

人工智能安全知识图谱构建技术研究

沈晓晨¹, 葛寅辉¹, 陈波¹, 于冷²

(1. 南京师范大学计算机与电子信息学院/人工智能学院, 江苏 南京 210023; 2. 南京师范大学数学科学学院, 江苏 南京 210023)

摘要: 人工智能技术在飞速发展的同时引发众多安全隐患, 现阶段人工智能安全数据来源广泛、种类复杂并且缺乏规范描述, 为此提出了一种人工智能安全知识图谱构建方法, 利用知识图谱对当前多源异构数据进行分析与整合, 将复杂关联的数据进行科学表示, 挖掘潜在价值并形成领域知识库。针对人工智能安全领域概念的多样性和关联性, 提出一种分层结构人工智能安全本体, 使本体结构更加多元化和扩展化, 为知识图谱构建过程提供规则约束, 并基于此形成人工智能安全知识库; 为了有效利用特征信息及减少噪声干扰, 采取基于双向长短时记忆网络-条件随机场 (BiLSTM-CRF) 的命名实体识别算法和基于卷积神经网络-注意力机制 (CNN-ATT) 的关系抽取算法进行信息抽取, 利用构建的人工智能安全数据集证明算法性能; 以提出的安全本体为知识体系, 以 3D 效果展现人工智能安全知识图谱多层次可视化结果, 有效关联多源安全数据信息。实验结果表明, 信息抽取算法性能良好, 取得了比传统方法更好的效果; 构建的人工智能安全知识图谱直观展示了层次结构及相互关系, 符合准确性、一致性、完整性、时效性维度的知识图谱多维度评估标准, 能够为人工智能安全研究提供支持。

关键词: 人工智能安全; 知识图谱; 本体构建; 信息抽取; 可视化

中图分类号: TP391

文献标志码: A

DOI: 10.11959/j.issn.2096-109x.2023030

Research on construction technology of artificial intelligence security knowledge graph

SHEN Xiaochen¹, GE Yinhui¹, CHEN Bo¹, YU Ling²

1. School of Computer and Electronic Information/School of Artificial Intelligence, Nanjing Normal University, Nanjing 210023, China

2. School of Mathematical Science, Nanjing Normal University, Nanjing 210023, China

Abstract: As a major strategic technology, artificial intelligence is developing rapidly while bringing numerous security risks. Currently, security data for artificial intelligence is collected from disparate sources and lacks

收稿日期: 2022-07-09; 修回日期: 2022-12-25

通信作者: 陈波, bchen@njnu.edu.cn

基金项目: 教育部科技司赛尔网络下一代互联网技术创新项目滚动项目 (NGIICS20190504); 江苏省“十四五”教育科学规划重大课题 (A/2021/05)

Foundation Items: The CERNET Next Generation Internet Technology Innovation Project under the Ministry of Education's Science and Technology Department (NGIICS20190504), The Education Scientific Planning Project of Jiangsu Province during the "14th Five-Year Plan" (A/2021/05)

引用格式: 沈晓晨, 葛寅辉, 陈波, 等. 人工智能安全知识图谱构建技术研究[J]. 网络与信息安全学报, 2023, 9(2): 164-174.

Citation Format: SHEN X C, GE Y H, CHEN B, et al. Research on construction technology of artificial intelligence security knowledge graph[J]. Chinese Journal of Network and Information Security, 2023, 9(2): 164-174.

standardized description, making it difficult to integrate and analyze effectively. To address this issue, a method for constructing an artificial intelligence security knowledge graph was proposed. The knowledge graph was used to integrate the current multi-source heterogeneous data, scientifically represent complex relationships of the data, mine potential value and form a domain knowledge base. In view of the diversity and correlation of concepts in the field of artificial intelligence security, a hierarchical structure of artificial intelligence security ontology was proposed to make the ontology structure more diversified and extensible, provide rule constraints for the process of knowledge graph construction, and form an artificial intelligence security knowledge base. To effectively utilize feature information and reduce noise interference, named entity recognition algorithm based on BiLSTM-CRF and relationship extraction algorithm based on CNN-ATT were adopted for information extraction. The constructed artificial intelligence security dataset was then used to verify the performance of the algorithm. Based on the proposed ontology, the multi-level visualization results of the artificial intelligence security knowledge graph were presented in 3D effect, effectively connecting the multi-source security data information. The experimental results show that the constructed knowledge graph meets the multi-dimensional evaluation criteria of accuracy, consistency, completeness, and timeliness, providing knowledge support for artificial intelligence security research. Overall, the proposed method can help address the complexity and heterogeneity of security data in artificial intelligence and provide a more standardized, integrated approach to knowledge representation and analysis.

Keywords: artificial intelligence security, knowledge graph, ontology construction, information extraction, visualization

0 引言

科技是一把双刃剑, 人工智能技术在推动各领域前进的同时, 也引发了安全新挑战。人工智能安全问题本质是其安全赋能及伴生效应造成的, 根据安全效应不同可以将人工智能安全问题分为衍生安全、助力安全和内生安全^[1]。人工智能安全数据信息来源广泛、种类复杂并且迭代迅速, 并未形成标准信息库; 针对这些海量的碎片化信息进行整合处理, 挖掘潜在价值并形成人工智能安全知识库, 为人工智能安全防御提供知识支撑是目前科学界面临的重大问题。能够对多源异构数据进行关联挖掘的知识图谱, 为解决当前人工智能安全问题提供了一种思路。

知识图谱旨在实现高效搜索^[2], 是一种更为高效的管理和利用海量数据的方式, 近年常用于智能问答、情报分析、语义搜索等领域。与此同时, 知识图谱的数据结构与人工智能领域许多技术任务所基于的数据一脉相承, 可以为后续的智能搜索和知识推理任务提供强有力的支持。虽然目前还没有出现人工智能安全领域的知识图谱, 但与其具有相互依赖性的网络安全知识图谱已经应用于网络安全分析、网络空间态势感知等领域, 为人工智能安全知识图谱构建提供了借鉴。

