计算机科学
COMPUTER SCIENCE

面向流程工业控制的双安融合知识图谱研究

王敬, 张淼, 刘杨, 李昊霖, 李昊天, 王佰玲, 魏玉良

引用本文

王敬, 张淼, 刘杨, 李昊霖, 李昊天, 王佰玲, 魏玉良. 面向流程工业控制的双安融合知识图谱研究[J]. 计算机科学, 2023, 50(9): 68-74.

WANG Jing, ZHANG Miao, LIU Yang, LI Haoling, LI Haotian, WANG Bailing, WEI Yuliang. Study on Dual-security Knowledge Graph for Process Industrial Control[J]. Computer Science, 2023, 50(9): 68-74.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.230500233

面向流程工业控制的双安融合知识图谱研究

王 敬 张 淼 刘 杨 李昊霖 李昊天 王佰玲 魏玉良
哈尔滨工业大学(威海)计算机科学与技术学院 山东 威海 264209
(wangjing_122@126.com)

摘 要 随着工业控制系统不断走向现代化和智能化,工业控制系统的安全问题日益凸显。然而,传统的工业控制系统往往仅关注信息安全或生产安全,不能同时兼顾两方面的安全问题。知识图谱作为一种结构化的数据表现形式,能够存储领域知识并建模知识之间的因果关系。现有研究大多使用知识图谱解决网络安全问题,鲜有研究将知识图谱用于解决工业控制系统的信息与生产安全问题。文中提出了一种面向流程工业控制系统的双安融合知识图谱构建方法,通过基于BERT的命名实体模型和图对齐等技术,有效地从工控领域网络安全数据库和实际化工生产相关文档中提取了实体和关系,并构建了流程工业双安融合知识图谱。该知识图谱融合了化工生产流程特征和网络攻击行为特征,能通过两种特征知识间的耦合关系为工控系统提供综合的网络安全和生产安全保障。

关键词:知识图谱;工业控制系统;双安融合;知识图谱构建;网络安全;生产安全
中图法分类号 TP391

Study on Dual-security Knowledge Graph for Process Industrial Control

WANG Jing,ZHANG Miao,LIU Yang,LI Haoling,LI Haotian,WANG Bailing and WEI Yuliang
School of Computer Science and Technology,Harbin Institute of Technology(Weihai),Weihai,Shandong 264209,China

Abstract With the development of industrial control systems,security issues in these systems have become increasingly important. However,traditional industrial safety systems usually focus on either information security or production safety,thus failing to consider both issues at the same time. As structured representation of data,knowledge graph(KG) is capable of hosting domain-specific knowledge and modeling causal relationships among knowledge. However,most studies leverage KG to handle cybersecurity,while rarely pay attention to information security and production safety problems in industrial control systems. This paper proposes a set of construction methods for dual-security KG for process industrial control systems. Using the techniques of named entity recognition and relation extraction,it builds a large number of dual-security knowledge triples from a real-world production corpus. The built KG incorporates both features of chemical industry production process and potential network security flaws,providing comprehensive security guarantee for industrial control system.

Keywords Knowledge graph,Industrial control system,Dual security,Knowledge graph construction,Cybersecurity,Production safety

1 引言

随着信息化和工业化的深度融合,现代工业控制系统正逐步向互联、开放和智能化的模式转变。这种转变在提升生产效率的同时,也伴随着更多信息接口的开放,使得工业控制系统连接到内部IT网络和外部互联网的能力增强。然而,这也为工业系统带来了新的安全挑战[1]。

在现代流程工业控制系统中,安全防护的重要性日益凸显[2]。传统的工业安全防护体系大多基于通用安全威胁模型实现异常检测和边界防护等安全机制,缺乏对流程工业生产过程的特征挖掘,无法全面分析和防御涉及信息安全和生产安全的攻击。同时,工业控制系统具备典型的信息物理融合系统[3](Cyber-Physical System,CPS)特征,需要综合考虑信息安全和物理安全。对于工业控制系统的安全问题,安全知识图谱[4]是一种有效的解决方法。安全知识图谱是基于知识图谱技术的人工智能分支,能够描述客观世界中各种实体和概念之间的关系。它将隐形的安全关系转化为可计算的数据,并以可视化的形式展示出来,为决策制定、风险预判、事故分析和攻击识别等提供智能化辅助和自动化处理。

为了建立面向工业控制系统的安全知识图谱,需要同时融合现代工业控制系统中网络信息安全和工业设备的生产特征。通过挖掘流程工业生产控制过程中潜在的特征及其变化规律,并将工业特征与信息安全相融合,形成流程工业控制双安融合知识图谱,是构建现代流程工业控制防护体系的重要

到稿日期:2023-05-31 返修日期:2023-07-08
基金项目:国家重点研发计划(2021YFB2012400)
This work was supported by the National Key R&D Program of China(2021YFB2012400).
通信作者:魏玉良(wei.yl@hit.edu.cn)

