基于路径推理的网络空间资产归属分析方法

施凡，窦开洋，李晔，屠铖，许成喜，薛鹏飞，李阳
（国防科技大学电子对抗学院，安徽 合肥 230031）

摘要：针对网络空间资产归属分析中面临的数据碎片化、归属关系模糊及现有方法可解释性不足等挑战，提出了一种基于知识图谱路径推理的网络资产归属分析方法。首先，通过融合WHOIS、安全套接层（SSL）证书及网页文本特征等多源异构数据，构建了面向网络空间测绘领域的资产知识图谱，从结构上打破数据孤岛；其次，设计了规则匹配与大语言模型协同的属性补全机制，初步判定缺乏显性特征资产的归属；最后，综合利用节点PageRank值、关系自信息量及路径跳数奖励因子，提出了一种融合剪枝策略的深度优先搜索算法与基于随机游走的路径评分模型，将归属预测转化为知识图谱中的路径推理过程。实验结果表明，该方法在中国教育网（CERNET）的资产归属分析任务上的精确率达到98%，显著优于基于WHOIS查询及网络测绘平台查询的方法，有效解决了网络空间资产归属数据碎片化以及可解释性不足的问题，提升了网络空间资产归属判定的准确性和鲁棒性。

关键词：网络资产归属；知识图谱；路径推理；可解释性
中图分类号：TP309
文献标志码：A
doi：10.11959/j.issn.1000–436x.TXXB260230

Cyberspace assets ownership analysis method based on path reasoning

Shi Fan, Dou Kaiyang, Li Ye, Tu Cheng, Xu Chengxi, Xue Pengfei, Li Yang
College of Electronic Engineering, National University of Defense Technology, Hefei 230031, China

Abstract: To address the challenges of data fragmentation, ambiguous attribution relationships, and lack of interpretability in cyberspace asset attribution analysis, a cyberspace asset attribution analysis method based on knowledge graph path reasoning was proposed. Firstly, a domain-specific cyberspace asset knowledge graph was constructed by fusing multi-source heterogeneous data, including WHOIS, secure sockets layer (SSL) certificates, and web textual features, thereby structurally breaking down data silos. Secondly, an attribute completion mechanism coordinating rule matching with large language models was designed to achieve preliminary attribution for assets lacking explicit features. Furthermore, a depth-first search algorithm integrating pruning strategies and a path scoring model based on random walks were proposed. By comprehensively utilizing node PageRank values, relational self-information, and path hop reward factors, the attribution prediction was transformed into a quantifiable and visualizable evidence chain search process. Experimental results based on real-world data from the China Education and Research Network (CERNET) show that the proposed method achieves an attribution accuracy of 98%, significantly outperforming traditional WHOIS queries and mainstream cyberspace mapping platforms. This method effectively resolves the attribution of implicit assets in multi-hop association scenarios and enhances the robustness and interpretability of attribution analysis.

Key words: cyberspace asset ownership, knowledge graph, path reasoning, interpretability

---

收稿日期：2026–04–27；**修回日期：**2026–06–16
通信作者：李阳，liyanghf@nudt.edu.cn
基金项目：国家自然科学基金资助项目（No.62372456, No.62276001, No.62136008）
Foundation Items：The National Natural Science Foundation of China (No.62372456, No.62276001, No.62136008)

0　引言

网络空间已成为继陆、海、空、天之后的“第五空间”，其安全性是国家安全的重要组成部分。随着全球信息化进程加速，网络空间资产数量呈爆发式增长，各类组织与其所属资产之间的关系日益错综复杂。这种复杂性导致海量资产在面临安全威胁时难以准确归属至具体责任主体，严重制约了网络安全事件的应急响应效率。因此，研究高效、准确的网络空间资产归属分析方法，对于厘清虚拟资产与现实社会组织之间的归属关系、明确各类组织的资产边界与责任范围，进而提升国家整体网络安全防御水平具有重要的战略意义和现实价值。

当前网络资产归属分析主要依赖规则匹配与多模态命名实体识别等方法。基于规则匹配的方法以已确定归属关系的资产[1]为基准，通过手工定义的匹配规则（例如IP同C段、同域名）来判断其他资产的归属关系；而基于多模态命名实体识别的方法[2]通过提取图像的多尺度特征，并采用两步门控协同注意力机制，先融合图像与域名得到跨模态特征，再与文本进行交互，进行所有者实体识别。尽管上述方法在一定程度上解决了资产归属的问题，但在复杂网络环境下仍面临严峻挑战：一方面，基于规则匹配的方法虽然可靠性高、可解释性强，但高度依赖数据源（如WHOIS数据库）的质量，受限于数据稀疏问题，资产覆盖率较低，且该方法需手工定义规则，扩展性较差；另一方面，基于命名实体识别的方法通过网页内容和图形数据进行实体识别，泛化能力较强，但其归属结果缺乏可解释性，且往往忽略了网络空间资产的拓扑结构与连接信息。

事实上，上述两种方法具有显著的互补性：规则匹配侧重于挖掘资产间的多维拓扑关系，而多模态方法则聚焦于解析资产的内容语义信息。为实现两者的有机融合，本研究提出一种基于知识图谱路径推理的网络资产归属分析方法。首先，融合多源数据构建网络资产知识图谱，将离散的资产特征转化为实体间的语义连接，从而缓解数据碎片化的问题；其次，设计规则与大模型协同的归属分析算法，综合利用字段匹配、正则表达式及大语言模型，从证书、HTML源码及邮箱等信息中提取名称信息，初步判断网络资产的归属；最后，基于剪枝策略的深度优先搜索算法，在知识图谱中挖掘目标资产至潜在组织的连通路径，将归属分析转化为“证据链”溯源过程，增强资产归属过程的可解释性；在此基础上，结合节点权重、关系权重及路径跳数构建评分模型，对候选路径进行量化评分并选择路径评分最高的组织作为归属结果，实现对目标资产归属的准确预测。

本文以中国教育自治域下的网络资产数据为基础，通过多源数据融合与清洗，构建了一个包含15万个节点与20万条关系的网络资产知识图谱数据集。实验结果表明，本文基于知识图谱路径推理的方法在归属分析任务上表现优异，其归属分析准确率达到98%，不仅显著优于基于WHOIS数据库查询及现有网络空间引擎查询的归属判定方法，而且实现了预测过程的白盒化，有效提升了资产归属分析结果的可解释性。

本文的主要贡献总结如下。

（1）构建了面向网络空间测绘领域的知识图谱数据集。针对网络资产数据多源异构、碎片化严重的特点，整合WHOIS、域名系统（DNS）解析记录、安全套接层（SSL）证书等多维数据，构建了一个包含15万个节点与20万条关系的中国教育领域网络资产知识图谱，填补了该领域结构化基准数据集的空白，为挖掘资产间的深层关联关系奠定了数据基础。

（2）提出了一种基于知识图谱路径推理的归属分析方法，有效增强了网络资产归属分析结果的可解释性。针对现有采用深度学习进行归属分析缺乏解释性的局限性，设计了采用剪枝策略的深度优先搜索算法，将归属预测转化为寻找证据链的过程；并结合随机游走算法对推理路径进行置信度量化与排序，有效解决了复杂网络环境下的资产归属模糊问题。

（3）基于真实网络资产数据进行实验，验证了本文方法的有效性与先进性。基于中国教育自治域真实数据的对比实验结果表明，本文方法在保证高准确率的同时，能够提供更直观的归属分析路径，在应对数据缺失和多跳关联场景时表现出比现有方法更强的鲁棒性。

1　相关工作

1.1　网络安全知识图谱本体与构建

知识图谱是一种有组织的大规模语义数据库，

起源于2012年谷歌公司的倡议并在之后得到了迅速的发展[3]。近年来，知识图谱在复杂异构数据关联与语义推理方面具有独特优势，使其在网络安全、金融风控、生物医药等多个领域得到了广泛应用，网络空间测绘同样逐渐依赖知识图谱对网络实体及其关系进行高效组织和语义关联。通过将分散的网络资源、资产信息和多元外部数据以结构化和语义化的方式进行融合，显著提升了网络空间测绘的数据整合与挖掘能力。在知识图谱中，本体定义了实体的类型、实体的属性以及实体之间关系的类型，是对某一领域知识的形式化表达。网络资产知识图谱构建的数据基础是资产探测得到的IP、端口、服务、域名、源码等数据以及相关的漏洞数据。这些数据多源异构、概念纷繁、关联复杂，包含网络空间、地理空间和社会空间等多方面信息，需要进行抽象并定义规范化的本体模型，统一知识的表达形式。

1.2 网络资产归属分析方法

传统的网络资产归属分析高度依赖于显式的注册信息，如IP Whois和域名Whois记录。然而，随着GDPR等隐私法规的实施，Whois数据的隐私化导致直接归属方法的有效性大幅下降[1]。研究表明，传统的Whois反查方法在隐私保护下的F1值仅为0.54，而结合被动DNS、TLS证书指纹等多源被动数据的关联方法可将F1值提升至0.940[1]。此外，利用历史Whois数据进行时间回溯分析也成为突破隐私壁垒的重要手段[4]。当前的趋势是从单一的静态规则匹配转向基于多源异构数据（如证书、Banner信息、DNS解析行为）的动态关联分析。

大语言模型在网络安全命名实体识别任务中展现了显著优势，尤其是在识别非标准化的组织机构名称方面。传统的BiLSTM-CRF模型在面对复杂的安全文本时泛化能力有限，而基于BERT或微调后的大语言模型（如CyberBERT）能够更准确地从Banner信息、SSL证书字段及网络威胁情报（cyber threat intelligence，CTI）中提取组织机构实体[5]。大语言模型不仅用于抽取，还被应用于知识图谱的增强构建[6]。通过设计提示词，大语言模型可以从非结构化文本中直接推理出实体间的隐含关系（如organization→owner→domain），并用于补全图谱中的属性缺失。

