计算机科学
COMPUTER SCIENCE

一种面向开源异构数据的网络安全威胁情报挖掘算法

魏涛, 李志华, 王长杰, 程顺航

引用本文

魏涛, 李志华, 王长杰, 程顺航.一种面向开源异构数据的网络安全威胁情报挖掘算法[J]. 计算机科学, 2023, 50(6): 330-337.

WEI Tao, LI Zhihua, WANG Changjie, CHENG Shunhang. Cybersecurity Threat Intelligence Mining Algorithm for Open Source Heterogeneous Data [J]. Computer Science, 2023, 50(6): 330-337.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.220700073

一种面向开源异构数据的网络安全威胁情报挖掘算法

魏 涛 李志华 王长杰 程顺航
江南大学人工智能与计算机学院 江苏 无锡 214122
(6201924168@stu.jiangnan.edu.cn)

摘 要 针对如何从开源网络安全报告中高效挖掘威胁情报的问题,提出了一种基于威胁情报命名实体识别(Threat Intelligence Named Entity Recognition,TI-NER)算法的威胁情报挖掘(TI-NER-based Intelligence Mining,TI-NER-IM)方法。首先,收集了近10年的物联网安全报告并进行标注,构建威胁情报实体识别数据集;其次,针对传统实体识别模型在威胁情报IoC攻击指示器挖掘领域的不足,提出了基于自注意力机制和字符嵌入的威胁情报实体识别(Threat Intelligence Entity Identification based on Self-attention Mechanism and Character Embedding,TIEI-SMCE)模型,该模型融合字符嵌入信息,再通过自注意力机制捕获单词间潜在的依赖权重、语境等特征,从而准确地识别威胁情报IoC实体;然后,基于TIEI-SMCE模型,提出了一种威胁情报命名实体识别算法;最后,集成上述模型和算法,进一步提出了一种新的威胁情报挖掘方法。TI-NER-IM方法能实现从非结构化、半结构化网络安全报告中自动挖掘威胁情报IoC实体。实验结果表明,与BERT-BiLSTM-CRF模型相比,TI-NER-IM方法的F1值提升了1.43%。

关键词:威胁情报挖掘;自然语言处理;实体抽取;攻击指示器(IoC)
中图法分类号 TP393.08

Cybersecurity Threat Intelligence Mining Algorithm for Open Source Heterogeneous Data

WEI Tao,LI Zhihua,WANG Changjie and CHENG Shunhang
School of Artificial Intelligence and Computer Science,Jiangnan University,Wuxi,Jiangsu 214122,China

Abstract To address the problem of how to efficiently mine threat intelligence from open source network security reports,a TI-NER-based intelligence mining(TI-NER-IM) method is proposed. Firstly,the IoT cybersecurity reports of nearly 10 years are collected and annotated to construct a threat intelligence entity identification dataset. Secondly,in view of the lack of performance of traditional entity recognition models in the field of threat intelligence mining,a threat intelligence entity identification based on self-attention mechanism and character embedding(TIEI-SMCE) model is proposed,which fuses character embedding information. The potential dependency weights between words,contexts and other characteristics are then captured through self-attention mechanism to accurately identify threat intelligence entities. Thirdly,a threat intelligence named entity recognition(TI-NER) algorithm based on TIEI-SMCE model is proposed. Finally,a TI-NER-based intelligence mining(TI-NER-IM) method is designed and proposed. TI-NER-IM method enables automated mining of threat intelligence from unstructured and semi-structured security reports. Eexperimental results show that compared with the BERT-BiLSTM-CRF model,TI-NER-IM's F1 value increases by 1.43%.

Keywords Threat intelligence mining,Natural language processing,Entity extraction,Indicators of compromise

1 引言

网络安全威胁情报由多个维度的攻击指示器(Indicators of Compromise,IoC)构成。攻击指示器通常包括攻击者、漏洞、攻击活动、恶意软件等信息。一条完整的威胁情报的攻击指示器几乎覆盖了攻击活动有关的技术、战术和方法的全部内容。从本质上讲,网络安全威胁情报是一种数据驱动的安全威胁发现和预警技术,主要应用于信息情报交流和共享,同样也适用于传统网络安全领域中不适用"封、堵、查、杀"措施的安全需求应用场景,如物联网和物联网应用领域。当前,网络安全威胁情报主要用于监视、预警和发现网络攻击事件的发生。网络安全威胁情报的来源比较丰富,如网络安全领域的专业人员通过分析新发现、新发生的网络安全事件后所撰写的分析报告等,类似的分析报告广泛存在于互联网中。

到稿日期:2022-07-07 返修日期:2022-09-06
基金项目:工业和信息化部智能制造项目(ZH-XZ-180004);中央高校基本科研业务费专项资金(JUSRP211A41,JUSRP42003)
This work was supported by the Intelligent Manufacturing Project of the Ministry of Industry and Information Technology(ZH-XZ-180004) and Fundamental Research Funds for the Central Universities of Ministry of Education of China(JUSRP211A41,JUSRP42003).
通信作者:李志华(jswxzhli@aliyun.com)

这些报告中通常包含丰富的各种攻击指示器(IoC)，可以为构建威胁情报提供丰富的内容。如何从这些安全分析报告中高效地挖掘 IoC 并构建威胁情报，是学术界面临的一个新挑战[1]。近年来，面向互联网开源异构大数据的威胁情报挖掘已成为学术界的研究热点之一[2]，其中面向自然语言处理的命名实体识别和“实体-关系-实体”挖掘技术，由于其不仅可以高效地挖掘 IoC 信息，而且还可以用来挖掘实体之间的关系，因而倍受学术界关注[3]。遗憾的是，威胁情报的多维属性以及不同属性之间的异构性，导致传统的实体识别模型在网络安全威胁情报挖掘领域表现一般，还有待进一步提高。并且，当前的实际情况是，国际上也缺乏统一、标准的网络安全威胁情报实体数据样本集，而实体挖掘技术又有助于学术界和工业界从互联网上广泛存在的非结构化、半结构化报告文本中有效且高效地挖掘 IoC 信息，以此来生成网络安全威胁情报。