对信息安全领域中知识图谱的研究主要集中

在网络安全、威胁情报、安全态势感知等方向, 如秦娅等^[3]针对网络安全知识图谱构建进行研究, 在复用经典实体的基础上对海量碎片化的多源异构数据进行细粒度的分析与挖掘; Jia 等^[4]提出一种基于机器学习的网络安全知识图谱构建方法与基于五元组模型的推演规则, 将本体构建分为模式层与数据层, 但本体构建核心模式层知识范围不够全面; 王通等^[5]对威胁情报进行本体建模, 提出一种监督性的深度学习模型, 该模型对威胁情报的实体和实体关系进行自动化抽取并可视化展示。对于知识图谱构建的关键层次本体构建, 当前信息安全领域本体研究范围较为广泛且未形成统一标准, 尚无人工智能安全本体可供借鉴, 领域内应用较为广泛的本体有: Iannacone 等^[6]提出的网络安全通用本体 STUCCO, 涵盖 15 个实体类和 115 个属性; Qamar 等^[7]提出针对威胁情报的本体, 实现了情报信息的融合。以上工作在本体构建部分大多采用了模式层与数据层两层结构的表达和关联, 知识表达能力不足, 影响后期知识图谱的实用价值。

实际应用于信息安全领域的知识图谱项目更少, 主要有 MITRE 公司针对图模型研究的 CyGraph 原型系统^[8], 该系统将来自众多数据源的数据关联到一个规范模型中, 利用 Neo4j 构建持久数据存储, 对网络资产进行高性能分析并推

理网络攻击关系；国内的绿盟科技针对网络安全领域的多源异构数据提出智慧安全知识图谱，来辅助安全事件分析、进行 APT 追踪等。

人工智能安全作为信息安全领域较新的研究方向，其知识图谱的研究还处于起步阶段，相关研究少，还没有针对人工智能安全领域的知识图谱应用项目。虽然欧洲网络与信息安全局 (ENISA)^[9]对人工智能网络安全生态系统及威胁图谱概念进行了描述，定义 AI 安全生命周期并给出具体威胁分类，但并未从知识图谱角度对本体结构、可视化方法给出完整框架。因此，本文围绕人工智能安全知识图谱的构建，主要贡献如下。

1) 提出了一种针对人工智能安全领域的多层次人工智能安全本体结构 (MLAISO, multi-layer artificial intelligence security ontology)。本文整合分析了人工智能安全领域的多源异构数据及元素间的关系，包括攻击信息、数据信息和其他安全数据等，针对人工智能安全知识图谱构建存在的实体概念关联复杂问题，将本体结构分为原理层、问题层与措施层，使人工智能安全本体更加多元化和扩展化，措施层的增加为后期知识推理阶段提供更为准确的安全决策推理。

2) 提出了一种人工智能安全知识库构建方法。依据 MLAISO 对当前结构化以及多源异构数据进行整合，按照其规则描述来约束人工智能安全数据，形成针对命名实体识别和关系抽取任务的数据集 AISKE (artificial intelligence security knowledge

extraction)。继而采取基于 BiLSTM-CRF 的模型进行命名实体识别，有效利用前后特征信息和句子级别标记信息；采取基于 CNN-ATT 的远程监督关系抽取方法进行关系抽取，在充分提取特征的同时解决噪声数据干扰和长距离依赖问题，实验证明 AISKE 数据集表现良好。

3) 实现了人工智能安全知识图谱的多维度可视化。通过使用 Vue 框架结合 3d-force-graph.js 力导图工具绘制 3D 效果的“节点链接图”形式的知识图谱可视化界面，全面清晰地展示人工智能安全知识之间的关联，评估结果显示图谱的质量较高，能充分体现人工智能安全领域的最新研究成果，有效解决多源异构人工智能安全数据复杂关联的科学表示问题。

1 人工智能安全知识图谱构建框架

人工智能安全知识图谱构建框架如图 1 所示，该框架包括以下 3 个部分。

1) 本体构建：整个构建过程的基础，贯穿全程^[10]，针对人工智能安全领域知识进行形式化定义，对人工智能安全本体进行准确描述；随着人工智能安全领域发展对本体实例及时迭代更新，如近年出现的标签翻转对抗攻击、快速梯度下降法 (FGSM, fast gradient sign method) 对抗训练等。

2) 知识抽取：又称信息抽取，是构建框架中的核心过程。来自标准数据库的结构化数据以及聚焦爬虫获取的多源异构数据通过数据清洗形成

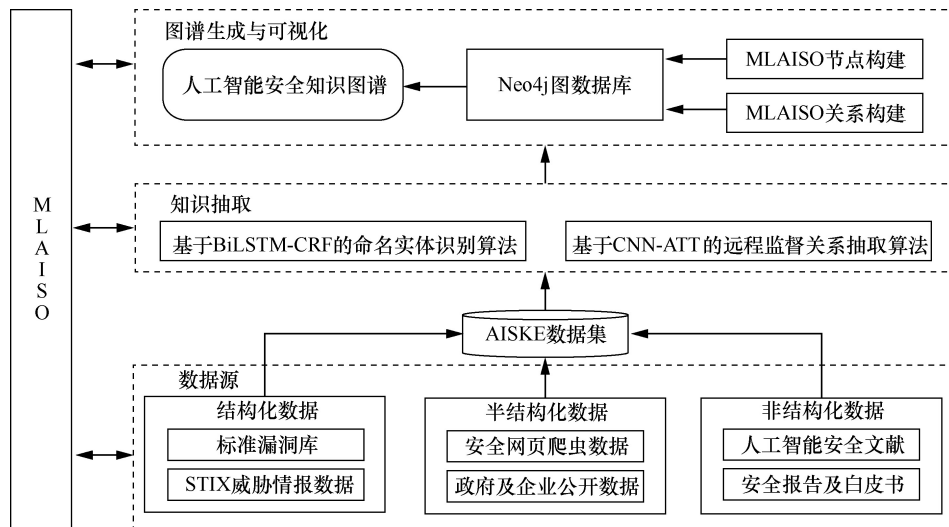


图1 人工智能安全知识图谱构建框架

Figure 1 The framework of artificial intelligence security knowledge graph

AISKE 数据集,运用深度学习方法自动化地、高效地进行实体抽取和关系抽取。

3) 图谱生成与可视化:采取图数据库 Neo4j 的知识图谱三元组存储方法,利用 JavaScript、Vue 等开发技术和框架实现安全知识图谱的构建与 3D 展示。

2 分层人工智能安全本体

本体是一个通用数据模型,其中包含了某个领域内的基本术语和术语之间的关系等。安全本体作为人工智能安全知识图谱构建的关键基础,实现由实际数据向知识转化的过程,为后续知识推理过程提供规则约束^[11]。

