

基于知识图谱的网络安全事件数据推荐算法

祝现威, 刘伟, 刘自豪, 顾泽宇

(中国人民解放军 61660 部队, 北京 100080)

摘要: 针对网络安全运维人员分析网络安全事件时难以及时准确找出所需的数据问题, 提出基于知识图谱的网络安全事件数据推荐算法, 利用网络威胁框架 ATT&CK 构造本体模型, 根据本体模型建立网络威胁知识图谱, 将攻击技术、漏洞、防御措施等离散的安全数据提炼为互相关联的安全知识。基于知识图谱提取实体数据, 通过 TransH 算法进行实体向量化, 并利用实体向量计算网络威胁数据实体间数据相似性。获取网络安全事件处置文献中网络安全数据实体作为运维人员处置行为, 构造处置行为矩阵, 通过行为矩阵实现网络威胁数据向量化表示, 计算基于处置行为的网络威胁数据实体相似性。将网络威胁数据实体的相似度与基于处置行为的网络威胁数据实体的相似度进行融合, 形成面向网络安全事件的数据推荐列表, 实现基于用户行为的网络威胁领域间的关联。实验分析表明, 在融合权重 $\alpha=7$ 和推荐数据量 $K=5$ 时, 所提算法最优, 召回率和精确率分别为 62.37% 和 68.23%。所提算法在数据相似度的基础上加入了处置行为相似度, 更接近事实处置行为, 与其他算法相比, 所提算法的召回率具有较大优势, 精确率在推荐数据量小于 10 的范围内具有较大优势。

关键词: 网络威胁数据; 网络安全事件; 知识图谱; 相似度; 事件处置行为; 数据推荐

中图分类号: TP393

文献标志码: A

DOI: 10.11959/j.issn.2096-109x.2023087

Data recommendation algorithm of network security event based on knowledge graph

ZHU Xianwei, LIU Wei, LIU Zihao, GU Zeyu

The 61660 Unit of PLA, Beijing 100080, China

Abstract: To address the difficulty faced by network security operation and maintenance personnel in timely and accurate identification of required data during network security event analysis, a recommendation algorithm based on a knowledge graph for network security events was proposed. The algorithm utilized the network threat framework ATT&CK to construct an ontology model and establish a network threat knowledge graph based on this model. It extracted relevant security data such as attack techniques, vulnerabilities, and defense measures into interconnected security knowledge within the knowledge graph. Entity data was extracted based on the knowledge graph, and entity vectors were obtained using the TransH algorithm. These entity vectors were then used to calculate

收稿日期: 2023-08-25; 修回日期: 2023-11-10

通信作者: 祝现威, zhu13939055330@163.com

引用格式: 祝现威, 刘伟, 刘自豪, 等. 基于知识图谱的网络安全事件数据推荐算法[J]. 网络与信息安全学报, 2023, 9(6): 116-126.

Citation Format: ZHU X W, LIU W, LIU Z H, et al. Data recommendation algorithm of network security event based on knowledge graph[J]. Chinese Journal of Network and Information Security, 2023, 9(6): 116-126.

data similarity between entities in network threat data. Disposal behaviors were extracted from literature on network security event handling and treated as network security data entities. A disposal behavior matrix was constructed, and the behavior matrix enabled the vector representation of network threat data. The similarity of network threat data entities was calculated based on disposal behaviors. Finally, the similarity between network threat data and threat data under network security event handling behavior was fused to generate a data recommendation list for network security events, which established correlations between network threat domains based on user behavior. Experimental results demonstrate that the algorithm performs optimally when the fusion weight $\alpha=7$ and the recommended data volume $K=5$, achieving a recall rate of 62.37% and an accuracy rate of 68.23%. By incorporating disposition behavior similarity in addition to data similarity, the algorithm better represents factual disposition behavior. Compared to other algorithms, this algorithm exhibits significant advantages in recall rate and accuracy, particularly when the recommended data volume is less than 10.

