软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2022,33(9):3453−3469 [doi: 10.13328/j.cnki.jos.006364]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

社交网络中负责隐私协商的智能体行为追责*

古天龙[1,2], 郝峰锐[2], 李 龙[1,2], 李晶晶[1], 常 亮[2]

[1](暨南大学 信息科学技术学院/网络空间安全学院, 广东 广州 510632)

[2](广西可信软件重点实验室 (桂林电子科技大学), 广西 桂林 541004)

通信作者: 李龙, E-mail: lilong@guet.edu.cn

摘 要: 隐私协商可以协助社交网络用户在信息分享前建立隐私保护共识, 具有一定的隐私泄露的预先防护作用. 可追责是行为或后果的责任主体可以被追究的属性, 是透明、可解释人工智能应用的一个重要方面. 社交网络中隐私协商过程的可追责, 对于提升应用平台或系统的透明、可解释性具有重要的意义. Kekulluoglu 等人提出了基于智能体的互惠隐私协商体系, 但尚缺乏针对智能体行为的追责研究. 以此为基础设计实现了用于社交网络隐私协商、具有定性追责和定量追责的智能体行为追责系统, 并提出了追责要求及实现追责的行为指标, 其中, 定性追责方法可以准确判断隐私协商智能体是否存在不当行为并能够精准锁定不当行为具体发生位置; 定量追责包含简单量化、加权马氏距离和改进 Minhash 这 3 种方法, 能够量化智能体不当行为的严重程度. 实验数据表明了所提出系统及方法的有效性和合理性.

关键词: 社交网络; 隐私保护; 隐私协商; 智能体; 追责

中图法分类号: TP306

中文引用格式: 古天龙, 郝峰锐, 李龙, 李晶晶, 常亮. 社交网络中负责隐私协商的智能体行为追责. 软件学报, 2022, 33(9): 3453–3469. http://www.jos.org.cn/1000-9825/6364.htm

英文引用格式: Gu TL, Hao FR, Li L, Li JJ, Chang L. Behavior Accountability of Agents Responsible for Privacy Negotiation in Social Networks. Ruan Jian Xue Bao/Journal of Software, 2022, 33(9): 3453–3469 (in Chinese). http://www.jos.org.cn/1000-9825/6364.htm

Behavior Accountability of Agents Responsible for Privacy Negotiation in Social Networks

GU Tian-Long[1,2], HAO Feng-Rui[2], LI Long[1,2], LI Jing-Jing[1], CHANG Liang[2]

[1](College of Information Science and Technology/College of Cyber Security, Jinan University, Guangzhou 510632, China)

[2](Guangxi Key Laboratory of Trusted Software (Guilin University of Electronic Technology), Guilin 541004, China)

Abstract: Privacy negotiation performs a pre-protective role against privacy disclosure as it can assist social network users to build a consensus on privacy protection before information sharing. Accountability is an attribute that a subject is responsible for an action or consequence, and it is an important aspect of transparent and explainable artificial intelligence applications. Accountability in the privacy negotiation process in social networks is of great significance for improving the transparency and explainability of application platforms or systems. Although Kekulluoglu *et al*. proposed an agent-based reciprocal privacy negotiation system, the accountability for the behaviors of agents was not discussed. For this reason, a novel system for agent behavior accountability during privacy negotiation in social networks is designed and implemented, and qualitative and quantitative accountability methods are developed. Moreover, requirements and behavior indicators are also proposed to achieve accountability. Specifically, the qualitative accountability method can accurately determine whether a privacy negotiation agent has misbehavior and pinpoint the specific location of the misbehavior. The quantitative accountability methods include simple quantification, weighted Mahalanobis distance, and improved Minhash and can quantify the severity of the agent's misbehavior. The experimental data demonstrate the validity and rationality of the proposed system and methods.

Key words: social networks; privacy protection; privacy negotiation; agents; accountability

* 基金项目: 国家自然科学基金 (U1711263, U1811264, 61966009, 61961007, 61862016); 广西自然科学基金 (2019GXNSFBA245049, 2018GXNSFDA281045)

收稿时间: 2020-12-18; 修改时间: 2021-02-05, 2021-03-14, 2021-04-13; 采用时间: 2021-04-29; jos 在线出版时间: 2022-07-15

快速发展的社交网络一定程度上改变了人们之间传统的沟通交流方式, 促进了信息分享的广泛发生, 但也带来了严重的隐私泄露问题. 如 2019 年 12 月发生的 Facebook 泄漏事件, 导致 2.67 亿名用户的用户名、手机号等信息泄露 (https://www.bankinfosecurity.com/267-million-facebook-user-records-for-sale-on-dark-net-a-14158); 2020 年 3 月, 5.38 亿条微博用户数据在暗网出售, 其中 1.72 亿条数据中包含用户名、用户性别、已发布微博数、粉丝数、地理位置等信息 (https://securityboulevard.com/2020/03/stolen-data-of-538-million-weibo-users-for-sale-on-the-dark-market/). 此类事件的发生彰显了隐私保护的重要性, 使得社交网络中的隐私保护技术研究受到了广泛关注.

除去传统的 *k*-匿名、*l*-多样化、*t*-接近性等隐私保护技术[1], 基于对抗生成网络[2]、神经网络[3]等人工智能技术的隐私保护方法日益兴起. 此类方法将人工智能技术与隐私保护机制相结合, 解决模型不公平[4]、隐私侵犯[5]、模型安全[6]等问题, 达到高效保护用户分享信息 (如照片)[7]、位置信息[8]等目的. 如 Kekulluoglu 等人[9]基于描述逻辑建立了基于智能体的互惠隐私协商体系, 以协助社交网络用户在信息分享前达成隐私保护共识, 起到提前避免隐私泄露的效果.

在借助智能体进行隐私协商的同时, 能够对智能体的行为进行追责[10,11], 对于提升系统的透明[12]、可解释性[13]具有重要的意义. 但目前关于追责的研究主要聚焦于电子投票[14]、安全协议[15]等方面, 较少关注智能体领域[16,17]. 因此, 本文基于 Kekulluoglu 等人[9]的研究工作, 针对社交网络中负责隐私协商工作的智能体 (以下简称隐私协商智能体), 设计并实现了智能体追责系统, 在保护社交网络中隐私信息的同时, 借助于定性追责和定量追责两种方案实现对隐私协商智能体行为的追责. 其中, 定性追责是将智能体隐私协商过程的真实数据与隐私保护的预定阈值进行对比, 对智能体是否存在不当行为进行判定; 定量追责分别借助于简单量化方法、加权马氏距离方法和改进 Minhash 方法, 对智能体的不当行为进行量化.

本文的主要贡献有以下几点: 1) 基于 Kekulluoglu 等人提出的互惠隐私协商体系, 设计并实现了隐私协商智能体行为追责系统, 能够提高隐私协商系统的透明性、安全性和可信性. 2) 提出了定性追责方法, 可以准确判断隐私协商智能体是否存在不当行为并在存在不当行为时精准锁定其具体发生位置, 通过实验证明了定性追责方法的有效性. 3) 提出了简单量化、加权马氏距离、改进 Minhash 共 3 种定量追责方法, 能够求得隐私智能体的责任量化值, 从而对不当行为的严重程度进行量化; 实验证明量化结果准确、可信, 即定量追责方法是有效的.

本文第 1 节介绍近几年隐私保护领域, 尤其是追责问题方面的相关工作. 第 2 节介绍追责系统框架及其功能, 并提出追责要求及行为指标. 第 3 节介绍智能体行为的定性追责过程及算法. 第 4 节介绍定量追责过程及 3 种定量追责方法. 第 5 节给出了实验验证及结果分析. 第 6 节总结全文, 并对未来值得关注的研究方向进行初步探究.

1 相关工作

照片分享时的隐私保护是社交网络隐私保护研究关注的主要问题之一. 主要研究有: Amon 等人[18]从用户行为角度出发, 提出了针对用户照片分享行为的干预方法. 通过调查影响用户在社交媒体上进行照片分享的因素, 以行为干预的方式协助用户避免隐私泄露. 由于未考虑到的影响因素较多, 该方法实现效果还有较大的提升空间. Pensa 等人[19]为了协助用户主动进行隐私保护, 提出了一种针对社交网络隐私保护的自我评估框架, 能够测量用户隐私泄露风险、对已发生的隐私泄露进行告警以及半自动地设置隐私保护策略. 该框架功能全面, 具有一定的实际应用价值, 但是隐私策略的设置粒度较粗, 尚需细化研究. Hasan 等人[7]从照片内容识别的角度, 提出了用于识别照片主体及旁观者 (如, 无关路人) 的图像检测技术, 以达到保护旁观者隐私的目的. 该技术侧重于训练用于照片人物识别的机器学习模型, 可用于保护图片形式数据中的隐私, 但无法保护文本、音频等其他类型数据中的隐私. Yang 等人[8]针对照片中的位置隐私保护, 首先借助于数据聚合证明了部分现有位置隐私保护方案仍然存在隐私泄露风险, 继而基于混合整数线性规划提出了一种效果较为显著的位置隐私保护方案.