随着人工智能技术的飞速发展,人工智能生命周期趋于复杂化和高级化,各个维度的概念趋于多样性和关联性,单层的人工智能领域或网络安全领域的本体已经难以涵盖目前人工智能安全数据源的全部概念,这种局限性会限制知识图谱的作用。为了更好地对多源人工智能安全数据进行处理与整合,考虑人工智能安全当前研究热点及发展方向,本文将本体分为 3 个层次,MLAISO 表达为 $MLAISO = \{Entity, Relation, Layer\}$,其中 $Entity = \{\text{威胁主体, 目标客体, 攻击活动, 攻击工具, 攻击模式, 恶意程序, 漏洞, 攻击结果, 威胁, 安全事件, 事件报道, 影响, 防御措施, 人因, 技术, 设施}\}$,共 16 种实体类型; $Relation = \{\text{发起, 使用, 针对, 利用, 作用, 包含, 导致, 破坏, 引发, 来源, 采取}\}$,共 11 种关系类型; $Layer = \{\text{原理层 (T 层, theory layer), 问题层 (P 层, problem layer), 措施层 (M 层, measure layer)}\}$ 。多层次人工智能安全本体结构如图 2 所示。

参考 STIX2.1^[12]和我国国家标准《信息安全技术网络安全威胁信息格式规范》(GB/T 36643—2018),不同层次定义如下。

1) 原理层:是整个本体模型的核心,描述人工智能安全领域相关威胁信息和攻击信息,展现了不同人工智能安全实体间的语义关系。 $Entity_T = \{\text{威胁主体, 目标客体, 攻击活动, 攻击工具, 攻击模式, 恶意程序, 漏洞, 攻击结果}\}$, $Relation_T = \{\text{发起, 使用, 针对, 利用, 作用, 包含, 导致, 破坏}\}$ 。MLAISO 原理层实体描述如表 1 所示。

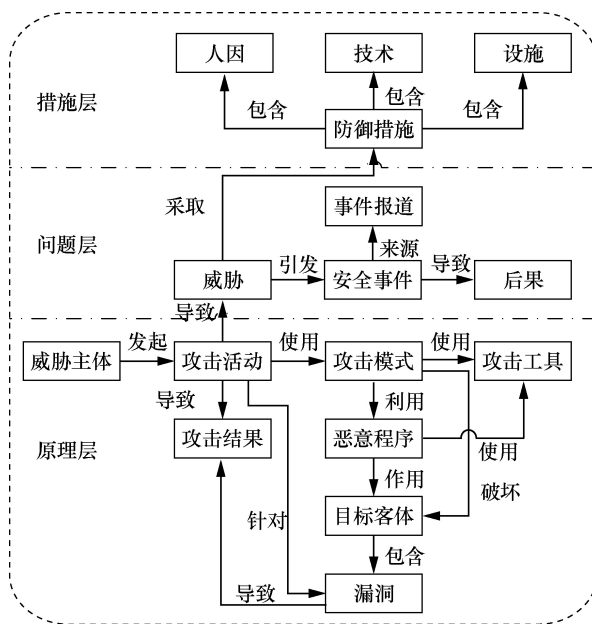


图2 多层次人工智能安全本体结构
Figure 2 Multi-layer artificial intelligence security ontology

表1 MLAISO 原理层实体描述
Table 1 Entity description of MLAISO theory layer

实体类型	描述
威胁主体	攻击活动发起者, 可以为个人、团体或组织
目标客体	攻击目标资产, 包括设备、数据等
攻击活动	针对目标客体的一系列攻击行为
攻击工具	威胁主体在攻击活动中常用的工具, 包括防御响应中所使用的工具
攻击模式	威胁主体试图攻击目标客体的方式, 如 SQL 注入等
恶意程序	未经授权或者破坏系统的代码, 目的为破坏相关数据、程序或者其他资产的机密性、完整性和可用性
漏洞	威胁主体或目标客体容易遭受威胁的特征
攻击结果	攻击活动的影响

2) 问题层:定义人工智能安全领域所存在的各种待解决问题,威胁是本层次的核心实体。 $Entity_P = \{\text{威胁, 安全事件, 事件报道, 影响}\}$, $Relation_P = \{\text{引发, 来源, 导致, 采取}\}$ 。

威胁通常指获取和使用人工智能安全赋能 AI 安全系统、设备等目标客体的安全行为,充分利用威胁信息能够进一步提高主动防御性能。威胁会“引发”安全事件,其中包含所发生事件的大量信息,相比 P 层的攻击活动更为详细地描述了人工智能安全的完整信息;这些安全事件会“导致”不同程度的影响,如滥用、故障、物理攻击等^[13]; MLAISO 结构中的安全事件均“来源”于

事件报道,如安全新闻网站报道、人工智能安全数据标准、相关期刊文献等。

3) 措施层:从不同角度给出人工智能安全领域针对不同安全事件的防御措施,从人因、技术和设施3个角度展开防御措施的介绍。 $Entity_M=\{\text{防御措施, 人因, 技术, 设施}\}$, $Relation_M=\{\text{包含}\}$ 。

人因指通过人为因素实施的防御措施,如加强信息安全意识教育、合理化安全管理制度等;技术指通过技术角度执行的防御措施,如安全事件智能化检测、优化算法模型等;设施是从基础资产角度进行防御,如智能化检测基础架构、管理服务资源等。

3 基于深度学习的信息抽取算法

对于信息抽取中的核心任务命名实体识别(NER, named entity recognition)和关系抽取,本文针对人工标注的 AISKE 数据集,给出了一种基于神经网络模型 BiLSTM-CRF^[14]的人工智能安全命名实体识别方法,以及一种基于 CNN-ATT^[15]的远程监督关系抽取方法。

3.1 命名实体识别

命名实体识别属于自然语言处理中的序列标注任务,旨在从文本中识别出特定命名指向的词,并对其进行分类,即输入自然语言序列,能够给出对应的标签序列。本文采取监督学习方法进行命名实体识别,当应用监督学习方法时 NER 便被建模为序列标注问题^[16],步骤可分为数据预处理和模型训练,满足评估有效结果的语料转换为 csv 格式即可作为知识图谱可视化源数据,实现对语料库的可视化数据更新,人工智能安全命名实体识别流程如图3所示。

3.1.1 数据预处理

由于完整的 MLAISO 结构所涉及的实体类型过多,且部分实体数据缺乏广泛性并存在中英文混杂问题,故本文以一个基于 MLAISO 的子集 Sub_MLAISO 来说明方法,以构造一个人工智能安全子图谱。

