

基于知识图谱的隐私政策合规性检测与分析

张西珩, 李昕, 唐鹏, 黄锐奇, 何渊, 邱卫东

(上海交通大学网络空间安全学院, 上海 200210)

摘要:《中华人民共和国个人信息保护法》(以下简称“个保法”)作为保护个人信息权益的重要法律,对个人信息处理者收集、存储、使用、分享等信息处理活动提出明确规范,并要求个人信息处理者在所提供服务的隐私政策中予以说明。这意味着任何公司在国内提供服务时,需首先提供符合“个保法”要求的隐私政策。为了实现自动分析面向“个保法”的隐私政策合规性,提出了基于知识图谱构建隐私政策合规性智能检测方法。首先,对“个保法”进行全面分析并构建相应的多级隐私政策知识图谱,涵盖需要在隐私政策中予以说明的信息保护相关概念。然后,构建半自动化的隐私政策收集方法并收集400份中文App的隐私政策,对其中100份基于知识图谱进行交叉标注后形成首个面向“个保法”的中文隐私政策语料库APPCP-100,使用剩余300份隐私政策构建中文概念分类器模型CPP-BERT,实现高效的隐私政策合规性智能检测。最后,应用知识图谱对隐私政策进行全面的合规性分析,结果显示当前中文App隐私政策对“个保法”中细粒度概念的合规性仍有待提高。

关键词: 中华人民共和国个人信息保护法; 隐私政策; 隐私保护; 合规性检测

中图分类号: TP309.2

文献标志码: A

DOI: 10.11959/j.issn.2096-109x.2024087

Privacy policy compliance detection and analysis based on knowledge graph

ZHANG Xiheng, LI Xin, TANG Peng, HUANG Ruiqi, HE Yuan, QIU Weidong

School of Cyber Science and Engineering, Shanghai Jiao Tong University, Shanghai 200210, China

Abstract: The personal information protection law (PIPL) of the People's Republic of China served as an important legal framework for safeguarding personal information rights. It established clear regulations for personal information controllers in their activities involving the collecting, storing, using, and sharing of personal information. It also required that these controllers provide explanations within their privacy policies for the services they offered. This meant that any company providing services in China must first offer a privacy policy that complied with the requirements of the PIPL. Therefore, in order to analyze the compliance of privacy policies with respect to the PIPL, an intelligent method was presented for assessing privacy policy compliance based on a knowledge graph. First, a comprehensive analysis of the PIPL was conducted, and a multi-level privacy policy knowledge graph was pro-

收稿日期: 2024-01-19; 修回日期: 2024-03-11

通信作者: qiuwd@sjtu.edu.cn

引用格式: 张西珩, 李昕, 唐鹏, 等. 基于知识图谱的隐私政策合规性检测与分析[J]. 网络与信息安全学报, 2024, 10(6): 151-163.

Citation Format: ZHANG X H, LI X, TANG P, et al. Privacy policy compliance detection and analysis based on knowledge graph[J]. Chinese Journal of Network and Information Security, 2024, 10(6): 151-163.

posed that covered concepts related to information protection that needed to be explained in privacy policies. Next, a semi-automated method was built for collecting privacy policies and collected the privacy policies of 400 Chinese Apps. 100 policies were cross-annotated based on the knowledge graph, resulting in the creation of the first Chinese privacy policy corpus tailored to the PIPL called APPCP-100 (APP-privacy-policy-corpus-for-PIPL-100). Using this corpus, a Chinese concept classifier model CPP-BERT was constructed to achieve efficient detection of privacy policy compliance. Finally, the knowledge graph was applied to conduct a comprehensive compliance analysis of privacy policies, and the results indicate that the current compliance of Chinese App privacy policies with the fine-grained concepts of the PIPL still needs improvement.

Keywords: personal information protection law of the People's Republic of China, privacy policy, privacy protection, compliance check

0 引言

随着智能手机的普及与通信网络的广泛覆盖, 互联网服务已经成为现代生活的一部分, 为用户提供了个性化和便捷的服务。然而, 服务商提供服务的同时往往会收集和用户的个人信息, 信息泄露、滥用、盗用等问题逐渐凸显, 个人隐私安全遭到威胁。在此背景下, 为规范个人信息处理活动, 限制网络服务侵犯用户隐私, 我国于 2021 年 8 月 20 日出台了《中华人民共和国个人信息保护法》(以下简称“个保法”)。“个保法”规定, 个人信息定义为“以电子或者其他方式记录的与已识别或者可识别的自然人有关的各种信息, 不包括匿名化处理后的信息”, 在一定程度上扩大了个人信息的范围。此外, “个保法”赋予个人对个人信息处理的应有权利, 包括个人信息查询权、更正权、删除权、自动决策权等具体权利, 还赋予抽象的权利(如对个人信息处理的知情权、决定权、限制或拒绝处理个人信息等), 从而建立起合法公开的个人信息权益体系, 为个人隐私权利的保护提供法律保障。

为了满足“个保法”的要求, 应用程序(App)或者在线服务(Web)需要提供隐私政策, 告知用户其如何收集和使用个人信息, 帮助用户对个人信息数据进行控制和行使个人信息权利。但由于隐私政策通常篇幅较长且包含复杂的专业术语与措辞, 普通用户常难以理解其内容而选择直接跳过。因此, 专业研究人员有必要分析隐私政策, 以帮助用户了解个人信息数据的使用方式, 同时帮助在线服务建立合规的个人信息收集与处理流程。然而, 考虑到现有在线服务数量

与隐私政策文本长度, 单纯依赖专业人员进行人工分析显然是不现实的。因此, 隐私政策合规性自动化检测和分析成为隐私保护领域的研究方向之一。

本文调研了现有的“个保法”和隐私政策合规性研究工作, 发现面向“个保法”的中文隐私政策分析领域仍有较大空白。为了推动该领域的发展, 本文的研究人员与相关法律专家合作构建了面向“个保法”的隐私政策知识图谱, 其中包含隐私政策所应明确声明的“个保法”相关概念和通用条款。此知识图谱可以作为“个保法”隐私政策分析的基础, 如构建隐私政策语料库、设计概念分类器以及分析“个保法”下隐私政策合规性。本文还在收集的语料库上应用了此知识图谱, 展示了知识图谱在实际研究中的应用, 为未来该领域的研究提供参考。

本文的主要贡献概述如下。

1) 提出首个面向“个保法”的分级隐私政策知识图谱, 可用于中文隐私政策合规性自动化分析。该知识图谱参考了法律专家的意见与《信息安全技术 个人信息安全规范》的指导, 由三层节点组成, 包括了中文隐私政策需声明的所有相关概念。

