

面向 GDPR 隐私政策合规性的智能化检测方法

李昕¹, 唐鹏¹, 张西珩¹, 邱卫东¹, 回红²

(1. 上海交通大学网络空间安全学院, 上海 200240; 2. 上海交通大学网络安全技术研究院, 上海 200240)

摘要: 欧盟《通用数据保护条例 (GDPR, general data protection regulation)》自 2018 年施行以来, 已开出罚单 300 多起, 其中不乏谷歌这类知名企业未能提供透明易懂的隐私政策而遭受巨额处罚。这项严格的数据保护法律使得各国企业在提供跨境服务特别是向欧盟地区提供服务时变得尤为谨慎。同时其管辖范围规定, GDPR 适用于任何为欧盟公民提供服务的企业, 无论其是否在欧盟境内注册, 这意味着世界各地涉及海外业务的企业都要考虑其隐私政策面向 GDPR 的合规性, 国内企业也不例外。面向这一需求, 构建了一套智能化检测方法, 自动提取各在线服务企业的隐私政策, 并采用机器学习和自动化技术, 将其转化为具有结构层次的标准格式。之后进行基于自然语言处理的文本分类, 识别其中涵盖的相应的 GDPR 概念, 并以搭建的 GDPR 知识图谱为依据, 检验隐私政策是否缺少部分 GDPR 要求披露的概念, 从而实现面向 GDPR 的隐私政策合规性智能化检测, 为国内企业向欧盟用户提供跨境服务提供支撑。对话料库中样本的分析结果进一步揭示了主流在线服务企业普遍未达到 GDPR 合规要求的现状。

关键词: 通用数据保护条例; 隐私政策; 层级结构; 合规性检测

中图分类号: TP393

文献标志码: A

DOI: 10.11959/j.issn.2096-109x.2023088

GDPR-oriented intelligent checking method of privacy policies compliance

LI Xin¹, TANG Peng¹, ZHANG Xiheng¹, QIU Weidong¹, HUI Hong²

1. School of Cyberspace Security, Shanghai Jiao Tong University, Shanghai 200240, China

2. Institute of Cyber Science and Technology, Shanghai Jiao Tong University, Shanghai 200240, China

Abstract: The implementation of the EU's General Data Protection Regulation (GDPR) has resulted in the imposition of over 300 fines since its inception in 2018. These fines include significant penalties for prominent companies like Google, which were penalized for their failure to provide transparent and comprehensible privacy policies. The GDPR, known as the strictest data protection laws in history, has made companies worldwide more

收稿日期: 2023-04-06; **修回日期:** 2023-07-04

通信作者: 邱卫东, qiuwd@sjtu.edu.cn

基金项目: 国家自然科学基金 (61972249); 国家重点研发计划 (2023YFB3106500)

Foundation Items: The National Natural Science Foundation of China (61972249), The National Key R&D Program of China (2023YFB3106500)

引用格式: 李昕, 唐鹏, 张西珩, 等. 面向 GDPR 隐私政策合规性的智能化检测方法[J]. 网络与信息安全学报, 2023, 9(6): 127-139.

Citation Format: LI X, TANG P, ZHANG X H, et al. GDPR-oriented intelligent checking method of privacy policies compliance[J]. Chinese Journal of Network and Information Security, 2023, 9(6): 127-139.

cautious when offering cross-border services, particularly to the European Union. The regulation's territorial scope stipulates that it applies to any company providing services to EU citizens, irrespective of their location. This implies that companies worldwide, including domestic enterprises, are required to ensure compliance with GDPR in their privacy policies, especially those involved in international operations. To meet this requirement, an intelligent detection method was introduced. Machine learning and automation technologies were utilized to automatically extract privacy policies from online service companies. The policies were converted into a standardized format with a hierarchical structure. Through natural language processing, the privacy policies were classified, allowing for the identification of relevant GDPR concepts. In addition, a constructed GDPR taxonomy was used in the detection mechanism to identify any missing concepts as required by GDPR. This approach facilitated intelligent detection of GDPR-oriented privacy policy compliance, providing support to domestic enterprises while they provided cross-border services to EU users. Analysis of the corpus samples reveals the current situation that mainstream online service companies generally fail to meet GDPR compliance requirements.

Keywords: GDPR, privacy policy, hierarchical structure, compliance checking

0 引言

为了向用户提供个性化的服务,在线服务企业通常会收集用户的个人数据。对于通过互联网提供的在线服务,其网站上的隐私政策仍然是在线用户了解服务提供商如何收集其个人数据的主要文件类型。典型的隐私政策描述了在线服务收集和处理个人数据的不同方面,如收集哪些个人数据、如何收集这些数据、为什么收集这些数据、如何保护这些数据、如何存储这些数据,以及与第三方共享哪些数据。虽然开始使用在线服务时,接受隐私政策的内容通常是强制性的,但大多数用户倾向于忽略阅读隐私政策,因为它们通常太长,而且使用了正式的措辞,很难理解^[1]。尽管许多国家有数据保护的法律法规,但人们发现,服务提供商的隐私政策往往不完全符合这些法律法规,导致网络用户、研究人员和隐私倡导者的担忧^[2]。例如,2019年,谷歌未能在其数据同意政策中提供透明和可理解的信息而被法国数据保护机构罚款5 000万欧元^[3]。

欧盟《通用数据保护条例(GDPR, general data protection regulation)》于2016年通过,并于2018年5月25日在整个欧盟和欧洲经济区生效。GDPR协调了欧盟、欧洲经济区、英国的数据隐私法律,被广泛视为全球数据保护立法的基准。GDPR为数据控制方和数据处理方定义了许多原则和要求,以便符合法律要求。例如,根据GDPR的透明度原则,数据主体有权获知其个人数据的收集和处理情况。满足这一要求的常用方法是向

数据主体提供一份隐私政策文件,告知数据主体根据GDPR有权了解的所有重要信息。GDPR要求每个国家都有一个国家权威机构来负责GDPR的执行,这些机构经常向数据控制者和数据处理者发布关于如何符合GDPR的指南。例如,信息专员办公室(ICO, Information Commissioner's Office)作为负责GDPR执行的英国国家数据保护机构,提供了一份关于GDPR的一般指南和模板,供数据控制者用作GDPR合规的隐私政策参考。为了履行其在GDPR方面的法律义务,许多机构更新了其隐私政策以符合GDPR。例如,《纽约时报》在GDPR生效的前一天,即2018年5月24日更新了其隐私政策,加入了关于国际数据传输的条款。