本文对所收集的近 10 年的物联网安全分析报告进行标注，根据 STIX(Structured Threat Information Expression)2.1 标准[4]构建了 4 类威胁情报实体，即网络安全威胁情报 IoC，包括攻击者信息、恶意软件名称、漏洞和地理位置。以它们为例进行讨论，并进一步构建了关于这 4 类实体的威胁情报数据样本集。为了提高威胁情报实体的识别效率，提出了一种基于自注意力机制和字符嵌入的威胁情报实体识别(TIEI-SMCE)模型。该模型融合了单个字符的编码向量，能更加完整地表示文本信息，并借助自注意力机制关注到不同单词在句子中的重要性，可以获得较好的实体识别精度。本文进一步设计了一种基于 TIEI- SMCE 模型的威胁情报命名实体识别(TI-NER)算法，该算法能对文本进行命名实体识别。最后，集成上述算法，提出了一种基于 TI-NER 实体识别算法的开源威胁情报挖掘(TI-NER-IM)方法。TI-NER-IM 方法能从非结构化、半结构化网络安全报告中实现攻击指示器 IoC 的自动化挖掘。

2　相关工作

通常，网络安全报告中包含各种命名实体(Named Entity)，如组织名、人名、恶意软件名称、漏洞名称等。命名实体是网络安全报告中最具价值的信息，是构建威胁情报的 IoC 的重要来源。命名实体识别(Named Entity Recognition，NER)可以从非结构化、半结构化网络安全报告中提取 IoC 实体，生成结构化的威胁情报。

早期自然语言处理领域的命名实体识别主要是基于规则的方法，这类方法需要依靠大量人工设计的规则，而众多实体规则的制定依赖领域专家，不同领域的规则不方便复用。随着自然语言处理技术的发展，基于神经网络的模型被广泛用于命名实体识别，文献[5-7]借助神经网络可以自主学习样本数据的特征，避免了大量的特征工程训练，具有很好的复用性。例如，借助 BERT(Bidirectional Encoder Representation from Transformers)[6]等预训练语言模型或其改进的模型，可以在下游任务中取得良好的效果，避免了从零开始训练新模型的不足。文献[8]将长短期记忆网络与 CRF(Conditional Random Field)相结合对文本进行实体识别，该模型在不同语言的实体识别中都取得了较好的识别效果。文献[9]使用字符嵌入作为输入，将 BiLSTM(Bidirectional Long Short-Term Memory)作为语义编码层，借助 BiLSTM 捕获文本的上下文信息，同时使用条件随机场作为标签解码层，对文本序列中的人名、地理位置等实体进行识别，实验证明 BiLSTM-CRF 识别性能优于 BiLSTM。文献[10]使用 Word2vec 获得输入文本序列的编码向量，再将编码向量送入双向长短期记忆网络和条件随机场进行序列标注，结果表明 Word2vec 能有效提升序列标注的性能。文献[11]提出了一种迁移学习模型，该模型利用 BERT 编码器获得丰富的文本表征，并用双向长短期记忆网络和条件随机场进行编码和解码，与 BiLSTM-Attention-CRF 相比，其实体识别性能得到了很大提高。文献[12]使用 ALBERT(A Lite Bidirectional Encoder Representation from Transformers)预训练模型获得文本表征，以此来加速模型的训练过程，使用双向 LSTM 和条件随机场实现对命名实体的准确识别，实验结果表明，ALBERT-BiLSTM-CRF 模型的实体识别性能优于 BiLSTM-CRF 模型。

不同于其他领域，网络安全威胁情报领域存在公开数据集偏少的问题，且网络安全威胁情报 IoC 实体的专业性较强，具有强烈的专业领域背景，需要结合具体的语境来理解。由于网络安全威胁情报领域存在大量的专业词汇，传统的实体识别模型在网络安全威胁情报领域表现较差，无法直接应用于网络安全威胁情报之 IoC 挖掘领域。

3　数据样本集构建

标准数据样本集是其他学科与人工智能学科进行交叉、样本训练的基础，但目前，网络安全威胁情报领域缺乏大规模的实体识别标准数据样本集。另外，网络安全领域报告的实体数据多源异构，存在大量 PDF 和 HTML 文档等。因此，网络安全威胁情报实体识别数据样本集的构建显得尤为重要。本文的实验数据集是从开源互联网数据中自动爬取的 227 篇专门的物联网安全博客和文章，其格式主要是 HTML 和 PDF。网络安全威胁情报实体识别标准数据样本集的构建流程如图 1 所示。

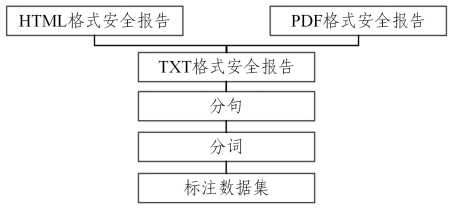


图 1　标准数据样本集构建

Fig. 1　Construction of standard data set

3.1　数据预处理

这些与物联网安全话题相关的 227 篇经典安全报告的内容包括：对典型恶意软件的分析、对物联网攻击活动的分析等。

首先，借助 python 的 Beautiful Soup[14]编写处理程序将所有 HTML 格式的网络安全报告转换为 TXT 文本。由于 HTML 页面中包含有许多噪声信息，如广告、无关链接等，因此在处理程序中同样需要过滤这些噪声数据，以生成高质量

TXT 格式的文本。PDF 格式的网络安全报告通过改进的 PDFMiner[14] 进行转换，同样通过处理程序过滤 PDF 中的富文本内容，如图片等，最终转换成 TXT 格式的文本。

然后，为了实现实体的最终标注，对文本进行分句和分词处理。本文方法是根据特定的标点符号“.”“?”“!”对 TXT 格式文本进一步分句，划分为单句的格式。

最后，通过改进 BERT 中的 BertTokenizer[7] 对分句结果进行分词操作，完成分词处理。这样，就将单句文本划分为了多个单词，方便进行后续的命名实体标注。

3.2 命名实体标注

根据 STIX 2.1 标准[4] 选取 4 项网络安全威胁情报命名实体进行标注，包括攻击者信息、恶意软件名称、漏洞、地理位置，这些都是构建威胁情报的主要攻击指示器(IoC)。按照 BIO 标注规范[15] 对处理后的 TXT 文本进行标注。标注策略如表 1 所列。

表 1 标注样例

Table 1 Example of labeling

处理后的 TXT 文本	标注标签
Emails	O
from	O
IXESHE	B-ThreatActor
APT	I-ThreatActor
Campaign	I-ThreatActor
were	O
sent	O
from	O
the	B-Location
United	I-Location
States	I-Location

