren J. An Algebra for Recommendations: Using Reader Data As a Basis for Measuring Document Proximity[EB/OL]. 1990.

[5] Resnick P, Iacovou N, Suchak M, et al. GroupLens: An Open Architecture for Collaborative Filtering of Netnews[C]. *The 1994 ACM conference on Computer supported cooperative work*, 1994: 175-186.

[6] Balabanović M, Shoham Y. Fab[J]. *Communications of the ACM*, 1997, 40(3): 66-72.

[7] Linden G, Smith B, York J. Amazon.com Recommendations: Item-to-Item Collaborative Filtering[J]. *IEEE Internet Computing*, 2003, 7(1): 76-80.

[8] Funk S. Netflix Update: Try This at Home. https://sifter.org/~simon/journal/20061211.html. Dec.2006.

[9] Salakhutdinov R, Mnih A, Hinton G. Restricted Boltzmann Machines for Collaborative Filtering[C]. *The 24th international conference on Machine learning - ICML '07*, 2007: 791-798.

[10] Mnih A, Salakhutdinov R R. Probabilistic Matrix Factorization[C]. *Advances in neural information processing systems*, 2008: 1257-1264.

[11] Ma H, Yang H X, Lyu M R, et al. SoRec: Social Recommendation Using Probabilistic Matrix Factorization[C]. *The 17th ACM conference on Information and knowledge management*, 2008: 931-940.

[12] Zhao T, McAuley J, King I. Leveraging Social Connections to Improve Personalized Ranking for Collaborative Filtering[C]. *The 23rd ACM International Conference on Conference on Information and Knowledge Management*, 2014: 261-270.

[13] Jamali M, Ester M. *TrustWalker*: A Random Walk Model for Combining Trust-Based and Item-Based Recommendation[C]. *The 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2009: 397-406.

[14] Zhou G R, Zhu X Q, Song C R, et al. Deep Interest Network for Click-through Rate Prediction[C]. *The 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2018: 1059-1068.

[15] Covington P, Adams J, Sargin E. Deep Neural Networks for YouTube Recommendations[C]. *The 10th ACM Conference on Recommender Systems*, 2016: 191-198.

[16] Cyberspace Administration of China. Provisions on Ecological Governance of Network Information Content. http://www.cac.gov.cn/2019-12/20/c_1578375159509309.htm. 2019.
(国家互联网信息办公室. 网络信息内容生态治理规定. 2019)

[17] Wang Q, Xu Y, Zhang X R, et al. Review of User Profile Research Progress[J]. *Modern Computer*, 2020(24): 60-63.
(汪倩, 徐勇, 张心蕊, 等. 用户画像研究进展综述[J]. *现代计算机*, 2020(24): 60-63.)

[18] Mooney R J, Roy L. Content-Based Book Recommending Using Learning for Text Categorization[C]. *The fifth ACM conference on Digital libraries*, 2000: 195-204.

[19] Breese J S, Heckerman D, Kadie C. Empirical Analysis of Predictive Algorithm for Collaborative Filtering[EB/OL]. 2013.

[20] Burke R. Knowledge-based Recommender Systems[J]. *Encyclopedia of Library and Information Systems,* 2000, 69(Supplement 32): 175-186.

[21] Guo H F, Tang R M, Ye Y M, et al. DeepFM: A Factorization-Machine Based Neural Network for CTR Prediction[C]. *The Twenty-Sixth International Joint Conference on Artificial Intelligence*, 2017: 1725-1731.

[22] Gong Y, Zhang Q. Hashtag Recommendation Using Attention-Based Convolutional Neural Network[C]. *The 25th International Joint Conference on Artificial Intelligence*, 2016: 2782-2788.

[23] He X N, Liao L Z, Zhang H W, et al. Neural Collaborative Filtering[C]. *The 26th International Conference on World Wide Web*, 2017: 173-182.

[24] Lian J X, Zhang F Z, Xie X, et al. CCCFNet: A Content-Boosted Collaborative Filtering Neural Network for Cross Domain Recommender Systems[C]. *The 26th International Conference on World Wide Web Companion*, 2017: 817-818.

[25] Alashkar T, Jiang S, Wang S, et al. Examples-rules Guided Deep Neural Network for Makeup Recommendation[C]. *The AAAI Conference on Artificial Intelligence*, 2017, 31(1).

[26] Wu C Y, Ahmed A, Beutel A, et al. Recurrent Recommender Networks[C]. *The Tenth ACM International Conference on Web Search and Data Mining*, 2017: 495-503.

[27] Huang P S, He X D, Gao J F, et al. Learning Deep Structured Semantic Models for Web Search Using Clickthrough Data[C]. *The 22nd ACM international conference on Information & Knowledge Management*, 2013: 2333-2338.