同时,GDPR的管辖范围规定,GDPR适用于任何为欧盟公民提供服务的公司,无论其是否在欧盟境内注册。因此,对于开展了海外业务的国内企业,其向欧盟用户提供的隐私政策必须满足GDPR的各项要求,这意味着对国内企业的隐私政策进行面向GDPR的分析是有意义的。它可以帮助增强用户对自身重要数据保护问题的意识,还可以帮助各企业改进其隐私政策,更好地履行其在GDPR方面的法律义务。这种分析显然可以由经验丰富的GDPR专家进行定性分析,但在许多应用程序中,为了减少所花费的时间和成本,自动化分析更受欢迎。许多研究者研究了如何增强面向GDPR合规性的隐私政策分析的自动化,但本文注意到一个主要差距:过去的大多数研究将隐私政策视为一个线性的句子和段落集合,

其中的层次信息没有得到充分考虑。

针对这一情况,本文通过将隐私政策中的层次信息添加到面向GDPR合规性的隐私政策分析中,显著拓展了之前的工作。本文的工作是从搭建GDPR知识图谱开始的,该知识图谱旨在用GDPR中的各项概念分析隐私政策。本文开发了一个面向GDPR的有层次的隐私政策收集框架(APPC, architecture of privacy policy collection),收集样本并构建了一个涵盖150份隐私政策的新语料库,其中包含人工验证的层次信息和人工标注的GDPR概念。本文依据该语料库设计和训练了相应的GDPR概念分类器,用于处理隐私政策中的标题和段落,以自动检索相关的GDPR概念。本文对语料库中的样本进行了针对各GDPR概念的涵盖分析,以及参考ICO模板的合规性分析,即语料库中的150份隐私政策如何涵盖ICO推荐的更重要的GDPR概念。

本文的主要贡献概括如下。

1) 提出了一个分层的知识图谱,以促进面向GDPR的隐私政策分析。GDPR知识图谱来源于ICO提供的隐私政策模板以及GDPR法规原文。

2) 构建了一个面向GDPR相关研究的层次化隐私政策语料库,并开发了APPC,以进一步扩充相关的语料库。本文的语料库包括从顶级网站排名中收集的150份隐私政策。该语料库保存了所有隐私政策的人工验证过的层次信息,并包含基于GDPR知识图谱的人工标注的GDPR概念标签。APPC具有高水平的自动化,有助于隐私政策的大规模自动收集。

3) 开发了相应的GDPR概念分类器。使用新构建的语料库,本文开发和训练了相应的GDPR概念分类器。由于考虑了从隐私政策的层次结构中提取的特征,分类器的测试性能较以往工作有显著提升。

4) 通过分析语料库样本,本文对隐私政策的GDPR相关行为有了新的认识,包括总体上未能达到GDPR合规要求的新证据,以及GDPR知识图谱中的概念如何被隐私政策提及。

1 相关工作

本节介绍4个密切相关领域的相关工作:GDPR要求分析、隐私政策语料库、面向GDPR的隐私政

策合规性分析,以及隐私政策分析的其他方面。

(1) GDPR要求分析

Tom等^[4]提出了初步的GDPR模型,简单直观地概述了GDPR中不同概念之间的关系。Lippi等^[5]为适用GDPR的隐私政策定义了33种元数据类型,并采用基于人工规则和基于机器学习的方法进行隐私政策的歧义检测。Torre等^[6]做了一项更全面的工作,提出了55种元数据类型,涵盖了Lippi等工作中的所有类型,并利用自然语言处理(NLP, natural language processing)和机器学习的组合,研究了隐私政策完整性的自动检测。Amaral等^[7]进一步改进了隐私政策合规性检测系统,使用了相同的GDPR知识图谱和分类器,这是本文GDPR概念分类器的基准。

(2) 隐私政策语料库

2016年,Wilson等^[8]聘请3名法律专家进行人工标注,创建了名为“在线隐私政策”(OPP-115)的公共隐私政策语料库。2017年,Sathyendra等^[9]基于OPP-115构建了一个更细粒度的语料库,重点关注了隐私政策中的“选择-退出”概念。2019年,Zimmeck等^[10]选择了谷歌应用商店中最受欢迎的350个应用程序的隐私政策,创建了App-350隐私政策语料库。2020年,Srinath等^[11]创建了一个包含超过100万份英文网站隐私政策的自动构建语料库PrivaSeer。Kuznetsov等^[12]在2022年使用一种识别物联网设备隐私政策统一资源定位符(URL, uniform resource locator)的技术,构建了一个包含592个隐私政策的新语料库。

(3) 面向GDPR的隐私政策合规性分析

Tesfay等^[13]提出了基于机器学习的方法,使用预定义的关键字对多类隐私政策内容进行分类。Nejad等^[14]提出了一个名为KnIGHT的工具,利用单词之间的语义相似性将隐私政策中的句子与GDPR文本中的相关段落关联起来。Torre等^[15]使用统一建模语言和对象约束语言构建GDPR的统一建模语言表示。Hamdani等^[16]在数据供应链中实现了以文档为中心的遵从性检查,并开发了具体方法来自动检查隐私政策的GDPR合规性。

(4) 隐私政策分析的其他方面

为了提高隐私政策的透明度,Zimmeck等^[17]

提出了 Privee, 基于众包和自动分类技术对基本隐私政策术语进行分析。Harkous 等^[18]提出了一种支持多维隐私政策分析的综合系统 Polisis, 结合使用 NLP 和深度学习技术从隐私政策中提取片段并添加一组描述数据处理行为的标签, 为隐私政策分析提供在线服务。Ayala-Rivera 等^[19]提出了一种基于模型的方法, 以帮助在线服务理解 GDPR 的数据保护义务。Degeling 等^[20]对 6 759 个网站的隐私政策和 Cookie 同意通知进行了纵向分析。Linden 等^[21]对比了在线服务在 GDPR 生效前后提供的不同隐私政策的快照, 发现在 GDPR 生效后, 欧盟内外的在线服务提供的隐私政策文本发生了显著变化, 表明 GDPR 已成为隐私政策重大变化的催化剂。

2 面向 GDPR 的隐私政策语料库

本节详细介绍如何搭建面向 GDPR 的语料库, 包括作为基础的 GDPR 知识图谱和实现隐私政策样本半自动化收集的 APPC。

2.1 GDPR 知识图谱

为了识别隐私政策中的各种 GDPR 概念, 同时了解各种概念被提及的情况, 本文参考 ICO 提供的面向 GDPR 的隐私政策模板以及 GDPR 法规原文, 形成一个涵盖与 GDPR 相关的隐私政策概念集的知识图谱。