3.3 数据样本集

所选用的 227 篇文档详细的数据和实体统计信息如表 2、表 3 所列。数据集包含 4 类实体：攻击者信息(ThreatActor)、恶意软件名称(Malware)、漏洞(Vulnerability)、地理位置(Location)。进一步将标注后的数据样本集按照 8∶1∶1 的比例进行随机抽取，划分成训练集、验证集和测试集。各实验数据样本子集的构成如表 2、表 3 所列。

表 2 数据集统计

Table 2 Dataset statistics

数据集类别	文档数量	句子数量	实体数量
训练集	181	7071	9290
验证集	23	1044	1364
测试集	23	1008	1322

表 3 实体数量分布

Table 3 Distribution of the number of entities

实体类型	训练集	验证集	测试集
ThreatActor	3584	570	772
Vulnerability	463	63	45
Malware	3754	554	397
Location	1489	177	108

4 TI-NER 实体识别算法

4.1 TIEI-SMCE 模型

为了提升网络安全威胁情报 IoC 实体识别的精准度和准确率，本文提出了一种基于自注意力机制和字符嵌入的威胁情报实体识别(TIEI-SMCE)模型，借助模型抽取网络安全报告中的威胁情报 IoC 实体，并进行命名实体识别。TIEI-SMCE 命名实体识别模型如图 2 所示。

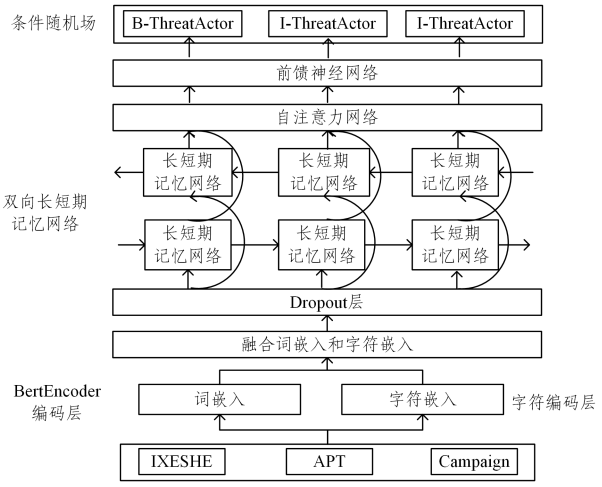


图 2 TIEI-SMCE 命名实体识别模型

Fig. 2 TIEI-SMCE named entity recognition model

在 TIEI-SMCE 网络安全威胁情报 IoC 实体识别模型中，针对网络安全威胁情报句子序列，模型综合考虑了词特征、字符特征、词字的位置特征等信息，并融合了词特征与字符特征。其中字符嵌入是基于单个字符的编码向量，使用每个单词的单个字符作为输入。

词嵌入是通过 BERT Encoder 编码层编码得到的文本表征，BERT Encoder 编码层将 WordPiece 嵌入、位置嵌入、分割嵌入作为输入，通过 Transformer 编码器提取具有上下文语义的文本向量。通过 BERT Encoder 编码层和字符编码层获得词嵌入和字符嵌入后，将词嵌入和字符嵌入融合作为整个句子的文本向量。又因为 TIEI-SMCE 参数规模较大，存在过拟合倾向，因此加入 Dropout 层来阻止过拟合现象的发生。在实验过程中发现，当丢弃率设置过大时，有大量神经网络单元的信息丢失，模型无法很好地拟合训练数据；当丢弃率过小时，模型不会丢失神经网络单元的信息，但是容易出现过拟合现象。经过多次实验比较，最终将 Dropout 层的丢弃率设置为经验值 0.3。完成 Dropout 层的处理后，利用 BiLSTM 对文本向量进行语义编码。然后使用自注意力网络学习语义编码，计算并输出新的自注意力加权的编码向量。为了提高条件随机场 CRF 的解码速度，在条件随机场之前加入前馈神经网络层，自注意力网络的输出向量经过前馈神经网络输入 CRF 解码器。这里，CRF 解码器的输入维度由前馈神经网络输出层的神经单元数量决定。这样处理的主要好处是可以降低条件随机场 CRF 解码器输入向量的维度。

4.2 Char Encoder 字符编码层

字符编码层的主要功能是学习每个单词的字符级别的信息，便于输出文本的字符嵌入。借助双向长短期记忆网络捕获字符的上下文信息，可以学习到单词内部的字符关系，从而丰富语义信息。字符编码器使用每个单词的单个字符作为输入，使用字符嵌入技术将单词中的每个字符映射到低维向量空间中，再通过 BiLSTM 网络捕获字符的上下文信息，从而

得到每个单词基于字符的512维编码向量。LSTM使用门的结构去除或增加信息到记忆单元的状态变量，通过遗忘门、输入门、输出门控制信息的记忆和遗忘，从而解决实体的长期依赖问题。

Char Encoder编码层中，将文本的字符序列 $\boldsymbol{X}^t=(X^{(1)},X^{(2)},\cdots,X^{(n)})$ 作为双向长短期记忆网络的输入。在长短期记忆网络中，$\boldsymbol{X}^{(t)}$ 表示文本字符序列中第 t 个字符，其中长短期记忆网络按如下公式计算：

$$\boldsymbol{i}^{(t)}=\sigma(\boldsymbol{W}^{(i)}[\boldsymbol{H}^{(t-1)},\boldsymbol{X}^{(t)}]+\boldsymbol{b}^{(i)}) \tag{1}$$

$$\boldsymbol{f}^{(t)}=\sigma(\boldsymbol{W}^{(f)}[\boldsymbol{H}^{(t-1)},\boldsymbol{X}^{(t)}]+\boldsymbol{b}^{(f)}) \tag{2}$$

$$\boldsymbol{C}^{(t)}=\boldsymbol{f}^{(t)}\cdot C^{(t-1)}+i^{(t)}\cdot\tanh(\boldsymbol{W}[\boldsymbol{H}^{(t-1)},\boldsymbol{X}^{(t)}]+\boldsymbol{b}) \tag{3}$$

$$\boldsymbol{O}^{(t)}=\sigma(\boldsymbol{W}^{(o)}[\boldsymbol{H}^{(t-1)},\boldsymbol{X}^{(t)}]+\boldsymbol{b}^{(o)}) \tag{4}$$

$$\boldsymbol{H}^{(t)}=\boldsymbol{O}^{(t)}\cdot\tanh\boldsymbol{C}^{(t)} \tag{5}$$