[28] Zhang S, Yao L N, Sun A X, et al. Deep Learning Based Recommender System[J]. *ACM Computing Surveys*, 2019, 52(1): 1-38.

[29] Institute of Security, China Academy of Information and Communications Technology. Artificial Intelligence Security White Paper (2018). 2018.
(中国信息通信研究院安全研究所, 人工智能安全白皮书(2018), 2018)

[30] Council A C M U S P P. Statement on Algorithmic Transparency and Accountability[J]. *Commun, ACM*, 2017.

[31] The European Commission's High-level Expert Group on Artificial Intelligence. Draft Ethics Guidelines for Trustworthy AI[S]. 2018.

[32] The State Council of the People's Republic of China. Notice of the State Council on Issuing the Development Plan on the New Generation of Artificial Intelligence. http://www.gov.cn/zhengce/ content/2017-07/20/content_5211996.htm. 2017.
(中华人民共和国国务院. 新一代人工智能发展规划. 2017)

[33] Standardization administration, Cyberspace Administration of China, National Development and Reform Commission, Ministry of Science and Technology of the People's Republic of China, Ministry of Industry and Information Technology. Guidelines for the construction of a new generation of national artificial intelligence standard system. http://www.gov.cn/zhengce/zhengceku/2020-08/09/content_5533454.htm. 2020.
(国家标准化管理委员会, 国家互联网信息办公室, 国家发展和改革委员会, 科学技术部 工业和信息化部. 国家新一代人工智能标准体系建设指南. 2020)

[34] Miller T. Explanation In Artificial Intelligence: Insights from the Social Sciences[J]. *Artificial Intelligence*, 2019, 267: 1-38.

[35] Zhang Y F, Lai G K, Zhang M, et al. Explicit Factor Models for Explainable Recommendation Based on Phrase-Level Sentiment Analysis[C]. *The 37th international ACM SIGIR conference on Research & development in information retrieval*, 2014: 83-92.

[36] Zhang Y F, Chen X. Explainable Recommendation: A Survey and New Perspectives[J]. *Foundations and Trends® in Information Retrieval*, 2020, 14(1): 1-101.

[37] Saltelli A. Sensitivity Analysis for Importance Assessment[J]. *Risk Analysis*, 2002, 22(3): 579-590.

[38] Ancona M, Ceolini E, Öztireli C, et al. Towards Better Understanding of Gradient-Based Attribution Methods for Deep Neural Networks[EB/OL]. 2018.

[39] Molnar C. Interpretable Machine Learning[M]. Lulu.com, 2020.

[40] Park W, Kim D, Lu Y, et al. Relational Knowledge Distillation[C]. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019: 3962-3971.

[41] Ghazimatin A, Pramanik S, Roy R S, et al. ELIXIR: Learning from User Feedback on Explanations to Improve Recommender Mod-

els[C]. *The Web Conference 2021*, 2021: 3850-3860.

[42] Balog K, Radlinski F, Karatzoglou A. On Interpretation and Measurement of Soft Attributes for Recommendation[C]. *The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2021: 890-899.

[43] Zhan H L, Zhang H N, Chen H S, et al. User-Inspired Posterior Network for Recommendation Reason Generation[C]. *The 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2020: 1937-1940.

[44] Ma W Z, Zhang M, Cao Y, et al. Jointly Learning Explainable Rules for Recommendation with Knowledge Graph[C]. *WWW '19*: *The World Wide Web Conference*, 2019: 1210-1221.

[45] Tran K H, Ghazimatin A, Roy R S. Counterfactual Explanations for Neural Recommenders[C]. *The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2021: 1627-1631.

[46] China Academy of Information and Communications Technology. Artificial Intelligence Security Framework (2020). 2020. (中国信息通信研究院. 人工智能安全框架(2020). 2020)

[47] Burke R. Multisided Fairness for Recommendation[EB/OL]. 2017.

[48] Abdollahpouri H, Burke R. Multi-Stakeholder Recommendation and Its Connection to Multi-Sided Fairness[EB/OL]. 2019: arXiv: 1907.13158[cs.IR]. https://arxiv.org/abs/1907.13158.

[49] Burke R, Sonboli N, Ordonez-Gauger A. Balanced Neighborhoods for Multi-Sided Fairness in Recommendation[C]. *Conference on Fairness, Accountability and Transparency* PMLR, 2018: 202-214.

[50] Lu H Y, Zhang M, Ma S P. Between Clicks and Satisfaction: Study on Multi-Phase User Preferences and Satisfaction for Online News Reading[C]. *SIGIR '18*: *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, 2018: 435-444.

[51] Beutel A, Chen J L, Doshi T, et al. Fairness In Recommendation Ranking through Pairwise Comparisons[C]. *The 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2019: 2212-2220.