2) 构建了中文隐私政策语料库与隐私政策合规检测方案。构建半自动化的工具收集国内 400 个热门 App 的隐私政策, 形成首个面向“个保法”的中文隐私政策语料库 APPCP-100(APP-privacy-policy-corpus-for-PIPL-100), 并基于此语料库构建中文隐私政策合规性检测模型 CPP-BERT (Chinese-privacy-policy BERT), 实现高效隐私政策合规性智能化检测, 为该领域未来的研究工作

提供基准参考。

3) 应用“个保法”知识图谱对隐私政策进行合规性分析,并发现了当前中文隐私政策中普遍存在的一些问题。使用提出的知识图谱对所收集的应用程序隐私政策进行覆盖率分析,发现特定类别和细粒度的概念节点覆盖率较低,应当引起个人信息处理者对相关信息保护工作的重视。

1 相关工作

1.1 “个保法”相关研究

当前,对“个保法”的研究整体仍处于起步阶段,多集中于法规解读探讨或特定场景下的问题分析。Determann等^[1]总结了“个保法”的主要要求,并就其与GDPR (general data protection regulation) 和美国各州隐私保护法案的异同点进行了讨论。Calzada^[2]对“个保法”的政策背景进行了分析,并提供了关于隐私的文献综述,还特别阐明了社会信用体系 (social credit system, SCS) 的影响。Liu等^[3]运用文献分析法分析了人工智能和物联网背景下隐私和个人信息保护相关的法律法规,特别讨论了“个保法”对智能音箱等物联网 (internet of things, IoT) 设备的潜在影响。Tan等^[4]讨论了GDPR和“个保法”(以及适用时的数据安全法 (data security law, DSL) 之间的重大差异,并创建了一个路线图,以实现与“个保法”和DSL的合规性。Li等^[5]对“个保法”的实施与企业处理用户数据合规性的影响进行了研究,并对企业日常合规管理个人数据提出了一些建议。

1.2 数据隐私知识图谱构建

知识图谱是记录实体间关系的语义知识库,可用于实体识别等自动化任务。数据隐私领域的知识图谱构建大多需要人工介入,以确保准确性和灵活性。Skinner等^[6]提供了一种协作环境 (CE) 中信息隐私的知识图谱,该知识图谱提供的知识可以用于制定更好的信息隐私政策、实践和隐私增强技术。Leone等^[7]展示了数据隐私领域的专有词汇表 (DPV),用于在隐私政策中发现与词汇表本身建模概念相关的信息,并提供了该信息的机器表示,将非结构化文本信息与正式知识图谱联系起来。Chawla等^[8]开发了一种表示隐私和数据控制信号的知识图谱,并应用于文档摘录以验证知识图谱的作用。Benjumea等^[9]提出

了一种简单的知识图谱,根据组件评估对健康领域应用程序实施隐私评估。Torre等^[10]针对GDPR构建了一个包含55个概念节点的知识图谱,并使用机器学习相关技术进行辅助分类和完整性评估。Amaral等^[11]利用模型驱动工程构建GDPR概念模型,描述GDPR中与数据处理协议 (DPA, data processing agreement) 相关的要求,还建立了DPA合规性标准供机器分析。Sanga-roonsilp等^[12]对GDPR等4个隐私保护法规进行了综合分析,提取其中蕴含的隐私需求并提出了一个知识图谱,最后在大型开源软件项目上应用知识图谱并给出分析报告。

1.3 隐私政策语料库

隐私政策语料库是进行自动化分析的前提,Wilson等^[13]以法律专家人工标注的方式创建了公开隐私政策语料库OPP-115,包含115个英文在线网站的隐私政策文档。Sathyendra等^[14]在OPP-115的基础上进一步细化,整理出关于“选择”实例的表述。Arora等^[15]对隐私政策语言版本间“披露”概念的差异展开研究,创建了英语-德语隐私政策语料库MAPP。Zimmeck等^[16]构建了APP-350隐私政策语料库,包含谷歌应用商城热门的350个App,并对文档涉及的隐私惯例进行标注。Ravichander等^[17]聚焦于用户关注的问题本身,提出了1750个隐私相关问题的隐私政策问答语料库PrivacyQA。

1.4 隐私政策分析

(1) 手工分析

传统上,隐私政策分析由专业研究人员人工实施,应用多种文本分析方法 (如可视化分析法^[18]、内容分析法^[19]等),分析周期长、成本高。Caramujo等^[20]提出了用于隐私感知的UML概要文件,并对Facebook和LinkedIn的隐私政策进行分析,展示文本中的隐私需求。Winkler等^[21]收集了4个流行平台的隐私政策,分析比较了各平台信息收集情况和文本可理解性,最后提出了用户隐私解决方案。Perez等^[22]针对IoT设备隐私政策进行审查和比较,并验证两种IoT设备行为符合其隐私政策。Parker等^[23]对61个知名心理健康App的隐私政策进行内容分析,发现近一半App未告知用户个人信息收集情况,呼吁加

强对移动 APP 隐私实践的监管。

(2) 自动化分析

利用机器学习技术进行自动化分析可以大大节省人力、物力,有着广阔的实际应用场景。Zimmeck 等^[24]提出了基于众包和自动分类技术的 Privee 架构,以浏览器扩展的方式在线分析隐私政策,达到 90% 的总体 $F1$ 分数。Torre 等^[10]应用 GDPR 知识图谱和 AI 辅助方法对基金行业的 234 份隐私政策进行分析,达到 85% 的精确率和 96% 的召回率。Hamdani 等^[25]实现了数据供应链中的合规性检查框架,并开发自动化隐私政策 GDPR 检查方法,实现端到端的合规性评估。Amaral 等^[26]使用自然语言处理与监督学习技术对 GDPR 下的隐私政策完整性进行自动化分析,达到 92.9% 的精确率和 89.8% 的召回率。

2 “个保法”隐私政策知识图谱

知识图谱是一种将信息组织成图状结构,以表示实体之间的关系和属性的方法。根据领域、数据类型、任务需求的不同,知识图谱的构建方式有领域本体构建、文本挖掘、半结构化数据转化等。本文与隐私保护领域法律专家合作,使用基于领域本体的方式构建了首个面向“个保法”的分级隐私政策知识图谱。在构建过程中,首先参考“个保法”原文、典型隐私政策样本和相关规范,然后通过分析法规文本、提取概念实体、抽取实体间关系,将法律领域的知识组织成一个分级图谱结构。构建面向“个保法”的分级隐私政策知识图谱的具体步骤如下。