其中，$\boldsymbol{H}^{(t-1)}$ 表示第 $t-1$ 个字符的隐藏层状态，$\boldsymbol{i}^{(t)}$ 表示第 t 个字符记忆门的值，$\boldsymbol{f}^{(t)}$ 表示第 t 个字符遗忘门的输出，$\boldsymbol{O}^{(t)}$ 表示第 t 个字符输出门的值，$\boldsymbol{C}^{(t)}$ 为第 t 个字符的记忆单元的状态变量，$\boldsymbol{C}^{(t-1)}$ 为第 $t-1$ 个字符记忆单元的状态变量，σ 和 tanh 为激活函数，$\boldsymbol{W}^{(i)}$ 和 $\boldsymbol{b}^{(i)}$ 为输入门的权重矩阵和偏置向量，$\boldsymbol{W}^{(f)}$ 和值$^{(f)}$ 为遗忘门的权重矩阵和偏置值，$\boldsymbol{W}^{(o)}$ 和 $\boldsymbol{b}^{(o)}$ 为输出门的权重矩阵和偏置向量，$\boldsymbol{W}$ 和 $\boldsymbol{b}$ 分别为临时记忆单元的权重矩阵和偏置向量。

4.3　BERT Encoder 编码层

BERT Encoder编码层的主要功能是学习文本单词级别的信息，便于得到文本的词嵌入。在此，借助双向Transformer编码器学习单词的上下文语境，以期获得丰富的单词级别特征信息，从而有效解决一词多义问题。BERT是采用双向Transformer编码器的语言表征模型[16]，通过对大规模语料进行无监督学习，生成深层的双向语言向量表征。输入为网络安全威胁情报文本序列中的每个单词。通过多层双向Transformer编码器的训练，最终得到文本的词嵌入。此外，TIEI-SMCE模型使用BERT-Medium预训练模型进行编码，经过多次实验比较，将其隐藏层数和注意力头数均设置成8。与注意力头数为2的BERT-Tiny预训练模型相比，注意力头数为8的BERT-Medium具有更强的文本编码能力，实验结果也说明其具有更好的实体识别效果。

4.4　自注意力网络

自注意力网络的主要功能是关注句子中的重要单词，为实体识别的关键词分配更高的权重，从而注意到不同单词的差异，便于得到注意力机制加权的新向量。自注意力(Self-attention)网络可以为输入特征分配不同的权重，重点注意关键的特征[17]。自注意力网络工作的具体步骤为：对于自注意力网络的输入向量 $\boldsymbol{I}$，首先按照式(6)－式(8)分别计算查询向量 $\boldsymbol{Q}$、关键向量 $\boldsymbol{K}$、值向量 $\boldsymbol{V}$。

$$\boldsymbol{Q}=\boldsymbol{W}^{q}\boldsymbol{I} \tag{6}$$

$$\boldsymbol{K}=\boldsymbol{W}^{k}\boldsymbol{I} \tag{7}$$

$$\boldsymbol{V}=\boldsymbol{W}^{v}\boldsymbol{I} \tag{8}$$

其中，$\boldsymbol{W}^{q}$，$\boldsymbol{W}^{k}$，$\boldsymbol{W}^{v}$ 分别为查询向量、关键向量和值向量的权重矩阵，利用自注意力机制训练权重矩阵，对输入的向量 $\boldsymbol{I}$ 进行计算，得到查询向量 $\boldsymbol{Q}$、关键向量 $\boldsymbol{K}$、值向量 $\boldsymbol{V}$。

然后，将查询向量 $\boldsymbol{Q}$ 与关键向量的转置 $\boldsymbol{K}^{\mathrm{T}}$ 进行点积运算，并除以 $\sqrt{D_K}$，再通过Softmax激活函数计算句子中每个单词的权重，得到每个单词的注意力分数。其中，D_K 为关键向量 $\boldsymbol{K}$ 的维度，除以 $\sqrt{D_K}$ 是为了在训练过程中保持梯度值稳定。最终，将激活函数输出的单词注意力分数与值向量 $\boldsymbol{V}$ 相乘，得到注意力机制加权的新向量 $\boldsymbol{A}$。计算过程如式(9)所示：

$$\boldsymbol{A}=\mathrm{Softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{D_K}}\right)\boldsymbol{V} \tag{9}$$

4.5　Feed Forward 前馈神经网络

如图2所示，Feed Forward前馈神经网络包括输入输出层、隐藏层，通过输出层可以控制CRF层输入向量的维度。又因为当CRF层输入向量维度过大时，会增加标签解码的计算量和模型的复杂度，从而需要消耗大量的时间进行标签解码，所以需要先对向量进行降维，再由CRF层进行标签解码。为了加快条件随机场CRF的解码速度，在此借助Feed Forward前馈神经网络进行特征抽取，把自注意力网络的输出向量降到50维。Feed Forward层的计算式如式(10)所示：

$$\begin{cases}\boldsymbol{Z}^{(l)}=\boldsymbol{W}^{(l)}\boldsymbol{X}^{(l-1)}+\boldsymbol{b}^{(l)}\\ \boldsymbol{X}^{(l)}=f(\boldsymbol{Z}^{(l)})\end{cases} \tag{10}$$

其中，$\boldsymbol{W}^{(l)}$ 和 $\boldsymbol{b}^{(l)}$ 表示 l 层的权重矩阵和偏置项，$\boldsymbol{X}^{(l-1)}$ 表示 $l-1$ 层传入的值，f 表示 l 层的激活函数。

4.6　CRF 标签解码层

CRF标签解码层的主要功能是对输入向量进行解码，预测单词的实体标签。条件随机场CRF可以感知相邻实体标签的依赖关系，确保最终预测标签的合理性，从而提升识别的准确率。网络安全威胁情报IoC实体标签通常具有严格的规则，标签顺序不能随意排列，因此，本文借助CRF对标签进行解码，CRF通过学习标签约束规则来确保最终预测序列的合法性[18]。如图2所示，Feed Forward层的输出是CRF解码层的输入，CRF通过计算Feed Forward层输出序列的得分矩阵来预测标签，以此获取IoC实体的标签预测值。对于输入CRF层的向量 $\boldsymbol{h}=\{h_1,h_2,\cdots,h_n\}$，其得分矩阵的计算式如式(11)所示：

$$S(\boldsymbol{h},\boldsymbol{y})=\sum_{i=1}^{n}P_{i,y_i}+\sum_{i=1}^{n}E_{y_i,y_{i+1}} \tag{11}$$