[52] Sacharidis D. Top-N Group Recommendations with Fairness[C]. *The 34th ACM/SIGAPP Symposium on Applied Computing*, 2019: 1663-1670.

[53] Kaya M, Bridge D, Tintarev N. Ensuring Fairness In Group Recommendations by Rank-Sensitive Balancing of Relevance[C]. *RecSys '20*: *Fourteenth ACM Conference on Recommender Systems*, 2020: 101-110.

[54] Li Y Q, Chen H X, Fu Z H, et al. User-Oriented Fairness In Recommendation[C]. *The Web Conference 2021*, 2021: 624-632.

[55] Abdollahpouri H, Mansoury M, Burke R, et al. The Unfairness of Popularity Bias In Recommendation[EB/OL]. 2019: arXiv: 1907.13286[cs.IR]. https://arxiv.org/abs/1907.13286.

[56] Collins A, Tkaczyk D, Aizawa A, et al. Position Bias in Recommender Systems for Digital Libraries[C]. *International Conference on Information*, 2018: 335–344.

[57] Khenissi S. Modeling and Counteracting Exposure Bias In Recommender Systems.[D]. University of Louisville, 2019. DOI:10.18297/etd/3182.

[58] Chen J W, Dong H D, Wang X, et al. Bias and Debias In Recommender System: A Survey and Future Directions[EB/OL]. 2020.

[59] Wu Y, Cao J, Xu G D, et al. TFROM: A Two-Sided Fairness-Aware Recommendation Model for both Customers and Providers[C]. *The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2021: 1013-1022.

[60] Patro G K, Biswas A, Ganguly N, et al. FairRec: Two-Sided Fairness for Personalized Recommendations In Two-Sided Platforms[C]. *WWW '20*: *The Web Conference 2020*, 2020: 1194-1204.

[61] Cyberspace Administration of China. List of sources of Internet news information releases. http://www.cac.gov.cn/2016-08/08/c_1119356489.htm, Aug.2016(国家互联网信息办公室. 互联网新闻信息稿源单位名单. 2016-08)

刘总真 于 2015 年在中国科学院大学计算机技术专业获得硕士学位。现在中国科学院大学计算机软件与理论专业攻读博士学位, 现任中国科学院信息工程研究所助理研究员。研究领域为网络空间安全。研究兴趣包括: 互联网数据分析、智能算法安全等。Email: liuzongzhen@iie.ac.cn

张潇丹 于 2012 年在中国科学院大学计算机系统结构专业获得博士学位, 现任中国科学院信息工程研究所副研究员。研究领域为新型网络技术测量分析与评估。研究兴趣包括: 区块链、互联网数据分析、计算传播等。Email: zhangxiaodan@iie.ac.cn

郭涛 中国科学院信息工程研究所博士生导师, 研究员, 工学博士。研究领域为网络空间安全。研究兴趣包括: 网络空间安全、漏洞分析与风险评估。Email: guotao@iie.ac.cn

周熙 于 2017 年在北京师范大学通信与信息系统专业获得硕士学位, 现为中国科学院信息工程研究所助理研究员, 研究领域为网络安全, 研究兴趣包括: 区块链、网络传播等。Email: zhouxi@iie.ac.cn

陈家玓 于 2015 年在北京邮电大学通信与信息系统专业获得博士学位。现在中国科学院信息工程研究所任助理研究员。研究领域为新型互联网架构、区块链监管与评测、区块链技术应用等。Email: chenjiadi@iie.ac.cn

林俊宇 于 2014 年在哈尔滨工程大学计算机应用技术专业获得博士学位, 现任中国科学院信息工程研究所高级工程师。研究领域为网络安全。研究兴趣包括: 未来网络, 信息安全, 人工智能等。Email: linjunyu@iie.ac.cn

葛敬国 于 2003 年获中国科学院计算技术研究所计算机系统结构专业获得博士学位。现任中国科学院信息工程研究所研究员, 中国科学院网络安全学院教授。研究领域是计算机网络架构, 软件定义网络, 网络虚拟化, 网络安全和移动通信网络。Email: gejingguo@iie.ac.cn

王宇航 于 2016 年在北京交通大学软件工程专业获得硕士学位, 现在中国科学院大学网络空间安全专业攻读博士学位, 现任中国科学院信息工程研究所助理研究员。研究领域为网络空间安全。研究兴趣包括: 信息内容安全、区块链信息服务安全、机器学习。Email: wangyuhang@iie.ac.cn

吕红蕾 于 2007 年在中国科学院计算机网络信息中心计算机应用技术专业获得工学硕士学位, 现任中国科学院信息工程研究所高级工程师、硕士生导师。研究领域包括互联网舆情分析和传播引导、自然语言处理、人工智能等。Email: lvhonglei@iie.ac.cn