1) 领域定义和范围明确。确定知识图谱范围,即个人信息保护法和规范的相关内容。

2) 实体提取。从法律文本中识别并提取出关键概念,如个人信息种类、权利与义务、数据主体等。

3) 关系抽取。分析法律文本上下文,提取实体之间的关系(如数据主体享有的权利、个人信息处理者的义务等)。

4) 本体构建。基于提取的实体和关系,构建关于隐私政策的领域本体,定义个人信息控制者需在隐私政策中声明的概念实体。

下面将对提出的“个保法”知识图谱进行详细介绍,针对其中的重要概念节点,参照法律原

文与相关法律专家意见进行解释。

考虑到隐私政策常按照一定的顺序分级叙述,每一章节叙述特定相关的术语,本文提出了一种三级知识图谱,对应“个保法”概念间层级关系。

“个保法”第 5-7 条规定了个人信息处理应遵循的原则,其中一些应该在隐私政策中声明,如个人信息处理的明确合理目的和数据最小化。因此,本文将节点“个人信息处理原则”划分为两个子节点“明确合理目的”和“数据最小化”。

“个保法”第 51 条要求个人信息处理者采取若干措施,防止未经授权的访问和披露、篡改和丢失个人信息,这些措施包含在节点“个人信息安全”中。“个保法”第 52 条规定处理个人信息达到一定数量的,应指定个人信息保护负责人,这对应“处理代表”和“个人信息保护负责人”节点。通常来说,处理者可以任选一个节点进行声明。

对于“个保法”未涉及的概念,本文为知识图谱设置了两个一级节点:“非‘个保法’要求”和“其他”。“非‘个保法’要求”是指与其他法规相关的概念,主要包含其他国家或地区的 COOKIE 法规和 数据保护法,即节点“COOKIES”和“其他法律法规”。“其他”包括所有不能被标记的情况。

本文还参考了国家信息安全标准化技术委员会发布的《信息安全技术 个人信息安全规范》,该规范为 App 服务商提供了隐私政策模板。参照模板中“响应您的请求”部分,本文在“投诉”节点下添加了 3 个子节点,分别为响应时间、所需材料和拒绝情形。

基于以上考虑,本文得到了包含 88 个概念节点的“个保法”隐私政策知识图谱,如图 1 所示。

3 隐私政策合规性检测

本节介绍了“个保法”知识图谱在隐私政策自动化检测领域的应用,实现了由语料库构建、文本特征提取和概念分类器 3 部分组成的隐私政策自动化合规性检测方案(如图 2 所示),构建了中文隐私政策合规性检测模型 CPP-BERT: 首先设计了一个半自动的隐私政策收集方法,在华为应用商城收集了来自 16 个分类的 400 个热门中文 App 的隐私政策,并抽取 100 份在段落级

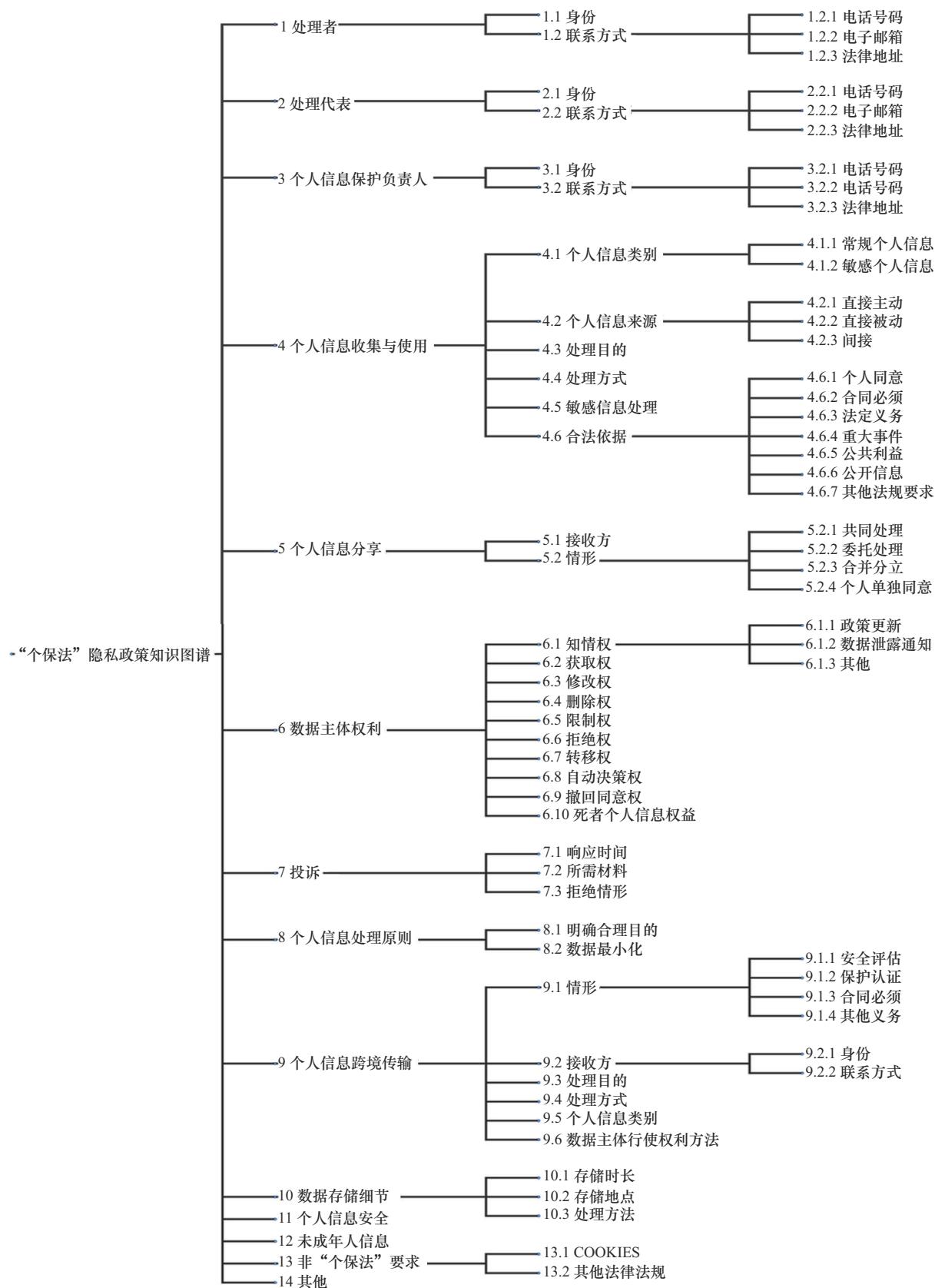


图1 “个保法”隐私政策知识图谱

Figure 1 PIPL-based privacy policy knowledge graph

别标注, 根据段落是否涉及某一知识图谱概念决定是否为该段落标注相应的概念标签, 形成首个面向“个保法”的中文隐私政策语料库 APPCP-100; 然后基于 APPCP-100 构建隐私政策合规性检测模型, 结合 TF-IDF (term frequency - inverse document frequency) 等模型提取的多维文本特征以及关键字、标题等特征训练分类模型, 对比了多种分类方法。结果表明, 基于本文提出的 APPCP-100 语料库实现的 CPP-BERT 模型表现最好, 可作为未来研究的基准参考。