$Sub_MLAISO = \{Sub_Entity, Sub_Relation, Layer\}$, 其中, $Sub_Entity=\{\text{攻击活动 (CAM, campaign), 防御措施 (MS, measure), 目标客体 (OBJ, object), 威胁 (THR, threat), 安全专}$

业术语 (RT, relevant term) $\}$; $Sub_Relation=\{\text{针对, 包含, 导致, 破坏}\}$; $Layer=\{\text{原理层, 问题层, 措施层}\}$ 。

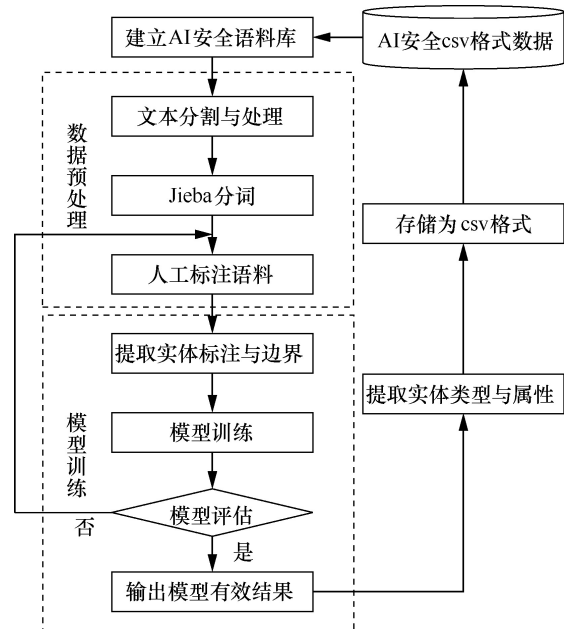


图3 人工智能安全命名实体识别流程
Figure 3 Named entity recognition process in artificial intelligence security

实验语料依据 MLAISO 对当前结构化以及多源异构数据进行整合,包括来自标准漏洞库以及 STIX^[12]威胁情报的结构化数据;通过聚焦爬虫方法获取来自安全网站、人工智能安全文献以及相关报告、白皮书等的半结构化和非结构化数据,部分半结构化数据来自政府和安全企业的公开数据。

在数据预处理模块,首先对已获取的人工智能安全领域内原语料进行文本分割预处理,防止长短句问题对模型结果产生影响;再进行 Jieba 分词和人工标注,针对 Sub_Entity 包含的5类人工智能安全实体,采取 BIO 标注方法,B代表实体开始,I代表实体中间,O表示非实体;通过 word2vec 进行词向量训练后,将标注好的数据集划分成训练集、验证集、测试集之后放入模型训练。

3.1.2 模型训练

本文采取基于深度学习的 NER 方法中的经典模型 BiLSTM-CRF,该模型主体由双向长短期记忆网络(BiLSTM)和条件随机场(CRF)组成,

模型输入是字符特征，输出是每个字符对应的预测标签。输入句子先转换为词向量序列，BiLSTM进行特征提取，接收每个字符的 embedding，并预测每个字符对每一个标注标签的概率，分别计算前向以及后向隐特征并进行拼接。将 BiLSTM 层的预测概率作为 CRF 层输入，利用 CRF 层进行约束，类别序列中分数最高的序列即最终预测结果，BiLSTM-CRF 模型结构如图 4 所示。

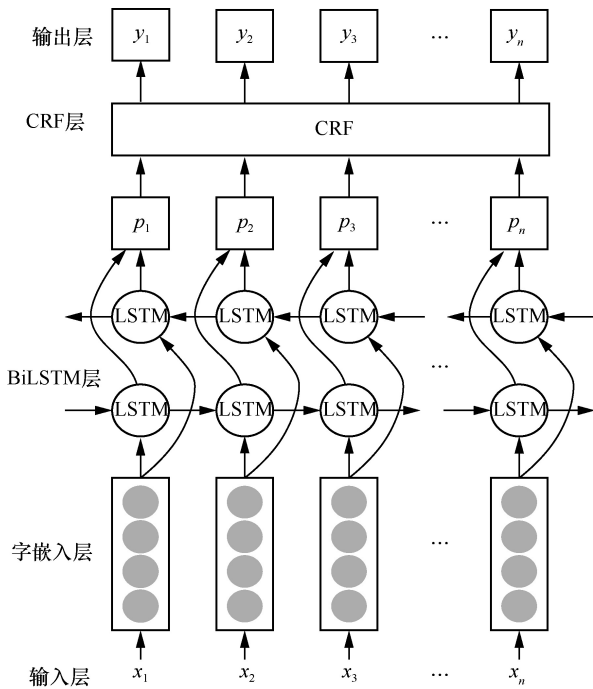


图4 BiLSTM-CRF 模型结构
Figure 4 BiLSTM-CRF model structure

对于 BiLSTM-CRF 模型，将含有 n 个字的句子记作 $\mathbf{x}=(x_1, x_2, \dots, x_n)$ ，在字嵌入层经 word2vec 进行词向量训练后，根据对应的嵌入矩阵实现给定句子的逐字映射，形成融合特征的字向量表示；在 BiLSTM 层提取句子特征，将各个编码序列作为输入，再将前向 LSTM 隐状态序列与后向 LSTM 隐状态序列进行位置拼接，得到完整的隐状态序列，再接入一个线性层，将隐状态向量从 n 维映射到 k 维， k 为标注数据总标签数，得到特征矩阵 $\mathbf{P}=(\mathbf{p}_1, \mathbf{p}_2, \dots, \mathbf{p}_n) \in \mathbb{R}^{n \times k}$ ，将 $\mathbf{p}_i \in \mathbb{R}^k$ 的每一维 p_{ij} 看作字 x_i 划分到第 j 个标签的概率，该特征矩阵 $\mathbf{P}$ 即 BiLSTM 层输出；在 CRF 层进行序列标注，初始化转移矩阵 $\mathbf{A}$ ， A_{ij} 表示从第 i 个标签到第 j 个标签的转移概率，由于添加 start 与 end

标签划分句子边界，在训练过程中逐渐更新的转移矩阵 $\mathbf{A}$ 的实际维度为 $k+2$ 。对于输出序列 $\mathbf{y}=(y_1, y_2, \dots, y_n)$ ，模型对于句子 $\mathbf{x}$ 的输出序列 $\mathbf{y}$ ，定义分数为

$$s(\mathbf{x}|\mathbf{y}) = \sum_{i=1}^n \mathbf{P}_{i, y_i} + \sum_{i=0}^n \mathbf{A}_{y_i, y_{i+1}} \quad (1)$$

