基于知识图谱的渗透路径推荐系统

司红星　秦少华　孙文凯　颜诗羚
四维创智（北京）科技发展有限公司　北京　100085

摘　要：针对渗透测试中漏洞信息的复杂性与多样性问题，本文提出了一种基于知识图谱构建与链路预测的渗透路径推荐系统。该系统利用自然语言处理（NLP）技术构建网络安全渗透测试知识图谱，并采用Node2Vec算法进行节点嵌入表征。结合A*算法与Adamic-Adar算法，利用节点Embedding向量与预设渗透规则进行链路预测，从而为安全工程师提供高效、精准的渗透路径建议。实验结果在仿真靶场环境中验证了系统在渗透路径推荐准确性及整体性能方面的显著优势。

关键词：渗透测试；知识图谱；链路预测；Node2Vec；A*算法；Adamic-Adar算法；节点嵌入；渗透路径推荐

引言

随着信息技术的快速发展，网络安全形势日益严峻，传统防御措施在应对智能化、自动化的攻击手段时暴露出诸多不足。渗透测试作为一种模拟真实攻击、评估系统安全性的关键技术，正受到广泛关注。知识图谱凭借其对多源异构数据的深度融合与复杂关系的直观表达，在安全威胁情报共享、漏洞关联分析等领域展现出巨大潜力。将知识图谱与图算法结合，能够实现漏洞的自动关联与渗透路径的智能推荐，提升渗透测试的自动化与智能化水平。

一、渗透路径推荐和网络安全知识图谱研究现状

渗透路径推荐：渗透路径推荐是自动化渗透测试的核心环节，通过构建目标网络中漏洞间的关联模型，并采用图算法预测潜在攻击链路。具体实现上，首先构建包含漏洞、攻击工具及安全事件的图结构，依据漏洞共现和攻击链关系建立边权重，并通过节点嵌入获取低维向量，随后利用余弦相似度、欧氏距离等指标进行链路预测[1][2]。

网络安全知识图谱：网络安全知识图谱通过多源安全数据的抽取、融合与结构化，实现漏洞、攻击手法和威胁情报的关联，为渗透测试提供数据支撑，针对复杂攻击场景，利用图数据库与增量更新机制实现新漏洞与攻击信息的动态捕捉[3][4]。

二、系统实现

（一）渗透测试知识图谱的构建

透测试知识图谱构建包括两个过程：（1）漏洞和渗透测试工具信息采集与分析，采用智能化爬虫、信息抽取、文本解析等技术采集多维漏洞和渗透测试工具数据；（2）渗透测试专家知识图谱的构建，利用采集的相关数据进行渗透测试专家知识图谱的构建。

1、多源漏洞数据的捕获与分析

a、采用自动化爬虫技术，从NVD、CNVD等漏洞库及Exploit-DB等漏洞利用代码库抓取非结构化漏洞信息，结合文本解析与数据整理，提取漏洞知识、软件组件信息，为漏洞分析奠定数据基础。b、针对软件库依赖与组件复用问题，通过收集历史资产信息（IP、服务器、组件等）并解析pom.xml等依赖文件，提取软件依赖关系，识别潜在漏洞影响范围；c、全面收集信息收集、漏洞扫描、利用、密码破解、流量嗅探等关键渗透测试工具，并以实体节点形式嵌入知识图谱，形成标准化管理体系，提升渗透测试过程的智能化分析与决策能力。

2、渗透测试专家知识图谱的构建

渗透测试专家知识图谱的实体范式如下图所示，主要包含漏洞本体和渗透动作本体两大本体，不同实体之间通过不同的关联关系建立连接。漏洞本体中包含漏洞基本信息、漏洞利用Exploit、影响到的软件、影响到的软件版本、漏洞补丁等实体结构。渗透动作本体包含渗透工具、漏洞参数、攻击类型、软件脆弱点、PoC等实体结构

实体抽取和关系挖掘：实体抽取和关系挖掘，借助于NLP中的实体识别技术，利用正则规则和机器学习模型实现，针对固有的特定表征的实体预设正则规则，如：

IP：\b((25[0-5]|2[0-4][0-9]|1?[0-9]{1，2})\.){3}(25[0-5]|2[0-4][0-9]|1?[0-9]{1，2})\b