1.3 知识图谱推理算法

网络资产归属分析在图谱层面本质上是一个链接预测问题，即预测网络资产与组织机构之间缺失的关系。根据推理逻辑的不同，可将其分为基于表示学习、基于路径和基于传播三大类。基于表示学习的推理方法，是指将知识图谱嵌入低维向量之中从而使其可计算，主要分为两类方法：①翻译距离模型，通过定义基于距离的评分函数来量化计算不同三元组间的相似程度；②语义匹配模型，主要包括基于张量分解和基于神经网络两种模型。其中基于翻译的模型主要有TransE（translating embedding）[7]、RotatE[8]等，利用翻译的思想将三元组中的实体和关系转换成欧氏空间中的多维向量，从而实现对知识图谱的降维。基于翻译的方法虽然大大提高了结果的准确性，但是却无法对多跳关系作出解释，因此在应用于较为复杂的大型知识图谱推理时缺乏可解释性。基于路径的推理方法，最先由Yu等[9]提出，将异构信息网络集成到矩阵分解中，以进行个性化实体推荐。该方法采用时序神经网络架构处理用户行为序列数据，充分发挥了该类模型在分析时序信息方面的优势。特别地，通过引入门控机制，实现了对知识图谱结构化特征的有机融合。其他工作如KSR（knowledge-enhanced sequential re-commender）算法[10]、KPRN（knowledge-aware path recurrent Network）算法[11]等在其特定领域或数据集上都有着非常不错的表现。基于传播的推理方法，采用嵌入传播的思想，通过聚合知识图谱中多跳邻居节点的信息，来丰富实体的表示，并用于用户偏好预测。例如，Wang等[12]提出的RippleNet模型引入了偏好传播的概念，能够自动挖掘用户在不同层级的潜在兴趣；Wang等[13]设计的KGCN模型结合了知识图谱与图卷积神经网络，有效捕捉了邻域信息，并通过引入邻居节点权重提升了推荐效果。这些方法通常是在推荐出相关项目后，通过软匹配的方式生成解释路径，因此本质上属于事后解释。

2 基于知识图谱推理的网络资产归属分析方法

2.1 多源异构的网络资产知识图谱构建

知识图谱本体构建采用融入决策过程的斯坦福

七步法。基于网络空间测绘平台的原始数据，结合自顶向下与自底向上两种设计方式构建本体。首先根据专家知识明确领域范围，设计基本实体与关系，并评估现有本体的复用性；随后，对来自不同渠道的网络空间资产探测数据进行分析，从中抽取领域术语，分别构建各子模块的本体模型，明确定义其类层次结构、对象属性、数据属性及相关约束规则；最后融合为统一的层次化本体模型。整个过程在专家指导下通过决策进行迭代。“七步法”本体构建原理如图1所示。

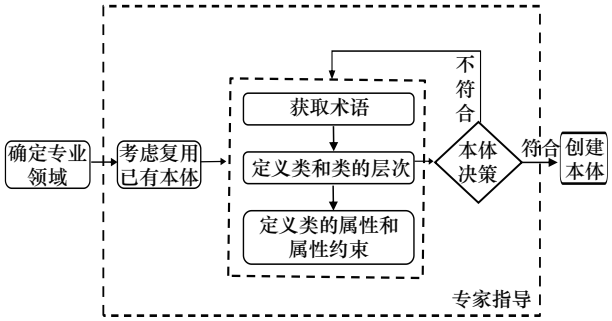


图1　“七步法”本体构建原理

按照上述流程，本文首先确定本体构建的领域范围为网络空间资产，即部署在网络中具有唯一标识性的软硬件、数据及服务等资源，包括但不限于主机、服务器、网站、子域名、业务系统、软件服务、中间件、API接口、虚拟机、云资源、应用账号、邮件服务器、SSL证书及CDN节点等。每项资产均可能属于特定组织或个人，并与其他资产存在直接或间接关联。在构建本体模型时，经评估发现，当前网络安全领域主要存在面向特定数据结构（如漏洞信息）和特定应用场景（如入侵检测）的两类本体，但在网络资产归属分析领域尚无成熟的可复用模型。因此，本文基于网络空间测绘数据，分析资产相关概念并结合归属关联分析任务，提取归属推理过程中的关键实体，推理所需要的网络空间资产本体实体描述如表1所示。

此外，网络资产实体的语义关系也十分丰富。在定义实体的基础上，本文进一步梳理了实体间的语义关联，网络空间资产本体关系描述如表2所示。

由此，可以得到网络资产本体，其结构示意如图2所示。

表1　推理所需要的网络空间资产本体实体描述

实体名称	说明	标识
网络资产	用于标识网络空间中各类软硬件实体、服务与数据资源的可寻址实体	IP-PORT
域名	用于在互联网上标识和定位计算机或资源的人类可读的文字标识符	Domain
IP	用于标识和定位互联网设备（如计算机、服务器、路由器、智能手机等）的唯一数字地址	IP
邮箱	互联网电子邮箱地址	Email
组织机构	公司/企业/单位	ORG
地理位置	地点	Location
城市	城市	City
经纬度	经纬度	Rpt
C段	每个IP都有其对应的C段	IP_C
证书	用于确保网站数据传输安全性的SSL证书或TLS证书	Subject

表2　网络空间资产本体关系描述

关系名称	说明	标识
位于	组织机构或者IP所在的城市	LOCATED_AT
拥有	表示实体间的归属、包含与管理关系	HAS
归属于	网络资产所归属的组织是	BELONGS_TO
经纬度是	某个IP的经纬度坐标	RPT_IS
域名是	IP对应的域名	DOMAIN_IS
C段是	IP对应的C段是	IPC_IS
邮箱是	域名注册的邮箱是	EMAIL_IS
SSL证书是	组织机构或者域名拥有的证书是	SSL_IS
是某IP的C段	某C段对应的IP是	IS_IPC_OF
是某IP的域名	某域名对应的IP是	IS_DOMAIN_OF

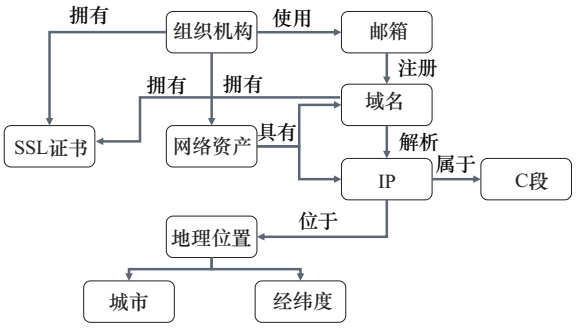


图2　网络空间资产本体结构示意

在完成上述本体构建后，本文开展知识图谱的构建工作。因为测绘数据中只包含部分关系，后续的工作就是将未能直接实例化的关系进行推理补全，因此通过 Cypher 语句，依据预设本体依次完成节点、关系及属性的实例化。具体节点类型包括域名（Domain）、IP C 段（IP_C）、邮箱（Email）、城市（City）及经纬度（Rpt）；关系类型涵盖位于（LOCATED_AT）、属于（BELONGS_TO）、经纬度是（RPT_IS）、域名是（DOMAIN_IS）、C 段是（IPC_IS）及邮箱是（EMAIL_IS）。在明确节点类型与关系语义后，将数据以（头节点，关系，尾节点）三元组形式进行存储。知识图谱节点连接方式如表 3 所示。

表 3　　知识图谱节点连接方式

头节点类型	关系	尾节点类型
IP	IPC_IS	IP_C
IP	DOMAIN_IS	Domain
IP	LOCATED_AT	City
IP	EMAIL_IS	Email
IP	RPT_IS	Rpt
IP_C	IS_IPC_OF	IP
Domain	IS_DOMAIN_OF	IP

至此，利用 Neo4j 数据库成功构建了涵盖 15 万个节点与 20 万条关系的中国教育网有向网络资产知识图谱。该图谱将原始、杂乱的网络资产数据进行了结构化处理与清晰呈现。保存在 Neo4j 中的知识图谱局部示例如图 3 所示。

2.2　规则与大模型协同的网络资产归属分析

从测绘平台获取的原始数据往往未经处理，导致信息利用率较低，且单项信息的可用数量较少，准确度难以保证。因此，本文综合利用证书信息、HTML 页面源码及邮箱信息，通过规则与大模型协同的方式确定网络资产归属，实现知识图谱的初步补全。

2.2.1　基于字段匹配的归属组织提取

网页证书中的组织信息源于证书申请者在申请 SSL/TLS 证书时，向证书颁发机构（CA）提交的组织资料。颁发机构在签发证书前会对申请者的组织身份进行严格审查，因此，在多种组织归属信息来源中，证书信息的置信度最高。基于此，本研究将利用证书确定组织归属的优先级设为首位。图 4 展示了中国 XX 大学官网的证书信息，其中可以明确识别其组织归属详情。



图 4　中国 XX 大学 SSL 证书

网络空间测绘平台获取的证书数据通常呈现海量且半结构化的特征。平台通过全球扫描，将捕获的原始证书以 JSON 格式存储。在提取组织机构信

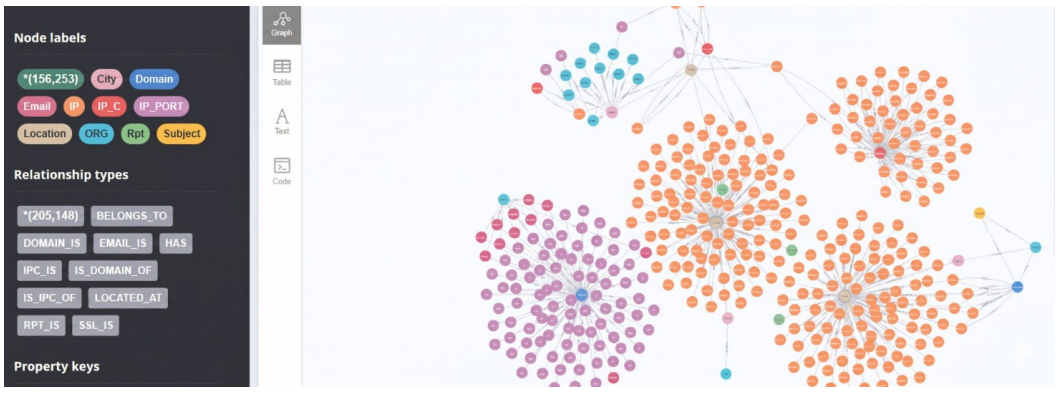


图 3　网络资产知识图谱

息时，首先需从复杂的嵌套字段中解析出证书内容，随后精确定位并提取组织机构这一关键字段。最后，对提取的组织名称进行去空格、去引号等规范化处理，以获得统一的格式化结果。