基于智能体的隐私协商是近期较受关注的隐私保护方法. 如: Such 等人[20]提出了首个隐私策略冲突自动检测方法, 并基于智能体提出了用于缓解冲突的协商机制. 该方案能够在一定程度上解决隐私泄露问题, 但智能体间的协商机制尚有待扩展, 如纳入所代表用户间的社交关系. Kökciyan 等人[21]开发了名为 PriGuard 的隐私保护系统,

该系统使用描述逻辑对基于智能体的复杂社交网络场景进行了刻画, 并通过建立智能体与系统间的承诺管理用户的隐私要求. PriGuard 可以检测及推理社交网络中多种形式的隐私侵犯行为, 但未涉及智能体间的隐私协商机制. Kekulluoglu 等人[9]基于语义隐私规则与效用函数提出了一种智能体间的互惠隐私协商体系. 该体系使用积分(point, 可理解为网络论坛中积分、点券等形式的报酬、奖赏, 所有智能体拥有相同的初始积分) 表示用户愿意承担的社会责任, 并以此为基础提出了 RGEP (reciprocal good-enough-privacy) 协商策略. 智能体借助 RGEP 协商策略便可代理用户完成隐私协商. 虽然以上基于智能体的隐私保护/协商方法具有一定的成效, 但是此类方法默认智能体是完全诚实可信的, 这与实际情形可能存在出入, 如智能体存在信息隐瞒等不当行为, 因此针对智能体 (及用户) 的追责研究也受到了关注, 通过智能体的追责方案, 检测并约束智能体行为, 进一步提高智能体的安全性.

关于追责的研究主要集中在电子投票、大数据交易等领域, 如: Baldoni 等人[22]2016 年研究了智能体交互系统中的责任衡量及计算问题, 并于 2017 年对该工作进行了进一步扩展[23], 提出了针对多智能体系统的追责协议. 此两研究均为理论研究, 未针对具体场景进行实用性分析. Küsters 等人[14]首先提出了符号追责、计算追责的定义, 通过揭示追责与可验证性间的密切关系, 进一步提出了符号验证、计算验证的定义. 通过将以上定义应用到电子投票中, 验证电子投票协议的可用性, 并针对不可用的协议提出了修复方案. 参照 Küsters 等人的研究工作, Jung 等人[15]提出了针对大数据交易的 AccountTrade 体系, 以完成对不诚实用户的追责. AccountTrade 对符号追责和计算追责的定义进行了优化, 并进一步提出了多个支持追责的大数据交易协议, 以检测/追责大数据交易过程中用户的不当行为. 这类研究对象主要关注的是用户, 采用追责协议实现对用户的追责, 而针对智能体的研究还处于理论研究阶段, 因此还需要关注智能体追责方法的实现.

本文受到以上研究工作的启发, 针对社交网络中的隐私保护及追责问题进行了研究, 提出了针对社交网络中隐私协商智能体的行为追责框架及方法. 本文方法与文献 [14,15] 中所提方法在应用领域、追责/隐私保护功能及应用对象等方面存在一定区别, 如表 1 所示.

表 1 方案比较

方案	应用领域	追责功能	隐私保护	应用对象
PriGuard[21]	社交网络	无	有	智能体
RGEP[9]	社交网络	无	有	智能体
Accountability[14]	电子投票	有	无	用户
AccountTrade[15]	大数据交易	有	无	用户
本文	社交网络	有	有	智能体

2 追责系统模型及指标

2.1 隐私协商智能体行为追责系统

本文使用基于 RGEP 协商策略的协商架构[9]作为智能体追责的基础, 构建了图 1 所示的隐私协商智能体行为追责系统. 该系统中主要包含发起者智能体 (initiator agent)、协商者智能体 (negotiator agent)、追责服务器 (accountability server) 及观众 (audience) 这 4 类实体. 其中, 发起者智能体作为信息所有者 (即用户 A) 的代理, 发起信息分享; 协商者智能体作为可能遭受隐私泄露者 (即用户 B) 的代理, 与发起者智能体进行隐私协商; 追责服务器作为仲裁机构负责对智能体的不当行为进行追责; 观众为分享信息的接收者. 在本系统中, 发起者智能体和协商者智能体均可能存在不当行为, 如将信息分享给非观众列表中的观众、隐藏隐私协商过程信息、支付虚假积分、虚报协商要求等, 因此均被视为是不可信的; 追责服务器是完全可信的; 观众不存在追责问题.

社交网络智能体隐私追责系统主要实现隐私协商、行为追责两大功能.

1) 隐私协商. 其实现过程如下: ① 发起者智能体将包含隐私信息的信息分享请求 (request) 发送给协商者智能体, 从而发起隐私协商. ② 协商者智能体接收到信息分享请求后, 进行效用值计算, 并给出同意/拒绝该请求的回应. 若协商者智能体给出拒绝回应, 需同时给出拒绝理由, 包括引起隐私损害的观众列表及所期望得到的积分.

③ 发起者智能体根据拒绝理由更新信息分享请求并再次发起隐私协商, 直至完成协商或因双方无法调和的需求而终止协商. 更新信息分享请求的方式为: 如果所持有积分无法满足协商者智能体所期望积分, 则按照隐私损害程度递减删除观众. ④ 协商完成或终止后, 协商过程数据被存储到追责服务器.

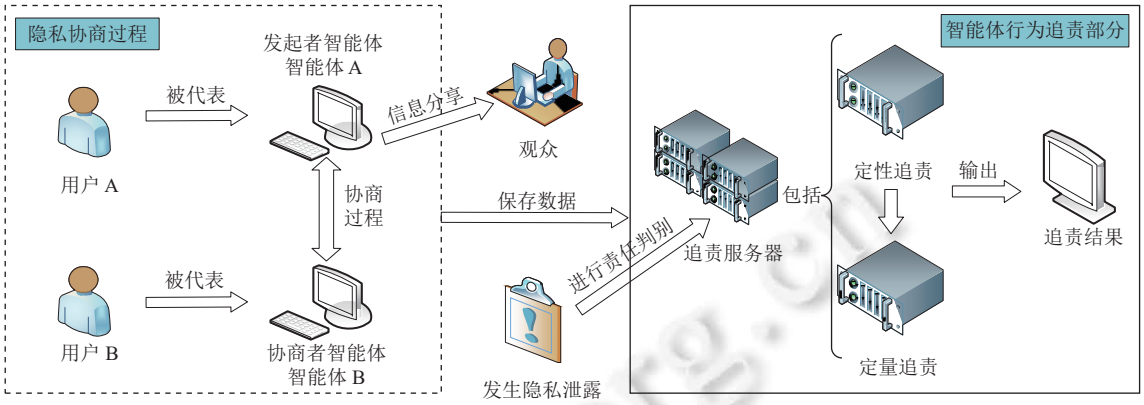


图 1 隐私协商智能体行为追责系统框架

2) 行为追责. 本研究中对追责的定义是, 在发生隐私泄露后对隐私协商智能体进行行为分析, 确定存在不当行为的智能体并进行追责. 如图 2 所示, 发生隐私泄露后, 可以依次进行定性追责、定量追责. 定性追责过程中, 确定隐私协商智能体, 并从追责服务器中提取协商过程数据后, 依次完成以下步骤: ① 正向模拟协商过程, 即模拟智能体完全诚实可信情况下的协商过程, 获得所需协商过程数据并将其设置为正向阈值; ② 逆向复现协商过程, 即从协商结果出发, 逆向计算各个智能体应有的行为数据并将其设置为逆向阈值; ③ 借助正向阈值及逆向阈值判断智能体是否存在不当行为并锁定不当行为发生的具体位置. 定性追责结束后, 可进一步进行定量追责: ① 简单量化方法, 用于验证定性追责结果是否准确、有效; ② 加权马氏距离方法, 基于协商过程中数值型数据进行责任量化; ③ 改进 Minhash 方法, 基于协商过程中文本型及混合型数据 (数值型数据、文本型数据并存) 进行责任量化.

2.2 追责要求及行为指标

文献 [14,15] 对电子投票及大数据交易中的安全协议追责模型进行了研究, 本文由此得到启发, 提出了针对社交网络隐私协商智能体进行行为追责的要求:

1) 公平性: 无不当行为的实体不会受到责备.

2) 完整性: 如果发生隐私泄露, 能够对整个协商过程进行分析, 并对所有执行过不当行为的智能体追责.

3) 具体性: 针对执行过不当行为的智能体, 能够查找出所有不当行为及发生位置.

4) 透明性: 追责过程及结果是完全透明的.

追责过程针对发起者智能体和协商者智能体设置了不完全相同的行为指标.

针对发起者智能体设置的行为指标如下:

1) 观众列表 (*Au*): 可以接收到分享信息的用户名单, 具体包含观众用户数量 Au_1、观众用户姓名集合 Au_2, 均为正向阈值指标.

2) 积分 (*Pin*): 发起者智能体为完成协商愿意支付的积分, 具体包含正向阈值指标 Pin_1 及逆向阈值指标 Pin_2、Pin_3.

3) 效用值 (*Uin*): 发起者智能体对自身利益满足程度的度量, 具体包含正向阈值指标 Uin_1 及逆向阈值指标 Uin_2、Uin_3.

4) 协商结果 (*Cr*): 协商者智能体对信息分享请求的回应.

为后续使用方便, 将以上指标归类为: ① 正向阈值指标集合 $INDEX_{in}^{+}=\{Au_1, Au_2, Pin_1, Uin_1\}$, ② 逆向阈值指标集合 $INDEX_{in}^{-}=\{Pin_2, Pin_3, Uin_2, Uin_3\}$.