前提。例如，一个现代工业互联网网络一般由工业设备安全网络与云端业务系统共同组成，其示意图如图 1 所示。在该网络中，工业生产设备运行过程中产生的各种数据如系统运行日志、操作记录和各种设备状态变更记录等信息被现场仪表记录，并传输到 PLC(Programmable Logic Controller)中。而云端业务系统可以远程操作工业生产设备，通过工业生产设备网络防火墙，与数据交换机加密通信数据流，实现外网与内网的数据传递。交换机是该网络的信息交换枢纽，上位机对工业生产设备的控制信息流、PLC 传输的工业生产设备运行信息，以及外网的云端数据通过交换机进行通信并交换数据。交换机则将系统日志、工业设备数据流量以及网络数据等的镜像数据传输到存储服务器中。

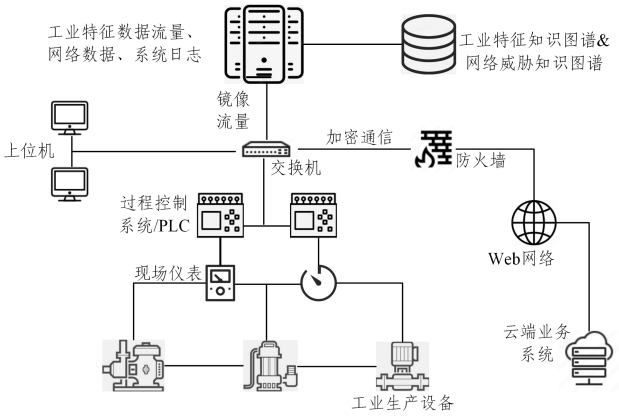

图 1　工业互联网网络拓扑图

Fig. 1　Industrial Internet network topology diagram

本文旨在研究在流程工业控制领域中构建双安融合的工业系统防护体系的关键技术——双安融合工业安全知识图谱。具体来说，本文从数据库中获取既有的网络数据与工业生产设备数据流量，应用知识抽取技术从非结构化的日志流量中提取实体与关系三元组，并构建内嵌工业特征安全与网络威胁知识的双安融合知识图谱。

2　相关工作

在流程工业控制系统领域，有一系列面向特定生产过程的工业知识图谱的构建方法和实例被陆续提出。Chen 等[5]提出了一种基于本体论和知识图谱的多源数据建模方法，用于解决冷轧过程中的带钢断裂问题。该方法通过融合异构数据源，并将其转换为知识图谱形式，利用嵌入方法进行断裂建模，为冷轧生产提供了新的解决方案。

Liang 等[6]采用一种基于自顶向下的方法来构建电力知识图谱，结合了领域专家知识和自然语言处理技术，提高了知识图谱的准确性和完整性；并且还提出了基于知识图谱的故障诊断方法，该方法能够快速准确地诊断发电机组故障。

Wang 等[7]提出利用危害和可操作性分析报告(Hazard and Operability Analysis，HAZOP)来构建工业安全知识图谱，该方案结合数据科学和工业设计，为不同书写形式的 HAZOP 报告提供了统一的建模方案，实现了工业安全知识的整合。将构建方案划分为模式层、数据层和存储层，能够有效地从分析报告中抽取信息以构建知识图谱。

EIBECK 等[8]提出了一个基于本体论的平台，用于处理跨领域场景中的过程工业问题。该平台具备动态更新知识库、实现应用程序间互操作性和集成实时数据的能力，且能够在过程工业领域实现更高效、更可靠的数据管理和决策支持。

也有学者在网络安全知识图谱构建方面展开相关研究。Jia 等[9]提出了一种基于知识图谱的方法，利用机器学习、自然语言处理和图论等技术，自动化地从网络安全文本中提取实体和关系，构建结构化的网络安全知识库。同时，他们通过基于路径排序算法的推理方法，从知识图谱中推导新的安全规则，提升了网络安全的检测和预测能力。

Li 等[10]基于知识图谱构建了标准化的网络安全本体，并从知识获取、知识融合/抽取、知识存储、知识推理、知识更新 5 个阶段介绍了网络安全知识库(Cyber Security Knowledge Base，CSKB)的实现过程，实验证明，CSKB 中存储的知识能够有效实现安全数据的规范化和集成化。

3　知识图谱构建技术概述

知识图谱(Knowledge Graph，KG)是结构化的语义知识库，用于描述概念及其物理世界中符号形式的相互关系[11]。形式上，知识图谱通常可以被定义为 $G=(E,R,T)$。其中 G 是一个带标记和有向的多重图；$E=\{e_1,e_2,\cdots,e_{|E|}\}$，$R=\{r_1,r_2,\cdots,r_{|R|}\}$ 分别是实体和关系的集合，$|E|$ 和 $|R|$ 分别代表知识图谱中实体与关系的数量[12]；$T=\{(e,r,e')\mid e,e'\in E,r\in R\}$ 代表三元组集合，其中每一个三元组表示关系的事实 r 来自主实体 e 到尾部实体 e'。作为 KG 基本要素的实体主要包括集合、类别、对象类型、事物类别(如组织、软件、技术、工具)，而关系将实体连接起来形成一个图形结构[13]。