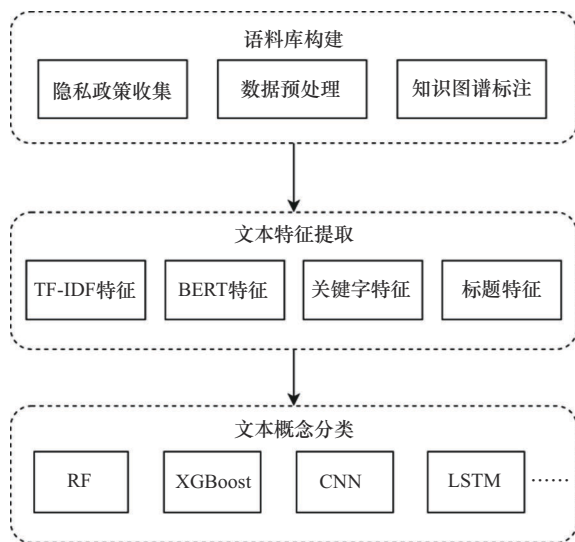


图2 隐私政策自动化合规性检测方案

Figure 2 Automated compliance checking solution for privacy policy

3.1 语料库构建

本文设计了一种高效可靠的基于 Selenium 的隐私政策半自动收集方法, 大大减少人工工作量, 并应用于华为应用市场 16 个分类共 400 个热门 App 的隐私政策收集。首先, 本文利用半自动收集方法处理嵌套 iframe、网页动态加载和反爬虫机制等复杂情况, 并根据关键字自动识别应用程序隐私政策页面。App 隐私政策收集流程如图 3 所示。

然后, 本文对收集的 HTML 文档进行数据预处理和数据结构化。数据预处理的具体方式为对 DOM 树进行标签过滤, 去除多媒体元素、嵌入非文本对象和其他语义无关的元素。为确保语料库结构的一致性, 引入了一个标准的 XML 结构, 它只包含标题和段落两种类型的节点。这种结构

有助于维护文档的原始段落结构, 为异质的 HTML 文档到通用的结构化语料库的转换提供了可靠而有效的方法。经过上述预处理流程, 隐私政策 HTML 文档得到充分的数据清洗和过滤, 得到 400 份 XML 文件组成的空白语料库。

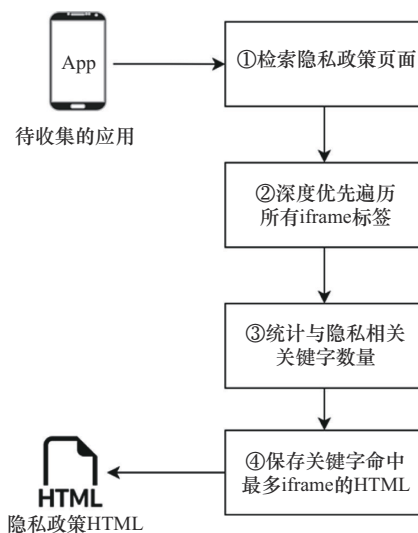


图3 App 隐私政策收集流程

Figure 3 Privacy policy collection process for APP

最后, 本文抽取 100 份 XML 文件在文本段落级别进行标注, 按照 8: 2 划分训练集与测试集, 形成“个保法”中文隐私政策语料库 APPCP-100 (APPCP-100 各级节点统计如表 1 所示)。为了确保注释过程的有效性和相关性, 应用提出的“个保法”隐私政策知识图谱来标注隐私政策的每个节点。其中, 文档的一个标题或段落作为单个样本, 可能对应知识图谱层次结构中的一个或多个节点, 故节点总数大于样本总数。为确保标注过程的准确性和可靠性, 本文参考了《信息安全技术 个人信息安全规范》提供的隐私政策模板, 并咨询了具有隐私法专业知识的法律专家。另外, 将初始标注结果分发给其他标注者以统一意见, 从而识别和纠正标注过程中的错误和标准差异, 为后续的隐私政策分析等提供可靠的基础。

表1 APPCP-100 各级节点统计

Table 1 Node statistics at all levels of APPCP-100

数据集	样本总数	一级节点	二级节点	三级节点
训练集	15 280	13 736	16 183	4 619
测试集	4 156	3 897	4 199	1 222

3.2 文本特征提取

在完成隐私政策语料库的构建后，通过文本特征提取将文本信息映射到特征空间，作为下一步概念分类器的输入。本文对比了多种特征提取方法，包括TF-IDF特征、CPP-BERT特征和基于关键字和标题的特征，得出了最优方案。

TF-IDF模型^[27]是一种自然语言处理技术，通过测量单词在文档中出现的频率（TF）和单词在语料库中的罕见度（IDF），然后将TF和IDF相乘，得到一个词在文档中的权重。

$$TF_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \quad (1)$$

$$IDF_i = \log \frac{|D|}{|j:t_i \in d_j|} \quad (2)$$

$$TFIDF_{i,j} = TF_{i,j} \cdot IDF_i \quad (3)$$

其中， $n_{i,j}$ 为单词 t_i 在文档 d_j 中的出现次数， $|D|$ 为总文档数， $|j:t_i \in d_j|$ 为包含单词 t_i 的文档数。

对于TF-IDF向量化，首先创建包含隐私政策文档集合中所有唯一单词词汇表；然后计算词频和逆文档频，将TF-IDF模型应用于文本样本；最后根据TF-IDF分数为每条文本生成一个向量，表示每个单词的得分。在训练集上拟合并在整个语料库上实施TF-IDF向量化，得到一个包含所有文本嵌入的TF-IDF矩阵，作为分类器的输入特征。

BERT模型^[28]是谷歌提出的预训练语言模型，基于Transformer^[29]编码器结构，使用自注意力机制来捕捉输入序列中的长距离依赖关系，可用于文本分类等下游任务。其具有双向性，同时考虑某单词前后上下文，以更好地理解单词在句中的含义。使用时，首先在大规模文本语料上进行预训练，通过遮盖单词预测任务学习单词间关系，然后根据具体的下游任务执行微调，在标注数据集上有监督学习，适应特定任务要求。具体实现步骤如下。

1) 预训练模型选择。本文选择谷歌通用中文预训练模型BERT-Base-Chinese，其在各个领域有着较好的泛化能力。

2) 数据预处理。使用BERT提供的中文词表分词，并添加特殊的[CLS]标记，用于下游文本

分类任务。

3) 领域迁移。在300份隐私政策语料上继续训练5轮，使用Adam优化器^[30]，设置超参数 $batchsize = 8$ ， $lr = 10^{-5}$ ，得到隐私政策领域迁移模型CPP-BERT（Transfer）。