对于超出 GDPR 的概念, 本文在知识图谱中设置了两个一级节点: NON-GDPR 和 OTHERS。NON-GDPR 指的是与其他国家或地区的数据保护法律相关的概念, 如美国的《加州消费者隐私法》、巴西的《一般个人数据保护法》, 以及中国的《个人信息保护法》, 这不仅有利于探究不同国家或地区的数据保护条例间的差异, 也为未来将工作扩展到其他国家或地区的数据保护法规提供了帮助。以此, 本文得到了一个面向 GDPR 的隐私政策概念知识图谱, 如表 1 所示, 本文的 GDPR 知识图谱是一个三级结构, 从左到右分别为 GDPR 中与隐私政策相关的 3 层概念, 第一列为隐私政策中与 GDPR 相关的一级概念节点, 这些节点为 GDPR 对隐私政策内容的顶层要求, 涉及隐私政策中需要提到的各个方面; 第二列的二级概念节点为各一级概念节点下的分支, 对各一级节点进行了进一步的细化; 第

三列的三级概念节点作为知识图谱的叶子节点, 用于标识各个隐私政策概念的具体细节。其中“/”指不存在这一位置对应的节点。

2.2 语料库搭建

本文使用 Alexa 官网提供的顶级网站排名来收集隐私政策样本, 以保证语料库中的内容更全面、更专业, 并开发了隐私政策收集框架(如图 1 所示)来构建语料库, 促进其在未来的进一步扩展。隐私政策收集框架的目的是支持从预定义的 URL 列表中自动化地识别隐私政策网页, 提取隐私政策相关内容, 并将相应的超文本标记语言(HTML, hypertext markup language)元素转换为遵循定义的可扩展标记语言(XML, extensible markup language)模式的层次结构。隐私政策收集框架有助于减少人工操作, 可用于帮助研究人员进行自动隐私政策分析。

2.2.1 识别隐私政策页面

为了确保语料库的隐私政策样本与 GDPR 更相关, 本文使用了位于伦敦的代理服务器来收集在线服务的隐私政策。本文开发了一个半自动的工具, 可以从预定义的网站列表中识别和下载隐私政策, 最主要的任务是自动识别给定网站上的相关隐私政策网页。该工具按照下面描述的步骤工作。

步骤 1 该工具访问每个网站并搜索(不区分大小写)一个网页元素, 其文本内容包括预定义的代表隐私政策的关键词(privacy policy、privacy notice 和 privacy terms)之一。如果只找到一个链接, 则单击链接进入隐私政策页面, 跳转到步骤 4, 否则跳转到步骤 2。

步骤 2 该工具试图通过用户注册页面识别隐私政策网页, 用户注册页面通常包含跳转到隐私政策的链接。为此, 它不区分大小写地搜索关键词(create account、register、sign up、sign-up)。如果找到任何关键词, 它就单击链接并转到用户注册页面。然后, 该工具将步骤 1 应用到当前页面, 查找并访问隐私政策页面, 转到步骤 4。如果未找到隐私政策链接, 则转到步骤 3。

步骤 3 该工具寻求人工干预来识别隐私政策的链接。

步骤 4 该工具将识别的隐私政策页面保存为包含所有元素的本地 HTML 文件。

表 1 面向 GDPR 的隐私政策知识图谱
Table 1 GDPR-oriented privacy policy taxonomy

一级概念节点	二级概念节点	三级概念节点
CONTROLLER96.6%	IDENTITY31.5%	REGISTER NUMBER1.3%
	CONTACT93.9%	/
CONTROLLER REPRESENTATIVE18.7%	IDENTITY2.6%	REGISTER NUMBER0%
	CONTACT14.1%	/
DPO42.9%	IDENTITY0.6%	REGISTER NUMBER0%
	CONTACT40.3%	/
PD CATEGORY99.3%	TYPE4.0%	THIRD-PARTY0%
	SPECIAL2.0%	PUBLICLY0%
		CONDITION1.3%
PD ORIGIN100%	DIRECT ACTIVE95.3%	/
	DIRECT PASSIVE95.9%	COOKIE6.0%
	INDIRECT74.4%	THIRD-PARTY2.0%
		PUBLICLY0%
PROCESSING PURPOSES99.3%	/	/
DATA SHARING98.6%	RECIPIENTS92.6%	/
	CONDITION93.9%	/
LAWFUL BASIS90.6%	CONTRACT48.9%	/
	PUBLIC TASK12.7%	
	LEGITIMATE INTEREST58.3%	
	VITAL INTEREST12.0%	
	CONSENT51.0%	
PD STORAGE DETAILS89.2%	LEGAL OBLIGATION52.3%	/
	TIME67.1%	
	LOCATION26.8%	
	DISPOSAL METHOD46.9%	
DATA SUBJECT RIGHT100%	INFORMATION97.9%	POLICY CHANGE79.8%
		DATA BREACH NOTIFICATION9.3%
		OTHERS84.5%
	ACCESS91.9%	/
	RECTIFICATION95.3%	
	ERASURE88.5%	
	RESTRICTION71.1%	
	OBJECTION75.1%	
	PORTABILITY61.7%	
	AUTO DECISION MAKING14.7%	
	WITHDRAW CONSENT89.2%	
COMPLAINT75.1%	SA1.0%	/
NON-GDPR79.8%	COOKIE54.3%	/
	OTHER LEGISLATIONS45.6%	/
OTHERS	/	/

2.2.2 隐私政策网页预处理

识别隐私政策网页后，本文需要对网页进行预处理，移除 3 种类型的文档对象模型（DOM，

document object model）元素：① 对分析网页中包含的隐私政策文本内容没有帮助的多媒体相关元素，如<picture>、、<video>和<audio>；

② 嵌入非文本对象或外部网页的元素, 如<applet>、<embed>、<object>和<iframe>; ③ 语义上与网页主体无关或可以作为网页主体一部分出现的元素, 如<footer>和<nav>。