在所有可能的标签序列上利用 Softmax 得到归一化后的序列概率。

$$p(\mathbf{y}|\mathbf{x}) = \frac{e^{s(\mathbf{x}|\mathbf{y})}}{\sum_{\tilde{\mathbf{y}} \in \mathbf{Y}_{\mathbf{x}}} e^{s(\mathbf{x}|\tilde{\mathbf{y}})}} \quad (2)$$

其中， $\mathbf{Y}_{\mathbf{x}}$ 代表句子 $\mathbf{x}$ 出现的所有可能标签序列；输出目标是输入 $\mathbf{x}$ 使对应的序列 $\mathbf{y}$ 概率最大，在训练过程中利用最大化似然函数表示如下。

$$\log(p(\mathbf{y}|\mathbf{x})) = s(\mathbf{x}|\mathbf{y}) - \log\left(\sum_{\tilde{\mathbf{y}} \in \mathbf{Y}_{\mathbf{x}}} e^{s(\mathbf{x}|\tilde{\mathbf{y}})}\right) \quad (3)$$

在解码时，使用基于动态规划的 Viterbi 算法来求解最优路径^[16]。

$$\mathbf{y}^* = \arg \max s(\mathbf{x}|\mathbf{y}) \quad (4)$$

以下为一个简单的以字为基本单元 BIOES 标注中文命名实体识别的示例，输入序列为 $X=<x_1, x_2, \dots, x_{13}>$ ，输出序列为 $Y=<y_1, y_2, \dots, y_{13}>$ 。若原语料的语句为“数据投毒攻击破坏数据完整性”，在输入序列和输出序列对应如下： $x_1=$ “数”， $x_2=$ “据”， $x_3=$ “投”， $x_4=$ “毒”， $x_5=$ “攻”， $x_6=$ “击”， $x_7=$ “破”， $x_8=$ “坏”， $x_9=$ “数”， $x_{10}=$ “据”， $x_{11}=$ “完”， $x_{12}=$ “整”， $x_{13}=$ “性”； $y_1=$ “B-CAM”， $y_2=$ “I-CAM”， $y_3=$ “I-CAM”， $y_4=$ “I-CAM”， $y_5=$ “I-CAM”， $y_6=$ “I-CAM”， $y_7=$ “O”， $y_8=$ “O”， $y_9=$ “B-OBJ”， $y_{10}=$ “I-OBJ”， $y_{11}=$ “B-RT”， $y_{12}=$ “B-RT”， $y_{13}=$ “I-RT”。BiLSTM-CRF 算法如下。

算法1 BiLSTM-CRF 算法

输入 $\mathbf{x}=(x_1, x_2, \dots, x_n)$

输出 $\mathbf{y}=(y_1, y_2, \dots, y_n)$

- 1) 对于每个 epoch 循环：
- 2) 对于每个 batch 循环：
- 3) LSTM 模型前向传递，自动学习提取特征；
- 4) 向后状态的前向传递；
- 5) 向前状态的前向传递；

6) CRF 向前和向后传递, 计算序列全局的似然概率;

7) LSTM 模型后向传递;

8) 向后状态的后向传递;

9) 向前状态的后向传递;

10) 更新参数;

11) 结束 batch 循环;

12) 结束 epoch 循环

3.2 关系抽取

关系抽取的目的是挖掘实体之间存在的语义关系, 其结果是关系实例, 构成知识图谱中的边。关系抽取产生的结果为一个三元组 $\langle s, p, o \rangle$, 其中, s 为 subject (主实体), o 为 object (客实体),

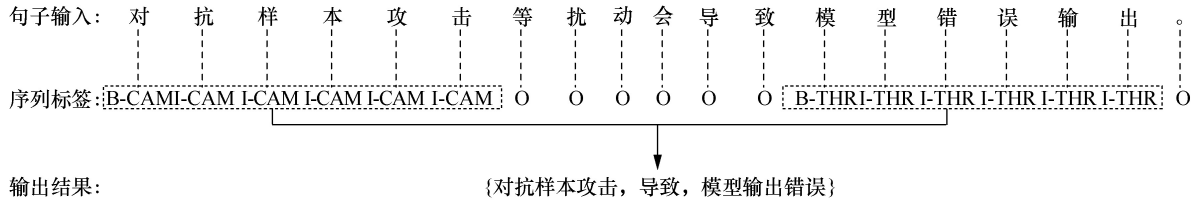


图5 关系抽取输入输出示例

Figure 5 Relation extraction input and output examples

本文采取卷积神经网络 (CNN, convolutional neural network) 结合注意力 (Attention) 机制, 即 CNN-ATT 的远程监督方法进行关系抽取, 在充分提取特征的同时解决噪声数据干扰和长距离依赖问题。在卷积层加入了门控线性单元 (GLU)^[17], 在考虑字词的语义信息的基础上结合词之间的位置信息, 结合残差结构使得信息在多通道传输。GLU 相比 LSTM 模型更加简单, 并且无须依赖前一个时刻的隐藏层状态, 能够实现并行化; 在关系分类时使用 sigmoid 函数而不是 Softmax 函数激活, 有效降低了梯度弥散。GLU 将卷积和门控单元包装在一个预激活的残差块中, 通过宽度为 k 的卷积核平移操作捕捉上下文信息; 通过门结构控制不同层次传递的信息, 相邻层次激活情况不同, 故不能共享参数。卷积层输出 $h_1(\mathbf{x})$ 的计算如式(5)所示。

$$h_1(\mathbf{x}) = (\mathbf{X} * \mathbf{W} + \mathbf{b}) \otimes \sigma(\mathbf{X} * \mathbf{V} + \mathbf{c}) \quad (5)$$

其中, $\mathbf{W} \in \mathbb{R}^{k \times m \times n}$, $\mathbf{b} \in \mathbb{R}^n$, $\mathbf{V} \in \mathbb{R}^{k \times m \times n}$, $\mathbf{c} \in \mathbb{R}^n$ 为要学习的参数, $\mathbf{x}$ 为初始词序列向量, n 为词序列的长度, m 为词向量维度, k 为卷积核大小, σ 为 sigmoid

p 为 predicate (两个实体之间的关系)。例如, $\langle$ 对抗样本攻击, 导致, 模型输出错误 $\rangle$ 中, 对抗样本攻击为主实体, 属于原理层的攻击活动实体类型; 模型输出错误是客实体, 属于问题层的威胁实体类型; 对抗样本攻击的出现导致模型输出错误。