2.2.2　基于正则表达式的归属组织提取

超文本标记语言（hyper text markup language，HTML）是构建网页的基础语言，用于定义网页的结构与主要内容。通过解析网页的 HTML 源码，可以提取其包含的标题、段落、链接及多媒体资源等原始数据。调研发现，在高校等机构的门户网站中，网页标题常包含显著的组织名称。通过检索其 HTML 数据中的<title>标签内容，能够准确判定该网页的组织归属。中国 XX 大学官网的 HTML 信息如图 5 所示，HTML 源码中<title>标签内明确标注该网站的归属组织。

为了从非结构化的 HTML 源码中精准提取出组织实体信息，本文针对网页结构的不同层级，分别为网页标题、Meta 属性以及 ICP 备案主体，设计了 3 类具有高鲁棒性的正则表达式规则，如表 4 所示。网页标题是 HTML 文档中语义最为明确的标识，本文采用基于零宽断言的匹配模式，提取标签内部的纯文本内容，避免了传统匹配模式需要二次剔除标签的计算冗余。再使用非贪婪匹配，确保在复杂的单行源码中，一旦遇到第一个闭合标签即停止匹配，有效规避因 HTML 格式不规范导致的跨标签过度匹配风险。针对现代网页利用元数据（metadata）声明版权或实体的习惯，本文设计了关联匹配规则，能够兼容不同开发框架生成的代码风格，从元数据中过滤描述、关键词等无关信息，确保仅当等效身份相关字段关联时才触发匹配。在中国的机构网站中，ICP 备案主体信息具有极高的法律权威性。本文设计的正则表达式以“备案号”或“ICP 备”作为扫描锚点，快速定位至网页页脚（footer）区域的法定公示块，然后分离备案号数字串与组织全称，在确保数据提取完整性的同时，排除了导航栏链接或声明文字的干扰。

表 4　　3 类提取目标的正则表达式模式

提取目标	正则表达式模式
网页标题	(?<=<title>).+?(?=<\/title>)
Meta 属性	content\s*=\s*["']([^"']+)["'](?=.*?(?:author|copyright|og:site_name))
ICP 备案主体	(?<=备案号|ICP 备).+?[:：]?\s*([\u4e00-\u9fa5]{4,40}(?:有限公司|集团|中心|学校|大学|学院))

2.2.3　基于大语言模型的归属组织判定

通过分析数据中的证书 CN 字段与邮箱信息，可以提取出关键的域名信息。本文通过提取此类信息，构建了一种基于大语言模型的自动化归属组织判定架构，旨在通过多阶段的处理逻辑实现对域名

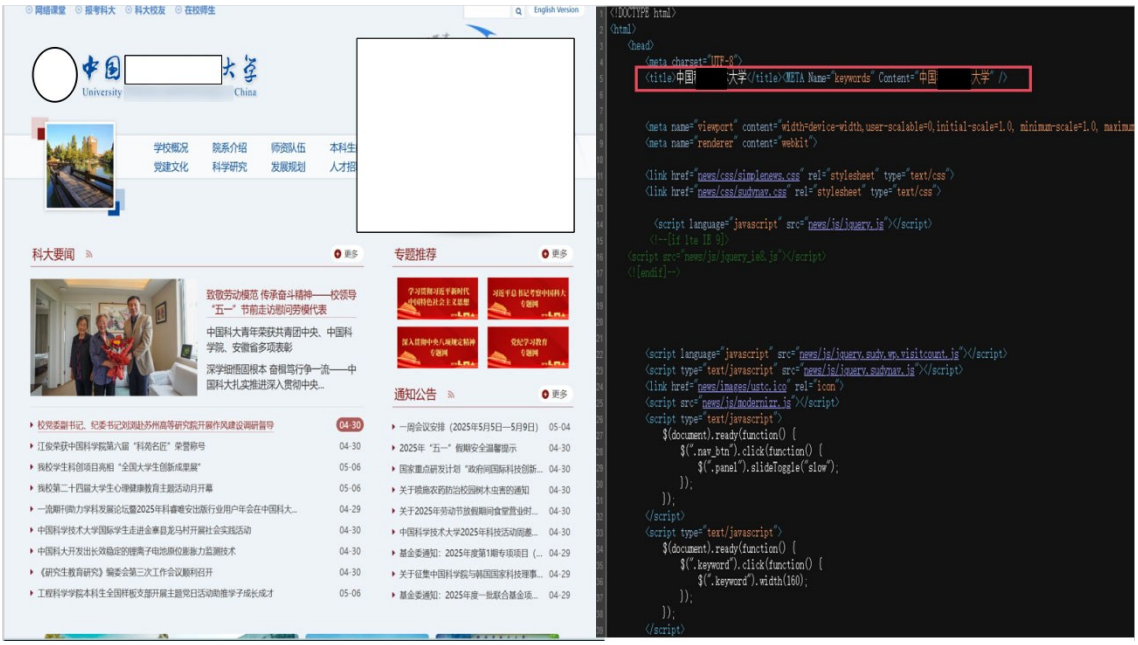


图 5　中国 XX 大学官网的 HTML 信息

资产所属实体的精准识别。具体实现时，采用通义千问模型（Qwen-Turbo）判定资产的组织归属，其中 API 的每秒请求数（query per second，QPS）为 4。为引导大语言模型高质量完成归属判定任务，本文针对性地设计了提示词策略，具体步骤如下。

（1）明确任务目标。核心任务是根据输入的邮箱及证书 CN 域名信息，推断网络资产的归属组织。输入与输出的核心要素分别为域名信息与组织名称。

（2）场景设计与角色设定。为使大语言模型以专业视角分析问题，将其身份设定为“经验丰富的网络安全分析师，擅长通过域名信息判定资产归属”，从而将语境限定在网络资产关联分析领域。

（3）明确任务约束。规定大语言模型的具体分析逻辑与预期输出。提示词设计为：“请基于以下域名相关信息，结合证书 CN 和注册邮箱，判断该域名或其指向资产的归属组织。”

（4）规范输入输出格式。根据数据特征，设定输入格式为{IP：xxx.xxx.xx.x；CN：xxx.xx.edu.cn；Email：xxx@xxx.xx.edu.cn}，并指定输出格式为{IP：xxx.xxx.xx.x|组织名称}。

基于大语言模型的归属判定流程如图 6 所示，首先始于对目标域名相关原始信息的广泛收集，这为后续的判定提供了必要的数据基础。在获取原始数据后，系统进入信息预处理与要素提取阶段，通过对证书 CN 信息或邮箱信息进行清洗，抽取出能够表征组织身份的关键特征维度。为了有效激活大语言模型的语义理解能力，提取出的关键要素被进一步封装并构建为标准化的提示词，以确保输入模型的信息具有高度的结构化与任务导向性。随后，标准化提示词被送入大语言模型内核进行深度推理，由模型基于其丰富的知识库和上下文理解能力生成归属判定结果。判定流程的核心在于后续的可靠性校验机制。系统对大语言模型输出的判定结果进行解析，并根据预设的评估标准进入决策分支。若判定结果表现出较高的置信度且逻辑自洽，则被视为可靠输出，直接进入最终的判定结果存储与入库环节；反之，若检测到判定结果存在逻辑冲突、信息置信度不足或多源数据不一致等不可靠特征，流程则会自动流转至异常处理节点。在此环节，系统引入人工审查机制或触发补充信息采集逻辑，通过引入更高维度的外部证据对争议资产进行二次研判，并将修正后的信息重新反馈至模型输入端进行重判。通过“自动化判定–闭环校验–人工协同”的迭代过程来保证域名归属分析在复杂网络环境下的准确性与稳健性。

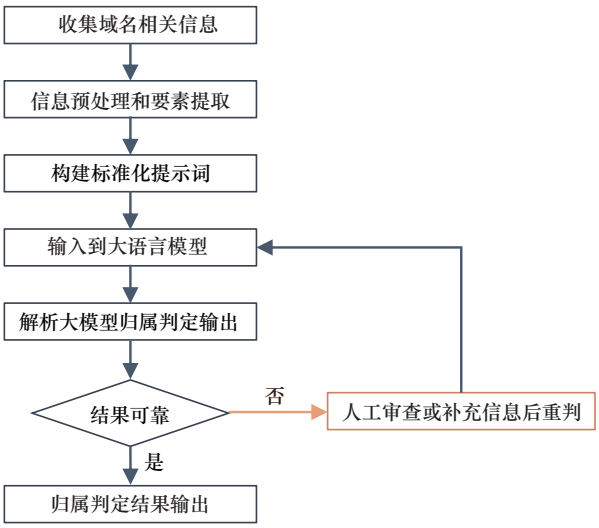


图 6　基于大语言模型的归属判定流程

基于字段匹配、正则表达式以及大语言模型方法，从证书、HTML 源码及邮箱信息完成归属组织提取后，本文依据置信度优先级确立最终的组织归属结果。具体逻辑为：若证书中的组织信息存在，则优先将其认定为该资产的组织归属；若证书组织信息缺失，则采用从 HTML 源码提取的组织信息；若 HTML 数据亦不可用，则利用证书 CN 字段或邮箱信息，结合通义千问大模型进行归属判定。对于无法通过上述多源信息直接确定的资产，将通过知识图谱推理完成预测。基于原始数据统计与置信度分析，本文将按数据来源可靠性划分为 3 类，组织归属信息来源可靠性排序如表 5 所示。