知识图谱构建有自底向上和自顶向下两种方法。自顶向下的方法需要先构建知识图谱的本体，本体相当于知识图谱的模板，由它定义图谱需要的实体类型，以及实体之间的关系。然后根据本体中确定的实体类型和关系生成知识图谱，这样构建的知识图谱冗余较少[14]，但结构较为单一。

与一般 KG 构建流程类似，安全知识图谱[15](Cybersecurity Knowledge Graph，CKG)的流程和框架遵循一般的 KG 结构。多源分散的领域数据可以在特定框架或预先设计下被聚集成多领域的网络安全本体，然后利用信息提取和实体对齐技术将实体和关系从原始网络安全数据中分离出来。知识推理技术可以基于现有的 KG 生成新的知识以提供支持预测和推理任务。KG 的一般构建流程如图 2 所示。

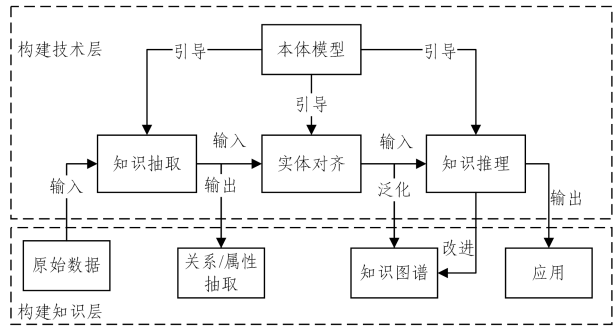

图 2　知识图谱构建一般化流程

Fig. 2　Generalized process for knowledge graph construction

4 流程工业安全知识图谱构建方法研究

在本文构建的面向流程工业控制的安全知识图谱中，需要融合网络安全和工业设备安全两方面的数据，并且这两部分数据往往来源广泛，例如，网络安全数据不仅可以从开源的攻击技术库、漏洞库等结构化数据中获取，一些半结构化/非结构化数据如文档和攻击报告中也蕴含了丰富的网络安全知识。如果采用自顶向下方法，这些知识将变得难以抽取。而采用自底向上方法构建知识图谱，可以融合机器学习、NLP等领域的实体抽取、关系抽取等方法，从不同数据源中抽取实体以及关系，方法较为灵活，构建的知识图谱语义较为丰富。

为了构建同时融合网络安全知识与工业设备知识的安全知识图谱，本文首先需要从多元异构数据中抽取相关实体与关系，并自底向上构建双安融合知识图谱。

4.1 安全知识图谱实体抽取

命名实体是一个词或短语，可以清楚地从一组具有相似特征的其他语句中被分辨出来。在一般领域，命名实体有组织、人物和地点名称等。在特定领域如网络安全领域，命名实体有技术、漏洞、工具和设备等。NER(Named Entity Recognition)[16]是在文本中定位实体和将命名实体分类到预定义的实体类别的过程。给定一系列输入值 $s=\langle w_1, w_2, \cdots, w_N\rangle$，NER 需要输出一个元组列表$\langle I_s, I_e, I_t\rangle$，其中的每一项都代表在 s 中出现的一个实体。$I_s \in [1, N]$和$I_e \in [1, N]$是一个命名实体开始和结束的下标，t 是预先定义的实体种类。图 3 展示了一个 NER 系统从一个语句中识别出 3 个命名实体的例子。

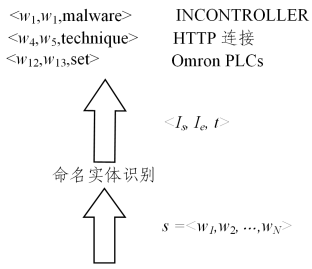

图 3 命名实体识别示例

Fig. 3 Example of named entity recognition

具体来说，首先对文本数据进行句子分割和字符分解，然后手动进行命名实体标注，构建一个包含已标注数据的语料库。最后，使用这个语料库中带有标签的数据对深度学习模型进行训练。

在选择合适的模型时，基于 BERT 的模型对于命名实体识别任务非常有效[17]。BERT 是一个基于 Transformer[18]的预训练语言模型，在实体抽取任务中，其常常与训练模型如 Bi-LSTM[19]，BERT+CRF[20]等模型结合起来运用。

首先进行网络安全图谱实体抽取。该部分主要需要从大量的非结构化文本中提取出带有网络安全语义知识的实体，并进行知识对齐。

为获得与流程工业工控网络相关的网络安全知识，首先需要获取一定量的带有安全知识语义的数据，如已知和新发现的漏洞、弱点、威胁和攻击模式等，以此来构建安全知识图谱。构建安全知识图谱的数据来源一般有漏洞数据库(CVE，CWE，CPE 等)，以及攻击者常见的的攻击模式枚举和分类(CAPEC)，这样的数据一般由研究机构、政府机构以及行业专家，例如计算机应急响应小组(CERT)和 MITRE 来收集和整理。本文通过对比研究，发现代表攻击者的攻击行为数据语义信息较为丰富。如网络威胁情报(Cyber Threat Intelligence，CTI)，同时包含了攻击方法、漏洞信息、弱点和资产等信息，是构建安全知识图谱比较理想的数据来源[21]。图 4 给出了从一段节选报告中提取攻击知识的示例。但从 CTI 报告这种非结构化数据中自动提取知识有两个关键问题：