其中 n 表示待处理文本的输入序列长度；P_{i,y_i} 表示第 i 个单词 h_i 的标签值是 y_i 的概率；$E_{y_i,y_{i+1}}$ 表示当第 i 个单词 h_i 的标签值是 y_i 时，第 $i+1$ 个单词的标签值是 y_{i+1} 的概率。

4.7　TI-NER 实体识别算法

基于3.1节的TIEI-SMCE模型，本节进一步提出网络安全威胁情报IoC命名实体识别(TI-NER)算法。TI-NER算法的主要功能是对文本中的IoC进行实体识别。

TI-NER算法中 $\boldsymbol{E}$ 为文本的单词输入，$\boldsymbol{C}$ 为文本中单词的字符输入，$\boldsymbol{y}$ 为TI-NER算法最终预测的实体标签。单词输入 $\boldsymbol{E}$ 借助BERT Encorder编码为词嵌入 $\boldsymbol{W}$，单词的字符输入 $\boldsymbol{C}$ 借助Char Encoder编码为字符嵌入 $\boldsymbol{T}$。然后，融合词嵌入 $\boldsymbol{W}$ 和字符嵌入 $\boldsymbol{T}$，得到融合后的文本向量 $\boldsymbol{X}$。对于 $\boldsymbol{X}$ 中的每个单词，逐个按照式(1)－式(5)计算LSTM的输入门、遗忘门、记忆单元的状态变量、输出门、隐藏状态，最终计算出所有隐藏状态 $\boldsymbol{H}$。完成隐藏状态计算后，借助自注意力网络

学习单词之间的依赖权重，并按照式(9)计算自注意力加权的新向量 $\boldsymbol{A}$，再根据式(10)，将自注意力加权向量 $\boldsymbol{A}$ 与权重矩阵 $\boldsymbol{W}_f$ 相乘，并加上偏置 b_f，通过激活函数得到向量 $\boldsymbol{h}$，即 Feed ForWard 层的输出向量。最后，利用式(11)计算所有可能的标签序列的得分，得到最高得分的标签序列 $\boldsymbol{y}$，并将其作为预测的实体标签。

TI-NER 算法的伪代码如算法 1 所示。

算法 1 TI-NER 算法

Input: $\mathbf{E}=E_1,E_2,\cdots,E_n$, $\mathbf{C}=C_1,C_2,\cdots,C_n$ // * **E** 为文本的单词输入，**C** 为文本中单词的字符输入

Output: **y** // * TI-NER 算法预测的 IoC 实体标签

1. **W** ← BERT Encoder($\mathbf{E}_1,\mathbf{E}_2,\cdots,\mathbf{E}_n$) // * 编码得到词嵌入
2. **T** ← Char Encorder($\mathbf{C}_1,\mathbf{C}_2,\cdots,\mathbf{C}_n$) // * 编码得到字符嵌入
3. **X** ←Concatenate(**W**,**T**) // * 融合词嵌入和字符嵌入
4. for t=1,2,…,n
5. Begin
6. 使用式(1)计算 t 时刻的输入门
7. 使用式(2)计算 t 时刻的遗忘门
8. 使用式(3)计算 t 时刻的记忆单元的状态变量
9. 使用式(4)计算 t 时刻的输出门
10. 使用式(5)计算 t 时刻的隐藏状态
11. End
12. 使用式(9)计算自注意力网络的输出
13. 使用式(10)计算 Feed Forward 层输出
14. 使用式(11)计算 CRF 层的得分矩阵 S(**h**,**y**)
15. 挑选得分最高的 y 作为最终预测的 IoC 实体标签

算法 1 的时间开销主要来自于长短期记忆网络的计算。假设长短期记忆网络的输入维度为 n，隐藏层的大小为 m，则长短期记忆网络的权重矩阵 $\boldsymbol{W}$ 的大小均为 $n*(n+m)$，偏置向量 $\boldsymbol{b}$ 的长度都为 n，那么计算门结构的时间复杂度为 $O(n*(n+m))$，计算记忆单元中状态变量的时间复杂度为 $O(n)$，计算隐藏状态的时间复杂度 $O(n)$，因此，总的时间复杂度为 $O(3n*(n+m)+n+n)$，即 $O(n^2)$。

5 TI-NER-IM 方法

TI-NER-IM 算法的伪代码如算法 2 所示。

算法 2 TI-NER-IM 算法

Input: reports // * 网络安全报告

Output: STIXOutput // * STIX 格式威胁情报

1. for t=1,2,…,n
2. Begin// * 数据预处理
3. if(report is html) then
4. Begin// * 过滤噪声数据，将 HTML 转为文本格式
5. txt ← Beautiful Soup(report)
6. End
7. else
8. Begin// * 过滤富文本，将 PDF 转为文本格式
9. txt ← PDFMiner(report)
10. End
11. txts[i++] ← txt
12. End
13. for txt ∈ txts
14. Begin
15. result ←call TI-NER(txt) // * 调用算法 1(TI-NER 算法)抽取 IoC 实体
16. STIXOutput[j++]←生成 STIX 格式威胁情报 // * 威胁情报标准化
17. End
18. 输出 STIX2.1 威胁情报 STIXOutput

进一步地，本文提出了基于 TI-NER 实体识别算法的网络安全威胁情报 IoC 挖掘(TI-NER-IM)方法。TI-NER-IM 方法的主要目的是从非结构化、半结构化的网络安全报告中自动挖掘 IoC 攻击指示器，并根据挖掘的 IoC 攻击指示器，生成标准格式的网络安全威胁情报。

TI-NER-IM 方法首先将 PDF 格式和 HTML 格式的网络安全报告自动转换为统一的 TXT 格式文本，并清洗掉文本中的噪声数据；其次对清洗完成的文本进行分句、分词处理；然后借助 TI-NER 实体识别算法进行实体识别，对网络安全威胁情报 IoC 实体信息进行抽取，即抽取各种 IoC 攻击指示器；最后，根据算法的识别结果生成符合 SITX2.1 规范的标准格式的网络安全威胁情报。

6 实验结果及分析

6.1 实验环境与实验参数

本文的实验环境如表 4 所列，GPU 使用 NVIDIA GeForce GTX 1660 SUPER，CPU 为 Intel Core i7-9700，内存大小为 32 GB，在 Windows 10 平台上使用 Python3.6 和 Tensorflow2.1.0 进行实验。