CVE：CVE-\d{4}-\d{4，7}

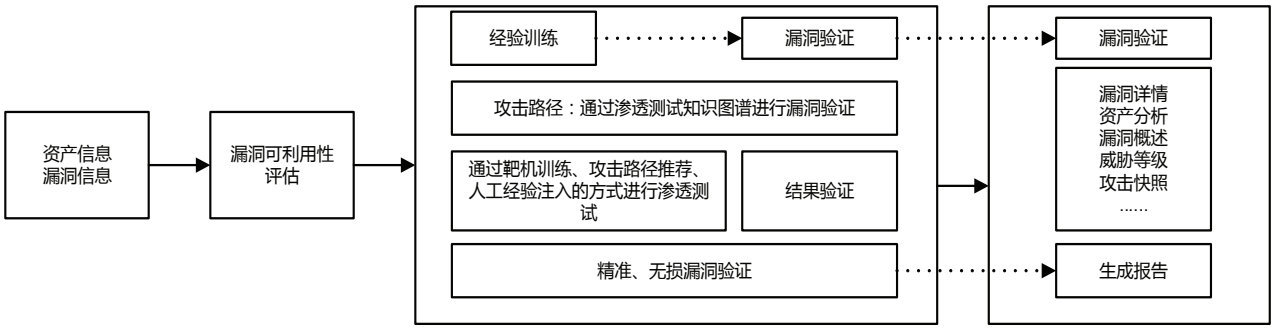

图1　渗透测试系统实现

机器学习：针对文本的嵌入表征，利用Fasttex模型和Bert模型来分别实现，并借助于Bi-LSTM模型和CRF模型来实现实体识别任务和关系识别任务。实现的方案包括：正则规则、正则规则+Fasttext+CRF、Bert+Bi-LSTM+CRF、正则规则+Bert+Bi-LSTM+CRF。

（二）基于知识图谱的自动化渗透测试系统构建

以资产识别和漏洞识别信息为输入，将整体漏洞利用流程分为经验训练、漏洞验证、漏洞告警等三个阶，系统各个阶段介绍和框架如图1所示。

（三）基于Node2Vec和图论的渗透测试路径推荐

1、Node2Vec算法

Node2Vec是基于随机游走（Random Walk）策略构建的图谱实体嵌入表征，其核心在于定义了两种搜索参数：返回参数p和搜索参数q。算法核心如下：

Step1：预设算法参数，如：嵌入的向量维度、滑动窗口上下文等；

Step2：以实体转移概率矩阵为基准对图谱中的实体进行采样，采样过程如下：

a、设定一个游走起点v

b、选择下一个节点x时，根据前一个访问的节点t、调整从v到x的转移概率：如果x=t（返回上一节点），则概率较高（受p控制）；如果x与t直接相连，则概率中等如果x离t较（非直接邻居），则概率较低（受q控制）。

c、这种方式使得游走可以在局部和全局结构之间平衡，其中跳转概率公式为：

$$P(v \to x)=\frac{\pi_{vx}}{Z}$$

$\pi_{vx}=\frac{1}{p}$若x为上一节点，$\pi_{vx}=1$若x当前节点的邻居，$\pi_{vx}=\frac{1}{q}$若x是二阶邻居，Z是归一化因子。

Step3：采用Skip-Gram训练Word2Vec模型，目标函数是：

$$max_{\varphi}\sum_{v\in V}\sum_{u\in N(v)}logP(u\mid\varphi(v))$$

$\varphi(v)$是节点v的嵌入向量，$P(u\mid\varphi(v))$由Skip-Gram模型计算。

Step4：将每个节点的Embedding向量以属性的方式添加到每个实体的属性中，属性名称：node_embedding。

2、基于图论的路径发现算法

A*算法：A*（A-star）算法是一种启发式搜索算法，用于在渗透测试图谱中找到渗透测试的寻找最优路径。其核心思想是通过评估函数$f(n)$来选择下一个要渗透测试的节点，其中

$$f(n)=g(n)+h(n)$$

$g(n)$：起点到当前节点n的实际代价；$h(n)$：从当前节点n到目标节点的启发式估计代价；$f(n)$：综合评估路径的总代价，A*会优先选择$f(n)$最小的节点进行扩展。

Adamic Adar算法：其核心思想是，两个渗透测试节点如果有共同邻居，并且这些共同邻居相对稀少，那么这两个节点之间的连接可能性较大，算法公式如下：

$$sim(u,v)=\sum_{w\in N(u)\cap N(v)}\frac{1}{\log(|N(w)|)}$$

其中u和v是渗透测试图谱中的两个节点、$N(u)$和$N(v)$是节点u和节点v的邻居集合。

3、渗透测试路径推荐实现

通过预设的渗透测试规则和NLP以及路径发现算法相结合的方式实现：

a、预设的渗透测试规则指定了渗透测试过程路径选择的有限性对渗透测试路径的选择进行收束：使用Nmap、Masscan等工具，优先探测高风险端口（如22、80、443、3389）。

b、根据NLP方法进行路径推荐：基于Node2Vec的节点嵌入表征，并利用余弦相似度算法对下一个渗透测

试节点推荐，优先选择余弦相似度高的渗透测试路径：

$$consin(p_s,p_t)=\frac{\sum_i p_s(\mathrm{i})p_t(i)}{\sqrt{\sum_i p_s(i)^2}\cdot\sqrt{\sum_i p_s(i)^2}}$$