表 5　组织归属信息来源可靠性排序

优先级	数据来源
1	证书组织信息
2	HTML 信息
3	证书 CN 信息与邮箱信息

2.3　基于路径评分的网络资产归属分析

组织机构所辖的网络资产，无论显性或隐性，普遍存在复杂的内在关联。因此，挖掘待判定资产与已知归属资产的拓扑近邻关系，可为归属推理提供关键依据。例如，若未知资产与已知资产位于同

一 IP C 段，二者归属于同一组织的概率显著提高；地理位置、域名结构等特征亦是重要的辅助判别指标。基于上述逻辑，在构建知识图谱并完成基于多源信息的初步补全后，针对仍缺乏显性特征的资产，本文进一步提出基于图谱路径的推理方法，旨在通过挖掘深层拓扑关联与语义路径，实现对目标资产归属的精准推断。

前述网络资产知识图谱形式化定义为 $G=(E,R)$，即三元组集合 $KG=\{(h,r,t)|h,t\in E,r\in R\}$，其中 E 为实体集，R 为关系集，(h,r,t) 表示头实体 h 通过关系 r 连接尾实体 t。令 IP - PORT $=\{i_1,i_2,\cdots,i_n\}$ 和 ORG $=\{o_1,o_2,\cdots,o_m\}$ 分别表示网络资产实体集与组织实体集。对于任意给定的网络资产–组织关联对 (i_m,o_n)（其中 $i_m\in$ IP - PORT，$o_n\in$ ORG），知识图谱中连接 i_m 与 o_n 的三元组序列 $\{(i_m,r_1,e_1),(e_1,r_2,e_2),\cdots,(e_{l-1},r_l,o_n)\}$ 构成一条语义路径 p。由此，对于关联对 $A_k=(i_m,o_n)$，所有符合条件的路径构成了路径集 $P_k=\{p_1,p_2,\cdots,p_z\}$。在实际归属预测任务中，由于目标组织实体未知，需将搜索的尾节点约束为组织类型实体。

2.3.1　基于深度优先搜索的路径提取

原始图谱中包含大量与源网络资产无关的冗余信息。为降低噪声干扰，本文采用深度优先搜索算法对图谱进行剪枝，即从源节点出发，检索所有可达的候选组织节点，构建用于归属推理的子图。在此过程中，同一源资产可能指向多个候选目标组织，从而形成多个网络资产–组织关联对 $A_{k1}(i_m,o_1)$，$A_{k2}(i_m,o_2),\cdots,A_{kn}(i_m,o_n)$，且每个关联对均对应一个特定的路径集 P_{ki}，预测路径示意如图 7 所示。

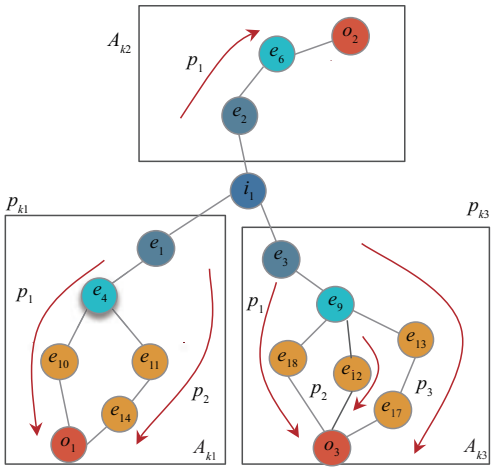


图 7　预测路径示意

深度优先搜索算法提取从源网络资产节点至目标组织节点的语义路径。算法从初始网络资产节点出发，递归遍历邻居节点并记录路径类型，直至到达 ORG 类型节点或满足终止条件。当某条路径搜索完毕后，执行回溯操作以探索其他潜在分支。此外，考虑到知识图谱规模庞大，且长路径会导致头尾实体间的语义相关性衰减，本文设置了最大路径跳数限制以过滤噪声路径，具体流程如算法 1 所示。

算法 1　基于深度优先搜索的路径抽取算法

输入　有向图 G，起始点 ip，最大跳数 max_hops

输出　所有从 ip 到 org 节点的路径集合 paths

```
1) 初始化路径集合 paths = {}
2) define DFS(current_node, path, depth):
3)     if depth > max_hops:
4)         return
5)     for G.neighbors(current_node) 中的每个 neighbor:
6)         edge_data ← G[current_node][neighbor]
7)         relation ← edge_data['relation']
8)         if neighbor 已在 path 路径节点集合:
9)             continue
10)        构造 new_path ← path + [(neighbor, relation)]
11)        neighbor_type ← G. nodes[neighbor].get('type', '')
12)        if neighbor_type == 'org':
13)            构造 path_entry = {
14)                'org': neighbor,
15)                'path': new_path,
16)                'hops': len(new_path) −1
17)            }
18)            添加 path_entry 到 paths
19)        else:
20)            调用 DFS(neighbor, new_path, depth+1)
21)    return
```

2.3.2　基于综合路径的预测路径评分

基于前述工作，本文已构建网络资产知识图谱，并实现了从目标网络资产至候选组织节点的全路径提取。分析表明，单一网络资产可能指向

多个候选组织，且每个“网络资产–组织”关联对下存在多条语义路径。因此，如何从候选集合中甄别出概率最高的归属组织，即量化目标网络资产与特定组织间的关联紧密度，成为亟待解决的核心问题。

为了解决上述问题，本文融合基于自信息量的边权计算、基于 PageRank 算法的节点重要性评估以及跳数因子，构建完整的路径评分体系，评估网络资产与组织机构的相关性。完整的评分体系结构如图 8 所示。

（1）基于自信息量的关系权值计算

为确定不同关系的权重，本文引入信息论中的自信息（self-information）概念。自信息用于衡量单个事件的不确定性与信息量，即事件发生的概率越低，其携带的信息量越大。其数学定义为：对于发生概率为$p(x)$的事件x，其信息量$I(x)$表示为：

$$I(x) = -\mathrm{lb}\,p(x) \tag{1}$$

其中，$p(x)(0 \leqslant p(x) \leqslant 1)$为事件$x$的发生概率。

在网络资产知识图谱中，节点间经由特定关系的跳转可视为彼此独立的随机事件。本文采用关系在图谱中的出现频率近似表征该关系的发生概率。根据自信息原理，关系出现的概率越低（即越稀有），其在组织归属推理中提供的区分度与信息量越大；反之，高频泛化关系的信息贡献度较低。基于此，本文设计关系权重计算式为：

$$W(r) = -10\mathrm{lb}f(r) \tag{2}$$

其中，$f(r)(0 \leqslant f(r) \leqslant 1)$为关系$r$在图谱中的相对频率。

经过统计，关系及其计算出的自适应权重如表 6 所示。

表 6　　关系及其自适应权重

关系	频率	权重
IPC_IS	0.489 66	10.301
IS_IPC_OF	0.489 66	10.301
DOMAIN_IS	0.008 11	69.456
IS_DOMAIN_OF	0.008 11	69.456
BELONGS_TO	0.004 45	78.107

（2）基于 PageRank 算法的节点计算

确定关系权重后，需对节点重要性进行量化评分。本文采用 PageRank 算法[14]计算节点的 PR 值，以此作为重要性度量的依据。该算法的核心思想为：节点的重要性取决于指向该节点的连边数量以及源节点的自身质量。即一个节点若被众多高权重的节点所指向，则其重要性越高。此外，本文引入个性化向量对标准算法进行改进（即 personalized PageRank，PPR），使随机游走模型更倾向于访问与目标节点（待预测网络资产）相关的路径。该机制使与目标网络资产拓扑邻近或结构相关的节点获得更高得分，从而实现对节点重要性的精准评估。

对于路径中节点的PR值，其具体计算步骤如下。

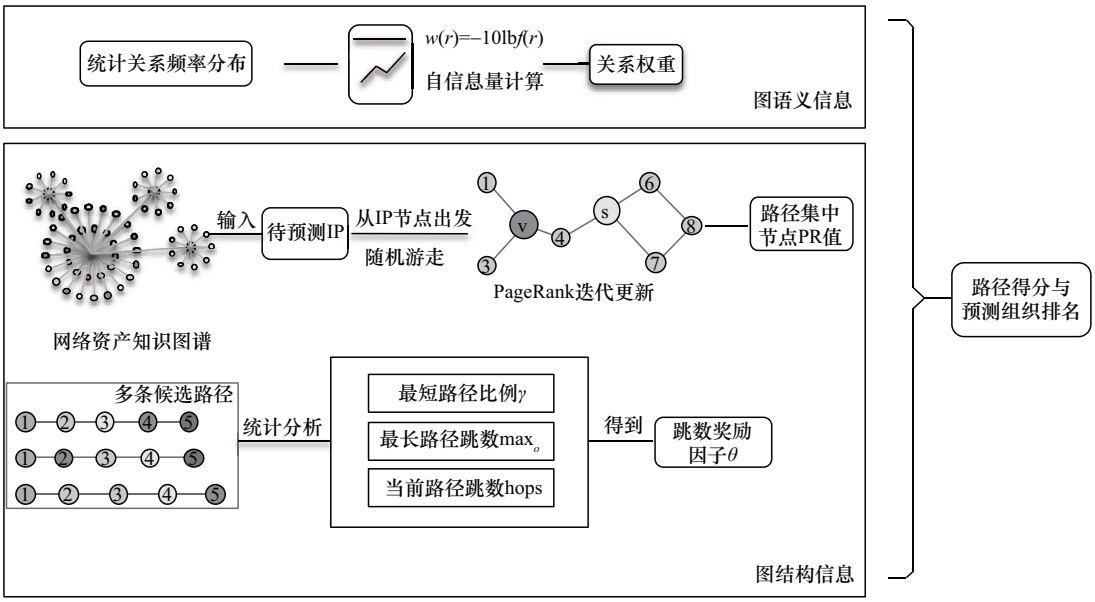


图 8　评分体系结构

赋初值。初始化阶段，算法将图谱中所有节点的PR值预设为0。随后，针对输入的待预测IP节点集合，将其初始PR值置为1。该初始化策略旨在引导随机游走过程聚焦于目标节点邻域，从而在降低计算开销的同时提升算法的收敛效率与预测精度。在后续计算中，节点PR值将随迭代次数t动态更新，其初始状态需满足如下条件：

$$\mathrm{PR}_i(0)=0$$
$$\sum_{j=1}^{N}\mathrm{PR}_j(0)=1 \tag{3}$$

其中，j代表输入的IP列表中的节点。

迭代更新。基于深度优先搜索（DFS）构建的路径子图，算法从目标IP节点出发执行随机游走。在每步状态转移中，游走者以概率α沿当前节点的连边移动，或以概率$1-\alpha$执行随机跳转。伴随每一次迭代，算法将动态更新节点的PR值。