表 4 实验环境参数

Table 4 Experiment environment parameters

参数	参数值
GPU	NVIDIA GeForce GTX 1660 SUPER
CPU	Intel Core i7-9700
内存/GB	32
操作系统	Windows 10
编程语言	Python3.6
深度学习框架	Tensorflow2.1.0
开发工具	PyCharm Community Edition 20202.1

另外，模型实验参数设置如表 5 所列。其中，vocab_size 表示训练数据的词典大小；*embedding_size* 表示 *Word2vec* 模型输出的词嵌入维度；*recurrent_dropout* 表示长短期记忆网络的丢弃率，丢弃率按照 0.4 的概率丢弃相邻的长短期记忆网络单元；为了防止出现过拟合现象，*Dropout* 层按照 0.3 的概率从网络中丢弃神经网络单元；*learning rate* 设置为 1×10^{-4}，表示模型的学习率，其决定了模型参数更新的步长；*batch_size* 设置为 32，表示模型每个批次训练的样本的数量为 32。

表 5 模型实验参数

Table 5 Model experimental parameters

参数	参数值
vocab_size	30000
embedding_size	300
recurrent_dropout	0.4
dropout	0.3
learning rate	1×10^{-4}
batch_size	32

6.2　评价指标

本文的实验仍然采用自然语言处理中命名实体识别常见的评价指标[19]来评估算法和方法的有效性，包括精确率 P（Precision）、召回率 R（Recall）和 $F1$ 值（F1-Score）。各评价指标分别按式(12)－式(14)计算：

$$P=\frac{TP}{TP+FP} \tag{12}$$

$$R=\frac{TP}{TP+FN} \tag{13}$$

$$F1=\frac{2\times P\times R}{P+R} \tag{14}$$

6.3　实验结果与分析

6.3.1　词嵌入模型比较

将 TIEI-SMCE 模型与主流的实体识别模型[10-12]进行比较，这 3 个模型都以词嵌入作为输入。其中，词嵌入处理分别采用 ALBert[12]，BERT[11]，Word2vec[10] 方法进行实验。BERT 输出的词嵌入维度为 512，ALBERT 输出的词嵌入维度为 768，Word2vec 输出的词嵌入维度为 300。语义编码层统一采用 BiLSTM 网络处理，标签解码层统一采用 CRF 实现。比较结果如表 6 所列。

表 6　词嵌入模型实验结果

Table 6　Word embedding model experimental results

模型框架	P	R	$F1$
ALBERT-BiLSTM-CRF[12]	0.7380	0.6653	0.6968
BERT-BiLSTM-CRF[11]	0.7968	0.7039	0.7420
Word2vec-BiLSTM-CRF[10]	0.7339	0.5473	0.6192
TIEI-SMCE	0.8012	0.7205	0.7563

由表 6 可知，TIEI-SMCE 模型的精确率、召回率、F1 值都最高，分别为 80.12%，72.05%，75.63%。这主要是因为 TIEI-SMCE 模型融合了 BERT 的文本表示和字符嵌入，捕获了完整的文本向量，同时自注意力机制可以学习句子中任意两个单词之间的依赖关系，获取句子的内部结构信息，使得网络安全威胁情报实体的识别效果更好。由于 Word2vec 训练出的词向量是静态词向量，无法分辨语境和上下文的差异，极易混淆相似的实体，如攻击者信息实体“Syrian Electronic Army”和地理位置实体“Syrian”，前者是攻击者信息实体的一部分，而后者是地理位置 Location 实体。因此，Word2vec-BiLSTM-CRF 模型对网络安全威胁情报实体识别效果较差，其精确率、召回率、$F1$ 值都为 4 个模型中最低的，特别是 $F1$ 值比 TIEI-SMCE 模型低 13.71%。BERT-BiLSTM-CRF 模型由于使用了预训练语言模型，结合了位置编码、句子特征、分词特征生成的文本表征，同时使用 BiLSTM 充分学习了文本的双向信息，因此 $F1$ 值达到了 74.20%，但是 BERT-BiLSTM-CRF 模型没有充分挖掘单个字符的信息，对于不在其词汇表中的单词，无法准确地对其进行编码，因此 F1 值比 TIEI-SMCE 模型低 1.43%。ALBERT 是 BERT 的轻量级版本，虽然优化了任务训练速度，但是其网络层参数规模较小，因此学习能力比 BERT 差，识别效果也弱于 BERT-BiLSTM-CRF 模型。当各算法均采用 BiLSTM-CRF 进行编码解码时，使用 ALBERT 的模型比使用 BERT 的模型在 $F1$ 值上低 4.52%，这主要是因为 ALBERT 的网络参数规模较小，在提高训练速度的同时降低了实体识别效果。

6.3.2　字符嵌入模型比较

由于 LSTM 可以灵活地选择记忆和遗忘信息，因此成为实体识别中常用的网络结构。使用字符嵌入作为输入，分别验证 LSTM-CRF[8] 和 BiLSTM-CRF[9] 的性能。实验结果如表 7 所列。

表 7　字符嵌入模型实验结果

Table 7　Character embedding model experimental results

模型框架	P	R	$F1$
LSTM-CRF[8]	0.6179	0.5385	0.5713
BiLSTM-CRF[9]	0.6762	0.5532	0.6081
TIEI-SMCE	0.8012	0.7205	0.7563

单词按照字符划分并编码为字符向量后，分别使用 BiLSTM-CRF 模型和 LSTM-CRF 模型进行实体识别。由于缺少分词信息，字符嵌入的模型表现较差，LSTM-CRF 和 BiLSTM-CRF 的 $F1$ 值分别仅有 57.13% 和 60.81%。由于单向 LSTM 无法获取到当前字符的下一个字符信息，但 BiLSTM 能学习文本的上下文信息，因此 BiLSTM 的识别效果比 LSTM 好。LSTM-CRF 和 BiLSTM-CRF 都使用 CRF 作为标签解码层时，BiLSTM 比 LSTM 的 $F1$ 值高 3.86%，这主要是因为 TIEI-SMCE 模型融合了词嵌入特征，可以更好地识别实体的边界，避免了将同一个单词的不同字符划分到不同实体，所以 $F1$ 值达到了 75.63%。

6.3.3　BERT-Tiny 与 BERT-Medium 的性能比较

为了验证不同参数配置对实体识别性能的影响，分别用 BERT-Tiny 与 BERT-Medium 替代本文提出的 TIEI-SMCE 模型中的 BERT Encoder 编码层进行实验比较。二者都采用双向 Transformer 编码器的语言表征模型，但参数配置不同。参数配置如表 8 所列，实验结果如表 9 所列。