本文将关系抽取的问题定义为: 给定句子 S , 从 S 中抽取其包含的所有三元组 $\langle s, p, o \rangle$ 。考虑 $P(s, p, o) = P(s)P(o|s)P(p|s, o)$, 结合 seq2seq 概率图思路, 先预测 s , 根据传入的 s 预测对应的 o 以及 p 。由于使用命名实体方法识别时, 可以挖掘出句子中包含的实体, 作为关系抽取的主体和客体, 从而所给的句子信息可以对目前已有的三元组进行填充。关系抽取输入输出示例如图5所示。

函数, $\otimes$ 为矩阵之间对应元素相乘, $*$ 为卷积核在不同特征层的卷积运算。式(5)的含义为, 在第 l 层, $\mathbf{x}$ 经参数为 $\mathbf{W}$ 与 $\mathbf{b}$ 的卷积处理后, 由参数为 $\mathbf{V}$ 与 $\mathbf{c}$ 的卷积处理并激活的输出门来控制最终的输出结果。

Attention 机制采取近年力推的 Self Attention^[18], 解决了循环神经网络 (RNN, recurrent neural network) 不能并行计算的问题, 采取 Multi-Head Attention 实现计算并行化, 有效捕捉特征相关性以解决长距离依赖问题; 在运行过程中, Attention 机制会针对文本中的内容自动标注, 减少噪声数据干扰。在具体应用时, 可以将注意力机制简单理解为对于一个已经给定的 query 查询, 计算 query 与各个 key 的相似性得到对应 value 的权重系数, 之后对 value 信息进行加权求和。MultiHead、 head_i 、Attention 的计算分别如式(6)、式(7)、式(8)所示, $\mathbf{Q}$ 、 $\mathbf{K}$ 、 $\mathbf{V}$ 为由输入经过线性变换后的矩阵。

$$\text{MultiHead}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = (\text{head}_1 \oplus \dots \oplus \text{head}_h) \mathbf{W}^O \quad (6)$$

$$\text{head}_i = \text{Attention}(\mathbf{Q} \mathbf{W}_i^Q, \mathbf{K} \mathbf{W}_i^K, \mathbf{V} \mathbf{W}_i^V) \quad (7)$$

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V \quad (8)$$

其中, W_i^Q 、 W_i^K 、 W_i^V 为参数矩阵, W^O 为权重矩阵, $\oplus$ 为拼接操作, h 为多头数目, d 为词向量维度。

对于模型整体结构, 编码阶段融入位置嵌入信息, GLU 进行特征提取, 且 Self Attention 中解码器中的自注意力子层能够防止当前位置对后续位置的关注, 保证位置 i 的预测依赖小于 i 位置的已知输出, CNN-ATT 算法如下。

算法 2 CNN-ATT 算法

输入 $x = (x_1, x_2, \dots, x_n)$

输出 $\tilde{R}$

- 1) 模型参数初始化;
- 2) 词向量和位置向量映射;
- 3) CNN 层从输入层一直到输出层进行前向传播计算, 提取语义特征, 获得共享编码序列;
- 4) 多实例权重计算, 利用 Attention 计算实例向量 a_i ;
- 5) 预测主实体 s ;
- 6) 对于每一个客实体 p , 利用在第 3) 步获得的共享编码序列进行多实例权重计算, 使用 attention 计算实例向量 a_i ;
- 7) 根据抽取的 s 和 p , 预测最终关系;
- 8) 重复 3)-7), 预测所有数据的最终关系

3.3 实验及结果分析

本文采取常用的准确率 (precision)、召回率 (recall) 以及 $F1$ 值 ($F1$ -score) 作为评价指标, 来衡量模型的性能。各个参数的公式计算如下。

$$\text{precision} = \frac{TP}{TP+FP} \quad (9)$$

$$\text{recall} = \frac{TP}{TP+FN} \quad (10)$$

$$F1 = \frac{2 \cdot \text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \quad (11)$$

其中, TP 为实体抽取或关系抽取能够成功识别出的数目, FP 为错误识别出的实体或关系数目, FN 为被遗漏识别的实体或关系数目。

本文在 Pytorch 平台实现人工智能安全信息

抽取实验, 选取的硬件平台为 Intel Core CPU, 实验数据为 AISKE。

3.3.1 命名实体识别实验

本实验采用的 3 种模型均设置了 6 种参数, 预处理时采用 pad_sequences 方法对语句进行预处理, 保证语句长度合适; BiLSTM 层设置了隐藏层单元维数为 128, 词向量维度为 256; 在训练过程中, 选取 Adam 优化器进行优化训练, 设置 batch size 为 64, 学习速率为 0.001, epoch size 为 50, 梯度裁剪为 5。AISKE 内的 NER 数据集中, 训练集实体实例共 2 461 个, 测试集实体实例共 214 个, 命名实体数据统计如表 2 所示。

根据 Sub_MLAISO, 不同模型之间的实验结果对比如表 3 所示。

表 2 命名实体数据统计
Table 2 Data statistic of named entities

实体类型	所属层次	高频实体实例	实体数目/个
攻击活动	原理层	后门攻击、投毒攻击等	157
目标客体	原理层	训练数据、目标模型等	989
威胁	问题层	模型投毒、数据破坏等	148
防御措施	措施层	模型繁华、噪声混淆等	321
安全专业术语	—	鲁棒性、可解释性等	1 056

表 3 不同模型之间的实验结果比较
Table 3 Comparison of experimental results in different models

模型	准确率	召回率	$F1$ 值
CRF	0.838 1	0.834 3	0.833 0
BiLSTM	0.831 0	0.824 2	0.823 1
BiLSTM-CRF	0.864 8	0.849 9	0.851 8

如图 6 所示, 从不同实体的召回率结果看, BiLSTM-CRF 虽然在对威胁的识别效果低于 CRF 模型, 但其他 4 种实体识别效果均好于另外两种模型。

3.3.2 关系抽取实验

关系抽取模型共设置了 4 种参数, 词向量维度为 128, 隐藏层单元维数为 64, batch size 为 64, epoch size 为 300。AISKE 内关系抽取数据集中, 训练集共 217 条, 测试集共 68 条。关系抽取实验结果如表 4 所示。