具体更新规则如下：

$$\mathrm{PR}^{(t+1)}(v)=(1-\alpha)p(v)+\alpha\sum_{u\in N(v)}\frac{\mathrm{PR}^{(t)}(u)}{d_u} \tag{4}$$

其中，$\alpha\in(0,1)$为阻尼因子，通常取值为0.85；$N(v)$为指向节点v的所有前驱节点；d_u为节点u的出度（即有多少条边从u出发）；$\mathrm{PR}^{(t)}(u)$为第t次迭代时，节点u的PR值；$\boldsymbol{p}(v)$为个性化向量，用于引导随机游走在跳转时优先指向目标节点集合，该向量的取值取决于待预测IP集合的规模。设输入IP集合为N，其中包含n个IP节点，则$\boldsymbol{p}(v)$定义为：

$$\boldsymbol{p}(v)=\begin{cases}\dfrac{1}{n}, & v\in N\\ 0, & v\notin N\end{cases} \tag{5}$$

PageRank算法的迭代更新流程示意如图9所示。

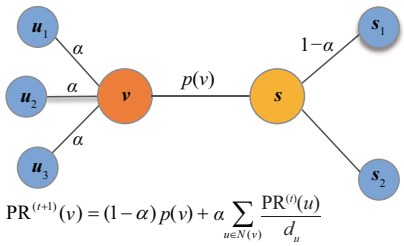


图9　PageRank算法迭代更新流程示意

迭代终止判定。算法持续循环执行上述迭代更新过程，利用式(4)计算并更新各节点的PR值，直至满足如下收敛条件：

$$\left\|\mathrm{PR}^{(t+1)}-\mathrm{PR}^{(t)}\right\|<\varepsilon \tag{6}$$

其中，收敛阈值设定为$\varepsilon=10^{-6}$。

（3）跳数奖励因子

实际搜索得到的组织归属路径集往往包含不同长度的路径。针对路径长度差异，研究表明，在“网络资产–组织”关联对中，路径越长，其解释归属关系的可信度越低；反之，路径越短意味着两者关联越直接，紧密度越高。此外，若路径集中短路径占比较低，则该类路径因其稀缺性而更具判别价值，应给予更高的权重奖励。因此，为量化路径长度对关联置信度的影响，本文引入跳数奖励因子β。具体计算式为：

$$\beta=(2-\gamma)^{\max_{\mathrm{hops}}-\mathrm{hops}} \tag{7}$$

其中，γ为归属到组织路径集中最短路径的比例，$\max_{\mathrm{hops}}$为归属到该组织路径集中最长路径跳数（跳数指从一个节点跳转到另一个节点的次数），hops为当前路径跳数。

（4）路径评分计算

综合上述基于权重与节点的评分规则，并引入跳数奖励因子，本文旨在充分利用知识图谱的语义与拓扑结构信息对推理路径进行量化评估。对于一条完整路径$p_i=i\{1\}\xrightarrow{r_1}e_1\xrightarrow{r_2}e_2\cdots\xrightarrow{r_i}o\{n\}$，其得分计算式定义为：

$$\mathrm{Score}(P)=\frac{1}{n-1}\sum_{i=1}^{n-1}w_r(e_i,e_{i+1})\mathrm{PR}(e_{i+1})\beta \tag{8}$$

其中，P为经搜索获取的完整语义路径，$w_r(e_i,e_{i+1})$为节点e_i与e_{i+1}之间关系边的权重，$\mathrm{PR}(e)$为节点e的PR值，β为路径跳数奖励因子。

在计算单条路径得分后，对指向同一候选组织的所有路径得分进行累加，从而得出目标网络资产归属该组织的最终量化评分。路径评分计算流程如算法2所示。

算法2　路径评分计算流程

函　数　Score_Paths(paths, pr_scores, rel_weights, beta_list):

1) 初始化scored_paths为一个空列表
2) 对于每一条路径 path 在 paths 中，顺次执行
3) 　初始化score = 0.000
4) 　对于path中的每一对连续节点(src, dst)
5) 　　取边 (src, dst) 的关系rel

6)　　取对应关系权重 weight=rel_weights[rel]
7)　　取目标节点 dst 的 PR 值 ppr = pr_scores[dst]
8)　　累加 score += weight × ppr
9)　　计算路径跳数 hops=len(path)-1
10)　　按式(8)对边权重累加值进行归一化：score=$\frac{\text{score}}{\text{hops}}$
11)　　如果 beta_list 存在
12)　　取 beta = beta_list[当前路径索引]
13)　　否则
14)　　beta = 1.0 # 默认无奖励或惩罚
15)　　计算最终得分 final_score = score × beta
16)　　构建路径评分对象（如字典），包含
17)　　- path: 当前路径节点序列
18)　　- score: final_score
19)　　- raw_score: score 未加 beta 奖励的分值
20)　　- beta: 当前路径的 beta 值
21)　　- hops: 跳数 = len(path) − 1
22)　　- org: path 最后一个节点（组织节点）
23)　　将评分对象加入 scored_paths
24)　　返回 scored_paths

3　实验与分析

3.1　实验设置

3.1.1　数据集

本文以中国教育自治域内的网络资产为研究对象，原始系统通过融合 WHOIS、DNS 解析记录、SSL 证书、网页文本、域名和 IP C 段等多源异构数据，构建了包含约 15 万个节点和 20 万条关系的网络资产知识图谱。在可复现实验中，进一步抽取 6 792 个具有明确组织归属标签且具备结构化特征的 IP 资产作为闭集评测基础。

为避免测试标签泄露，实验采用组织分层 held-out 策略划分数据，如表 7 所示。由表 7 可以看出，训练集包含 5 415 个 IP，测试集包含 1 377 个 IP，训练图谱中包含 854 个组织实体。测试 IP 的组织标签不参与训练图谱构建、路径搜索、候选组织生成和得分计算，仅在最终评价阶段作为 ground truth 使用。该设置比随机抽样测试更严格，能够降低归属边直接泄露对评价结果的影响，使测试结果更接近真实未知资产归属场景。

表 7　采用组织分层 held-out 策略划分数据

项目	数值
带标签 IP 资产数	6 792
训练集 IP 数	5 415
测试集 IP 数	1 377
训练图谱组织数	854
评价任务	闭集组织归属预测

除全量测试集外，为了能够更好地检验出算法各个模块的实际作用，本文进一步基于全量测试集构造多候选组织子集和长路径子集，用于观察不同模块在其目标场景中的作用。诊断式消融实验子集划分如表 8 所示，其中，多候选组织子集用于考察评分模块与节点权重在候选组织排序中的作用，长路径子集用于考察搜索深度增加后长路径噪声及跳数奖励的影响。

表 8　诊断式消融实验子集划分

诊断子集	样本数/个	占测试集比例
多候选组织子集	298	21.64%
长路径子集	1 362	98.91%

3.1.2　基准方法

为验证本文方法的有效性，本研究与以下 6 种主流的网络资产归属分析范式进行了对比实验。

（1）基于 IP-API 查询的方法。IP-API 提供免费的 IP 地理定位与归属查询 API，主要用于查询 IPv4/IPv6 地址的地理信息、运营商、自治域号等数据，可提供服务提供商、组织名称、自治域号、自治域名称、BGP 网段等数据。

（2）基于 WHOIS 数据库查询的方法。WHOIS 数据库存储了互联网号码分配机构（RIR）下发的 IP 地址块注册信息。通过索引该数据库，可追溯 IP 地址分配链条，从而确定其注册组织、ISP 或所属数据中心等法定责任主体。

（3）基于测绘平台查询的方法。本文选取了具

有代表性的主流测绘平台（包括 FOFA、Censys、Shodan 及 360quake 平台）进行对比测试。

（4）基于 Whosedomain 的方法[1]。Whosedomain 方法依托多源数据联动机制，首先采集目标域名或 URL 的 IP 地址及 DNS 解析记录；进而借助 VirusTotal 等威胁情报平台拓展子域名、关联 IP 及历史解析记录；最后通过加权分析与组织名相似度匹配算法，综合推断最终的组织归属。

（5）基于 Resource Allocation 的方法[15]。Resource Allocation 方法是一种基于局部拓扑相似性的链路预测方法，首先在网络资产知识图谱中提取待归属 IP 节点与候选组织节点之间的共同邻居节点；进而根据共同邻居的连接度为其分配权重，连接度越低的邻居节点通常具有更强的区分能力，连接度越高的公共节点则被赋予较低权重；最后将所有共同邻居贡献的资源分配得分进行累加，并选择得分最高的候选组织作为资产归属结果。