1)CTI 报告使用的是自然语言中的非正式格式，因此，识别结构化攻击行为需要分析非结构化 CTI 文本中的语义。

2)攻击知识分散在多个报告中。个别报告通常侧重于有限/不完整的攻击案例，因此很难获得全面的攻击知识。

(A) The threat actors sent the trojanized Microsoft Word documents, probably via email. Talos discovered a document named MinutesofMeeting-2May19.docx. Once the victim opens the document, it fetches a remove template from the actor-controlled website, hxxp://droobox[.]online:80/luncher.doc.
(B) Once the luncher.doc was downloaded, it used CVE-2017-11882, to execute code on the victim's machine. After the exploit, the file would write a series of base64-encoded PowerShell commands that acted as a stager and set up persistence by adding it to the HKCU\Software\Microsoft\Windows\CurrentVersion\Run Registry key.
(C) Once the evasion checks were complete, the threat actors used MSbuild to execute an actor-created file named "LOCALAPPDATA\Intel\instal.xml". Based on lexical analysis, we assess with high confidence that this component of the macro script was based on an open-source project called "MSBuild-inline-task."
(D) While this technique was previously documented last year, it has rarely been observed being used in operations. Talos suspects the adversary chose MSBuild because it is a signed Microsoft binary, meaning that it can bypass application whitelisting controls on the host when being used to execute arbitrary code.

图 4 CTI 报告中的网络攻击知识

Fig. 4 Network attack knowledge in CTI report

为解决上述在知识抽取中遇到的问题，本文使用一种结合攻击模板和图对齐算法的知识抽取方法。其中技术模板指使用单个技术从 MITRE 抓取的技术示例中提取生成的攻击图[22]，然后采用图对齐算法来识别从 CTI 报告中提取的具有预初始化模板的多技术图中的攻击技术。对齐多技术攻击图和技术模板后，可以使用模板中的攻击知识来增强对攻击知识的识别能力并利用来自 CTI 报告的丰富知识更新技术模板。最后方法会得到两个输出：

1)在技术层面上汇总和收集 CTI 报告中的攻击知识和模板。

2)在 CTI 报告中总结的完整攻击链和技术知识图谱。该方法允许模板中有一些实体含义不清晰的数据，即只要技术实际多次出现在不同的报告中，大多数报告均能被正确解析。

上述实体抽取方法架构图如图 5 所示。

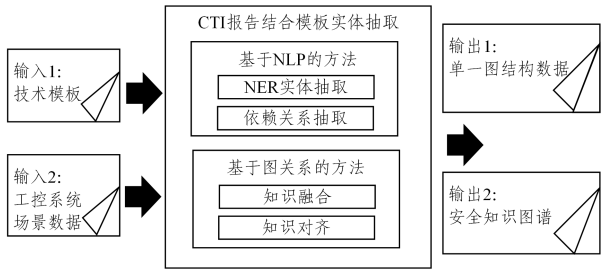

图 5 CTI 报告知识抽取流程

Fig. 5 Knowledge extraction process for CTI reports

其次需要进行工业生产设备安全知识抽取，该部分主要涉及一种低资源 NER 抽取方法。

在工业场景中，会出现数据获取困难、数据质量不高、数据关联性不明确等问题。因此本文使用一种轻量级的、低资源的 NER 模型来进行实体抽取的训练。该方法可以重新调节注意力并适应预先训练的权值，并且可以更好地跨领域传递知识。

现给出一个低资源的 NER 数据集 $L=\{(X_1^L,Y_1^L),\cdots,(X_r^L,Y_r^L)\}$，$r\ll R$。首先该方法基于 BART NER 的框架，输入一个句子，输出为实体的序列，每个实体包括：实体 span 在输入句子中的 start index，end index，以及实体类型（tag 或 index）。在 decoder 端，输出的 hidden state 与输入序列的编码和 tag 序列的编码做点积，并经过 softmax 函数，转换为输入句子词表以及 tag 词表中的概率分布，其中编解码器部分表达式如下：

$$H_{\text{en}}=Encoder(X),H_{\text{en}}\in\Re^{n\times d} \tag{1}$$

$$h_t=Decoder(H_{en},[y_{i=1}^{t-1}]) \tag{2}$$

4.2　安全知识图谱关系抽取

关系提取（Relation Extraction，RE）是提取基于其相关上下文的两个给定实体之间关系的技术，是信息抽取和知识图谱构建的基本任务。RE 数据集可以表示为 $D=\{X,Y\}$，其中 X 是实例列表，对于每个实例 $X=\{w_1,w_2,w_s,\cdots,w_o,\cdots,w_n\}$，RE 的目标是预测实体 w_s 和实体 w_o 之间的关系 $y\in Y$（w_s 和 w_o 简要代表所有实体）。