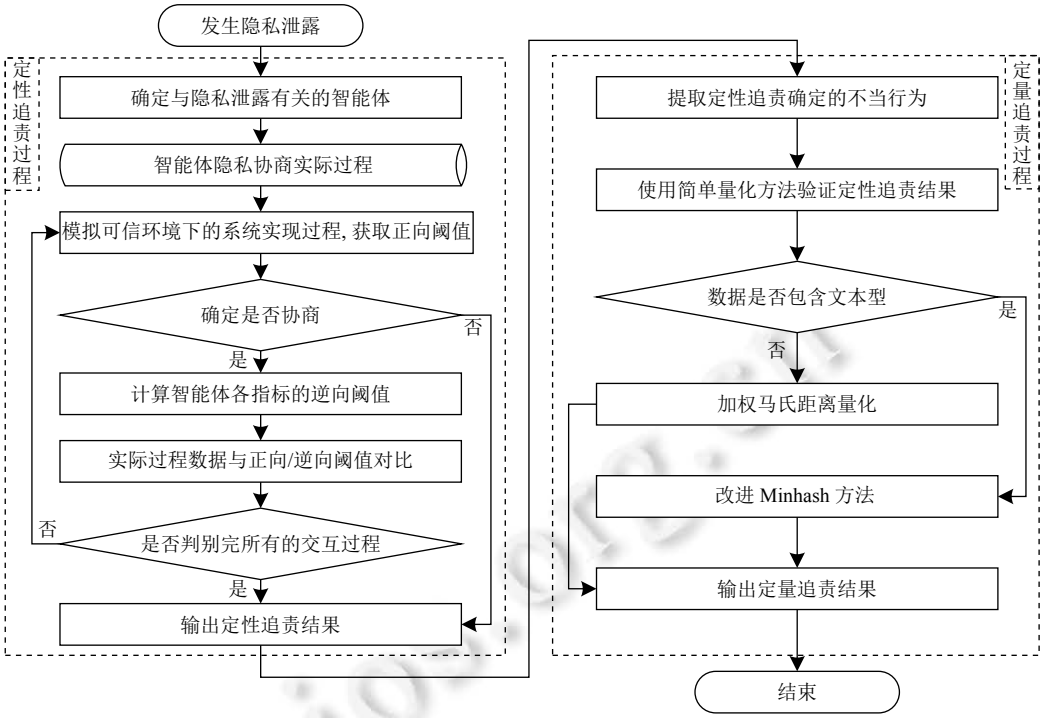


图 2 追责流程

针对协商者智能体设置的行为指标如下:

1) 敏感观众列表 (*Vng*): 协商者智能体根据隐私规则得出的可能会对其隐私造成损害的观众名单. 具体包含敏感观众数量 Vng_1、敏感观众姓名集合 Vng_2, 均为正向阈值指标.

2) 权重 (*W*): 敏感观众对协商者智能体的隐私造成损害的程度, 具体包含权重集合 W_1、按照权重大小排序后的敏感观众姓名集合 W_2、敏感观众数量集合 W_3, 均为正向阈值指标.

3) 效用值 (*Ung*): 协商者智能体对自身利益满足程度的度量, 具体包含正向阈值指标 Ung_1 及逆向阈值指标 Ung_2、Ung_3.

4) 积分 (*Png*): 协商者智能体同意信息分享请求时期望得到的积分, 具体包含正向阈值指标 Png_1 及逆向阈值指标 Png_2.

5) 协商结果 (*Cr*): 协商者智能体对信息分享请求的回应, 具体包含正向阈值指标 Cr_1 及逆向阈值指标 Cr_2.

为后续使用方便, 将以上指标归类为: ①正向阈值指标集合 $INDEX_{ng}^{+}=\{Vng_1, Vng_2, W_1, W_2, W_3, Ung_1, Png_1, Cr_1\}$, ②逆向阈值指标集合 $INDEX_{ng}^{-}=\{Ung_2, Ung_3, Png_2, Cr_2\}$.

3 定性追责

针对发起者/协商者智能体进行定性追责的具体过程如下: 首先, 提取发起者/协商者智能体在协商过程中的行为数据; 其次, 通过模拟智能体完全诚实可信情况下的协商过程获取正向阈值, 并根据阈值函数计算逆向阈值; 最后, 通过实际行为数据与正向阈值/逆向阈值间的对比, 判定是否存在不当行为, 并在存在不当行为时给出详细信息.

假设协商过程为 *N*, 针对发起者智能体 *in* 设立正向阈值集合 $C_{in\_1}=\{T_1^{in,i},\ldots,T_{|INDEX_{in}^{+}|}^{in,i}\}$、逆向阈值集合 $C_{in\_2}=\{t_1^{in,i},\ldots,t_{|INDEX_{in}^{-}|}^{in,i}\}$; 针对协商者智能体 *ng* 设立正向阈值集合 $C_{ng\_1}=\{T_1^{ng,i},\ldots,T_{|INDEX_{ng}^{+}|}^{ng,i}\}$、逆向阈值集合 $C_{ng\_2}=\{t_1^{ng,i},\ldots,t_{|INDEX_{ng}^{-}|}^{ng,i}\}$. 在以上各式中, *i* 为协商过程中的第 *i* 次交互, *n* 表示总的交互次数, |*|表示集合*中的指

标总数, T、t 分别表示正向阈值、逆向阈值.

3.1 针对发起者智能体的定性追责

假设 D_p 是实际协商过程中的行为数据集合, R_p 表示协商者智能体对信息分享请求的最终回应 (Y 或 N). 追责服务器对发起者智能体在协商过程中是否存在不当行为的判定公式如公式 (1) 所示:

$$Di_{in}\left(D_p\right)=\begin{cases}\text{No responsibility, if } \forall index_{in}^{+}\in INDEX_{in}^{+}, T_{index_{in}^{+}}^{in,i}=D_p \text{ and } \forall index_{in}^{-}\in INDEX_{in}^{-}, t_{index_{in}^{-}}^{in,i}=D_p \text{ and } R_p=\text{Y}\\ \text{Responsible and misbehavior, if } \exists index_{in}^{+}\in INDEX_{in}^{+}, T_{index_{in}^{+}}^{in,i}\neq D_p \text{ or } \exists index_{in}^{-}\in INDEX_{in}^{-}, t_{index_{in}^{-}}^{in,i}\neq D_p\\ \text{Continue distinguish, otherwise}\end{cases} \tag{1}$$

正向/逆向阈值是完成定性追责的关键因素, 各行为指标对应多个阈值.

1) 观众列表指标包含 2 个正向阈值: 第 i 次交互过程中的观众数量阈值 $T_{Au_1}^{in,i}$ 和观众姓名集合阈值 $T_{Au_2}^{in,i}$.

2) 积分指标包含 1 个正向阈值及 2 个逆向阈值: $T_{Pin_1}^{in,i}$ 为第 i 次交互过程中发起者智能体拟支付的积分, $t_{Pin_2}^{in,i}$ 为发起者智能体应支付的积分, $t_{Pin_3}^{in,i}$ 为协商者智能体持有积分的增加值.

$$t_{Pin_2}^{in,i}=\frac{\left[\left(u_{p_i}^{ng}\right)'-\left(u_{\max}-\frac{\sum_{k=1}^{K}u_{r_k}}{w_{\max}\times v_r}\right)\right]\times P_0}{w_P^{ng}} \tag{2}$$

$$u_{r_k}=w_{r_k}\times v_{r_k} \tag{3}$$

$$t_{pin_3}^{in,n}=\left|P_n^{ng}-P_0\right| \tag{4}$$

公式 (2) 参照文献 [9] 中效用函数得来, 其中 $\left(u_{p_i}^{ng}\right)'$ 为协商者智能体在第 i 次交互中关于信息分享请求的效用值, u_{max} 是协商过程中最大效用 (设置为 1), $w_{\max}$ 是敏感观众的最大权重 (设置为 10), v_r 是协商者智能体首次指出的敏感观众总数, K 为协商者智能体隐私规则的总数. u_{r_k} 为隐私规则 r_k 的效用值, 由公式 (3) 求得, 其中 w_{r_k} 表示取值区间为 [1, 10] 的权重, v_{r_k} 表示违反该隐私规则的敏感观众数. P_0 和 w_P^{ng} 分别代表智能体最初持有的积分及其权重, 分别设置为 5 和 0.5. 公式 (4) 中, P_n^{ng} 表示协商结束后协商者智能体持有的积分.

3) 效用指标包含 1 个正向阈值及 2 个逆向阈值: $T_{Uin_1}^{in,i}$ 为第 i 次交互发起者智能体关于信息分享请求的效用值; $t_{Uin_2}^{in,i}$ 表示发起者智能体根据修改后的分享请求计算出的效用值; $t_{Uin_3}^{in,i}$ 表示发起者智能体根据协商者智能体的回应计算出的效用值.

$$t_{Uin_2}^{in,i}=\begin{cases}t^{in}+\dfrac{P_i^{in}\times w_p^{in}}{P_0}, & u_{p_i'}^{in}\geqslant t^{in}\\ t^{in}-\dfrac{P_i^{in}\times w_p^{in}}{P_0}, & u_{p_i'}^{in}<t^{in}\end{cases} \tag{5}$$

$$t_{Uin_3}^{in,i}=u_{p_{i-1}}^{in}+\left[\frac{w_p^{in}\times\left(P_{i-1}^{in}-P_i^{ng}\right)}{P_0}\right] \tag{6}$$

公式 (5) 中, t^{in} 为发起者智能体效用值的最低要求 (设置为 0.7), P_i^{in} 为第 i 次交互过程中发起者智能体拟支付的积分, w_p^{in} 为积分权重 (设置为 0.5), $u_{p_i}^{in}$ 为第 i 次交互过程中发起者智能体根据修改后的分享请求计算出的效用值.