Keywords: network threat data, network security events, knowledge graph, similarity, event handling behavior, data recommendation

0 引言

随着网络攻击技术革新，传统的单步攻击方式逐步被多步复合攻击所取代，相应的安全防护和监测预警工具类型日趋复杂、数量日趋增多，造成网络安全事件相关判断数据的种类和数据量急剧上升，网络安全运维人员需要从多个维度对网络安全事件进行分析判断。因此，根据不同网络安全事件的特征和关注重点，将网络安全事件所需要的网络安全数据及时准确地向运维人员推荐成为重要问题。随着知识图谱的兴起，基于知识图谱的推荐方法越来越受到重视。知识图谱的概念是 Google 于 2012 年正式提出的，目的是解决信息关联查询问题，为数据联想提供可解释性。MITRE 公司分别构建了 ATT&CK 框架、CAPEC (common attack pattern enumeration and classification) 攻击模式、CWE (common weakness enumeration) 等知识库用于知识图谱关联。但是上述关联框架主要针对单一网络攻击进行建模，未对网络事件关联性进行分析，难以适应复杂的多步网络攻击和多维度网络安全事件处置。

为了解决运维人员在处置网络威胁过程中难以根据过往历史记录准确找到所需威胁数据的问题，本文提出基于知识图谱的网络安全事件数据推荐算法，利用 ATT&CK 构建本体抽取网络威胁数据，根据余弦相似度计算数据间的相似性，同时结合网络安全事件处置行为下的主题关联性构

建推荐数据列表，主要工作如下。

1) 提出基于 ATT&CK (adversarial tactic, technique, and common knowledge) 的网络威胁知识图谱，以 ATT&CK 的 TTP (tactic、technology、procedure) 威胁描述模式作为本体描述模型，在此基础上可进行实体扩展，该建模方式增加了本体模型的可扩展性。在提取的实体基础上，利用 TransH 算法实现网络威胁数据的向量化表示，并进行相似度计算，从数据角度给出数据推荐依据。

2) 提取网络安全事件记录和相关文献中实体，并将其作为运维人员处置行为，以此构建处置矩阵，提出一种基于处置行为的网络威胁数据实体向量化表示方法，从处置行为角度给出网络威胁数据实体的相似度计算，防止产生推荐数据在同一主题下聚集问题。

3) 提出基于处置行为和网络威胁 (DBNT, disposition behavior and network threat) 的数据推荐算法，该算法利用融合系数将网络威胁数据实体的相似度与基于处置行为的网络威胁数据实体的相似度进行融合，从而计算出不同主题下的数据相似度，在实现不同主题间数据关联前提下，提升推荐数据的广度。

1 相关工作

网络安全事件关联分析主要为了应对网络攻击，而将多种互补设备产生的信息进行收集融合，描述网络攻击行为，主要是通过事件属性进行相似性聚合，在聚合过程中通过算法去除噪声实现的。

例如,文献[1]通过告警验证、告警融合以及攻击场景构建3个步骤进行网络安全事件分析;文献[2]提出一种基于紧凑规划的网络攻击发现算法,利用攻击图对网络安全事件进行关联分析。上述方法受制于计算复杂度,网络规模和数据维度增加将大幅影响计算复杂度,且未考虑实体的状态信息。