表 8　BERT 模型参数

Table 8　BERT model parameters

模型框架	隐藏单元数目	注意力头数	参数规模
BERT-Tiny	128	2	4.3×10^6
BERT-Medium	512	8	40.8×10^6

表 9　Bert 参数规模的性能比较

Table 9　Performance comparison of Bert parameter scales

模型框架	P	R	$F1$
TIEI-SMCE(BERT-Tiny)	0.8051	0.7088	0.7515
TIEI-SMCE(BERT-Medium)	0.8012	0.7205	0.7563

由表 9 可知，在网络安全威胁情报实体识别中，BERT-Medium 比 BERT-Tiny 的效果更好，这说明在实体挖掘的过程中隐藏层单元数和注意力头数越多，对网络安全威胁情报实体识别的效果越好。同时，嵌入 BERT-Tiny 和 BERT-Medium 编码层的 TIEI-SMCE 模型比表 6 和表 7 中列出的词嵌入模型和字符嵌入模型表现都要好。这主要是因为仅使用字符嵌入的模型无法准确地识别实体的边界，而只使用词嵌入的模型无法有效编码不在词汇表中的单词，而本文提出的 TIEI-SMCE 模型融合了字符嵌入和词嵌入，并借助自注意力机制学习句子中单词的依赖权重，从而有效地克服了上述两点不足。这也从另一个方面充分说明 TIEI-SMCE 模型对

网络安全威胁情报领域实体识别的有效性。

6.3.4 加入 Dropout 的有效性验证

Dropout 指在网络训练过程中，对于神经网络单元，按照一定的概率将其从网络中丢弃，以此验证模型的鲁棒性。为了防止本文提出的 TIEI-SMCE 模型出现过拟合现象，本文加入了 Dropout 层来提升模型的鲁棒性和增强模型识别效果。将添加 Dropout 层和未添加 Dropout 层的 TIEI-SMCE 模型在构建的数据集上进行 IoC 实体识别对比实验。实验结果如表 10 所列。

表 10 添加 Dropout 与否的实验结果

Table 10 Experimental comparison of adding a dropout layer or not

模型框架	*P*	*R*	*F*1
TIEI-SMCE	0.7911	0.7181	0.7503
TIEI-SMCE$_{dropout}$	0.8012	0.7205	0.7563

表 10 表明加入 Dropout 层的 TIEI-SMCE 模型在 3 个指标上均表现更好，这主要是因为 TIEI-SMCE 模型的参数规模较大，添加 Dropout 层有效避免了模型出现过拟合现象，模型的鲁棒性更好。

6.3.5 TI-NER-IM 方法的有效性

针对网络安全威胁情报 IoC 的 TI-NER-IM 挖掘方法在 4 类威胁情报 IoC 实体上的有效性实验结果如图 3 所示。这 4 类威胁情报 IoC 实体包括攻击者信息(ThreatActor)、恶意软件名称(Malware)、地理位置(Location)和漏洞(Vulnerability)。

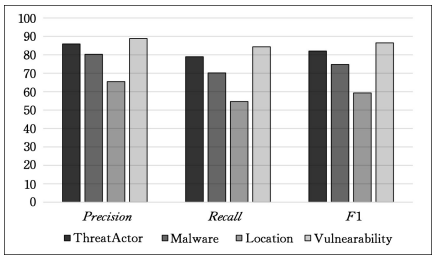


图 3 TI-NER-IM 方法实体挖掘结果

Fig. 3 Entity mining results of TI-NER-IM method

由图 3 可知，漏洞实体的识别效果最好。漏洞类实体用来描述系统漏洞信息和漏洞利用行为，如“OpenSSL Heartbleed Vulnerability”“Badlock Bug”等，上下文特征较为明显，虽然其在数据样本集中标注的数据较少，但是识别效果很好，*F*1 值达到了 86.47%。攻击者信息的识别效果仅次于漏洞实体，其 *F*1 值和精确率分别达到了 82.03%和 85.89%，这是由于攻击者信息在网络安全威胁情报数据样本集中数量较多，训练集中共有 3584 个攻击者信息实体，且 TI-NER-IM 方法融合字符向量的信息，能够学习到攻击者信息的字符特征，如“APT28”“APT23”这类攻击者名称，它们在字符级特征中较为明显。另外，攻击者信息在网络安全报告中出现频次也较高，关于攻击者的描述是网络安全报告的重要信息，也是网络安全威胁情报领域最常用的实体；恶意软件名称是数据样本集中出现频次最高的实体，但是恶意软件名称实体的特征较为复杂，对其名称的称呼容易受攻击者个人习惯的影响，这增加了识别难度，但是由于 TI-NER-IM 方法通过自注意力机制学习上下文语境和两个单词之间的依赖权重，其 *F*1 值和精确率依然达到了 74.72%和 80.29%；地理位置实体识别效果是 4 类实体中最差的，比恶意软件名称实体的 *F*1 值低 15.42%，这主要是因为地理位置信息在网络安全报告中极为珍贵，因为攻击者一般会隐藏自己的地理位置，所以网络安全报告中地理位置实体较少，数据集中仅有 1489 个地理位置实体，地理位置的精确率和召回率也较低，分别为 65.43%和 54.69%。

图 4 给出了使用 TI-NER-IM 方法从关于“Dropping Elephant”攻击活动的网络安全报告[20]中提取的网络安全威胁情报。图中的网络安全威胁情报表明此次攻击活动的攻击者为“Dropping Elephant”，该攻击者的地理位置为“India”，并且其在攻击过程中使用“backdoor”恶意软件。可见，虽然仅抽取了 3 类威胁情报 IoC 实体，却依然比较完整地描述出了这次网络攻击事件的总体轮廓。这从另一个角度说明了 TI-NER-IM 方法的有效性和一定的实用性。即它可以有效地从非结构化网络安全报告中自动挖掘网络安全威胁情报。

```
{
  "type":"bundle",
  "id":"bundle--f59abc20-6837-486a-951b-80216ade0662",
  "objects":[
    {
      "type":"location", //表明该攻击者的地理位置为"India"
      "spec_version":"2.1",
      "id":"location--03d001a3-641a-4705-ad5d-ee74a657c979",
      "created":"2021-08-26T12:28:38.167973Z",
      "country":"India"
    },
    {
      "type":"malware", //表明攻击者使用了"backdoor"恶意软件
      "id":"malware--3b4277da-6879-4fa9-8aaa-34626a0d6a72",
      "name":"backdoor",
      "is_family":false
    },
    {
      "type":"threat-actor", //表明攻击者为"Dropping Elephant"
      "id":"threat-actor--8e2e2d2b-17d4-4cbf-938f-98ee46b3cd3f",
      "name":"Dropping Elephant"
    }
}
```