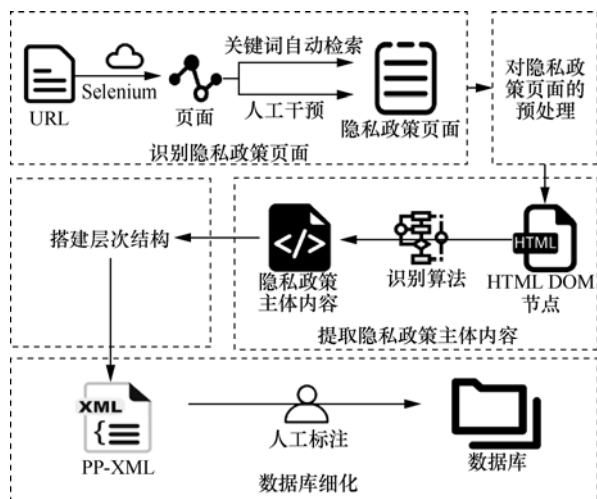


图1 隐私政策收集框架

Figure 1 Architecture of privacy policy collection

2.2.3 提取隐私政策主体内容

在对隐私政策网页进行预处理之后, 下一步是提取隐私政策的相关内容。由于网页的层次性和流行的网页结构均基于<div>元素, 因此隐私政策的相关内容总处于 HTML 元素下。本文需要识别包含隐私政策内容的单个 HTML 元素, 称之为隐私政策文本 (PPC, privacy policy content) 元素。

PPC 元素下的直接子元素在每个元素内的文本长度上整体比同级的非 PPC 元素更接近。此处“文本长度”指供读者使用的字符数 (而不是 HTML 代码中的字符), 反映了读者期望阅读的文本量。基于这一观察, 本文开发了查找 PPC 元素的算法 (如算法 1 所示)。

算法 1 查找 PPC 元素的算法

输入 HTML DOM 节点 node (通常为<body>)

输出 最可能为包含全部隐私政策相关内容的最底层 HTML 节点元素

定义 $R = \emptyset$, $r = \infty$, $PP_node = node$;

- 1) 循环
- 2) 若 PP_node 没有子节点, 则
- 3) 返回 PP_node ;

- 4) $s = \text{CSS value of } PP_node$;
- 5) 若 $\#R > 0$, 则
- 6) $r = s / \text{average}(R)$;
- 7) 若 $r < r_h$, 则
- 8) 返回 PP_node ;
- 9) 否则
- 10) $R = R \cup \{s\}$;
- 11) $PP_node = R$ 中具有最长可读文本的子节点;
- 12) 循环结束

该算法基于以下假设: ① 隐私政策的相关内容处于单个 PPC 元素之下; ② PPC 元素在所有兄弟元素中文本长度最长; ③ PPC 元素的子元素比非 PPC 元素的子元素有更多相似的文本长度。该算法使用子元素相似度评分 (CSS, child similarity score), 所有子元素文本长度的标准差来度量候选元素是 PPC 元素的可能性。由于 CSS 是一个相对的概念, 算法 1 将当前候选元素的 CSS 值与观察到的所有非 PPC 元素的 CSS 值的平均值进行比较, 如果比值低于阈值 r_h , 则认为该候选元素为 PPC 元素。

2.2.4 搭建层次结构

在提取隐私政策网页的相关内容之后, 本文将页面内容的基于 HTML 的 DOM 树转换为涵盖隐私政策语义内容的新层次结构, 该结构存储为 XML 文件, 遵循本文称为标准隐私政策 (SPP, standard privacy policy) 的特定 XML 模式。在这一模式下, 整份隐私政策有一个根节点<policy>, 其中包括隐私政策的许多语义段, 每个都被编码为<segment>元素。<segment>可以嵌套, 以允许隐私政策的内容具有层次结构。对于每个<segment>元素, 总有一个<title>元素表示对应段内容的语义标题, 以及有一个或多个<paragraph>元素代表对应段的内容。除了<segment>元素外, <policy>元素还包括独立的<paragraph>元素, 如在第一个<segment>元素之前的开头段落。<paragraph>元素可以包含一个或多个<list>元素作为它的子元素, <list>元素包含至少一个<item>元素作为它的子元素。<item>元素可以包含一个<list>元素, 允许在一个<paragraph>元素中嵌套列表。

因此, 在 SPP 中, 对于<title>元素, 处于同

一<segment>元素中的不同<title>元素互为兄弟节点,当前<segment>元素中的<title>元素与嵌套的<segment>元素中的<title>元素构成父子关系;对于<paragraph>元素,处于同一<segment>元素中的不同<paragraph>元素互为兄弟节点,当前<segment>元素中的<title>元素与<paragraph>元素构成父子关系。

为了将隐私政策中的 HTML 元素转换为基于 SPP 的新层次结构,本文遵循下面描述的 5 个步骤,并在图 2 (隐私政策层次结构还原) 步骤中说明。注意,在 HTML 文档的自动分割方面已经有一些相关的工作,如在文献[22]中提出的系统。然而,大多数此类工作只关注于识别顶层标题,这对本文的工作来说是不够的。

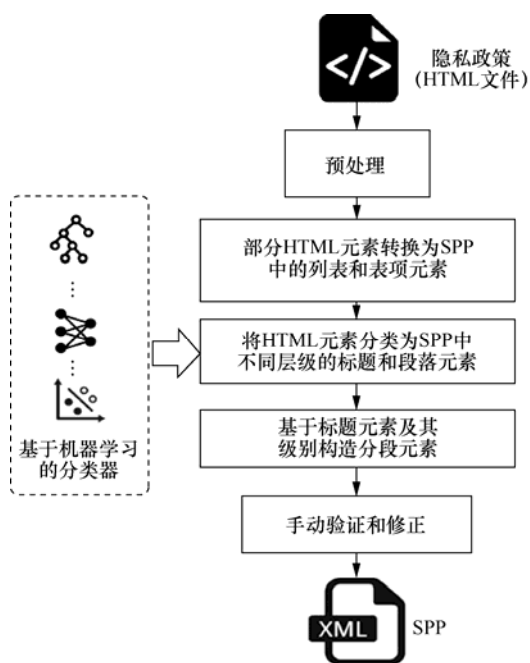


图2 隐私政策层次结构还原
Figure 2 Reduction of privacy policies' structure

步骤 1 预处理。虽然在预处理步骤中删除了许多 HTML 元素,但是一些特殊的内嵌元素,如出现在<p>元素中间的元素,在 SPP 中没有作用。这些内嵌元素在这一步骤中被删除。之后,本文将剩余的元素放入<policy>元素中。