（6）基于 Katz path index 的方法[16]。Katz path index 方法是一种基于多跳路径加权累积的链路预测方法，首先在网络资产知识图谱中搜索待归属 IP 节点与候选组织节点之间的不同长度路径；进而按照路径长度引入衰减系数，短路径被赋予较高权重，长路径因关联不确定性较强而被赋予较低权重；最后对所有可达路径的衰减得分进行累加，得到 IP 节点与候选组织节点之间的关联强度，并将关联强度最高的组织作为预测归属。

3.1.3　评估指标

本文将归属分析视作一个分类任务，采用分类任务常见的指标评估方法性能，评估指标计算公式及其意义如表 9 所示，其中 TP、FP、FN 和 TN 的含义如下。

（1）TP（true positive）：预测正确，即预测的组织与真实组织匹配。

（2）FP（false positive）：预测错误，即预测了组织但与真实组织不匹配。

（3）FN（false negative）：漏检，即应该预测出组织但没有预测（或预测错误）。

（4）TN（true negative）：正确拒绝（在本项目中通常不适用，因为每个 IP 都应该有组织归属）。

3.1.4　实验环境

所有实验均在高性能计算服务器上完成，为了保证实验的可复现性，详细的实验环境配置如表 10 所示。

表 10　　实验环境配置

类别	配置项	详细参数
硬件环境	CPU	Intel Xeon Gold 6230R (26 核心/52 线程，2.10 GHz)
	GPU	2*NVIDIA GeForce RTX 4090 (24 GB GDDR6X)
	内存	256 GB DDR4
	存储	4T HDD + 4T SSD
软件环境	操作系统	Ubuntu 22.04
	开发语言	Python 3.10
	neo4j	5.28.1 版本

3.2　结果分析

3.2.1　对比实验结果

闭集实验结果如表 11 所示。由表 11 可以看出，本文基于知识图谱推理的方法展现出明显的性能优势，其准确率、召回率以及 F1 值均高达 0.970。而 IP-API 方法的 F1 值为 0.080，基于 Whosedomain 的方法 F1 值为 0.930，性能均低于本文方法基于现有公开测绘平台难以获得准确归属结果。造成这一显著性能差异的主要原因在于方法底层逻辑的不同。传统的基准方法（如 IP-API、各类测绘平台）主要依赖于静态数据库的匹配、简单的规则查询或公开的 WHOIS 信息。在闭源或信息受限的数据集环境下，这些方法往往因为数据覆盖不全、缺乏深层关联挖

表 9　　评估指标计算公式及其意义

评价指标	计算公式	意义
精准率	$\text{Precision} = \frac{TP}{TP + FP}$	实际为正且预测为正的样本占所有被预测为正的样本的比例
准确率	$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$	所有预测正确的样本占总样本的比例
召回率	$\text{Recall} = \frac{TP}{TP + FN}$	实际为正且预测为正的样本占所有实际为正的样本的比例
F1 值	$F1 = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$	精确率和召回率的调和平均数，用于综合评估模型的性能

掘能力而导致查不到或查不准。本文方法基于知识图谱推理，能够利用图谱中实体与关系的拓扑结构，通过多跳推理挖掘出数据间隐含的语义关联，从而突破了单纯依赖表面特征匹配的局限性。实验结果有力地证明了基于知识图谱推理的方法在处理闭源数据集任务上的优越性。该方法不仅有效解决了传统工具在特定环境下召回率低、准确性差的问题，更展示了其在复杂数据关联分析中的鲁棒性，是优于现有商业测绘平台及API查询的高效解决方案。

表11　　闭集实验结果

方法	精确率	准确率	召回率	F1值
FOFA平台	—	—	—	—
360Quake平台	0	0	0	0
Censys平台	—	—	—	—
Shodan	—	—	—	—
Whosedomain	0.980	0.880	0.880	0.930
IP-API	0.080	0.080	0.080	0.080
Resource Allocation	0.970	0.960	0.950	0.960
Katz path index	0.950	0.940	0.940	0.950
知识图谱推理	**0.980**	**0.970**	**0.970**	**0.970**

本文方法在1 377个held-out测试IP上取得准确率为0.970、精确率为0.980、召回率为0.970、F1值为0.970的结果。与两个经典链路预测基线相比，本文方法高于Resource Allocation的F1值（0.960）和Katz path index的F1值（0.950）。这说明，简单共享邻居或固定路径衰减能够捕获一部分组织归属线索，但难以充分区分域名、IP C段、关系稀有度和路径长度等证据差异。

从实验结果还可以发现，FOFA、Shodan、Censys等主流测绘平台的实验结果为空值，360Quake平台因未能识别长尾资产而导致各项指标均为0。造成这种显著性能差异的主要原因在于，现有的通用网络测绘平台大多停留在粗粒度的自治域层面，无法深入具体的IP层级，或者在面对非热门、非标准化的长尾网络资产时存在严重的数据覆盖盲区。相比于传统的查询接口或商业测绘平台，基于知识图谱的推理方法在解决细粒度网络资产归属问题上具有极高的准确度与完整性，成功突破了现有工具在长尾资产识别上的技术瓶颈。

3.2.2　不同阶段网络资产归属情况

3个阶段资产数量及阶段增量如图10所示。可以看出，在网络资产知识图谱构建、规则与大模型协同分析、基于图谱路径推理分析3个阶段，资产信息数量呈现显著的阶梯式增长特征。网络资产知识图谱构建阶段完成初始积累，累计资产信息数量达2 513条，为后续分析奠定基础；引入规则与大模型协同分析机制后，本阶段新增资产信息1 118条，阶段增量占比达44.5%，累计总量提升至3 631条，验证了协同分析对资产扩充的有效性；采用基于图谱路径推理分析策略后，资产增量效应进一步凸显，本阶段新增资产信息2 010条，增量占比提升至55.4%，最终累计总量达5 641条。相较于前序阶段，基于图谱路径推理分析阶段展现出显著的性能优势，该阶段实现了2 010条的资产信息新增量，增量占比高达55.4%，不仅单阶段新增量远超规则与大模型协同分析阶段的1 118条，更推动资产信息总量从3 631条跃升至5 641条，实现了资产覆盖范围的突破性拓展。不同阶段资产占比如图11所示。以上结果表明，图谱路径推理能够充分挖掘网络资产知识图谱中实体间的关联关系，有效发现传统方法与基础协同分析难以覆盖的隐性资产关联与未知资产，展现了其在网络资产测绘任务中对资产提取能力的关键提升作用，为资产发现的完整性与全面性提供了强有力的技术支撑。

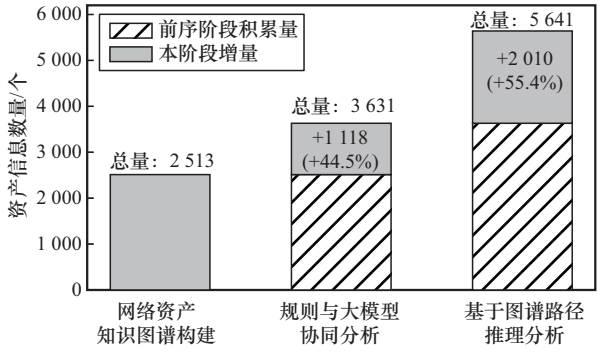


图10　3个阶段资产数量及阶段增量

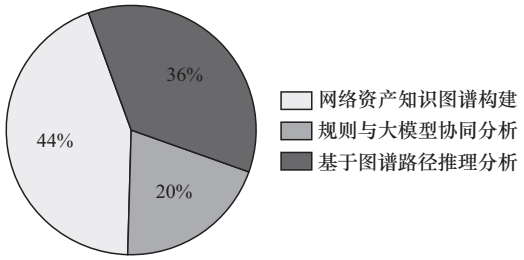


图11　不同阶段归属资产占比

3.2.3　评分模块消融实验

为明确评分模型中各组成部分的作用，本文将统一权重（所有路径证据都按相同分值处理，只看“有没有路径”，不区分路径类型、稀有程度和共享程度）、关系自信息量（根据不同关系类型在图谱中出现的频率赋权，越少见的关系类型，信息量越高；越常见的关系类型，信息量越低，其作用是区分不同边类型的重要程度，但只考虑关系类型的全局稀有度）、属性稀有度（根据具体属性的共享程度赋权）拆开比较。

在全量测试集上，统一权重F1值为0.952，加入属性稀有度后提升至0.973。在更能体现候选组织竞争关系的多候选组织子集上，统一权重F1值为0.819，仅使用关系自信息量后F1值为0.836，引入属性稀有度后F1值为0.916，评分模块消融实验结果如图12所示。

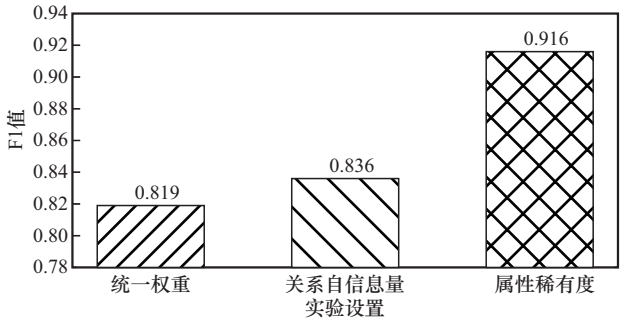


图12　评分模块消融实验结果

由图12可以看出，评分模块的主要作用并非简单调整全局分数，而是在多个候选组织同时存在时抑制弱证据累加。关系自信息量能够区分不同关系类型的全局稀有度，属性稀有度进一步刻画具体域名或C段被多少训练资产共享。当某个域名或IP C段连接大量资产时，其归属指示性下降；反之，低共享度证据通常更能指向特定组织。

3.2.4　节点权重与PageRank分析

节点权重主要用于多候选组织排序。为检验该模块在候选竞争场景中的作用，本文从测试集中选取包含多个候选组织的样本构成多候选组织子集。该子集共包含298个样本，平均候选组织数为6.13，能够较集中地反映候选组织得分接近时的排序能力。