图 4 IoC 构建威胁情报示例

Fig. 4 Example of building threat intelligence from IoC

结束语 本文提出了一种 TI-NER-IM 威胁情报挖掘方法，借助数据预处理、实体识别算法，从非结构化、半结构化的网络安全报告中挖掘威胁情报 IoC 实体，并标准化成 STIX2.1 格式的威胁情报。在进行 IoC 实体识别时，为了解决字符嵌入模型无法有效识别实体边界的问题，提出了基于自注意力机制和字符嵌入的 TIEI-SMCE 模型，提高了模型识别实体边界的精准度，从而有效提升了 IoC 实体的识别效果。在数据样本集上，将 TIEI-SMCE 模型与主流的实体识别模型进行对比，实验结果表明 TIEI-SMCE 模型的实体识别效果更优。另外，还对不同 BERT 参数规模进行了对比实验，证明了使用 BERT-Medium 编码层的 TIEI-SMCE 模型的 *F*1 值更高，达到了 75.63%。为了解决 TIEI-SMCE 模型参数规模

较大、易拟合的问题，给 TIEI-SMCE 模型加入了 Dropout 层，并进行对比实验，说明 Dropout 层对实体识别有提升效果。实验证明了 TIEI-SMCE 可以有效抽取网络安全报告中的各类威胁情报实体。

但是，目前的模型和算法局限于对英语网络安全威胁情报 IoC 实体进行识别和抽取，且网络安全威胁情报 IoC 实体比较多，而本文仅考虑了比较典型的 4 类实体。如何实现全语种和所有网络安全威胁情报 IoC 实体的有效且高效的识别，是下一步的主要研究工作。

参 考 文 献

[1] CASCAVILLAG, TAMBURRI D A, VAN DEN HEUVEL W J. Cybercrime threat intelligence: A systematic multi-vocal literature review[J]. Computers & Security, 2021, 105: 102258.

[2] BIANCHIG, CONTI M, DARGAHI T, et al. Editorial for the Special Issue on Sustainable Cyber Forensics and Threat Intelligence[J]. IEEE Transactions on Sustainable Computing, 2021, 6(2): 182-183.

[3] WU H, LI X, GAO Y. An effective approach of named entity recognition for cyber threat intelligence[C]// 2020 IEEE 4th Information Technology, Networking, Electronic and Automation Control Conference(ITNEC). IEEE, 2020, 1: 1370-1374.

[4] BARNUM S. Standardizing cyber threat intelligence information with the structured threat information expression(stix)[J]. Mitre Corporation, 2012, 11: 1-22.

[5] MOHIT B. Named entity recognition [M]// Natural Language Processing of Semitic Languages. Berlin: Springer, 2014: 221-245.

[6] LI J, SUN A X, HAN J L, et al. A survey on deep learning for named entity recognition[J]. IEEE Transactions on Knowledge and Data Engineering, 2020, 34(1): 50-70.

[7] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is allyou need[C]// Proceedings of the 2017 Advances in Neural Information Processing Systems. California, 2017: 5998-6008.

[8] LEE C. LSTM-CRF models for named entity recognition[J]. IEICE Transactions on Information and Systems, 2017, 100(4): 882-887.

[9] ARKHIPOV M Y, BURTSEV M S. Application of a Hybrid Bi-LSTM-CRF model to the task of Russian Named Entity Recognition[C] // Conference on Artificial Intelligence and Natural Language. Cham: Springer, 2017: 91-103.

[10] DASGUPTA S, PIPLAI A, KOTAL A, et al. A comparative study of deep learning based named entity recognition algorithms for cybersecurity[C]// 2020 IEEE International Conference on Big Data(Big Data). IEEE, 2020: 2596-2604.

[11] LIU S, YANG H, LI J, et al. Chinese Named Entity Recognition Method in History and Culture Field Based on BERT[J]. International Journal of Computational Intelligence Systems, 2021, 14(1): 1-10.

[12] HAO W, KEROU L, ZHEN M, et al. Identifying Multi-Type Entities in Legal Judgments with Text Representation and Feature Generation[J]. Data Analysis and Knowledge Discovery, 2021, 5(7): 10-25.

[13] THIVAHARAN S, SRIVATSUN G, SARATHAMBEKAI S. A survey on python libraries used for social media content scraping[C]// 2020 International Conference on Smart Electronics and Communication(ICOSEC). IEEE, 2020: 361-366.

[14] MIAHM S U, SULAIMAN J, SARWAR T B, et al. Sentence boundary extraction from scientific literature of electric double layer capacitor domain: tools and techniques [J]. Applied Sciences, 2022, 12(3): 1352.

[15] LIU X, CHEN H, XIA W. Overview of Named Entity Recognition[J]. Journal of Contemporary Educational Research, 2022, 6(5): 65-68.

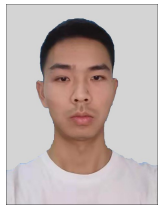
[16] KENTON J D M W C, TOUTANOVA L K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding [C]// Proceedings of NAACL-HLT. 2019: 4171-4186.

[17] NIU Z, ZHONG G, YU H. A review on the attention mechanism of deep learning[J]. Neurocomputing, 2021, 452: 48-62.

[18] YU B, FAN Z. A comprehensive review of conditional random fields: variants, hybrids and applications[J]. Artificial Intelligence Review, 2020, 53(6): 4289-4333.

[19] LI J, SUN A, HAN J, et al. A survey on deep learning for named entity recognition[J]. IEEE Transactions on Knowledge and Data Engineering, 2020, 34(1): 50-70.

[20] LI Z, CHEN Q A, YANG R, et al. Threat detection and investigation with system-level provenance graphs: a survey[J]. Computers & Security, 2021, 106: 102282.



WEI Tao, born in 1998, postgraduate. His main research interests include information system analysis and information security.

LI Zhihua, born in 1969, Ph.D, professor, master supervisor. His main research interests include the key technologies and information security of the end edge cloud, and its intersection with cutting-edge disciplines such as artificial intelligence.

（责任编辑：何杨）