知识图谱相较传统网络安全事件关联分析具有更好的可解释性和关联效率,因此基于知识图谱的网络安全事件关联分析被广泛应用,其主要涉及知识抽取、关系抽取、数据推荐。在知识抽取方面,文献[3]提出一种网络信息与网络态势感知统一的网络安全本体建模方法;文献[4]根据网络威胁 ATT&CK 中的攻击技术进行威胁实体抽取,并利用攻击技术关联性进行关系抽取,上述方法只简单利用威胁框架中的静态元素,扩展性较弱,本文在此基础上引入 TTP 描述模式作为本体构建范式,提升了本体建模的可扩展性。在关系抽取方面,文献[5]提出网络安全知识威胁框架,根据框架实现网络实体和网络关系抽取;文献[6]利用假设构建 Oday 攻击的位置属性,并将此作为实体间的关系,将攻击预测问题转化为知识图谱链接预测问题,以上通过人为规定或假设的方式作为实体关系抽取的补充具有一定的主观性,因此本文利用 TextRank 算法进行主题词抽取并进行训练,降低关系抽取的主观性。在数据推荐方面,文献[7]提出 RippleNet 关联分析模型,模型利用知识图谱中实体节点的属性构建特征关联度,实现实体属性的更新,由此进行关联数据推荐;文献[8]利用注意力机制计算各业务领域权重,从而获取用户、物品的表征。上述网络安全知识图谱只关注了威胁间的关联性,忽略了数据在应用过程中与安全事件场景和处置行为的关联性;文献[9]利用矩阵分解和贝叶斯算法实现基于用户习惯的数据推荐算法 BPR (Bayesian personalized ranking),文献[10]在 BPR 的基础上利用 TransR 算法实现 KG 增强推荐算法框架 CKE (collaborative knowledge base embedding)。文献[11]通过引入元图的概念来表示高级语义,利用“矩阵分解+分解机”的算法 FMG (factorization machine with group lasso) 来求解信息融合问题。

2 网络安全事件数据推荐算法

知识图谱作为一种符号化的知识特征表达方式,能够精准表示实体及特征之间的相关性^[5]。本文将网络威胁框架 ATT&CK 中的 TTP 作为本体构建范式进行本体建模,以此构建知识图谱,并利用 TransH 算法对实体数据进行向量化计算。同理,从网络安全事件处置文献中获取处置行为,并对其进行向量化,计算基于处置行为的网络威胁数据实体的相似度。将网络威胁数据实体的相似度和基于处置行为的网络威胁数据实体的相似度以不同权重进行融合,并利用重排序算法进行相似度排序形成数据推荐列表,网络安全事件数据推荐算法过程如图1所示。

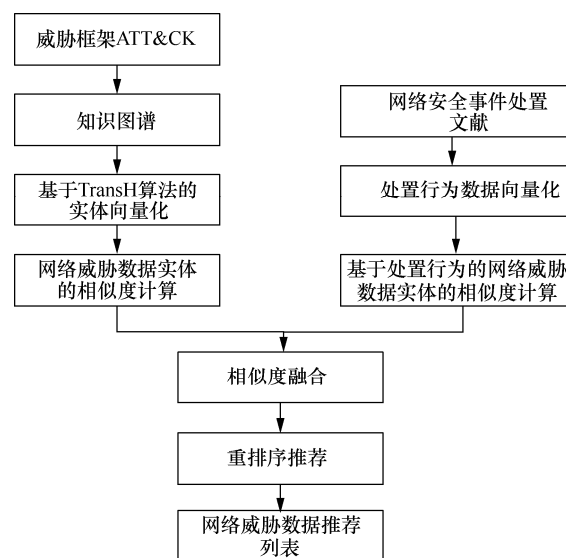


图1 网络安全事件数据推荐算法过程
Figure 1 Process of recommendation algorithm of network security event

2.1 架构构建

本文算法首先构建网络安全本体模型,然后利用多种知识数据抽取技术,从结构化和非结构化数据源中提取并建立知识数据,构建数据层,最后分别从数据相似度和处置行为相似度两方面进行相似度融合,算法构建架构如图2所示。

模式层以 ATT&CK 的 TTP 作为网络安全本体构建范式,为数据层知识建模提供模式定义。模式层定义了 Tactic、Technology、Procedure 这3个描述类及关系类型 exploit,数据层则在此基础上进行扩展,明确需要的数据实体并构建本体模型。