步骤 2 部分 HTML 元素转换为 SPP 中的列表和表项元素。在 SPP 中,某些 HTML 元素与<list>和<item>元素之间存在精确的映射,因此这些 HTML 元素可以直接转换。

步骤 3 将剩余 HTML 元素分类为 SPP 中不同等级的标题元素和段落元素。在隐私政策中,标题与段落往往有不同属性,本文选取以下 7 个可能有助于区分标题元素和段落元素的属性:① 文本长度;② 相对于<body>元素的基准字体大小;③ 相对于<body>元素的基准字体粗细;④ 文本是否为斜体;⑤ 文本是否带下划线;⑥ 在 HTML DOM 树中的节点深度;⑦ HTML 标签。除上述特性外,许多隐私政策的标题具有开头序号标签(如“1”“2.3”“a.”和“ii”),本文根据上述各特征的值域进行编码,构建了一个基于机器学习的分类器来将候选元素预测为以下 5 类之一:一至四级的标题和段落。根据对隐私政策的一般观察,本文选择 4 个级别的标题。本文测试了几种主流的机器学习模型,包括随机森林(RF, random forest)^[23]、支持向量机(SVM, support vector machine)(带有线性核和 RBF 核)^[24]、额外树(ET, extra tree)^[25]和 XGBoost^[26]。本文将所有 150 份隐私政策用于训练和测试,使用 5 倍交叉验证和 20% 作为测试集。其中 ET 分类器取得了最好的性能, F1 分数达到了 0.87。

步骤 4 基于标题元素及其级别构造分段元素。在确定了<title>元素及其级别之后,本文可以使用它们来构造适当嵌套的<segment>元素,以搭建隐私政策的层次结构。对于每个<title>元素,本文创建一个<segment>元素,在它之后包括<title>元素和所有非<title>元素,直到下一个<title>元素或 SPP 文档的末尾。<segment>元素的级别设置为与其包含的<title>元素相同。

步骤 5 手动验证和修正。让一名或多名专家手动检查生成的 SPP 文档和输入的 HTML 文档,以确认自动化结果并修复任何错误。

2.2.5 数据标注

通过实施 2.2.4 节中描述的步骤,本文处理了语料库中的 150 份隐私政策,最终获得了 150 份 SPP 文档。之后,本文根据扩展 GDPR 知识图谱,用相关的 GDPR 概念标注了 150 份 SPP 文档。本文对每个 SPP 文档中的每个标题和段落元素进行了标注,每个标记代表扩展 GDPR 知识图谱中唯一的 GDPR 节点。

对于标注工作,本文聘请了一家商业标注公

司,并遵循多步骤的质量保证过程。首先,笔者培训了4名标注员,这些标注员首先尝试标注20份SPP文档作为试点。接着,笔者检查试点标注结果并将效果反馈给4名标注员,以便进一步改进和对齐标注。之后,4名标注员标注所有剩余的SPP文档,每名标注员的标注结果的随机样本由另一名标注员检查。最后,验证和讨论标准结果,以达成一致意见。为了定量衡量4名标注员的工作质量,本文使用了Cohen's Kappa^[27]评估标注员之间的一致性,观察到平均评分为0.75,表明4名标注员之间的一致性相当高。在标注工作结束时,本文最终获得了一个遵循SPP和扩展GDPR知识图谱的完整标注的隐私政策语料库。

3 GDPR 概念分类器

过去关于隐私政策的自动GDPR合规性分析的研究主要集中在隐私政策中的长文本或段落,没有考虑层次结构,如以前的工作没有考虑隐私政策中标题的自动分析。在隐私政策中缺少这种层次信息会导致分析不完整和不准确。对于GDPR知识图谱中的59个概念,本文使用语料库中的数据来训练两类二分类器,即一个标题分类器Classifier4title和一个段落分类器Classifier4paragraph,Classifier4title通过二分类任务实现对标题所包含的概念标签的预测,Classifier4paragraph实现对段落所包含的概念标签的预测。例如,对于每个概念a,本文基于新语料库构建了每个分类器的训练集:那些带有a标签的节点为正样本,其他没有a标签的节点为负样本。注意,每个节点可以有多个GDPR概念标签。每个分类器以<title>或<paragraph> SPP元素作为输入,并生成二元决策,判断SPP元素是否包含特定的GDPR概念标签。

3.1 特征选取

对于Classifier4title和Classifier4paragraph两个分类器,本文考虑了3组特征:d_1特征通过句子向量化方法产生;d_2特征为同一份隐私政策中已经被处理的父节点和兄弟节点的句向量;d_3特征捕获与每个GDPR概念相关的一些预定义关键词的出现。第一组特征在类似的任务中很常见;第二组特征用于捕获当前节点与其父节点之间的

层次信息,以及当前节点与其同一父节点下的兄弟节点之间的上下文信息;第三组特征最初由Amaral等^[7]在GDPR概念分类器中提出。

对于句子向量化的第一组特征,本文测试了4种广泛使用的3种不同维度(100、200和300)的向量化方法:TF-IDF^[28]、GloVe^[29]、fastText^[30]和Word2vec^[31]。本文还测试了4种不同向量化方法的6对组合,每种方法产生一个100维的子向量,合并后为200维的向量。由于篇幅有限,本文展示部分具有代表性的结果,对于Classifier4title句向量统一取100维;对于Classifier4paragraph句向量统一取300维,将所有分类器的性能取平均进行比较,分类器性能通过3个常见指标——精确率(P, precision)、召回率(R, recall)和F1分数来评估,向量化方法性能对比如表2所示。

表2 向量化方法性能对比
Table 2 Comparison of vectorization methods performance

向量化 方法	Classifier4title			Classifier4paragraph		
	P	R	F1	P	R	F1
TF-IDF	0.92	0.84	0.88	0.92	0.73	0.81
GloVe	0.87	0.84	0.86	0.94	0.68	0.79
fastText	0.88	0.87	0.87	0.92	0.65	0.76
Word2vec	0.90	0.81	0.85	0.91	0.71	0.80

实验结果表明,采用100维的TF-IDF的Classifier4title分类器的性能最好,采用300维的TF-IDF的Classifier4paragraph分类器的性能最好,后续的实验也采用了这一特征。