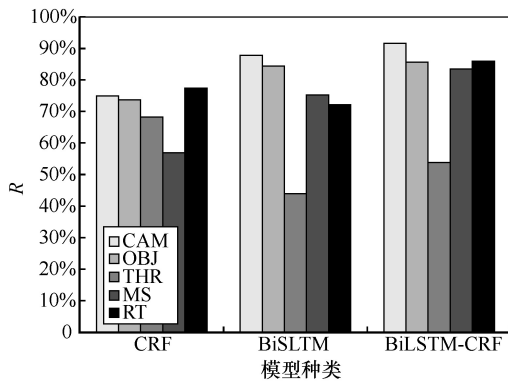


图 6 5 种安全实体的召回率对比

Figure 6 Recall value comparison of five security entities

表 4 CNN-ATT 关系抽取实验结果
Table 4 Experimental results of
CNN-ATT relation extraction

数据集	准确率	召回率	F1 值
AISKE	0.943 9	0.538 4	0.684 7

至此, 人工智能安全知识图谱的基本架构, 即网络图中的关键节点和节点之间的关系构建完毕。

4 可视化分析

经过知识抽取阶段得到人工智能安全数据, 这些数据呈现节点多、关系杂的特点, 而传统的关系数据库 (如 MySQL、Redis 等) 应用效率低, 并且节点边缘定义存在表示不明确的缺陷。因此本文采取 Neo4j 作为人工智能安全知识图谱数据库来存储数据, 使用 Vue 框架结合 3d-force-graph.js 力导图工具绘制 3D 效果的“节点链接图”形式的知识图谱可视化界面。

4.1 可视化解决方案

在知识图谱可视化的过程中, 本文解决了以下几个问题。

(1) 图结构选择

考虑到人工智能安全数据较为稀疏的特点, 本文使用了节点链接图法, 该结构展示法将本体表示为互连节点, 用点或圆等形状表示节点, 节点之间的关系用线表示, 比较适合网络结构数据的可视表达。

(2) 布局优化

节点链接图结构在进行渲染时往往会出现某一部分特别密集而其他地方特别稀疏的缺陷, 但

用户希望知识图谱在屏幕上呈现时尽量铺满整个屏幕, 节点之间距离合理。为了解决该问题, 使用了力导向图的绘图算法^[19], 它是建立在力学模型基础上的一种特殊图表, 可以完成很好的聚类, 方便用户看出节点之间的亲疏关系, 给予良好的可视体验。

(3) 多维度展示

随着图渲染技术的成熟, 2D 效果已经不能很好地满足用户对可视化观察的要求, 而 3D 结构效果更能拉近用户与知识图谱之间的距离, 以立体的、多角度的方式去观察数据之间的联系。为此, 本文利用 3d-force-graph.js 力导图 Web 组件, 使用强制导向的迭代布局来表示三维空间中的图数据结构, 使用 ThreeJS/WebGL 进行 3D 渲染。

4.2 可视结果展示

图 7 为人工智能安全 3D 知识图谱的整体展示, 用户可以使用鼠标左键拖动节点、使用右键改变视角以获得更好的交互视觉体验。

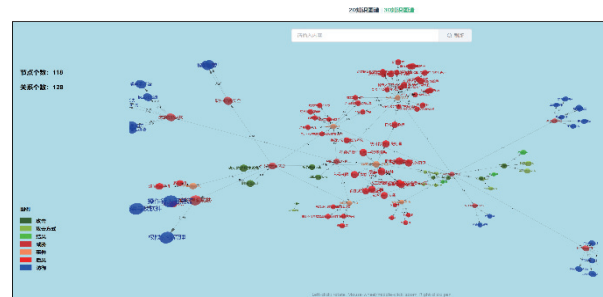


图 7 知识图谱的整体展示

Figure 7 Overall display of knowledge graph

图例部分使用 3 个色系标识 3 种安全数据图例, 绿色系为原理层, 其中深绿色表示攻击、墨绿色表示攻击方式、青绿色表示结果; 红色系表示问题层, 其中深红色表示威胁, 橙红色表示事件, 鲜红色表示后果; 蓝色系表示措施层, 其中深蓝色表示防御技术。用户点击节点可以查看更多信息, 问题层节点、措施层节点分别如图 8、图 9 所示, 左边弹出的节点属性卡显示此节点包含的详细信息。

相较于 2D 平面图, 3D 知识图谱可以更全面清晰地展示出人工智能安全数据之间的关联, 其能够为从事安全的科研人员提供更全面更便捷的视觉体验; 与此同时, 措施层节点同威胁以及攻

击活动关联，对于人工智能安全分析具有重大意义，能够作为图数据库为不同的安全场景提供检测和保护机制，挖掘不同安全实体之间的隐藏关系，推动人工智能安全智能化应用。

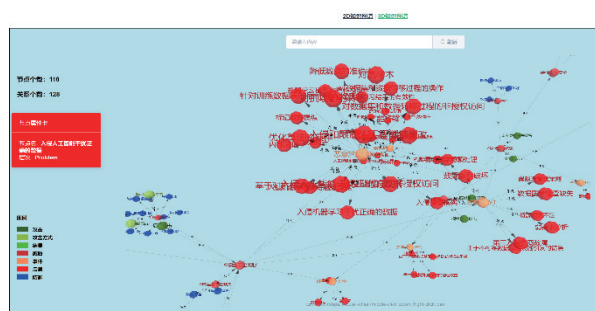


图8 问题层节点
Figure 8 Nodes in the problem layer

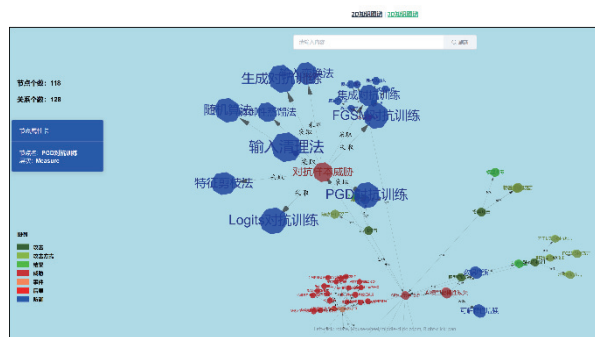


图9 措施层节点
Figure 9 Nodes in the measure layer

4.3 知识图谱质量分析

知识图谱的质量等同于知识图谱中知识的质量，其质量很大程度上决定它在具体应用场景的效用。本文依据肖仰华等^[20]提出的内检与外观质量控制思路对知识图谱的质量做出评价。