4) 协商结果指标: 通过分析发起者智能体在第 i 次交互过程中的行为 (如, 修改信息分享请求), 并与协商者智能体在 $i-1$ 次交互中的回应进行比较, 判定发起者智能体是否存在不当行为.

3.2 针对协商者智能体的定性追责

假设 D_p 是实际协商过程中的行为数据集合, $\{C_{ng\_1}, C_{ng\_2}\}$ 为针对协商者智能体 ng 设置的追责阈值集合. 追

责服务器对发起者智能体在协商过程中是否存在不当行为的判定公式如式 (7) 所示.

$$Di_{ng}\left(D_p\right)=\begin{cases}\text{No responsibility, if } \forall index_{ng}^{+}\in INDEX_{ng}^{+}, T_{index_{ng}^{+}}^{ng,i}=D_p \text{ and } \forall index_{ng}^{-}\in INDEX_{ng}^{-}, t_{index_{ng}^{-}}^{ng,i}=D_p\\ \text{Responsible and misbehavior, if } \exists index_{ng}^{+}\in INDEX_{ng}^{+}, T_{index_{ng}^{+}}^{ng,i}\neq D_p \text{ or } \exists index_{ng}^{-}\in INDEX_{ng}^{-}, t_{index_{ng}^{-}}^{ng,i}\neq D_p\\ \text{Continue distinguish, otherwise}\end{cases} \tag{7}$$

各行为指标对应多个阈值:

1) 敏感观众列表指标包含 2 个正向阈值: 第 i 次交互过程中的敏感观众数量阈值 $T_{Vng_1}^{ng,i}$ 和敏感观众姓名集合阈值 $T_{Vng_2}^{ng,i}$.

2) 权重指标包含 3 个正向阈值: $T_{W_1}^{ng,i}$ 为权重集合, 表示不同敏感观众对协商者智能体的隐私造成损害的程度各异, $T_{W_2}^{ng,i}$、$T_{W_3}^{ng,i}$ 分别表示按照权重大小排序后的敏感观众姓名集合、敏感观众数量集合.

3) 效用指标包含 1 个正向阈值及 2 个逆向阈值: $T_{Ung_1}^{ng,i}$ 为第 i 次交互协商者智能体关于信息分享请求的效用值, $t_{Ung_2}^{ng,i}$ 为不包含积分效用时协商者智能体的效用值, $t_{Ung_3}^{ng,i}$ 为包含积分效用时协商者智能体的效用值.

$$t_{Ung_2}^{ng,i}=\begin{cases}t^{ng}-\dfrac{p^{ng}\times w_p^{ng}}{P_0}, u_{p_i}^{ng}\leqslant t^{ng}\\ t^{ng}+\dfrac{p^{ng}\times w_p^{ng}}{P_0}, u_{p_i}^{ng}> t^{ng}\end{cases} \tag{8}$$

$$t_{Ung_3}^{ng,i}=t_{Ung_2}^{ng,i}+\left(\frac{w_p^{ng}\times P^{in}}{P_0}\right) \tag{9}$$

公式 (8) 中, t^{ng} 为协商过程中协商者智能体效用值的最低要求 (设置为 0.7), $u_{p_i}^{ng}$ 表示第 i 次交互时协商者智能体基于信息分享请求计算出的效用值.

4) 积分指标包含 1 个正向阈值及 1 个逆向阈值: $T_{Png_1}^{ng,i}$ 为第 i 次交互过程中协商者智能体期望得到的积分, $t_{Png_2}^{ng,i}$ 为协商者智能体期望得到的积分.

$$(P^{ng})'=\frac{\left[1-\dfrac{|a_0-a_i|}{|a_0|}-\left(u_{p_i}^{in}\right)'\right]\times P_0}{w_P^{in}} \tag{10}$$

公式 (10) 中, a_0 表示发起者智能体最初信息分享请求中包含的观众, a_i 表示第 i 次交互时发起者智能体信息分享请求中包含的观众.

5) 协商结果指标包含 1 个正向阈值及 1 个逆向阈值: $T_{Cr_1}^{ng,i}$ 为第 i 次交互协商者智能体的回应结果, $t_{Cr_2}^{ng,i}$ 为协商者智能体对发起者智能体请求的回应结果.

$$t_{Cr_2}^{ng,i}=\begin{cases}\text{Y, if} t_{Ung_3}^{ng,i}\geqslant t^{ng}\\ \text{N, if} t_{Ung_3}^{ng,i}< t^{ng}\end{cases} \tag{11}$$

3.3 算法实现

定性追责算法中使用的辅助函数包括:

- GetNegotiationProcess(*in*, *ng*) 获取智能体间的协商过程.
- GetTrustedNegotiationProcess(*in*, *ng*) 模拟智能体完全诚实可信情况下的隐私协商过程, 获得隐私协商过程数据.
- initQR() 创建空列表, 用于存放定性追责结果.
- initList() 创建新列表, 用于存放过程数据.
- calculateThreshold() 计算逆向阈值.
- calculateThresholdFin() 计算最后一次交互时的阈值.
- distinguish() 实现实际协商数据与各阈值间的对比, 并给出定性追责结果.

算法 1. 定性追责算法.

输入: *Pro,* Privacy disclosure related information;
输出: *Qr*, Qualitative conclusion.

```
1 in ← Pro.owner; ng ← Pro.negotiaer;
2 N ← GetNegotiationProcess(in, ng); Nt ← GetTrustedNegotiationProcess(in, ng);
3 iterlist ← N.iterList; itertrustedlist ← Nt.itertrustedList;
4 Qr← initQR();
5 subQr ← initList();
6 if iterList.size() = 0 then
7       Qr.agent ←in; Qr.reason ←No negotiation;
8 for i=1 to iterlist.size() do
9       if i = 1 then
10            in.i.threshold ←calculateThreshold(audience, point, recode);
11            ng.i.threshold ←calculateThreshold(violators, weight, recode);
12      else if i = iterlist.size() then
13            in.i.threshold ←calculateThresholdFin(audience, point, utility, recode);
14            ng.i.threshold ←calculateThresholdFin(violators, weight, utility, point, recode);
15      else
16            in.i.threshold ←calculateThreshold(audience, point, utility, recode);
17            ng.i.threshold ←calculateThreshold(violators, weight, utility, point, recode);
18 end for
19 for i=1 to iterlist.size() do
20            subQr.i.in ←distinguish (iterlist.get(i), itertrustedlist.get(i), in.i.threshold);
21            subQr.i.ng ←distinguish (iterlist.get(i), itertrustedlist.get(i), ng.i.threshold);
22            Qr.i ← subQr.i.in ∪ subQr.i.ng;
23 end for
24 return Qr;
```

算法 1 为智能体行为定性追责算法, 其输入为隐私泄露相关信息, 输出为定性追责结果. 算法开始时, 提取参与隐私协商过程的智能体 (行 1), 获取智能体的实际协商过程、模拟智能体完全诚实可信情况下的协商过程 (行 2), 将获取到的协商过程数据分别存储到 *iterlist* 和 *itertrustedlist* 中 (行 3). 协商数据获取完成后, 判断 *iterlist* 是否为空, 若为空, 说明没有协商过程, 直接判定发起者智能体存在不当行为, 原因是未进行隐私协商便分享了信息 (行 6、7); 否则, 计算第 1 次交互中的行为指标阈值 (行 9–11)、最后一次交互中的行为指标阈值 (行 12–14)、其余交互中的行为指标阈值 (行 15–17), 将实际协商过程数据与行为指标阈值进行对比 (行 19–21), 获得判定结果并将其存储到 *Qr* 中 (行 22).

4 定量追责

在定量追责的过程中加入了权重, 表示各行为指标对于隐私被泄露者的重要程度, 其取值是由遭受隐私泄露者确定, 以量化出更准确的结果, 使得责任判定更具有针对性和说服力.

4.1 简单量化方法

本文首先提出了简单量化方法, 不仅可以对定性追责结果进行验证, 还能以较高效率对智能体的不当行为进

行简单量化.

简单量化方法在考虑实际协商过程数据和对应行为指标阈值的同时, 计算智能体在协商过程中不当行为相关数据的变化区间, 并为不同行为指标设置不同权重.

假设智能体的实际协商过程数据为 $D^{ac}_{|INDEX|}$, 各指标阈值为 $Tt_{|INDEX|}$, 借助于简单量化方法对智能体进行定量追责的公式如下:

$$Q_{\text{en}}=\sum_{|INDEX|}\left(\frac{W_{|INDEX|}}{W}\times\frac{\left|D^{ac}_{|INDEX|}-Tt_{|INDEX|}\right|}{D_{\text{confine}}}\right) \tag{12}$$

公式 (12) 中, Q_{en} 表示对智能体不当行为的整体量化, $INDEX=INDEX^{+}_{in}\cup INDEX^{-}_{in}\cup INDEX^{+}_{ng}\cup INDEX^{-}_{ng}$, $W_{|INDEX|}$ 表示第$|INDEX|$个行为指标的权重, W 表示智能体行为权重之和, D_{confine} 为不当行为相关数据的变化区间边界值之差.