对于捕获层次信息和上下文信息的第二组特征,本文使用在同一隐私政策中已经处理的父节点和兄弟节点的句向量。考虑到当前节点为段落时,父节点通常为标题,文本较短,而兄弟节点通常为段落,文本较长,因此前者采用100维向量,后者采用300维向量较为适合。在当前节点为标题时,父节点和兄弟节点都是标题,文本均较短,100维向量就足够。采用的句子向量化方法仍然是TF-IDF,因为它在之前的实验中表现最好。这些特征帮助捕获当前节点及其父节点和兄弟节点之间的层次信息。实验表明这些特征提高了概念分类器的性能。

对于第三组特征，即关于预定义的关键词出现情况的特征，本文使用 59 维二进制向量，其中每一维表示是否至少有一个与特定 GDPR 概念相关的预定义关键词出现在候选元素中至少一次。例如，像 “When Google shares your information” 这样的句子将被标记为 “DATA SHARING CONDITION”。因为包含关键词 “share” 和 “when”。本文为每一个概念定义了相应的关键词列表，实验结果表明，这组基于关键词的特征只能帮助提高 Classifier4paragraph 分类器的性能，所以对于 Classifier4title 分类器，本文决定排除这组特征。

综上所述，每个 Classifier4title 分类器的输入向量有 $d_1 + d_2 + d_3 = 100 + 2 \times 100 + 0 = 300$ 维特征，每个 Classifier4paragraph 分类器的输入向量有 $d_1 + d_2 + d_3 = 300 + 100 + 300 + 59 = 759$ 维特征。

3.2 机器学习模型

为了构建 GDPR 概念分类器，本文测试了主流的机器学习方法，即 RF、SVM（RBF 核）、AdaBoost 和 XGBoost。其中 RF 设置参数 $\text{max_depth} = \text{None}$, $\text{random_state} = 50$, SVM（RBF 核）设置参数 $\text{random_state} = 50$, AdaBoost 设置参数 $\text{learning_rate} = 0.5$, $\text{n_estimators} = 6$, $\text{algorithm} = \text{'SAMME.R'}$, XGBoost 设置参数 $\text{max_depth} = \text{None}$, $\text{eta} = 0.5$, $\text{objective} = \text{binary:logistic}$, 其他参数均设置为默认参数。实验结果表明，在大多数分类器中，RF 分类器的性能最好，

具有更好的鲁棒性和可泛化性。因此，本文只报告 RF 的分类器结果。

同时，本文尝试了在自然语言处理领域热门的预训练语言模型 BERT^[32]，此次实验中采用的模型为 bert-base-uncased，即不区分大小写的基础模型。

3.3 评估

本文使用新语料库对所有 GDPR 概念分类器进行了训练和测试。本文使用 5 倍交叉验证来训练和测试分类器，所有样本的 20% 作为测试集。

在选取评估 GDPR 概念分类器性能的基准时，本文复现了与本文相似的方法——Torre 等的方法^[6]，包括他们的基于相似度、基于关键词和基于人工智能（AI, artificial intelligence）的分类器。由于他们没有公开数据集，本文通过将他们的分类器应用到本文的数据集上来比较性能。此处仍使用 P 、 R 和 $F1$ 来评估每个分类器的性能。篇幅所限，所选的 GDPR 概念分类器性能对比如表 3 所示，第 1 列对应表 1 的 GDPR 知识图谱中的各个概念节点，第 2 列为各个概念节点对应的样本数量，第 3~5、6~8、9~11 列分别为本文的 Classifier4paragraph 分类器、Torre 等的分类器以及训练语言模型 BERT 的性能指标。

表 3 中的实验结果表明，本文的分类器在所有情况下都明显优于 Torre 等提出的分类器。尽管 Torre 等的分类器在召回率方面表现普遍良好，但在精确率方面表现非常差，导致 $F1$ 分数

表 3 GDPR 概念分类器性能对比
Table3 Comparison of GDPR classifiers' performance

GDPR 概念	样本数量	Classifier4paragraph (TF-IDF + RF)			Torre's {GloVe + SVM(RBF 核)}			BERT (bert-base-uncased)		
		P	R	$F1$	P	R	$F1$	P	R	$F1$
CONTROLLER	522	1	0.41	0.58	0.01	1	0.01	0.79	0.69	0.74
CONTROLLER.CONTACT	443	0.95	0.65	0.77	0.05	0.91	0.1	0.79	0.68	0.73
PD CATEGORY	1 715	0.93	0.91	0.92	0.25	0.93	0.39	0.79	0.84	0.81
PD ORIGIN	1 710	0.97	0.65	0.78	0.06	0.9	0.11	0.77	0.69	0.72
PROCESSING PURPOSES	1 682	0.96	0.74	0.83	0.24	0.92	0.38	0.84	0.54	0.66
DATA SHARING.CONDITION	709	0.91	0.79	0.84	0.1	0.91	0.17	0.85	0.7	0.76
PD STORAGE DETAILS	441	0.99	0.55	0.71	0.03	0.88	0.06	0.89	0.76	0.82
DATA SUBJECT RIGHT	1 895	0.98	0.56	0.71	0.04	0.93	0.07	0.83	0.71	0.76
COMPLAINT	165	0.99	0.7	0.81	0.02	0.92	0.05	0.8	0.53	0.64

非常低，这可能是因为他们数据集和分类器没有考虑隐私政策中的层次信息。他们的分类器的糟糕结果意味着可能存在跨不同数据集的泛化问题。

而与预训练语言模型 BERT 相比，本文的分类器在总体上性能较优，但在个别类别上略逊于 BERT，其中不乏样本数量充足的类别。同时 BERT 模型的性能较为平均，虽然本文实验数据不够充足，但仍有不错的性能，表明预训练模型可能为本文未来改进工作的一个方向。

4 隐私政策合规性分析

本节展示了对语料库中隐私政策的分析结果。① GDPR 知识图谱中所有 GDPR 概念的涵盖分析。② 根据 ICO 提供的隐私政策模板，对新语料库中的 150 份隐私政策进行 GDPR 合规性分析。这两个示例分析都对不同的 GDPR 相关隐私政策行为产生了有益的新见解，并为本文的 GDPR 知识图谱、语料库的有效性提供了证据。

4.1 GDPR 概念的涵盖分析

新的语料库包含经过人工验证的层次信息和 GDPR 概念标签，使得本文能够分析其中的 150 份隐私政策，以了解不同的 GDPR 概念是如何被涵盖的。结果作为 GDPR 知识图谱的一部分显示在表 1 中，其中每个节点名称后面的百分比表示语料库中有多少隐私政策涵盖了相应的 GDPR 概念。