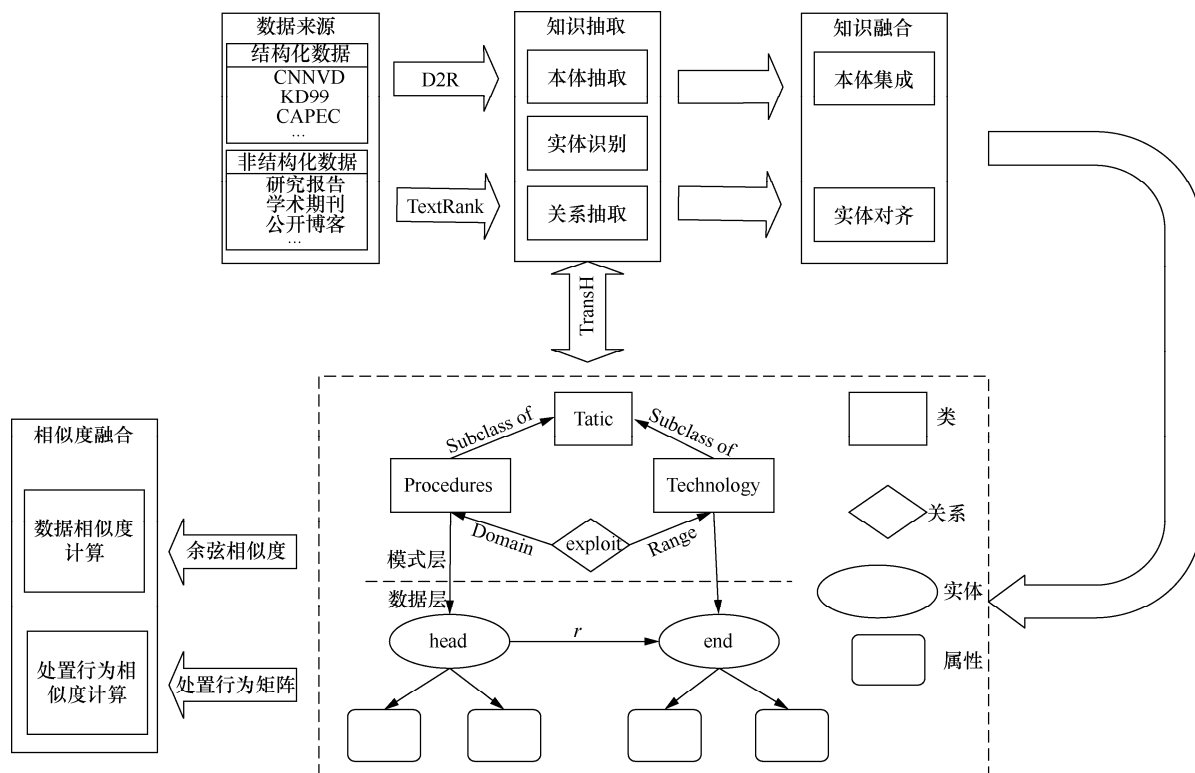


图2 算法构建架构

Figure 2 Algorithm construction architecture

完成本体模型后，数据层从结构化和非结构化的数据源中进行抽取、融合实现数据构建，针对结构化数据和非结构化数据，首先分别利用 D2R (relational database to RDF) 工具和 TextRank 算法实现知识抽取；随后一方面利用 TransH 算法进行实体向量迭代求解，将知识图谱中知识进行向量化表示，另一方面利用处置文献中实体出现频率构建处置行为矩阵，通过处置矩阵进行实体向量化表示；最后根据数据向量和处置行为向量给出实体间融合相似度。

2.2 知识图谱构建

知识图谱构建方式主要分为自顶向下和自底向上两种模式^[12]。自顶向下是预先制定本体模型与数据类型，根据预先定义的知识图谱属性进行构建，该方法适用于结构化知识；自底向上是利用知识抽取技术，从数据中抽取实体、关系和属性等信息，通过置信度筛选出合格数据信息，根据数据信息构建本体。