多候选组织子集上的PageRank排序效果如图13所示。Top-1 Accuracy指的是不加节点权重直接计算路径得分，并根据得分选取归属组织。组织频次根据训练集中某组织已有归属资产数量给候选组织一个轻量先验。训练集中出现越多的组织，其先验权重越高。全局PageRank指的是在训练知识图谱上构建包含IP、域名、C段、组织等节点的整体图，然后计算得到的PageRank值。组织节点的PageRank值越高，说明该组织在全局图结构中越中心、连接越稳定。用于多候选排序时，它可以帮助优先选择图结构中更可信、更中心的候选组织。无节点权重时Top-1 Accuracy为0.138，组织频次先验提升至0.818，全局PageRank进一步提升至0.951。同时，全局PageRank的MRR（mean reciprocal rank）为0.972，平均真实组织排名为1.056。该结果说明，在多个候选组织同时存在且路径证据相近的情况下，PageRank能够利用训练图中的组织中心性提供有效的排序先验，从而提高真实归属组织排在首位的概率。

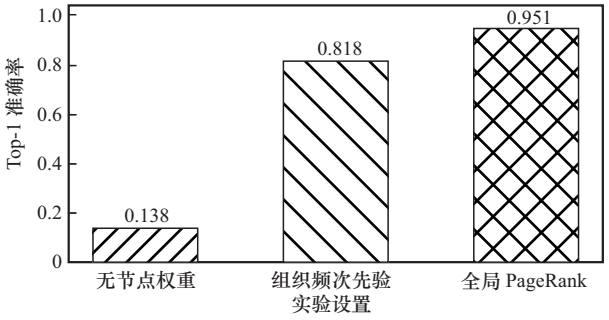


图13　多候选组织子集上的PageRank排序效果

3.2.5　跳数奖励、搜索深度与复杂度分析

本文方法采用带剪枝的深度优先搜索提取候选组织路径。DFS路径搜索在较大搜索深度下可能存在组合爆炸风险。为此，本文设置最大跳数与路径数量的敏感性实验。最大路径跳数对路径数量、耗时和F1值的影响如图14所示。当最大跳数从3增加到5时，平均路径数和运行时间明显上升，但总体F1值并未提升，说明主要有效证据集中在短路径中，盲目扩大搜索深度会引入弱关联路径和计算开销。

在长路径子集上进一步观察跳数奖励的作用，跳数奖励与搜索深度的诊断式消融结果如图15所示。最大跳数为5且不使用跳数奖励时F1=0.947，

加入跳数奖励后F1=0.973。这表明跳数奖励的主要价值在于当长路径被纳入候选集合后，降低长路径噪声对最终排序的影响。其合理定位不是鼓励无约束扩大DFS搜索深度，而是在受控搜索深度内提升路径评分的鲁棒性。

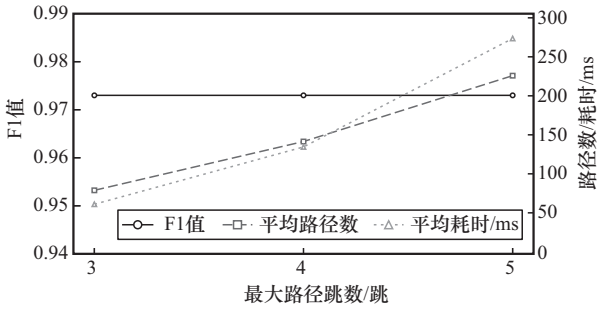


图14　最大路径跳数对路径数量、耗时和F1值的影响

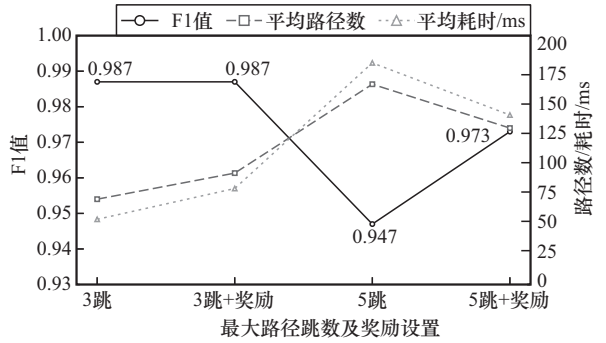


图15　跳数奖励与搜索深度的诊断式消融结果

3.2.6　大模型协同模块输入信号与独立调用分析

规则与大模型协同模块用于从证书组织字段、网页标题、证书CN、注册邮箱和域名等文本信号中识别或规范化组织名称。为避免大模型直接接触真实标签，本文在同一held-out测试集上首先对输入信号进行了可复现量化。规则与大模型协同模块输入信号定量结果如图16所示。证书组织字段的F1值为0.980，域名/CN词典代理的F1值为0.971，网页标题规则的F1值为0.130，多源融合上界的F1值为0.997。该结果说明，证书与域名类文本信号是当前教育网数据集中最稳定的组织归属依据，网页标题规则主要提供少量补充证据，但融合以后可以有效提取出目标资产的归属组织情况。

实验对50个样本进行3次重复调用，有效重复样本数为50个。调用成本方面，本次实验输入token为79 800，输出token为8 383，总token为88 183，估算成本为0.028 970元，平均时延为7.519 m/次。

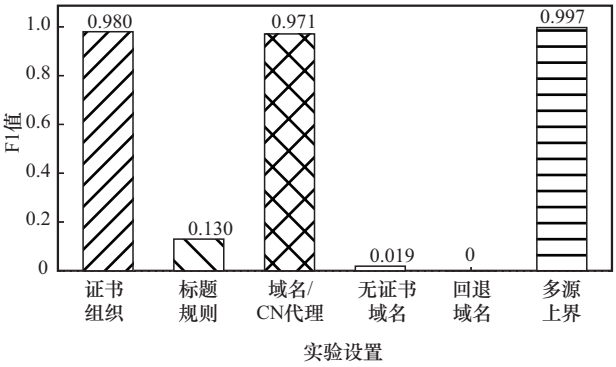


图16　规则与大模型协同模块输入信号定量结果

3.2.7　归属分析效率对比

不同归属分析方法的效率对比如图17所示。由图17可以看出，本文采用的知识图谱路径推理方法在归属速度与推理能力上实现了较好平衡，综合表现优于传统方法，具有明显的实用优势。基于Whois查询的方法完成100个组织的组织归属仅需0.730 s，耗时最短，但其依赖外部接口查询，存在数据滞后、解析不稳定及归属链路不完整等问题，难以支撑真实世界场景下网络空间资产的精准归属。Whosedomain方法分析100个网络资产的组织归属情况需要362.740 s，整体耗时较长，归属分析效率较低。本文基于知识图谱路径推理的方法在分析100个网络资产的归属组织时耗时180.100 s，效率显著优于Whosedomain方法，在保证处理速度的同时，能够通过语义关系实现多级、间接的资产归属推理，具备传统数据库匹配与单一接口查询所不具备的关联分析能力。实验结果表明，该方法在效率与推理能力之间取得了更优平衡，既能够满足规模化资产归属的时效性需求，又能实现更全面、更准确的归属判定，充分体现了本文方法在实际网络资产治理中的优越性。

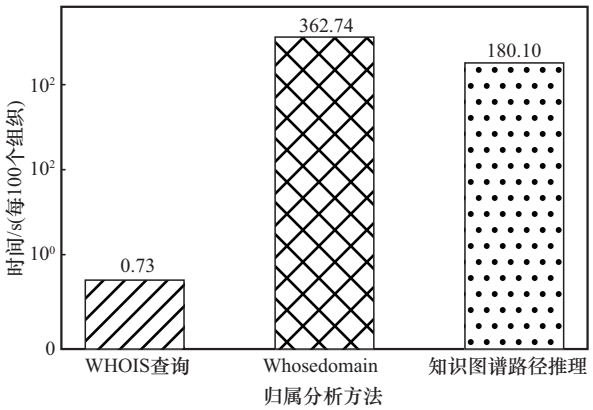


图17　不同归属分析方法的效率对比

3.2.8　案例分析

案例分析如表12所示。在针对210.XX.XX.179-80等4个典型待归属网络资产的测试中，模型输出的推理结果（分别为安徽XX大学、X州大学、广西XXX中心和X华大学）与专家人工校验的结果完全一致，展现出了极高的准确性。这一现象主要归因于系统采用了基于路径评分的知识图谱推理机制，该机制能够挖掘出得分最高的关键推理路径，将210.45.144.179-80归属至安徽XX大学路径示意如图18所示。由图18可以看出，将网络资产210.XX.XX.179-80通过中间域名节点“*.jwc.ahlyxy.cn”关联至安徽XX大学，或者利用同IP C段的规则将IP地址关联到一个有明确归属组织机构的IP地址，从而构建出逻辑严密且可视化的证据链条。综上所述，该实验有力地证明了基于路径评分的方法不仅能够实现精确的网络资产归属，更关键的是通过提供具体的推理路径，有效解决了传统归属算法中存在的黑箱难题，显著提升了判定的可解释性与可信度。

本文方法对10组网络资产进行了归属预测，开放数据案例分析如表13所示。由表13可以看出，本文方法成功识别出了包括“复X大学”“哈尔滨XX大学”在内的多个组织结构。经由专家校验，所有预测结果与实际情况相符合，显示出极高的准确性。与之形成鲜明对比的是，作为基准对比的360Quake测绘平台在所有列出的测试样本中均未返回有效结果。经过分析，造成上述结果差异的原因在于两类方法的机制不同。主流的网络空间测绘平台通常依赖于主动扫描或既有的指纹数据库匹配，对于未开放特定服务端口、处于静默状态或者未被数据库收录的网络资产，往往难以进行有效识