4.2 加权马氏距离

马氏距离 (Mahalanobis distance)(https://en.wikipedia.org/wiki/Mahalanobis_distance) 由 Mahalanobis 提出, 用于表示数据的协方差距离, 是一种有效的计算两个未知样本集相似度的方法. 马氏距离解决了欧式距离中各个维度尺度不一致但具有相关性的问题. 由于本文中协商过程数据存在维度间尺度差异较大的情况, 因此, 使用马氏距离计算数据之间的相似度具有较强的针对性, 可以得到比较准确的量化结果. 传统的马氏距离表示为:

$$d_{\text{Q}}=\sqrt{(\mathbf{x}_i-\mathbf{t}_i)^{\text{T}}\mathbf{C}^{-1}(\mathbf{x}_i-\mathbf{t}_i)} \tag{13}$$

鉴于传统的马氏距离无法计算包含带权重样本的集合间的距离, 而为多个行为指标设置不同的权重有助于区分用户对不同行为指标的关注程度, 从而使定量追责更具针对性, 本文对传统马氏距离进行了改进, 提出了加权马氏距离, 其计算公式如下:

$$\mathbf{X}=diag(\mathbf{W}(\mathbf{x}_{|INDEX|}-\mathbf{t}_{|INDEX|})) \tag{14}$$

$$d_{\text{Q}}'=\sqrt{\mathbf{X}^{\text{T}}\mathbf{C}^{-1}\mathbf{X}} \tag{15}$$

其中, $\mathbf{x}_{|INDEX|}$ 为向量形式表示的实际协商过程数据, $\mathbf{t}_{|INDEX|}$ 为向量形式表示的行为指标阈值, $\mathbf{W}=(W_{Au}, W_{Pin}, W_{Uin}, W_{Cr}, W_{Vng}, W_{W}, W_{Ung}, W_{Png}, W_{Cr})$ 为第 2.2 节中各行为指标的权重向量, $\mathbf{C}^{-1}$ 为协方差矩阵.

基于加权马氏距离进行定量追责的步骤如下:

1) 通过在限定区间 (与简单量化方法中不当行为相关数据的变化区间相同) 内随机模拟智能体行为, 生成用于训练协方差矩阵的训练集, 进一步获得协方差矩阵 $\mathbf{C}^{-1}$.

2) 借助于公式 (15), 求得加权马氏距离值 d_Q'.

3) 计算实际数据与其限定区间左右边界间的距离 d_1、d_2.

4) 借助于公式 (16), 求得责任量化值 (quantified values of responsibility, QVR) Q_m.

$$Q_m=\frac{d_Q'}{d_1+d_2} \tag{16}$$

4.3 改进 Minhash 方法

Minhash 是 Broder 提出的最小独立置换 (min-wise independent permutation) 概念的实现[24]. 给定通用集 U 及其子集 $S\subseteq U$, 假设有 D 个 hash 函数, 表示为 $\{h_d\}_{d=1}^{D}$, 其中, h_d: $U\to U$, S 的 Minhash 集合 (也称为签名矩阵) 是指每个 hash 函数中哈希值最小的元素组成的集合, 表示为 $\{\arg\min(h_d(S))\}_{d=1}^{D}$.

根据 Minhash 相关理论, 集合 $S\subseteq U$、$T\subseteq U$ 生成相同 Minhash 值的概率等于其 Jaccard 相似度:

$$\Pr(\min(h_d(S))=\min(h_d(T)))=Jaccard(S,T)=\frac{S\cap T}{S\cup T} \tag{17}$$

针对同时包含数值和文本的数据, 在使用传统 Minhash 方法 (https://okomestudio.net/biboroku/2014/04/near-

duplicate-detection-using-minhash-background/) 计算数据间的相似度时, 若不对数据进行划分, 将导致计算准确度差; 若使用 Shingling 方法对数据进行划分, 数值型数据被视作文本强行划分, 破坏数值型数据的本源意义; 同时 Minhash 方法中存在数据去重现象 (将多个相同数据并为单个数据进行处理), 将降低计算准确度. 为缓解这些问题, 本文提出了改进 Minhash 方法: 首先, 对数据进行分类, 以针对不同类型的数据分别进行 hash 计算, 并采用 Shingling 方法对文本型数据进行划分; 其次, 引入权重的概念, 区分不同数据的重要程度; 第三, 为避免数据去重的发生, 对重复数据进行了单一化处理.

改进 Minhash 方法的实现过程如图 3 所示. 第 1 步, 将数据分为文本型数据和数值型数据, 并采用 Shingling 方法将文本型数据划分为多个独立的文本块. 第 2 步, 对分类/划分后的数据赋予权重, 并根据权重进行数据的局部单一化处理. 单一化处理的基本思想是在保证数据可行且不影响实验结果的前提下, 修改重复数据: 对于数值型数据, 通过加 (或减) 适当的值来生成新的数据, 如: 初始数据为{10; 15}, 权重设置为{3; 2}, 加入权重及进行局部单一化处理后的数据为{10, 15, 20; 15, 20}; 对于文本型数据, 通过添加不同标记生成新的数据, 如数据 *John* 的权重为 3, 基于权重进行数据扩充及单一化处理后的结果为 *John*, *John'*, *John''*. 第 3 步, 对加入权重后的数据进行全局单一化处理, 其基本思想同数据的局部单一化处理. 如: 将{10, 15, 20; 15, 20}进行全局单一化处理后的结果为{10, 15, 20; 16, 21}. 需要注意的是, 在调整数值时要同步考虑两个数据集中的相同位置, 即如果对应位置的数据相同, 修改后也需要相同; 如果不同, 修改后则需不同. 第 4 步, 进行数据间的相似度计算.

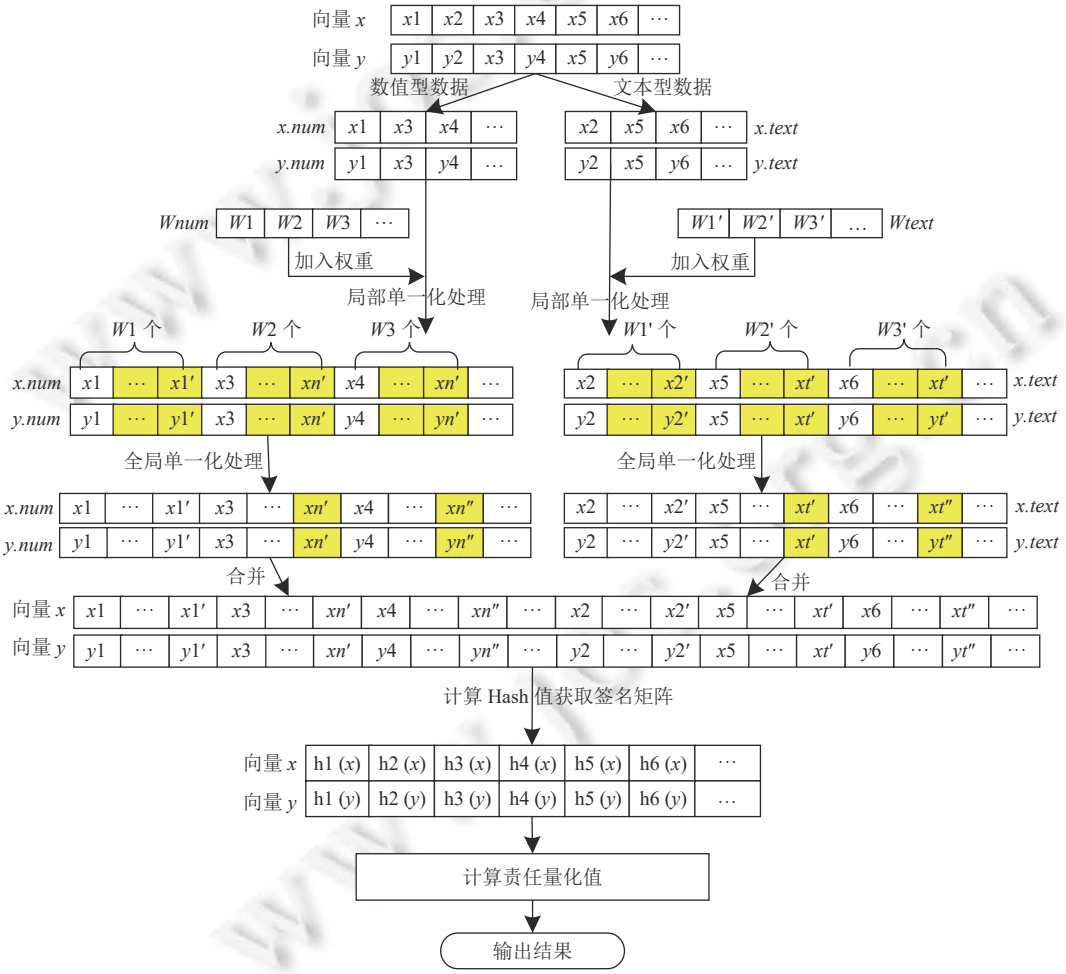


图 3 改进 Minhash 方法实现过程

假设数据向量为 $\mathbf{x}$, $\mathbf{W}_{|INDEX|}$为 $\mathbf{x}$ 中指标$\mathbf{x}_{|INDEX|}$的权重, 带权重的 hash 计算公式如下:

$$h_d(x) = \arg\min\ h_d(\mathbf{W}_{|INDEX|}, \mathbf{x}_{|INDEX|}) \tag{18}$$

公式 (18) 中, $1 \leqslant d \leqslant D$, D 为 hash 函数总量, D 个 hash 函数的最小值构成数据向量 $\mathbf{x}$ 的签名矩阵 $\{h_d(\mathbf{x})\}_{d=1}^{D}$. 同上, 数据向量 $\mathbf{y}$ 的签名矩阵为 $\{h_d(\mathbf{y})\}_{d=1}^{D}$.