4) 下游任务微调。将BERT模型最后一层的[CLS]标记隐藏状态池化输出连接至全连接层，分别对BERT-Base-Chinese和CPP-BERT（Transfer）预训练模型微调10轮，使用交叉熵损失函数，不改变其他配置。微调后得到最终的BERT文本特征提取模型CPP-BERT和CPP-BERT（Transfer）。

此外，本文还测试了一些手工特征，包括关键字特征和标题特征。根据标注人员标注语料库的经验总结，对每个概念节点都给出一个或多个关键字。例如，对“未成年人信息”节点设置关键字“未成年”和“儿童”。对每个文本节点，遍历各个类别的关键字列表进行匹配，得到88维的0-1向量，作为关键字特征。

考虑隐私政策的层级结构，标题多为段落的总结，建立的结构化语料库包含标题和段落两类节点。利用XML结构可以获取当前段落节点对应的标题节点，然后引入标题节点特征辅助判断段落节点类别。

3.3 概念分类器设计

在概念分类器部分，对现有的多种机器学习分类模型和深度学习方法进行了充分的评估，包括随机森林（RF）^[31]、极限树（ET）^[32]、XG-Boost^[33]、逻辑回归（LR）^[34]、朴素贝叶斯（GNB）^[35]、K-近邻（KNN）^[36]、支持向量机（SVC）^[37]、决策树（DT）^[38]、CNN^[39]和LSTM^[40]。实验设置上，传统机器学习方法使用Scikit-learn^[41]中的默认参数；对于深度学习方法，CNN为单层1维卷积网络，过滤器数量为64，卷积核尺寸为3，而双层LSTM包含64个隐藏单元，训练时均使用Adam优化器， $lr = 10^{-3}$ ， $batchsize = 128$ ，迭代10轮后收敛。

首先，为“个保法”知识图谱中的每个概念节点训练二分类器，将包含该节点的样本标记为正样本，其他节点标记为负样本，最后合并二分类预测得到多标签分类预测。训练阶段，由于正

负样本数量不均衡,为防止模型过拟合,应用欠采样和数据增强方法控制正负样本比例为 1:2,并保证总样本数大于 100 条。

3.4 评价指标

实验采用 marco-F1 分数 (F1-Score)、准确率 (Accuracy) 和 ROC 曲线下面积 (AUC) 作为模型性能评价指标。

$$F1\text{-Score} = \frac{1}{n} \sum_{k=1}^n \frac{2TP_k}{2TP_k + FP_k + FN_k} \quad (4)$$

$$Accuracy = \frac{1}{n} \sum_{k=1}^n \frac{TP_k + TN_k}{TP_k + TP_k + FN_k + TN_k} \quad (5)$$

$$AUC = \frac{1}{2n} \sum_{k=1}^n \sum_{i=1}^{m_k-1} (x_{k_{i+1}} - x_{k_i})(y_{k_{i+1}} + y_{k_i}) \quad (6)$$

其中, n 为总类别数, TP_k 、 FP_k 、 FN_k 、 TN_k 代表第 k 类的混淆矩阵指标, (x_{k_i}, y_{k_i}) 为第 k 类的 ROC 曲线上第 i 个样本点, m_k 为第 k 类的样本总数。

3.5 实验及结果分析

3.5.1 实验目的

本文实验旨在评估文本特征模型性能与概念分类器性能,证明本文提出的隐私政策合规性检测方案 CPP-BERT 的有效性。

3.5.2 数据集选择

实验使用的数据集为第 3.1 节中构建的 APPCP-100 语料库,共包括从 100 份隐私政策中抽取的 19 436 条标题或段落样本,均有人工标注的“个保法”知识图谱概念实体标签。数据集划分时随机抽取 20 份隐私政策作为测试集,其余的隐私政策作为训练集。

3.5.3 实验结果描述

特征模型评估实验统一使用 RF 分类器,将谷歌 BERT 预训练模型作为性能基线。特征模型评估结果如表 2 所示,TF-IDF 模型各指标略优于基线结果, CPP-BERT 模型表现优于 TF-IDF 模型和基线 ($F1$ 分数为 0.831 1,准确率达到 0.650 6),而 CPP-BERT (Transfer) 模型与 CPP-BERT 表现相近 ($F1$ 分数和 AUC 指标上略有提升,但准确率略有下降)。此外,在手工特征中,关键字特征分别为 TF-IDF 和 CPP-BERT 特征带来 0.027 5 和 0.004 8 的准确率提升,标题特征则使原有模型表现略有下降。

概念分类器评估实验在 TF-IDF 和 CPP-BERT 两个特征模型上进行,对比不同分类器的性能差异。概念分类器评估结果如表 3 所示,TF-IDF 下 ET 分类器的表现最好 ($F1$ 分数为 0.642 7),CPP-BERT 下 RF 分类器的表现最好 ($F1$ 分数为 0.831 1);而深度学习算法 CNN 和 LSTM 表现一般, $F1$ 分数和准确率在两种特征模型上都不如 LR 分类器。

3.5.4 实验结果分析

特征模型评估实验中, CPP-BERT 模型表现明显优于 TF-IDF 模型,这是因为 TF-IDF 难以捕捉单词间上下文关系, CPP-BERT 通过学习句子语义,可以更有效地提取隐私概念信息。关于手工设计特征,关键字特征能够有效地提升原特征模型的表示能力,但对 CPP-BERT 的提升小于对 TF-IDF 的提升,说明 CPP-BERT 已提取文本中隐含的与隐私相关的关键字信息,因此附加关键字

表 2 特征模型评估结果

Table 2 Evaluation results of text representation models

特征	F1 分数	Accuracy	AUC
TF-IDF	0.629 2	0.390 0	0.836 1
TF-IDF+Keyword	0.673 8	0.417 5	0.849 9
TF-IDF+Title	0.621 6	0.346 7	0.815 7
CPP-BERT	0.831 1	0.650 6	0.921 7
CPP-BERT+Keyword	0.833 1	0.655 4	0.922 6
CPP-BERT +Title	0.810 8	0.601 8	0.922 0
CPP-BERT (Transfer)	0.833 1	0.648 7	0.927 5
BERT (Baseline)	0.622 4	0.311 6	0.798 4
Keyword	0.512 5	0.181 2	0.825 1