对于本文所涉及的安全知识图谱关系抽取，可采取与上述实体抽取类似的低资源关系提取方法，即在学习中加入 Prompt-tuning 策略。该方法通过可学习的虚拟类型词和虚拟答案词进行知识注入的模板，来减少模板工程的工作量，并取代预定义规则。

5　流程控制工业双安融合知识图谱

5.1　网络安全知识图谱构建

网络安全知识图谱的构建数据来自 MITREATT&CK 数据库包含的 ATT&CK for ICS。

ATT&CK for ICS 是为了更好地理解、集中和传播关于工控技术领域攻击行为的知识的领域知识库。在 ICS 系统中，由于工业部门和供应商常常通过专有协议、硬件和软件等方式实现独有的部署和控制方式，使得与其相关的网络安全知识具有多样性和复杂性。ATT&CK for ICS 对于在工控领域影响高级系统的行为记录包括以下几个方面：1）基本过程控制系统；2）过程控制；3）操作界面和监控；4）事实和历史数据；5）告警信息；6）安全仪表系统和保护系统；7）工程和维护系统。

该网站收集和存储了大量和攻击技术相关的结构化数据/非结构化数据。其中结构化数据如表 1 所列，不仅包含一些实体类型，也包含关系类型。

构建 NLP 训练任务通常需要选取一个预训练模型，从而降低训练成本，提高训练效率。

本文采用 spaCy 作为框架进行实验环境的搭建。spaCy 是一个工业级自然语言信息处理的框架，由于 spaCy 内置了一些自然语言处理领域常用的预处理工具和训练模型，研究人员可以较快地获得一个基于大量中文/英文预训练的且效果不错的模型。

表 1　数据集中的实体与关系数量

Table 1　Statistics of datasets

名称	实体/关系	数量
intrusion-set	实体	387
Software	实体	774
Tool	实体	97
Technique	实体	1 008
use	关系	831
revoked-by	关系	60
alias	关系	494

基于上述实体抽取方法，对从 MITRE ATT&CK 网站的参考文献中收集的，如来自 Cisco Talos Intelligence Group，Microsoft Security Intelligence Center 等组织的部分 CTI 报告进行了知识抽取。同时从 MITRE ATT&CK 网站提到的与 ICS 相关的技术中抽取了一组技术实例，用来形成方法需要的基于知识的攻击技术模板。实验结果如表 2 和表 3 所列。

表 2　技术模板知识抽取结果

Table 2　Knowledge extraction results for technical templates

技术模板 Top	出现次数	可执行	网络	文件	注册表	漏洞
T0885-Commonly Used Port	1 006	10	450	371	99	8
T0864-Transient Cyber Asset	920	8	132	249	—	—
T0853-Scripting	862	7	140	255	56	6
T0838-Modify Alarm Settings	841	6	287	261	—	—
T0846-Remote System Discovery	780	—	—	220	67	—

表 3　CTI 报告知识抽取结果

Table 3　Knowledge extraction results for CTI report

CTI 报告	实体	依赖关系	技术
UNC4034/ZINC	13	12	15
UNC3890	6	8	9
IRIDIUM/Sandworm	17	8	15
OceanLotus(APT32)Campaign	5	17	5
EV-0530	12	9	12
Tropic Trooper	14	5	14
Woody Rat	17	6	8
Cloud Atlas/Inception	12	7	9
GreyEnergy	17	9	6
Energetic Bear	8	18	5
APT31 Campaign	12	5	6
APT41 Campaign	14	15	8

最终根据获取的结构化数据与抽取的非结构化数据知识，获得的网络攻击行为知识图谱如图 6 所示，可以看到从中抽取了丰富的网络安全知识，例如相关的攻击软件 DanBot、攻击组织 HEXANE 以及各个攻击软件与组织之间的关系。