进一步地, 计算 $\mathbf{x}$ 与 $\mathbf{y}$ 间的相似度. 本文中采用了两种相似度计算公式, 分别如公式 (19) 和公式 (20) 所示.

$$sim_{\text{Minhash}} = \frac{\sum_{d=1}^{D} \text{IF}\{h_d(x) = h_d(y)\}}{D} \tag{19}$$

其中, 当 $h_d(\mathbf{x}) = h_d(\mathbf{y})$ 成立时, $\text{IF}\{h_d(x) = h_d(y)\}=1$, 否则, $\text{IF}\{h_d(x) = h_d(y)\}=0$.

同时, 计算 $\{h_d(\mathbf{x})\}_{d=1}^{D}$、 $\{h_d(\mathbf{y})\}_{d=1}^{D}$的交集和并集, 并采用如下公式计算 Jaccard 相似度:

$$sim_{\text{Jaccard}} = \frac{\{h_d(\mathbf{x})\}_{d=1}^{D} \cap \{h_d(\mathbf{y})\}_{d=1}^{D}}{\{h_d(\mathbf{x})\}_{d=1}^{D} \cup \{h_d(\mathbf{y})\}_{d=1}^{D}} \tag{20}$$

最后借助公式 (21) 和公式 (22) 分别计算最终的责任量化值.

$$Q_{\text{Minhash}} = 1 - sim_{\text{Minhash}} \tag{21}$$

$$Q_{\text{Jaccard}} = 1 - sim_{\text{Jaccard}} \tag{22}$$

5 实验结果与分析

本文基于文献 [9] 中提出的社交网络隐私协商系统, 通过增加行为追责模块, 完成了隐私协商智能体行为追责系统的设计, 并基于 IntelliJ IDEA 开发软件完成了系统实现. 本节将基于文献 [9] 中使用的真实社交网络数据, 对本文所提出的定性/定量追责的有效性、合理性进行实验验证及分析.

5.1 不当行为及权重设置

(1) 不当行为设置

为验证本文所提方法的性能, 实验选取了包含 3 次交互的协商过程数据, 并随机设置了多种可能出现的不当行为, 如表 2.

表 2 不当行为设置情况表

智能体	交互情况	不当行为指标	实际数据设置	阈值设置
发起者智能体	第1次交互	观众列表姓名	增加3个用户: {user40, user41, user42};	无增添的用户
		观众列表数量	41	38
	第2次交互	观众列表数量	41	38
	第3次交互	观众列表姓名	增加3个用户: {user40, user41, user42};	无增添的用户
		观众列表数量	41	38
		效用	0.7	0.9
协商者智能体	第1次交互	权重敏感用户	改变用户列表	无改动
	第2次交互	效用	0.85665	0.68718
	第3次交互	敏感用户	改变敏感用户列表	无改动
		效用	160	122

(2) 权重设置

在实际的社交网络场景中, 用户对隐私保护的需求各不相同, 因此本章节设置了多种权重组合, 以全面地对本文中提出的定性/定量追责方法进行验证及分析. 表 3 为实验仿真中权重设置情况.

5.2 定性追责及验证

定性追责的实验结果如图 4 所示, 实线框内文本为智能体每次交互 (num1、num2、num3 分别代表第 1、2、

3 次交互) 中是否存在不当行为的判定结果, 虚线框内文本为综合判定结果. 实验结果表明, 发起者智能体和协商者智能体均存在不当行为, 且与表 2 相符, 说明定性追责方法是准确有效的.

表 3　权重设置情况表

权重组合	发起者智能体				协商者智能体				
	Au	*Pin*	*Uin*	*Cr*	*Vng*	*W*	*Ung*	*Png*	*Cr*
Com_1	3	2	4	1	4	3	5	2	1
Com_2	4	2	3	1	3	4	5	2	1
Com_3	3	4	2	1	5	2	4	3	1
Com_4	4	1	3	2	5	3	4	1	2
Com_5	2	1	4	3	2	5	4	1	3
Com_6	2	4	3	1	1	5	3	4	2
Com_7	1	3	2	4	2	1	3	4	5
Com_8	1	3	4	2	2	5	3	4	1

进一步使用简单量化方法对定性追责结果进行验证, 实验结果如图 5、图 6 所示. 图 5 统计了发起者智能体和协商者智能体在整个协商过程中各行为指标的综合责任量化值, 图 6 进一步详细统计了各次交互中行为指标的责任量化值. 从图 5 左图可以看出, 发起者智能体在观众列表指标方面的责任量化值为 0.2, 效用值指标方面的责任量化值为 0.4, 而积分指标和协商结果指标方面的责任量化值均为 0, 说明发起者智能体在观众和效用指标存在不当行为; 同理, 图 5 右图表明协商者智能体在敏感观众列表及其权重、效用值共 3 个指标方面存在不当行为.

从图 6 左图可以看出, 发起者智能体在第 1、2 次交互中的观众列表指标及第 3 次交互中的观众列表指标和效用值指标方面存在不当行为. 从图 6 右图可以看出, 协商者智能体在第 1 次交互中的敏感观众权重指标、第 2 次交互中的效用值指标和第 3 次交互中的敏感观众列表指标及效用值指标方面存在不当行为.

易见, 简单量化方法得出的责任量化值与实验前的不当行为设置及定性追责结果相符, 同时说明了定性追责方法及简单量化方法的有效性.

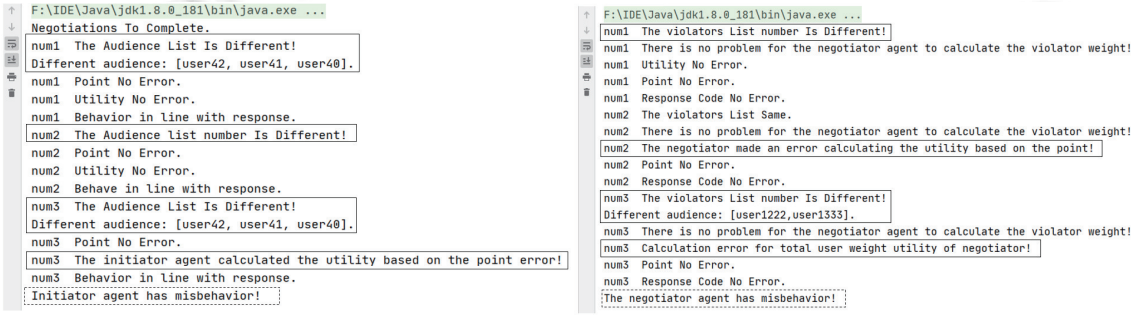
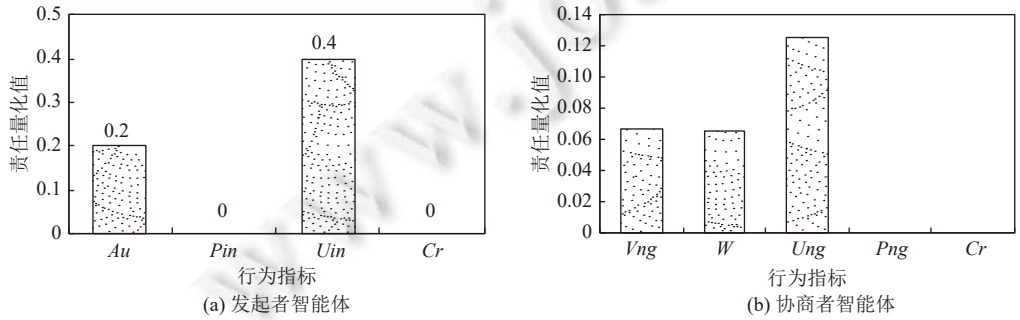


图 4　定性追责实验结果截图



(a) 发起者智能体　(b) 协商者智能体

图 5　基于简单量化方法的行为指标综合责任量化值

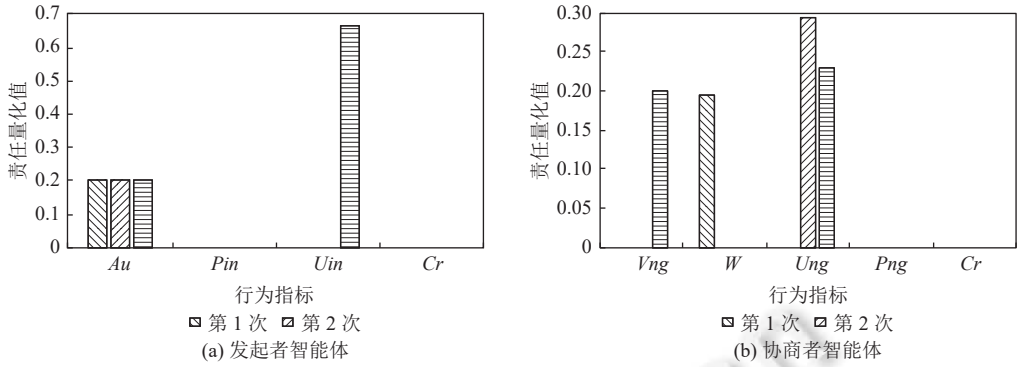


图 6 基于简单量化方法的各次交互中行为指标责任量化值

5.3 定量追责及分析

针对本文所提出的简单量化方法、加权马氏距离、改进 Minhash 方法, 图 7 和图 8 分别展示了不同权重设置下发起者智能体和协商者智能体不当行为的责任量化值. 关于改进 Minhash 方法, 在实验中使用了公式 (21) 和公式 (22) 两种方法计算相似度, 因此图 7 中展示了 4 种实验结果间的对比. 需要说明的是, 改进 Minhash 方法的实验结果借助不同数量的 hash 函数构建签名矩阵, 进一步通过相似度计算获得责任量化值, 并将其平均值作为实验结果进行展示.