表3 概念分类器评估结果

Table 3 Evaluation results of concept classifiers

特征	模型	F1-Score	Accuracy	AUC
TF-IDF	RF	0.629 2	0.390 0	0.836 1
	ET	0.642 7	0.406 6	0.836 7
	XGBoost	0.615 2	0.328 2	0.844 1
	LR	0.491 4	0.219 2	0.756 4
	SVC	0.414 8	0.300 0	0.800 1
	KNN	0.256 5	0.033 9	0.756 1
	GNB	0.158 0	0.036 1	0.741 3
	DT	0.438 1	0.126 6	0.806 7
	CNN	0.471 5	0.138 6	0.729 8
	LSTM	0.487 9	0.176 4	0.779 6
CPP-BERT	RF	0.831 1	0.650 6	0.921 7
	ET	0.829 3	0.646 3	0.922 9
	XGBoost	0.826 8	0.640 0	0.924 3
	LR	0.821 6	0.629 7	0.925 5
	SVC	0.821 3	0.625 1	0.926 9
	KNN	0.800 6	0.594 8	0.925 8
	GNB	0.794 5	0.515 6	0.923 3
	DT	0.784 7	0.545 7	0.913 2
	CNN	0.790 0	0.541 9	0.909 4
	LSTM	0.823 3	0.626 1	0.925 9

特征对性能提升的作用有限。标题特征无法带来更好的结果，这可能是因为段落与标题隐私分类不完全一致，段落中蕴含更细分概念，标题特征增加了细粒度概念的区分难度。

概念分类器对比实验中，RF、ET等集成方法通过组合基学习器得到更强的模型，泛化能力好，在TF-IDF和CPP-BERT下均表现出色；但深度学习方法中，CNN适用于空间结构特征，LSTM适用于序列特征，而TF-IDF和CPP-BERT特征均不属于上述类别，使用场景不匹配，表现一般。

3.6 小结

本节应用“个保法”知识图谱，建立从隐私政策语料库APPCP-100的收集、清洗与标注，到文本特征提取和概念分类的隐私政策自动合规性检测框架。

实验方面，首先研究对比TF-IDF、CPP-BERT两种特征表示模型，并引入标题特征、关键字特征作为辅助，在隐私政策概念分类任务上

评估各方案的表现。结果显示，CPP-BERT与关键字特征组合为最优方案（F1分数为0.833 1和准确率为0.655 4）。然后对比测试RF、ET、XGBoost、LR、GNB、KNN、SVC、DT这8种现有机器学习分类方法和CNN、LSTM两种深度学习模型，分别在TF-IDF和CPP-BERT特征表示下评估分类效果，发现CPP-BERT特征下各分类器表现普遍优于TF-IDF特征，而RF、ET和XGBoost分类模型的性能均高于其他模型。

4 隐私政策合规性分析

本节介绍提出的“个保法”知识图谱的实际应用，对语料库APPCP-100进行概念节点覆盖率分析，揭示当前众多隐私政策广泛存在的合规性问题。

4.1 各级节点覆盖率分析

依次统计各级“个保法”概念节点的覆盖率，得到如图4所示的条形图。其中，横坐标对应图1

知识图谱的概念节点号, 纵坐标代表覆盖此概念节点的隐私政策比例。对于一级节点, 合并了“处理代表”和“个人信息保护负责人”两个节点, 这是因为处理者可以任选其一进行声明。对于二、三级节点, 选择了具有代表性的子节点并给出覆盖率结果, 保证图表简洁同时保留关键信息。

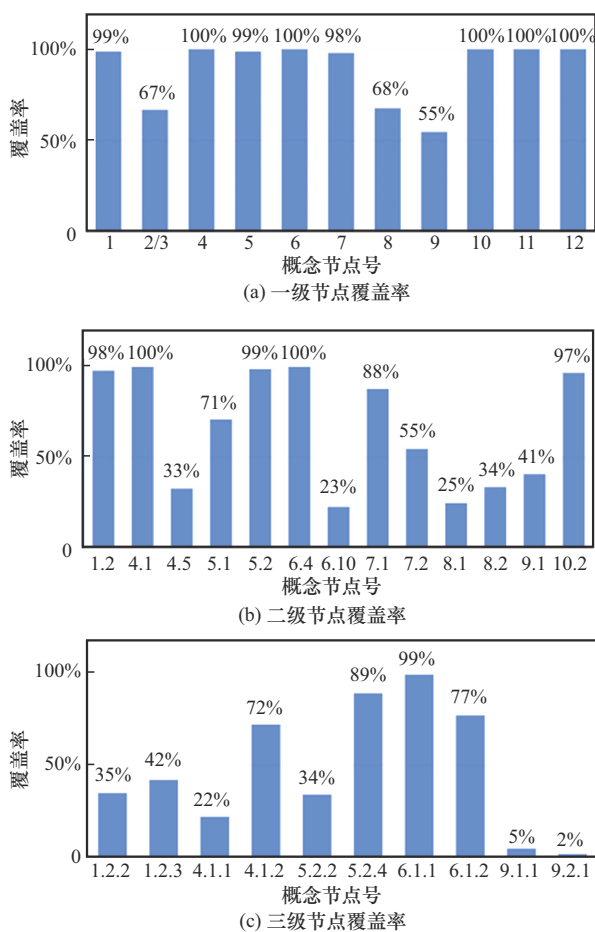


图4 热门中文App隐私政策覆盖率分析

Figure 4 Coverage of popular Chinese App' privacy policies

由图4(a)可知, 大多数一级节点在所有隐私政策有涉及, 但“处理代表”和“个人信息保护负责人”节点(2/3)约有1/3的隐私政策未进行声明, 这可能是因为部分应用未达到用户量巨大的标准, 暂时未设立专门的个人信息保护机构。此外, “个人信息处理原则”和“个人信息跨境传输”节点(8、9)的覆盖率也较低, 这说明目前一些处理者对个人信息处理的指导原则和可能出现的隐私信息跨境情况认识不足。

如图4(b)、图4(c)所示, 二、三级节点的覆

盖率分布差异非常显著, 部分节点在样本中被广泛忽视。例如, 在信息收集相关节点中, “个人信息类别”节点(4.1)在所有样本均有说明, 而“敏感信息处理”节点(4.5)覆盖率仅为33%, 这要求处理者重视个人敏感信息并作特别提示。值得注意的是, “个保法”第49条明确提到“死者个人信息权益”这一新概念, 但此概念节点(6.10)在样本中很少涉及。另外, 在投诉相关节点中, “所需材料”节点(7.2)是最容易被忽略的概念, 这会使用户投诉更加困难。

表5展示了语料库中各级节点的平均覆盖率, 从表中可以看出, 随着概念粒度的减小, 概念节点的平均覆盖率呈下降趋势。多数样本可以覆盖宽泛的一级概念节点, 但对更细粒度的“个保法”概念, 平均只有一半左右的节点会被提及。因此, 建议个人信息处理者注意完善“个保法”细粒度概念的要求并做好个人信息保护相关措施和说明。