在第一级概念中，“PD ORIGIN”和“DATA SUBJECT RIGHT”被所有隐私政策涵盖，而“PD CATEGORY”和“PROCESSING PURPOSES”两个概念仅被一份隐私政策遗漏，这表明本文的资料库涵盖的大多数网站认真考虑了 GDPR 这 4 个方面。

在第二级概念中，“DATA SUBJECT RIGHT.INFORMATION”和“PD ORIGIN.DIRECT PASSIVE”的涵盖率最高，分别在 146（97.9%）和 143（95.9%）份隐私政策中被提及。在低涵盖端，则是“PD CATEGORY.SPECIAL”和“COMPLAINT.SA”的覆盖率最低，仅为 1.0%，即仅被 3 份隐私政策提及。

在第三级概念中，只有“DATA SUBJECT

RIGHT.INFORMATION.POLICY CHANGE”的涵盖率相对较高，为 79.8%，在 119 份隐私政策中均有提及。所有其他第三级的概念均很少被提及。一种可能的解释是，在线服务可能认为隐私政策中这些低层级的细节是不必要的。

4.2 参考 ICO 模板的 GDPR 合规性

此前对隐私政策的 GDPR 合规性分析主要通过简单地检查一些特定的 GDPR 概念的存在与否来完成。而 Torre 等^[15]建议使用一组明确定义的标准来捕捉先决条件，这些条件决定了一个或多个特定的 GDPR 概念是否应该出现在隐私政策中。Torre 等设置的先决条件中部分是基于问题的，这些问题涉及进行合规性检测的网站的运营情况，由法律专家参考 GDPR 的官方文本制定，由提交网站的用户通过问卷给予回答。这一方法存在以下问题：首先，依赖用户回答问卷会降低分析的可靠性，用户无法保证提交答案的正确性（如用户无法得知数据控制者是否将与第三方共享数据）；其次，使用问卷调查意味着很难支持大规模的自动化分析；最后，虽然利用法律专家从 GDPR 文本中提炼标准可以确保完整性和严谨性，但提炼的标准无法反映国家数据保护部门（如 ICO）更实际的建议，因此有过于严格的倾向。

为了解决 Torre 等方法存在的问题，本文建议参考国家数据保护机构提供的符合 GDPR 的隐私政策官方模板，以得出一套更实用的标准。以英国的官方机构 ICO 为例，本文提炼出了 10 个标准，参考 ICO 模板的合规标准如表 4 所示。除了标准 C4，其先决条件“适用时”必须根据具体情况进行判断，所有其他 9 个标准都可以通过检查隐私政策中相关 GDPR 概念的存在而实现自动化。

本文应用了这些标准，对语料库 GoPPC-150 中的 150 份隐私政策进行了 GDPR 合规性分析。本文跳过了 C4，因为检查先决条件需要手动分析每个网站的数据收集行为，其中一些可能是不可见的。最终的 GDPR 合规性分析结果如表 5 所示。

结果表明，语料库中的样本对于某些标准的合规比例很高，表明在线服务普遍对这些标准中要求的隐私政策概念给予了关注。但对于某些标准（特别是 C7 和 C8）的合规比例很低，只有不

表 4 参考 ICO 模板的合规标准
Table 4 Compliance criteria based on the ICO template

标准	先决条件	后置条件
C1	—	CONTROLLER.CONTACT
C2	—	PD CATEGORY
C3	—	PD ORIGIN. {DIRECT ACTIVE 或 DIRECT PASSIVE} 和 PROCESSING PURPOSES
C4	适用时	PD ORIGIN.INDIRECT 和 PROCESSING PURPOSES
C5	—	DATA SHARING
C6	—	LAWFUL BASIS. {CONSENT, CONTRACT, LEGAL OBLIGATION, VITAL INTEREST, PUBLIC TASK 或 LEGITIMATE INTEREST}
C7	—	PD STORAGE DETAILS. {LOCATION, TIME 和 DISPOSAL METHOD}
C8	—	DATA SUBJECT RIGHT. {ACCESS, RECTIFICATION, ERASURE, RESTRICTION, OBJECTION 和 PORTABILITY}
C9	LAWFUL BASIS.CONSENT	DATA SUBJECT RIGHT.WITHDRAW CONSENT
C10	—	COMPLAINT

到一半的样本满足了这些标准。此外，在本文的语料库 GoPPC-150 中，只有 10（6.7%）份隐私政策同时包含了这 9 个标准所要求的隐私政策概念，表明整体的合规比例并不乐观。

表 5 GDPR 合规性分析结果
Table 5 Results of GDPR compliance analysis

标准	合规比例	标准	合规比例
C1	140/150 (93.3%)	C6	136/150 (90.6%)
C2	149/150 (99.3%)	C7	16/150 (10.7%)
C3	131/150 (87.3%)	C8	58/150 (38.7%)
C4	—	C9	59/62 (95.2%)
C5	148/150 (98.6%)	C10	113/150 (75.3%)

本文测试的 150 份隐私政策都来自成熟的（通常是跨国的）在线服务，在全球拥有庞大的用户基础，因此 GDPR 应该普遍适用。同时，本文考虑的标准是从 ICO 模板总结出的最低要求，但仍有不到 10%的合规率的结果进一步证明了在线服务普遍未达到合规要求的现状。

5 结束语

考虑到 GDPR 概念标注需要人工的努力，本文构建的语料库包含 150 份隐私政策，由于部分 GDPR 概念的样本数量较少，不足以训练部分 GDPR 概念分类器，因此只能通过扩充语料库的样本量来解决。

根据 GDPR 的管辖范围规定，GDPR 适用于任何为欧盟境内公民提供服务的企业，无论其是否在欧盟境内注册，这意味着世界各地涉及海外业务的企业都要考虑其隐私政策面向 GDPR 的合规性，国内企业也不例外。面向这一需求，本文构建了一套智能化检测方法，自动提取各在线服务公司的隐私政策且转化为具有结构层次的标准格式，并进行基于自然语言处理的文本分类，识别其中相应的 GDPR 概念，实现了面向 GDPR 隐私政策合规性的智能化检测。

本文实现的智能化检测工作，能够有效帮助国内企业对其隐私政策进行全面自动化审查，及时检索缺失的相应 GDPR 概念并改进，为其展开海外业务，尤其是向欧盟用户提供服务时提供法律支撑。未来，本文计划将该工作扩展到其他国家的数据保护法律和法规。考虑到不同国家和地区多个相关法律、法规和标准的并存，本文计划在更国际化的背景下研究隐私政策的法律合规性。