将本文所得出的知识图谱与外部知识源进行对比，并采用抽样检查的方式进行分析，得出以下结论：人工智能安全知识图谱的时效性高、完整性高、一致性高，能充分体现人工智能安全领域的最新研究成果，且很少出现知识相互矛盾、重要知识缺乏的情况；但准确性相比时效性、完整性等评估指标略低一些，主要原因是知识抽取环节仍然存在少数实体或三元组抽取不正确的情况，有待进一步提升。

5 结束语

本文提出了针对人工智能安全数据进行处理

与展示的知识图谱构建和可视化方法。建立的人工智能安全分层本体结构 MLAISO，在构建知识图谱表示模型中提供严格的约束规则，能够为后期知识推理阶段提供更为准确的安全决策推理，有效挖掘人工智能安全知识模式。依据 MLAISO 约束并标注 AISKE 数据集，进而采取 BiLSTM-CRF 算法对人工智能安全实体进行抽取，采取 CNN-ATT 算法对其关系数据进行抽取，该算法能够有效利用特征信息以及减少噪声干扰，构建了人工智能安全领域知识库。使用 Vue 框架结合 3d-force-graph.js 力导图工具绘制 3D 效果的“节点链接图”形式的知识图谱可视化界面，可以清晰地展示人工智能目前面临的攻击、威胁以及对应防御技术等信息，为人工智能安全防御提供重要的数据来源。后续研究一方面将进一步拓宽数据来源、调优抽取模型、丰富知识图谱；另一方面将使用该图谱推理和分析网络中存在的人工智能安全攻击事件，使其可视化结果能够作为图数据库为复杂的安全场景提供更好的检测和保护机制，为安全分析工作提供便利，形成良好的人工智能安全生态圈，促进人工智能安全技术的提高和发展。

参考文献：

- [1] 方滨兴, 顾剑铨, 崔翔, 等. 人工智能安全[M]. 北京: 电子工业出版社, 2020: 81-84.
FANG B X, GU Z Q, CUI X, et al. Artificial intelligence safety and security[M]. Beijing: Publishing House of Electronics Industry, 2020.
- [2] AMIT S. Introducing the knowledge graph[R]. America, Official Blog of Google, 2012.
- [3] 秦娅. 网络安全知识图谱构建关键技术研究[D]. 贵阳: 贵州大学, 2019.
QIN Y. Research on key technologies of network security knowledge graph construction[D]. Guiyang: Guizhou University, 2019.
- [4] JIA Y, QI Y L, SHANG H J, et al. A practical approach to constructing a knowledge graph for cybersecurity[J]. Engineering, 2018, 4(1): 53-60.
- [5] 王通, 艾中良, 张先国. 基于深度学习的威胁情报知识图谱构建技术[J]. 计算机与现代化, 2018, 128(12): 21-26.
WANG T, AI Z L, ZHANG X G. Knowledge graph construction of threat intelligence based on deep learning[J]. Computer and Modernization, 2018, 128(12): 21-26.
- [6] IANNAcone M, BOHN S, NAKAMURA G, et al. Developing an ontology for cyber security knowledge graphs[C]//Proceedings of the 10th Annual Cyber and Information Security Research Conference. 2015: 1-4.
- [7] QAMAR S, ANWAR Z, RAHMAN M A, et al. Data-driven analyt-

- ics for cyber-threat intelligence and information sharing[J]. Computers & Security, 2017, 67: 35-58.
- [8] NOEL S, HARLEY E, TAM K H, et al. CyGraph: graph-based analytics and visualization for cybersecurity[J]. Handbook of Statistics, 2016, 35: 117-167.
- [9] MALATRAS A, DEDE G. AI Cybersecurity challenges: threat landscape for artificial intelligence[R]. Greece: European Network and Information Security Agency, 2020.
- [10] CHEN B, LIU Y, ZHENG Y N. Research on ontology-based network security knowledge map[C]//Proceedings of 2018 International Conference on Cloud Computing, Big Data and Blockchain (ICCB). 2018: 1-7.
- [11] 董聪, 姜波, 卢志刚, 等. 面向网络空间安全情报的知识图谱综述[J]. 信息安全学报, 2020, 5(5): 56-76.
- DONG C, JIANG B, LU Z G, et al. Knowledge graph for cyberspace security intelligence: a survey[J]. Journal of Cyber Security, 2020, 5(5): 56-76.
- [12] OASIS cyber threat intelligence (CTI) TC. STIX™ version 2.1[EB].
- [13] 景慧昀, 魏薇, 周川, 等. 人工智能安全框架[J]. 计算机科学, 2021, 48(7): 1-8.
- JING H Y, WEI W, ZHOU C, et al. Artificial intelligence security framework[J]. Computer Science, 2021, 48(7): 1-8.
- [14] HUANG Z H, XU W, YU K. Bidirectional LSTM-CRF models for sequence tagging[J]. arXiv preprint: arXiv: 1508.01991, 2015.
- [15] LI G Z, CHEN X, WANG P, et al. FastRE: towards fast relation extraction with convolutional encoder and improved cascade binary tagging framework[C]//Proceedings of the 31st International Joint Conference on Artificial Intelligence. 2022: 4201-4208.
- [16] LAMPLE G, BALLESTEROS M, SUBRA-MANIAN S, et al. Neural architectures for named entity recognition[J]. arXiv preprint arXiv: 1603.01360, 2016.
- [17] DAUPHIN Y N, FAN A, AULI M, et al. Language modeling with gated convolutional networks[J]. arXiv preprint arXiv: 1612.08083, 2017.
- [18] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[J]. arXiv preprint arXiv: 1706.03762, 2017.
- [19] LIN C C, YEN H C. A new force-directed graph drawing method based on edge-edge repulsion[J]. Journal of Visual Languages & Computing, 2012, 23(1): 29-42.
- [20] 肖仰华, 徐波, 林欣等. 知识图谱概念与技术[M]. 北京: 电子工

业出版社, 2020.

XIAO Y H, XU B, LIN X, et al. Knowledge graph concepts and techniques[M]. Beijing: Publishing House of Electronics Industry, 2020.

[作者简介]



沈晓晨 (1998-), 女, 山东济宁人, 南京师范大学硕士生, 主要研究方向为人工智能安全、知识图谱。



葛寅辉 (1998-), 男, 江苏扬州人, 南京师范大学硕士生, 主要研究方向为人工智能安全。



陈波 (1972-), 男, 江苏南通人, 博士, 南京师范大学教授, 主要研究方向为人工智能安全和智慧教育。



于冷 (1971-), 女, 江苏常州人, 博士, 南京师范大学副教授, 主要研究方向为网络空间安全和智慧教育。