表5 各级节点的平均覆盖率

Table 5 Average coverage of nodes at all levels

节点	平均覆盖率
一级节点	88.08%
二级节点	58.23%
三级节点	49.09%

4.2 App类别覆盖率分析

不同类别App的用户群体可能有极大的差异, 在线服务提供者在个人信息保护力度上也存在着显著区别, 通过概念节点覆盖率对比可见一二。

不同类别App覆盖率分析如图5所示, 本文选取4个有代表性的二级概念节点, 统计16个类别App的覆盖率情况, 按照代表节点平均覆盖率降序排列。整体来看, 金融理财类和影音娱乐类App的平均覆盖率最高, 说明金融类服务对数据安全的要求普遍高于其他类服务; 而影音娱乐包括短视频、音乐、直播等诸多细分子类, 总体用户规模大, 应用使用率高, 要求企业更加重视个人信息保护。与之相对, 拍摄美化类和儿童类App的平均覆盖率最低。研究发现, 拍摄美化类App主要围绕滤镜、贴纸等多种相机功能打造服务, 使用场景单一, 且涉及的个人信息较少。儿童类App

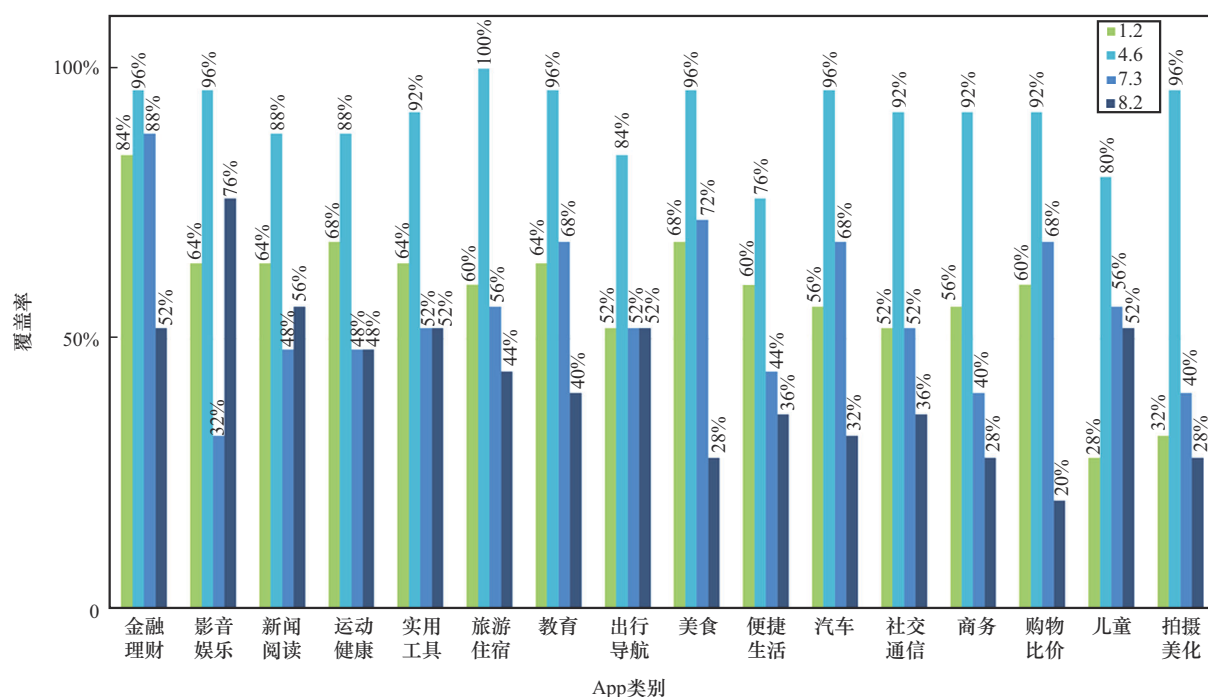


图5 不同类别App覆盖率分析

Figure 5 Coverage of App in different categories

情况类似，仅涉及儿歌、绘本等内容服务，用户规模相对较小，导致厂商对隐私政策不够重视。

5 结束语

本文通过调查现有的“个保法”和隐私政策合规性研究工作，提出了一种全面的面向“个保法”的三级隐私政策知识图谱，并与法律专家合作构建了相关的概念和通用条款。该知识图谱可作为隐私政策分析的基础，用于构建隐私政策语料库 APPCP-100、合规性自动化检测模型 CPP-BERT 和进行合规性分析。本文的主要贡献包括知识图谱的提出、中文 APP 隐私政策语料库 APPCP-100 的建立以及合规性检测模型 CPP-BERT 的构建，为未来该领域的研究提供了基准参考，促进了隐私政策合规性检测分析的发展。随着技术的不断进步和法律的完善，希望该领域的研究能够进一步推动个人隐私保护的发展，确保个人信息得到有效保护。

参考文献：

- [1] DETERMANN L, RUAN Z J, GAO T, et al. China's draft personal information protection law[J]. Journal of Data Protection & Privacy, 2021, 4(3): 235-259.
- [2] CALZADA I. Citizens' data privacy in China: the state of the art of the personal information protection law (PIPL) [J]. Smart Cities, 2022, 5(3): 1129-1150.
- [3] LIU Y L, HUANG L, YAN W, et al. Privacy in AI and the IoT: the privacy concerns of smart speaker users and the personal information protection law in China[J]. Telecommunications Policy, 2022, 46 (7): 102334.
- [4] TAN Z, ZHANG C. China's PIPL and DSL: is China following the EU's approach to data protection[J]. Journal of Data Protection & Privacy, 2021, 5(1): 7-25.
- [5] 李永红, 窦文珠. 试论《个人信息保护法》对企业合规的影响及建议[J]. 中外企业文化, 2022(8): 96-98.
- [6] LI Y H, DOU W Z. On the impact and suggestions of the personal information protection law on enterprise compliance[J]. Chinese & Foreign Corporate Culture, 2022(8): 96-98.
- [7] SKINNER G, HAN S, CHANG E. An information privacy taxonomy for collaborative environments[J]. Information Management & Computer Security, 2006, 14(4): 382-394.
- [8] LEONE V, DI-CARO L. The role of vocabulary mediation to discover and represent relevant information in privacy policies[J]. Legal Knowledge and Information Systems, 2020: 73-82.
- [9] CHAWLA K, HULSTIJN J. A taxonomy for the representation of privacy and data control signals[J]. Legal Knowledge and Information Systems, 2020: 23-32.
- [9] BENJUMEA J, ROPERIO J, RIVERA-ROMERO O, et al. Privacy