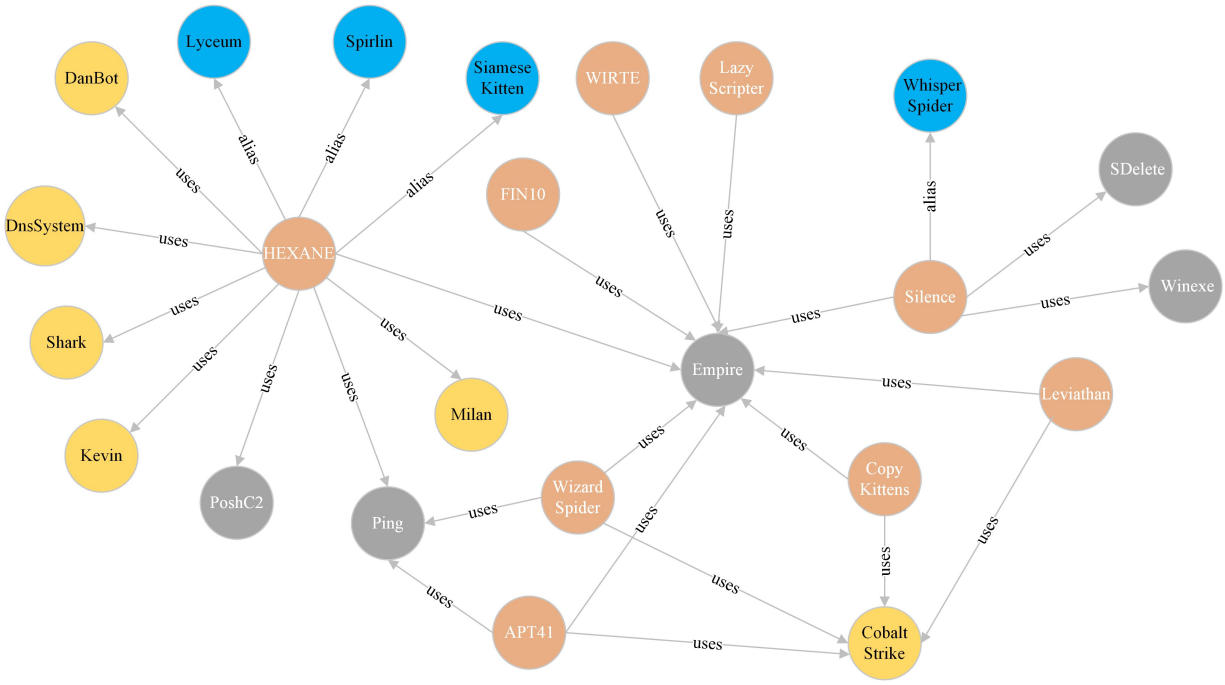

图 6 部分网络安全知识图谱

Fig. 6 Part of network security knowledge graph

5.2 工业生产安全知识图谱构建

本节对项目中取得的某描述煤气化生产过程的文档进行了实体抽取实验，实验中需要先对该文本资料进行数据标注，本节采用了 BIO 标注法来制作实验数据。

构建好数据集后，使用 4.1 节所述方法进行实体抽取实验，采用模型参数如表 4 所列。

表 4 模型训练参数

Table 4 Model training parameters

超参数	参数
BERT 词向量维度	768
BERT 学习率	5×10^{-5}
优化器	Adam
最大句子长度	350
最大训练次数	500
Bi-LSTM 隐层维度	200
实体标签 embedding 维度	20

根据表 4 所设置的模型参数，对该描述流程工业生成过程的文档数据集进行关系抽取实验，实验结果如图 7 所示。实验结果表明，在低资源场景下使用该方法进行实体抽取实验的准确率、召回率和 F1 分数均在 93%以上，说明可以有效地从文本数据中进行生产安全知识的实体抽取。最终抽取的安全实体统计结果如表 5 所列。

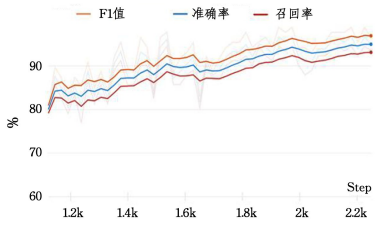

图 7 实体抽取训练结果

Fig. 7 Entity extraction training results

表 5 工业安全知识图谱实体统计

Table 5 Entity statistics of industrial safety knowledge graph

名称	数量
intrusion-set	497
Software	783
Tool	97
Technique	1 109
Group	389
传感器	12
反应装置	16
开关或阀	48
控制器	49
操纵变量	26
测量变量	25

类似地，对数据集中的关系进行 BIO 标注，对同一描述流程工业生成过程的文档数据集进行关系抽取实验，实验结果如图 8 所示，其中两条曲线分别表示训练损失率和验证过程中的 F1 值。

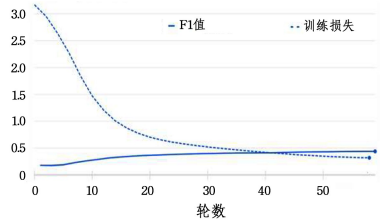

图 8 关系抽取训练结果

Fig. 8 Relation extraction training results

实验结果表明，本文方法的 F1 值达到了 60%左右，并且模型训练损失曲线平稳下降，说明抽取效果较为稳定，使用该方法可以有效地挖掘非结构化数据中的实体关系信息。最终从文档中抽取的安全知识关系统计结果如表 6 所列。

表 6　工业安全知识图谱关系统计

Table 6　Relation statistics of industrial safety knowledge graph

名称	数量
alias	509
mitigates	1 701
related-to	237
revoked-by	141
uses	12 286
包含	27
控制	56
测量	15
直接影响	44

通过实体抽取和关系提取，我们从化工领域文本数据中学习到了相关知识，并将其加入到知识图谱中，由此得到了内嵌工业特征和网络安全知识的双安融合工业安全知识图谱。

化工生产安全相关的部分知识如图 9 所示，可以发现所提方法成功抽取了从项目获得的典型生产过程——煤气化生产过程数据，获得了相关生产安全知识，如气化洗涤工艺中相关生产设备的实体和关系。