参考文献：

- [1] MCDONALD A, CRANOR L. The cost of reading privacy policies[J]. Journal of Law and Policy for the Information Society, 2008, 4: 543-568.
- [2] LI H X, ZHU H J, DU S G, et al. Privacy leakage of location sharing in mobile social networks: attacks and defense[J]. IEEE Transactions on Dependable and Secure Computing, 2016, 15(4):

- 646-660.
- [3] TUCKER S. Google fined \$57m by data protection watchdog over GDPR violations[EB].
 - [4] TOM J, SING E, MATULEVICIUS R. Conceptual representation of the GDPR: model and application directions[C]//Proceedings of 17th International Conference on Perspectives in Business Informatics Research(BIR). 2018: 18-28.
 - [5] LIPPI M, PALKA P, CONTISSA G, et al. CLAUDETTE: an automated detector of potentially unfair clauses in online terms of service[J]. *Artificial Intelligence and Law*, 2019, 27(2): 117-139.
 - [6] TORRE D, ABUALHAIJA S, SABETZADEH M, et al. An AI-assisted approach for checking the completeness of privacy policies against GDPR[C]//Proceedings of the 2020 IEEE 28th International Requirements Engineering Conference. 2020: 136-146.
 - [7] AMARAL O, ABUALHAIJA O, ABUALHAIJA S, et al. AI-enabled automation for completeness checking of privacy policies[J]. *IEEE Transactions on Software Engineering*, 2021, (N/A): 1-21
 - [8] 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(ACL). 2016: 1330-1340.
 - [9] 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.
 - [10] ZIMMECK S, STORY P, SMULLEN D, et al. MAPS: scaling privacy compliance analysis to a million apps[C]//Proceedings of the Conference on Privacy Enhancing Technologies. 2019: 66-86.
 - [11] SRINATH M, WILSON S, GILES C. Privacy at scale: introducing the PrivaSeer corpus of Web privacy policies[J]. *arXiv preprint arXiv: 2020.11131*.
 - [12] KUZNETSOV M, NOVIKOVA E, KOTENKO I, et al. Privacy policies of IoT devices: collection and analysis[J]. *Sensors*, 2022, 22(5): 1-23.
 - [13] TESFAY W, HOFMANN P, NAKAMURA T, et al. Privacy guide: towards an implementation of the EU GDPR on internet privacy policy evaluation[C]//Proceedings of the 4th ACM International Workshop on Security and Privacy Analytics. 2018: 15-21.
 - [14] NEJAD N, SCERRI S, LEHMANN J. KNIGHT: mapping privacy policies to GDPR[C]//Proceedings of 21st International Conference on Knowledge Engineering and Knowledge Management(EKAW 2018). 2018: 258-272.
 - [15] TORRE D, SOLTANA G, SABETZADEH M, et al. Using models to enable compliance checking against the GDPR: an experience report[C]//Proceedings of the 2019 ACM/IEEE 22nd International Conference on Model Driven Engineering Languages and Systems. 2019: 10-20.
 - [16] HAMDANI R, MUSTAPHA M, AMARILES D, et al. A combined rule-based and machine learning approach for automated GDPR compliance checking[C]//Proceedings of the 18th International Conference on Artificial Intelligence and Law. 2021: 40-49.
 - [17] ZIMMECK S, BELLOVIN S. Privee: an architecture for automatically analyzing web privacy policies[C]//Proceedings of the 23rd USENIX Security Symposium. 2014: 1-16.
 - [18] HARKOUS H, FAWAZ K, LEBRET R, et al. Automated analysis and presentation of privacy policies using deep learning[C]//Proceedings of the 27th USENIX Security Symposium. 2018: 531-548.
 - [19] AYALA-RIVERA V, PASQUALE L. The grace period has ended: An approach to operationalize GDPR requirements[C]//Proceedings of the 2018 IEEE 26th International Requirements Engineering Conference. 2018: 136-146.
 - [20] DEGELING M, UTZ C, LENTZSCH C, et al. We value your privacy now take some cookies: measuring the GDPR's impact on web privacy[C]//Proceedings of the 2019 Network and Distributed Systems Security Symposium. 2018: 1-14.
 - [21] LINDEN T, KHANDELWAL R, HARKOUS H, et al. The privacy policy landscape after the GDPR[C]//Proceedings on Privacy Enhancing Technologies, 2020(1): 47-64.
 - [22] ABHIJITHG, SHOMIR W, NORMAN S. Supervised and unsupervised methods for robust separation of section titles and prose text in web documents[C]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. 2018: 850-855.
 - [23] WILLIAMS B, HALLOIN C, LOBEL W, et al. Data-driven model development for cardiomyocyte production experimental failure prediction[C]//Proceedings of 30th European Symposium on Computer Aided Process Engineering. 2020: 1639-1644.
 - [24] MOFUERZA J, MUNOZ A. Support vector machines with applications[J]. *Statistical Science*, 2006, 21(3): 322-336.
 - [25] GEURTS P, ERNST D, WEHENKEL L. Extremely randomized trees[J]. *Machine Learning*, 2006, 63: 3-42.
 - [26] CHEN T, GUESTRIN C. XGBoost: a scalable tree boosting system[R]. 2016.
 - [27] COHEN J. A coefficient of agreement for nominal scales[J]. *Educational and Psychological Measurement*, 1960, 20(1): 37-46.
 - [28] SALTON G, BUCKLEY C. Term weighting approaches in automatic text retrieval[J]. *Information Processing & Management*, 1988, 24(5): 513-523.
 - [29] PENNINGTON J, SOCHER R, MANNING C. GloVe: global vectors for word representation[C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. 2014: 1532-1543.
 - [30] BOJANOWSKI P, GRAVE E, JOULIN A, et al. Enriching word vectors with subword information[J]. *Transactions of the Association for Computational Linguistics*, 2017, 5: 135-146.

[31] MIKOLOV T, CHEN K, CORRADO G, et al. Efficient estimation of word representations in vector space[J]. Computer Science, 2013, 5: 324-352.

[32] JACOB D, CHANG M W, KENTON L, et al. BERT: pre-training of deep bidirectional transformers for language understanding[R]. 2018.



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。

[作者简介]



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。



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，主要研究方向为密码分析/密码工程、人工智能安全、大数据隐私保护。



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



回红（1969- ），女，天津人，博士，上海交通大学副教授，主要研究方向为图像处理与模式识别、信息安全。