- assessment in mobile health Apps: scoping review[J]. *JMIR mHealth and uHealth*, 2020, 8(7): e18868.
- [10] TORRE D, ABUALHAIJA S, SABETZADEH M, et al. An AI-assisted approach for checking the completeness of privacy policies against GDPR[C]//2020 IEEE 28th International Requirements Engineering Conference (RE). 2020: 136-146.
- [11] AMARAL O, ABUALHAIJA S, SABETZADEH M, et al. A model-based conceptualization of requirements for compliance checking of data processing against GDPR[C]//2021 IEEE 29th International Requirements Engineering Conference Workshops. 2021: 16-20.
- [12] SANGAROONSILP P, DAM H K, CHOETKIERTIKUL M, et al. A taxonomy for mining and classifying privacy requirements in issue reports[J]. *Information and Software Technology*, 2023, 157: 107162.
- [13] WILSON S, SCHAUB F, DARA A, et al. The creation and analysis of a website privacy policy corpus[C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. 016: 1330 - 1340.
- [14] SATHYENDRA K, WILSON S, SCHAUB F, et al. Identifying the provision of choices in privacy policy text[C]// Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. 2017: 2774 - 2779.
- [15] ARORA S, HOSSEINI H, UTZ C, et al. A tale of two regulatory regimes: creation and analysis of a bilingual privacy policy corpus[C]// Proceedings of the Thirteenth Language Resources and Evaluation Conference. 2022: 5460-5472.
- [16] ZIMMECK S, STORY P, SMULLEN D, et al. MAPS: scaling privacy compliance analysis to a million Apps[C]// Proceedings on Privacy Enhancing Technologies, 2019:66-86.
- [17] RAVICHANDER A. Question answering for privacy policies: combining computational and legal[C]//Proceedings of the 2019 Conference on Empirical Methods in Natural. Association for Computational Linguistics, 2019: 4947-4958.
- [18] 唐家渝, 刘知远, 孙茂松. 文本可视化研究综述[J]. *计算机辅助设计与图形学学报*, 2013, 25(3): 273-285.
- TANG J Y, LIU Z Y, SUN M S. A survey of text visualization[J]. *Journal of Computer-Aided Design & Computer*, 2013, 25(3): 273-285.
- [19] KUCKARTZ U. Qualitative text analysis: a guide to methods, practice and using software[J]. *Qualitative Text Analysis*, 2013: 1-192.
- [20] CARAMUJO J, DA SILVA A M R. Analyzing privacy policies based on a privacy-aware profile: the Facebook and LinkedIn case studies[C]//2015 IEEE 17th Conference on Business Informatics. 2015: 77-84.
- [21] WINKLER S, ZEADALLY S. Privacy policy analysis of popular web platforms[J]. *IEEE Technology and Society Magazine*, 2016, 35(2): 75-85.
- [22] PEREZ A J, ZEADALLY S, COCHRAN J. A review and an empirical analysis of privacy policy and notices for consumer internet of things[J]. *Security and Privacy*, 2018, 1(3): e15.
- [23] PARKER L, HALTER V, KARLIYCHUK T, et al. How private is your mental health app data? an empirical study of mental health app privacy policies and practices[J]. *International Journal of Law and Psychiatry*, 2019, 64: 198-204.
- [24] ZIMMECK S, BELLOVIN S M. Privee: an architecture for automatically analyzing web privacy policies[C]//23rd USENIX Security Symposium (USENIX Security 14). 2014: 1-16.
- [25] HAMDANI R E, MUSTAPHA M, AMARILES D R, et al. A combined rule-based and machine learning approach for automated GDPR compliance checking[C]//Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law. 2021: 40-49.
- [26] AMARAL O, ABUALHAIJA S, TORRE D, et al. AI-enabled automation for completeness checking of privacy policies[J]. *IEEE Transactions on Software Engineering*, 2022, 48(11): 4647-4674.
- [27] RAMOS J. Using TF-IDF to determine word relevance in document queries[C]//Proceedings of the First Instructional Conference on Machine Learning: 2003: 29-48.
- [28] DEVLIN J, CHANG M W, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding[J]. *arXiv Preprint arXiv:1810.04805*, 2018.
- [29] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Advances in Neural Information Processing Systems. 2017.
- [30] KINGMA D P, BA J. Adam: a method for stochastic optimization[J]. *arXiv Preprint arXiv:1412.6980*, 2014.
- [31] HO T K. Random decision forests[C]//Proceedings of 3rd International Conference on Document Analysis and recognition. 1995: 278-282.
- [32] GEURTS P, ERNST D, WEHENKEL L. Extremely randomized trees[J]. *Machine Learning*, 2006, 63:3 - 42.
- [33] CHEN T, GUESTRIN C. XGBoost: a scalable tree boosting system[J]. *Proceedings of the 22nd International Conference on Knowledge Discovery and Data Mining*. 2016, 11:785-794.
- [34] DREISEITL S, OHNO-MACHADO L. Logistic regression and artificial neural network classification models: a methodology review[J]. *Journal of Biomedical Informatics*, 2002, 35(5-6): 352-359.
- [35] WEBB G I, KEOGH E, MIIKKULAINEN R. Naïve Bayes.[J]. *Encyclopedia of Machine Learning*, 2010, 15(1): 713-714.
- [36] FIX E, HODGES J L. Discriminatory analysis. nonparametric discrimination: consistency properties[J]. *International Statistical Review/Revue Internationale de Statistique*, 1989, 57(3): 238-247.

- [37] MOFUERZA J, MUNOZ A. Support vector machines with applications[J]. Statistical Science, 2006, 21(3):322 - 336.
- [38] QUINLAN J R. Induction of decision trees[J]. Machine Learning, 1986, 1(1): 81-106.
- [39] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. ImageNet classification with deep convolutional neural networks[J]. Communications of the ACM, 2017, 60(6): 84-90.
- [40] GREFF K, SRIVASTAVA R K, KOUTNÍK J, et al. LSTM: a search space odyssey[J]. IEEE Transactions on Neural Networks and Learning Systems, 2016, 28(10): 2222-2232.
- [41] PEDREGOSA F, VAROQUAUX G, GRAMFORT A, et al. Scikit-learn: machine learning in Python[C]//The Journal of Machine Learning Research. 2011: 2825-2830.

[作者简介]



张西珩（1999-），男，山东聊城人，上海交通大学硕士生，主要研究方向为联邦学习、隐私保护。



李昕（1999-），男，江苏宿迁人，上海交通大学硕士生，主要研究方向为自然语言处理、隐私保护。



唐鹏（1992-），男，江西抚州人，上海交通大学助理研究员，主要研究方向为人工智能安全、隐私保护。



黄锐奇（1999-），男，江苏张家港人，上海交通大学硕士生，主要研究方向为数据法、个人信息保护法。



何渊（1979-），男，浙江浦江人，法学博士，上海交通大学副教授，主要研究方向为数据治理与合规、个人信息保护、网络安全法及行政法等。



邱卫东（1973-），男，江西九江人，博士，上海交通大学教授、博士生导师，主要研究方向为计算机取证、密码分析、人工智能安全、大数据隐私保护。