表12　　案例分析

待归属网络资产	推理路径	推理结果	专家校验结果
210.XX.XX.179-80	210.XX.XX.179-80 -> 210.XX.XX.179 -> *.jwc.XX.cn-> 安徽XX大学	安徽XX大学	正确
202.XX.XX.212-80	202.XX.XX.212-80 -> *.zzu.XX.cn -> 202.XX.XX.214 -> X州大学	X州大学	正确
210.XX.XX.33-80	210.XX.XX.33-80 -> 210.XX.XX -> 210.XX.XX.63 -> 广西XXX中心	广西XXX中心	正确
202.XX.XX.206-80	202.XX.XX.206-80 -> *.xhu.XX.cn -> 202.XX.XX.207 -> X华大学	X华大学	正确

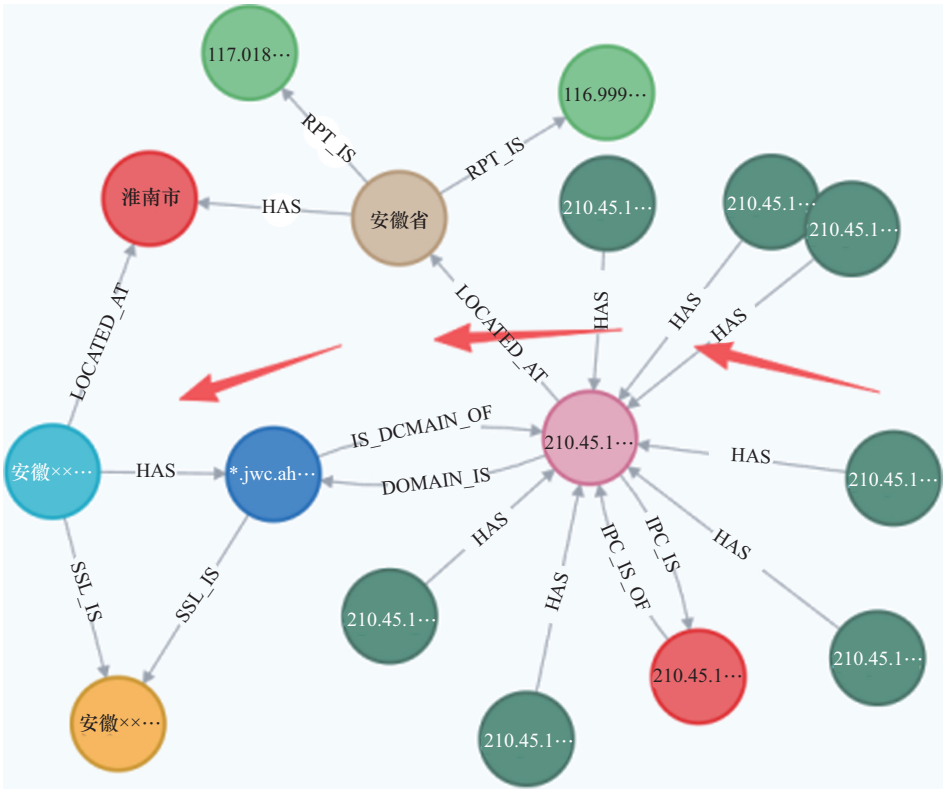


图 18　将210.45.144.179-80归属至安徽XX大学路径示意

表 13 开放数据案例分析

预测网络资产	预测结果	360quake	专家校验	是否准确
59.XX.XX.75-80	兰州XX职业技术大学	—	兰州XX职业技术大学	是
210.XX.XX.79-80	哈尔滨XX大学	—	哈尔滨XX大学	是
58.XX.XX.83-80	复X大学	—	复X大学	是
202.XX.XX.177-80	中国XX大学（武汉）	—	中国XX大学（武汉）	是
211.XX.XX.85-80	山东XX大学	—	山东XX大学	是
222.XX.XX.183.7-80	XX科技大学	—	XX科技大学	是
210.XX.99.53-80	安徽XX大学	—	安徽XX大学	是
218.XX.XX.228-80	黄河XX大学	—	黄河XX大学	是
42.XX.XX.228-80	新疆XX学院	—	新疆XX学院	是
219.XX.XX.29-80	广州XX学院	—	广州XX学院	是

别。而本文方法不单纯依赖直接的扫描结果，而是通过挖掘数据间的隐式关联和推理逻辑来确定资产归属，因此能够在传统测绘平台失效的情况下依然保持有效的识别能力。开放数据案例分析有力地说明了本文方法在实际应用场景中的有效性和准确性，该方法不仅在资产归属上有较高的准确率，更重要的是，它能够有效弥补主流测绘平台在资产归属上的不足，证明了其在处理开放、复杂网络环境下的资产归属任务中具有显著的优越性和实用价值。

4 结束语

针对网络空间资产归属分析中面临的数据碎片化、归属模糊及现有方法可解释性不足等问题，本文提出了一种基于知识图谱路径推理的归属分析方法。本研究首先通过融合 WHOIS、SSL 证书及网页文本特征，构建了面向特定领域的网络资产知识图谱，从数据层面打破了信息孤岛；其次，设计了规则匹配与大语言模型协同的初步归属补全机制，并进一步提出了一种融合剪枝策略的深度优先搜索算法与基于随机游走的路径评分模型，将归属预测转化为可量化、可解释的证据链搜索过程。基于中国教育自治域真实数据的实验结果表明，该方法在保证高准确率的同时，能够有效应对多跳关联与隐性资产归属难题，其性能显著优于传统的 WHOIS 查询及主流网络测绘平台。本文工作不仅验证了知识图谱技术在网络资产画像领域的有效性，也为网络安全防御中的资产攻击面梳理提供了新的技术思路。未来的研究工作将探索图谱的增量更新与动态演化机制，以适应网络资产高频变动的特性，实现资产归属的实时动态感知。

参考文献：

[1] Sebastián S, Diugan R G, Caballero J, et al. Domain and website attribution beyond WHOIS[C]//Proceedings of the Annual Computer Security Applications Conference. New York: ACM Press, 2023: 124-137.

[2] Ren Y M, Li H, Liu P P, et al. Owner name entity recognition in websites based on multiscale features and multimodal co-attention[J]. Expert Systems with Applications, 2023, 224: 120014.

[3] Introducing the knowledge graph: things, not strings[EB].[2012-05-16].

[4] Streibelt F, Lindorfer M, Gürses S, et al. Back-to-the-future whois: an IP address attribution service for working with historic datasets[C]//Passive and Active Measurement. Berlin: Springer, 2023: 209-226.

[5] Qiao Z, Zhang C, Du G. Improving cybersecurity named entity recognition with large language models[C]//Proceedings of the 2023 6th International Conference on Software Engineering and Computer Science (CSECS). Piscataway: IEEE Press, 2023: 1-6.

[6] Vizcarra J, Gempei Y, Wang Y, et al. Constructing cybersecurity knowledge graphs for hybrid LLM-graph reasoning on vulnerabilities[C]//Proceedings of the 24th International Semantic Web Conference, Berlin: Springer, 2025: 2-6.

[7] BORDES A, USUNIER N, GARCIA-DURAN A, et al. Translating embeddings for modeling multi-relational data[C]//Advances in Neural Information Processing Systems. New York: ACM Press, 2013: 2-6

[8] Sun Z Q, Deng Z H, Nie J Y, et al. RotatE: knowledge graph embedding by relational rotation in complex space[C]//Proceedings of the 7th International Conference on Learning Representations (ICLR 2019). Vancouver: ICLR, 2019: 1-18.

[9] Yu X, Ren X, Sun Y Z, et al. Personalized entity recommendation: a heterogeneous information network approach[C]//Proceedings of the 7th ACM International Conference on Web Search and Data Mining. New York: ACM Press, 2014: 283-292.

[10] Huang J, Zhao W X, Dou H J, et al. Improving sequential recommendation with knowledge-enhanced memory networks[C]//Proceedings of the 41st International ACM SIGIR Conference on Research & Development in Information Retrieval. New York: ACM Press, 2018: 505-514.

[11] Wang X, Wang D X, Xu C R, et al. Explainable reasoning over knowledge graphs for recommendation[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2019, 33(1): 5329-5336.

[12] Wang H W, Zhang F Z, Wang J L, et al. RippleNet: propagating user preferences on the knowledge graph for recommender systems[C]//Proceedings of the 27th ACM International Conference on Information and Knowledge Management. New York: ACM Press, 2018: 417-426.

[13] Wang H W, Zhao M, Xie X, et al. Knowledge graph convolutional networks for recommender systems[C]//Proceedings of the World Wide Web Conference. New York: ACM Press, 2019: 3307-3313.

[14] Brin S, Page L. The anatomy of a large-scale hypertextual Web search engine[J]. Computer Networks and ISDN Systems, 1998, 30(1-7): 107-117.
[15] Zhou T, Lyu L Y, Zhang Y C. Predicting missing links via local information[J]. The European Physical Journal B, 2009, 71(4): 623-630.
[16] Katz L. A new status index derived from sociometric analysis[J]. Psychometrika, 1953, 18(1): 39-43.

[作者简介]



施凡（1983– ），男，博士，国防科技大学教授、硕士生导师，主要研究方向为网络空间测绘和网络空间测量。



窦开洋（2003– ），男，国防科技大学硕士生，主要研究方向为网络空间测绘、边缘计算。



李晔（2002– ），男，国防科技大学博士生，主要研究方向为网络空间测绘和大语言模型。



屠铖（1998– ），女，国防科技大学博士生，主要研究方向为网络空间测绘、数据挖掘和大语言模型。



许成喜（1989– ），男，博士，国防科技大学副教授，主要研究方向为互联网基础设施安全和网络空间测绘。



薛鹏飞（1989– ），男，博士，国防科技大学副教授，主要研究方向为网络空间测绘与对抗和数据挖掘。



李阳（1990– ），男，博士，国防科技大学副教授，主要研究方向为漏洞发现、网络安全和计算机科学。