其中p_s表示的是节点s的node_embdding属性，即节点的嵌入表征，$p_s(i)$是向量的第i个值。

c、结合图论中的A*算法和Adamic Adar算法实现渗透路径的推荐

在渗透测试路径选择过程中利用预设的渗透测试规则进行收束，并结合前述NLP和图论算法对下一节点的进行细粒度的选择，具体的：

Step1：利用预设规则对渗透路径下一节点的选择进行收束。

Step2：优化下一节点的选择算法，具体公式如下：

$$Next(n)=\max(k*f(n)+j*sim(n,v)+(1-\mathrm{j})*consin(p_n,p_v))$$

其中$Next(n)$是选择下一节点的概率，j、k是权重系数，可以根据实际情况进行设置，或者使用网格搜索的方式选择适合当前环境效果最好的参数，选择最大的$Next(n)$进行推荐。

三、实验与测试

本系统涉及到知识图谱的构建和渗透测试路径推荐两个内容。

测试数据集设计：a、在知识图谱的构建过程中，通过人工打标签的方式构造实体识别模型的测试数据集；b、通过在虚拟环境中构建测试靶场（包含多个靶场 如DVWA、upload-labs、xss-labs、Vulhub等 覆 盖主流的CVE和CNNVD漏洞）进行渗透测试路径推荐的测试。

系统的评价指标：知识图谱：采用准确率、召回率和F1得分的评价指标，测试样本数量在10万+。针对渗透测试过程中的核心阶段，采用准确率的评价指标；针对整个测试全流程采用有效率的评价指标。

实验结果与分析：如表1、表2所示

表1　知识图谱构建中的实体识别方法效果对比表

实体识别方法	准确率（%）	召回率（%）	F1分数（%）
正则规则+Fasttext+CRF	85.2	80.5	82.8
Bert+Bi-LSTM+CRF	89.7	86.3	88.0
正则规则+Bert+Bi-LSTM+CRF	92.4	89.8	91.1

通过表1可以得出：正则规则+Bert+Bi-LSTM+CRF方法表现最佳，在准确率（92.4%）、召回率（89.8%）和F1分数（91.1%）方面均优于其他方法。

表2　渗透测试路径推荐效果对比表

渗透测试路径算法	信息收集推荐准确率（%）	漏洞扫描与利用准确率（%）	漏洞攻击准确率（%）	渗透测试流程有效率（%）
基于预设规则	78.5	75.2	74.3	72.5
基于预设规则+NLP	84.3	81.7	79.5	77.5
基于预设规则+NLP+图谱的路径推荐算法（A*算法和Adamic Adar算法）径推荐算法	91.2	88.9	89.5	86.3

通过表2得出：基于预设规则+NLP+图谱的路径推荐算法（A*算法和Adamic Adar算法）表现最佳，在信息收集推荐准确率（91.2%）、漏洞扫描与利用准确率（88.9%）、漏洞攻击准确率（89.5%）以及渗透测试流程有效率（86.3%）方面均领先。

结束语

本文提出了一种基于知识图谱与图算法的渗透路径推荐系统，系统通过以下几个方面的工作实现了对渗透路径的自动化推荐：（1）知识图谱的构建：构建图谱中的实体识别模型在准确率、召回率和F1分数上都取得了不错的效果；（2）自动化渗透测试系统的构建：系统结合了渗透测试工具和渗透测试图谱，实现了渗透测试的自动化；（3）基于图论中的相似度算法实现渗透路径的推荐：通过推渗透测试图谱中的节点进行Node2Vec并结合预设的链路预测和A*算法以及Adamic Adar算法，系统能够预测潜在攻击路径，在多个渗透测试过程和全流程渗透测试过程中取得了很好的效果。

参考文献

[1]黄为.计算机网络渗透测试技术探究[J].网络安全技术与应用，2021，(12)：8-9.

[2]周仕承，刘京菊，钟晓峰，等.基于深度强化学习的智能化渗透测试路径发现[J].计算机科学，2021，48(07)：40-46.

[3]孙澄，胡浩，杨英杰，等.基于网络防御知识图谱的0day攻击路径预测方法[J].网络与信息安全学报，2022，8(01)：151-166.

[4]王一凡，孙治，和达，等.基于知识图谱的网络系统安全风险评估方法[J].计算机应用与软件，2023，40(11)：312-320.