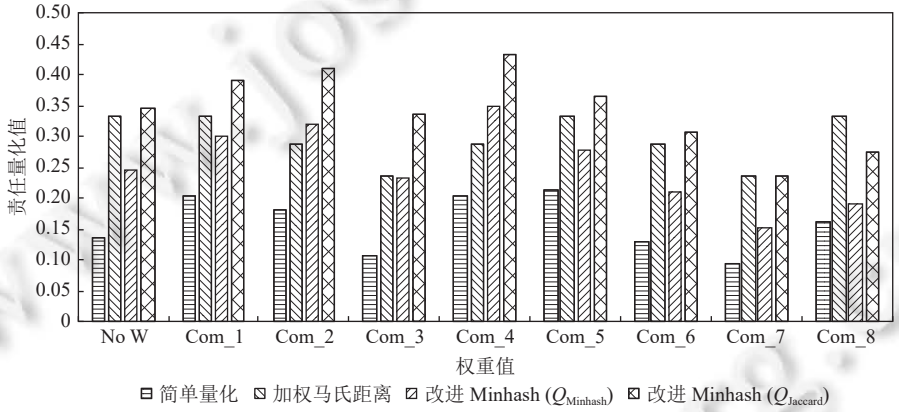


图 7 发起者智能体不当行为的责任量化值

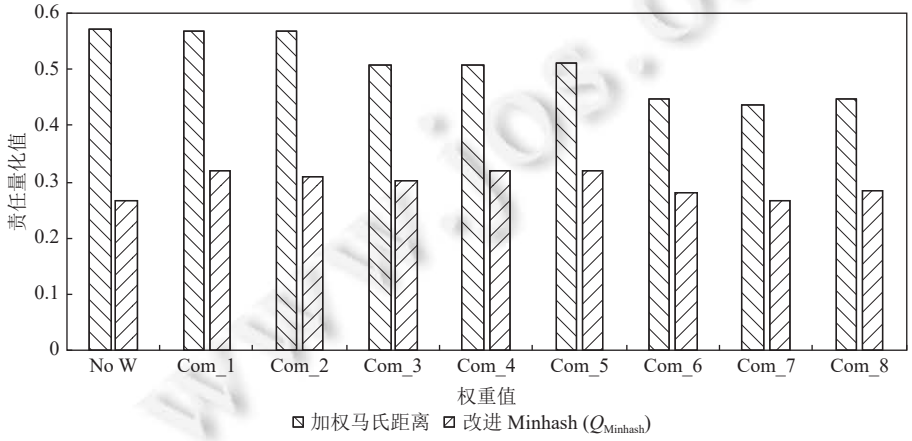


图 8 协商者智能体不当行为的责任量化值

根据实验前的不当行为设置, 发起者智能体在观众列表和效用值指标方面存在不当行为, 协商者智能体在敏感观众及其权重、效用值指标方面存在不当行为. 根据权重设置, 发起者智能体存在不当行为的指标 (*Au*、*Uin*) 在 Com_1、Com_2、Com_4 和 Com_5 等权重设置中占比较大, 所以从图 7 可以发现, 权重设置为 Com_1、Com_2、Com_4 和 Com_5 时的责任量化值明显大于权重设置为 Com_3、Com_6、Com_7、Com_8 时的责任量化值 (尤其是改进 Minhash 方法得出的结果更为突出), 且都大于 No W (无权重) 的责任量化值. 这一实验结果表明, 本文基于权重进行隐私协商智能体行为定量追责的研究是有效的, 为行为指标设置不同的权重有助于区分用户对不同行为指标的关注程度, 最终使得定量追责更具针对性.

图 7 表明简单量化方法得出的责任量化值相对较小, 结合实验数据分析发现主要由于本实验中数据集所含数据较为复杂, 且数值型数据在取值方面差距较大, 导致计算结果偏小, 因此在本实验中简单量化方法主要用来验证定性追责结果. 改进 Minhash 方法中 Q_{Jaccard} 方法责任量化值远大于 Q_{Minhash}, 主要原因是数据集中存在数值型数据; 经多次测试, Q_{Minhash} 方法的量化结果对包含数值型数据的数据集计算结果更准确. 从图中可以看出, 除 Com_2 和 Com_4 权重外, 其他权重的 Q_{Minhash} 方法责任量化值比加权马氏距离小, 主要是由于发起者智能体的协商数据中数值型数据占大部分, 且文本型数据存在较多相同字符 (结合表 2 也可以看出, 有多个不当行为指标仅包含数值型数据); 而 Com_2 和 Com_4 中 *Au* 指标的权重值最大, 且 *Au* 指标中包含文本型数据, 因而 Q_{Minhash} 方法责任量化值偏大.

图 8 表明基于加权马氏距离计算出的协商者智能体不当行为责任量化值较改进 Minhash 方法偏大, 原因是计算协商者智能体的责任量化值时所使用的行为指标中包含多个文本型数据, 而加权马氏距离仅能处理数值型数据, 因此计算出的责任量化值较大. 为了进一步对比加权马氏距离方法与改进 Minhash 方法在隐私协商智能体不当行为定量追责方面的表现, 通过提取行为指标中的数值型数据, 进行了进一步实验, 实验结果如图 9 所示. 可以看出, 当仅采用数值型行为指标进行定量追责研究时, 加权马氏距离方法与改进 Minhash 方法得出的结果较为相似, 说明此两种方法在定量追责方面具有较强的一致性, 印证了彼此在定量追责时的有效性.

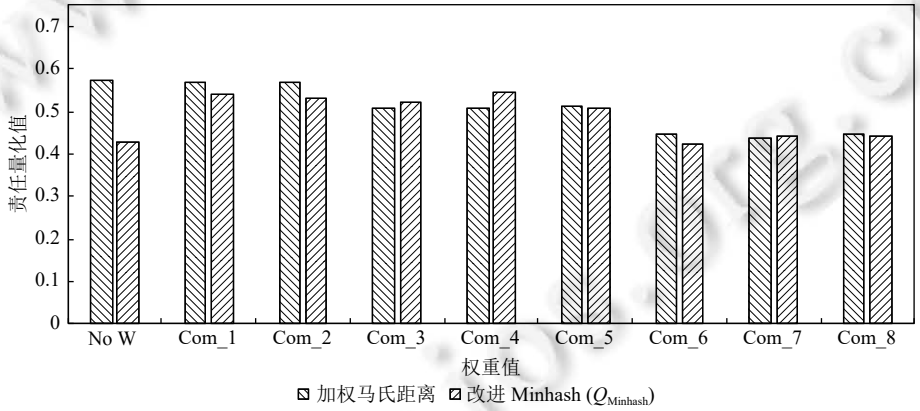


图 9　协商者智能体不当行为的责任量化值 (仅数值型数据)

为了说明改进 Minhash 方法在定量追责方面的优势, 选择 Q_{Minhash} 方法 (公式 (21)) 与传统 Minhash 方法进行了对比实验, 实验结果如图 10 所示. 与改进 Minhash 方法相比, 传统 Minhash 方法 (实验结果为图中 OR 折线) 对智能体不当行为的量化结果更大 (OR 折线具有更大的纵坐标取值), 原因是改进 Minhash 方法 (其实验结果为除 OR 折线之外的其他折线) 包含按数据类型进行数据分类、针对不同类型数据进行不同的 hash 计算、数据的局部/全局单一化处理等操作, 对智能体行为数据间的不同进行了充分区分, 提高了相似度度量值, 从而获得了较小的责任量化值. 换言之, 本文所提改进 Minhash 方法的计算结果更为准确合理.

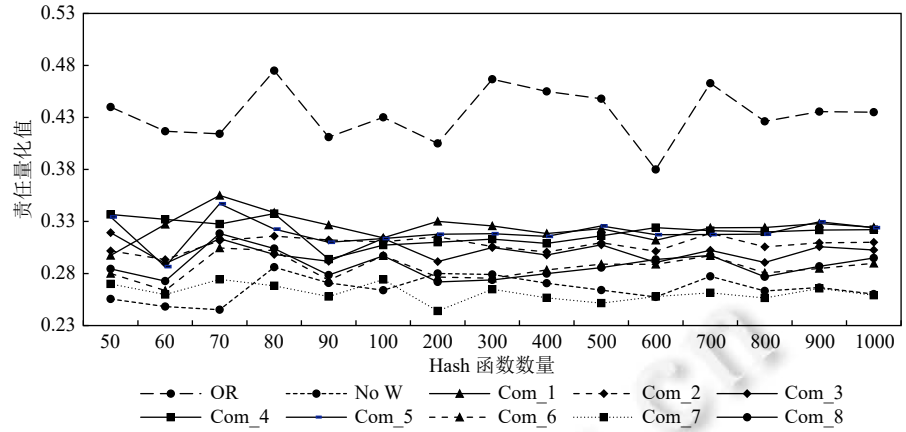


图 10 协商者智能体不当行为责任量化值 (仅 Q_{Minhash} 方法)

6 总 结

为进行社交网络中的隐私保护, 基于智能体的隐私协商受到了关注, 但是尚缺少与之相关的追责研究, 即无法对具有不当行为的隐私协商智能体进行行为追责. 本文基于文献 [9] 中提出的社交网络互惠隐私协商体系, 通过增加行为追责模块, 设计实现了具有定性追责和定量追责的隐私协商智能体行为追责系统, 并对其进行了详细的实验分析, 实验结果表明了系统和方法的合理性及有效性.