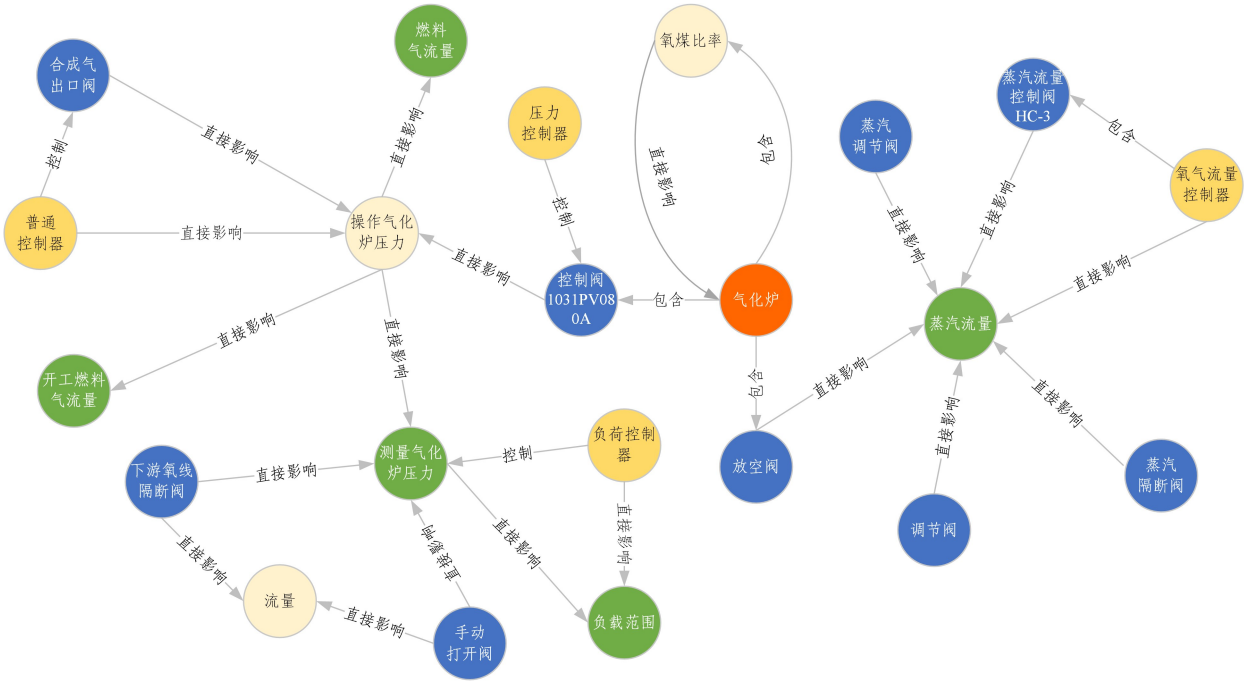

图 9　部分化工设备生产安全知识图谱

Fig. 9　Part of knowledge graph of production safety for chemical equipments

6　双安融合知识图谱推理挖掘分析

传统的工业安全防护体系存在多个缺陷，包括缺乏动态性、智能化、信息共享与全局决策等；并且，安全信息分散在不同系统和部门中，造成信息孤岛，阻碍了全面的安全分析。此外，传统防护体系局限于特定控制系统，缺乏对整个工业系统的全局视野。

针对上述问题，本文构建了双安融合知识图谱，它具有全面性和智能分析的优势，综合考虑了信息安全和物理安全两个层面的特征和关系，涵盖了工业控制系统的信息层面和物理层面。将这些信息融合在一个知识图谱中，能够提供更全面、综合的安全分析和决策支持。结合人工智能和知识图谱技术，可以进行智能分析和推理，从知识图谱中发现潜在的安全威胁、漏洞和攻击模式，并且可以帮助发现和应对安全威胁，更好地应对现代工业系统面临的复杂网络安全挑战。

以本文构建的双安融合知识图谱为例，查询在工控系统中与网络安全和功能安全相关的常用端口，可以通过 cypher 语句查询“match(n) where n.id contains ‘port’ return n”或“MATCH(n:Technique { name:‘Commonly Used Port’ }) RETURN n”来查询 ID 包含“port”的节点，以及含有属性“port”的节点，获得一系列节点相关信息。如与工控协议相关的端口 TCP:502(MODBUS)，TCP:102(S7comm/ISOTSAP)，TCP:20000(DNP3)，TCP:44818(Ethernet/IP)，以及设备层相关的端口，如节点“name:triton，description:Triton uses TriStations default UDP port,1502,to communicate with devices”。

基于类似的网络安全与工业安全知识之间的耦合关系，可以通过双安融合知识图谱构建更全面、智能的流程工业防护体系。本文构建的双安融合知识图谱存储了丰富的安全知识并具备了一定推理能力，一方面，可以利用 cypher 语句进行图数据库的快速检索，获取查询到的信息，帮助完善构建边界防护策略；另一方面，利用该知识图谱的推理能力，可以通过链接预测等方法辅助管理人员进行边界防护策略优化，挖掘潜在的网络安全和功能安全漏洞。

结束语　本文主要研究了使用自底向上方法构建的双安融合工业安全知识图谱。在网络安全知识方面，综合运用了来自 ATT&CK for ICS 等机构的结构化数据和半结构化/非结构化数据，抽取了与工业网络安全相关的实体和关系。在生产安全知识方面，结合自然语言处理技术对描述流程工业生产过程的文档进行低资源下的实体和关系抽取。最终将抽取的知识表示成图节点和边，存储在图数据库 Neo4j 中。该知识图谱将为后续的行为识别和异常判定等防护技术提供支撑。

未来，我们将改进知识抽取技术以获取更高质量的结构化数据，并进一步挖掘信息安全与物理安全之间的关系，构建更加健壮的用于流程工业防护体系的双安融合知识图谱。

参 考 文 献

[1] CONTI M, DONADEL D, TURRIN F. A survey on industrial control system testbeds and datasets for security research[J]. IEEE Communications Surveys & Tutorials, 2021, 23(4): 2248-2294.

[2] DING D, HAN Q L, XIANG Y, et al. A survey on security control and attack detection for industrial cyber-physical systems [J]. Neurocomputing, 2018, 275: 1674-1683.

[3] WOLF M, SERPANOS D. Safety and security in cyber-physical systems and internet-of-things systems[C]// Proceedings of the IEEE. 2017: 9-20.

[4] MAO S, ZHAO Y M, CHEN J H, et al. Development of process safety knowledge graph: a case study on delayed coking process [J]. Computers & Chemical Engineering, 2020, 143: 107094.

[5] CHEN Z Y, LIU Y, VALERA-MEDINA A, et al. Multi-sourced modelling for strip breakage using knowledge graph embeddings [J]. Procedia CIRP, 2021, 104: 1884-1889.

[6] LIANG H, PENG X J, ZHAO N, et al. An approach of top-down electric generation knowledge graph construction[J]. IOP Conference Series: Earth and Environmental Science, 2021, 661(1): 012021.

[7] WANG Z, ZHANG B, GAO D. A novel knowledge graph development for industry design: A case study on indirect coal liquefaction process[J]. Computers in Industry, 2022, 139: 103647.

[8] EIBECK A, LIM M Q, KRAFT M. J-Park Simulator: anontology-based platform for cross-domain scenarios in process industry [J]. Computers & Chemical Engineering, 2019, 131: 106586.

[9] JIA Y, QI Y, SHANG H, et al. A practical approach to constructing a knowledge graph for cybersecurity[J]. Engineering, 2018, 4(1): 53-60.

[10] LI K, ZHOU H, TU Z, et al. CSKB: A Cyber Security Knowledge Base Based on Knowledge Graph[C]// International Conference on Security and Privacy in Digital Economy. Singapore: Springer, 2020: 100-113.

[11] RYEN V, SOYLU A, ROMAN D. Building semantic knowledge graphs from (semi-) structured data: a review[J]. Future Internet, 2022, 14(5): 129.

[12] HOGAN A, BLOMQVIST E, COCHEZ M, et al. Knowledge graphs[J]. ACM Computing Surveys (CSUR), 2021, 54(4): 1-37.

[13] JI S, PAN S, CAMBRIA E, MARTTINEN P, et al. A survey on knowledge graphs: Representation, acquisition, and applications [J]. IEEE Transactions on Neural Networks and Learning Systems, 2021, 33(2): 494-514.

[14] EILICKE C, CHEKOL M W, RUFFINELLI D, et al. Anytime bot tom-up rule learning for knowledge graph completion[C]// Proceedings of the 28th International Joint Conference on Artificial Intelligence. 2019: 3137-3143.

[15] PIPLAI A, MITTAL S, JOSHI A, et al. Creating cybersecurity knowledge graphs from malware after action reports[J]. IEEE Access, 2020, 8: 211691-211703.

[16] AL-MOSLMI T, OCAÑA M G, OPDAHL A L, et al. Named entity extraction for knowledge graphs: A literature overview [J]. IEEE Access, 2020, 8: 32862-32881.

[17] DEVLIN J, CHANG M W, LEE K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[J]. arXiv: 1810. 04805, 2018.

[18] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]// Proceedings of the 31st International Conference on Neural Information Processing Systems. 2017: 6000-6010.

[19] HUANG Z, XU W, YU K. Bidirectional LSTM-CRF models for sequence tagging[J]. arXiv: 1508. 01991, 2015.

[20] SOUZA F, NOGUEIRA R, LOTUFO R. Portuguese named entity recognition using BERT-CRF[J]. arXiv: 1909. 10649, 2019.

[21] MILAJERDI S M, ESHETE B, GJOMEMO R, et al. Poirot: Aligning attack behavior with kernel audit records for cyber threat hunting[C]// Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security. 2019: 1795-1812.

[22] NADEEM A, VERWER S, MOSKAL S, et al. Alert-driven attack graph generation using s-pdfa[J]. IEEE Transactions on Dependable and Secure Computing, 2021, 19(2): 731-746.

WANG Jing, born in 2001, postgraduate. His main research interests include natural language processing and data mining.

WEI Yuliang, born in 1989, assistant researcher. His main research interests include natural language processing, industrial Internet security and data mining.

（责任编辑：何杨）