由于针对社交网络中隐私协商智能体的追责研究较少, 本文研究具有一定的探索意义, 但在技术成熟度方面尚有待改善. 后续研究将着重进行以下几方面工作: 1) 针对定量追责, 提出更加客观的评价标准, 为深入衡量追责方案的性能提供依据; 2) 扩展定性/定量追责技术的应用场景, 如电子病历、数据交易系统等, 验证其有效性、增强其实用价值; 3) 针对使用先进隐私保护技术 (如安全多方计算、差分隐私) 的人工智能系统以及具有较强自主决策能力的智能体, 提出追责机制, 解决其中可能存在的隐私泄露追责问题.

References:

[1] Tran HY, Hu JK. Privacy-preserving big data analytics a comprehensive survey. Journal of Parallel and Distributed Computing, 2019, 134: 207–218. [doi: 10.1016/j.jpdc.2019.08.007]

[2] Zhu TQ, Ye DY, Wang W, Zhou WL, Yu PS. More than privacy: Applying differential privacy in key areas of artificial intelligence. arXiv: 2008.01916, 2020.

[3] Zhao JW, Chen YF, Zhang W. Differential privacy preservation in deep learning: Challenges, opportunities and solutions. IEEE Access, 2019, 7: 48901–48911. [doi: 10.1109/ACCESS.2019.2909559]

[4] Chouldechova A, Roth A. The frontiers of fairness in machine learning. arXiv: 1810.08810, 2018.

[5] Tan ZW, Zhang LF. Survey on privacy preserving techniques for machine learning. Ruan Jian Xue Bao/Journal of Software, 2020, 31(7): 2127–2156 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/6052.htm [doi: 10.13328/j.cnki.jos.006052]

[6] Ji SL, Du TY, Li JF, Shen C, Li B. Security and privacy of machine learning models: A survey. Ruan Jian Xue Bao/Journal of Software, 2021, 32(1): 41–67 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/6131.htm [doi: 10.13328/j.cnki.jos.006131]

[7] Hasan R, Crandall D, Fritz M, Kapadia A. Automatically detecting bystanders in photos to reduce privacy risks. In: Proc. of the 2020 IEEE Symp. on Security and Privacy. San Francisco: IEEE, 2020. 318–335. [doi: 10.1109/SP40000.2020.00097]

[8] Yang JH, Chakrabarti A, Vorobeychik Y. Protecting geolocation privacy of photo collections. In: Proc. of the 35th AAAI Conf. on Artificial Intelligence. Palo Alto: AAAI, 2020. 524–531. [doi: 10.1609/aaai.v34i01.5390]

[9] Kekulluoglu D, Kokciyan N, Yolum P. Preserving privacy as social responsibility in online social networks. ACM Trans. on Internet Technology, 2018, 18(4): 42. [doi: 10.1145/3158373]

[10] European Commission's High-Level Expert Group on Artificial Intelligence. Ethics guidelines for trustworthy AI. Brussels: European Commission, 2019.

[11] China National New Generation Artificial Intelligence Governance Professional Committee. The new generation of artificial intelligence governance principles —Develop responsible artificial intelligence. 2019 (in Chinese). http://www.most.gov.cn/kjbgz/201906/t20190617_147107.htm

[12] Wu WJ, Huang TJ, Gong K. Ethical principles and governance technology development of AI in China. Engineering, 2020, 6(3): 302–309. [doi: 10.1016/j.eng.2019.12.015]

[13] Chen KR, Meng XF. Interpretation and understanding in machine learning. Journal of Computer Research and Development, 2020, 57(9): 1971–1986 (in Chinese with English abstract). [doi: 10.7544/issn1000-1239.2020.20190456]

[14] Küsters R, Truderung T, Vogt A. Accountability: Definition and relationship to verifiability. In: Proc. of the 17th ACM Conf. on Computer and Communications Security. Chicago: ACM, 2010. 526–535. [doi: 10.1145/1866307.1866366]

[15] Jung T, Li XY, Huang WC, Qiao ZY, Qian JW, Chen LL, Han JZ, Hou JH. AccountTrade: Accountability against dishonest big data buyers and sellers. IEEE Trans. on Information Forensics and Security, 2019, 14(1): 223–234. [doi: 10.1109/TIFS.2018.2848657]

[16] Gu TL, Li L. Artificial moral agents and their design methodology: Petrospect and prospect. Chinese Journal of Computers, 2021, 44(3): 632–651 (in Chinese with English abstract). [doi: 10.11897/SP.J.1016.2021.00632]

[17] Chen XP. Ethical system of artificial intelligence: Infrastructure and key issues. CAAI Trans. on Intelligent Systems, 2019, 14(4): 605–610 (in Chinese with English abstract). [doi: 10.11992/tis.201906037]

[18] Amon MJ, Hasan R, Hugenberg K, Bertenthal BI, Kapadia A. Influencing photo sharing decisions on social media: A case of paradoxical findings. In: Proc.of the 2020 IEEE Symp. on Security and Privacy. San Francisco: IEEE, 2020. 1350–1366. [doi: 10.1109/SP40000.2020.00006]

[19] Pensa RG, Di Blasi G. A privacy self-assessment framework for online social networks. Expert Systems with Applications, 2017, 86: 18–31. [doi: 10.1016/j.eswa.2017.05.054]

[20] Such JM, Rovatsos M. Privacy policy negotiation in social media. ACM Trans. on Autonomous and Adaptive Systems, 2016, 11(1): 4. [doi: 10.1145/2821512]

[21] Kökciyan N, Yolum P. PriGuard: A semantic approach to detect privacy violations in online social networks. IEEE Trans. on Knowledge and Data Engineering, 2016, 28(10): 2724–2737. [doi: 10.1109/TKDE.2016.2583425]

[22] Baldoni M, Baroglio C, May KM, Micalizio R, Tedeschi S. Computational accountability. In: Proc. of the AI*IA Workshop on Deep Understanding and Reasoning: A Challenge for Next-generation Intelligent Agents 2016 Co-Located with 15th Int'l. Conf. of the Italian Association for Artificial Intelligence (AIxIA 2016). Genova: AI*IA, 2016. 56–62.

[23] Baldoni M, Baroglio C, May KM, Micalizio R, Tedeschi S. Supporting organizational accountability inside multiagent systems. In: Proc. of the XVIth Conf. of the Italian Association for Artificial Intelligence. Bari: Springer, 2017. 403–417. [doi: 10.1007/978-3-319-70169-1_30]

[24] Wu W, Li B, Chen L, Gao JB, Zhang CQ. A review for weighted minhash algorithms. IEEE Trans. on Knowledge and Data Engineering, 2022, 34(6): 2553–2573. [doi: 10.1109/TKDE.2020.3021067]

附中文参考文献:

[5] 谭作文, 张连福. 机器学习隐私保护研究综述. 软件学报, 2020, 31(7): 2127–2156. http://www.jos.org.cn/1000-9825/6052.htm [doi: 10.13328/j.cnki.jos.006052]

[6] 纪守领, 杜天宇, 李进锋, 沈超, 李博. 机器学习模型安全与隐私研究综述. 软件学报, 2021, 32(1): 41–67. http://www.jos.org.cn/1000-9825/6131.htm [doi: 10.13328/j.cnki.jos.006131]

[11] 国家新一代人工智能治理专业委员会. 新一代人工智能治理原则——发展负责任的人工智能. 2019. http://www.most.gov.cn/kjbgz/201906/t20190617_147107.htm

[13] 陈珂锐, 孟小峰. 机器学习的可解释性. 计算机研究与发展, 2020, 57(9): 1971–1986. [doi: 10.7544/issn1000-1239.2020.20190456]

[16] 古天龙, 李龙. 伦理智能体及其设计: 现状和展望. 计算机学报, 2021, 44(3): 632–651. [doi: 10.11897/SP.J.1016.2021.00632]

[17] 陈小平. 人工智能伦理体系: 基础架构与关键问题. 智能系统学报, 2019, 14(4): 605–610. [doi: 10.11992/tis.201906037]

古天龙(1964－), 男, 博士, 教授, 博士生导师, CCF 高级会员, 主要研究领域为形式化方法, 可信人工智能, 伦理智能体设计, 人工智能伦理.

郝峰锐(1995－), 男, 硕士生, 主要研究领域为隐私保护, 人工智能伦理.

李龙(1989－), 男, 博士, 讲师, CCF 专业会员, 主要研究领域为人工智能安全, 逻辑程序设计.

李晶晶(1989－), 女, 博士, 讲师, 主要研究领域为阵列信号处理, 微弱信号检测, 公平机器学习.

常亮(1980－), 男, 博士, 教授, 博士生导师, CCF 高级会员, 主要研究领域为形式化方法, 知识表示与推理, 描述逻辑, 人工智能伦理.