由于网络威胁知识具有垂直性，本文选择自顶向下构建知识图谱本体。本文利用 ATT&CK 网

络威胁框架构建网络威胁实体及实体间关系，使用 TTP 三元组作为本体构建范式，在此基础上引入 8 类实体，其中 Tactic 对应的实体为 Attack-tactic，Technology 对应的实体为 Port、Vulnerability、Course-of-action、Attack-technology，Procedures 对应的实体为 Software、Malware，由于网络威胁数据主要通过威胁监测获取，故在 ATT&CK 网络威胁框架的基础上引入实体 Indicator 表示监测设备，具体实体类别如表 1 所示。

知识图谱^[4]利用知识化的特征表达方式，以图的形式表达实体与特征间的关系，网络安全本体模型如图 3 所示。方框表示概念，连线表示概念间的关系。由于 TTP 模型^[4]是一个过程描述模型，所以在确定实体后，根据过程描述确定实体关系，如从层次性方面，恶意代码和攻击技术共同组成攻击战术；从过程性方面，恶意代码利用攻击技术，通过端口进行攻击实施。缓解措施主要针对恶意代码、软件和漏洞进行修复消除隐患。攻击战术、恶意代码和漏洞被监测设备获取信息，因此三者是监测设备的子类。

表 1 实体类别
Table 1 Entity class

名称	含义	说明
Software	软件	作为工具的可运行代码程序
Malware	恶意代码	主要包括僵尸主机、木马及蠕虫病毒等
Indicator	监测设备	主要指用于监测网络威胁设备, 如流量分析、IPS/IDS、防火墙等
Port	端口	恶意代码或漏洞攻击利用的媒介
Vulnerability	漏洞	设备系统自身原因的脆弱性问题, 主要分为不可远程利用漏洞和可远程利用漏洞
Course-of-action	防御措施	可以阻止攻击或防止危害进一步扩散的操作和手段
Attack-tactic	攻击战术	ATT&CK 中攻击者为达到目的的一类手段
Attack-technology	攻击技术	ATT&CK 中实现战术目的的一类技术

2.3 网络威胁数据相似度计算

为获取网络威胁数据的相似度, 网络威胁知识图谱对网络威胁实体表示进行训练, 根据训练结果实现实体关系的向量化表示。为了便于表示网络威胁知识图谱实体间存在一对多、多对一和多对多的映射关系, 利用基于嵌入式的 TransH 算法进行实体关系表示, 利用该算法对实体进行向量化计算, 计算过程如下。构建本体三元组 $S = \{h, r, d\}$, h 为头实体, r 为实体关系, d 为尾实体, 其中 $h, t \in E, r \in R, E$ 为实体集合, R 为关系集合。同时负三元组集合 $S' = \{(h', r, d) \cup$

$(h, r, d')\}$, 其中 $h', t' \in E$ 。TransH 在 TransE 的基础上引入超平面^[13], 将关系定义在超平面 w_r 上, 并进行超平面投影。

定义 1 超平面。法向相量 w_r 为超平面的单位向量, 头实体和尾实体在超平面上的投影分别为 $h_{\perp}$ 和 $d_{\perp}$, 在超平面上按实体关系向量 r 平移得到, $h_{\perp} + r \approx d_{\perp}$, 投影公式如下。

$$h_{\perp} = h - w_r^T h w_r \quad (1)$$

$$d_{\perp} = d - w_r^T d w_r \quad (2)$$

定义 2 评分函数。三元组越接近事实时, 评分越低, 反之三元组与事实相差越大, 则距离越远、得分越高, 具体计算如下。

$$f(h, r, d) = \|h_{\perp} + r - d_{\perp}\|_2^2 = \|(h - w_r^T h w_r) + r - (d - w_r^T d w_r)\|_2^2 \quad (3)$$

模型通过扩大样本间距离进行训练, 利用最小化代价函数求解正负样本间的最大化距离。为了区分正三元组和负三元组, 将 margin-base ranking function 作为损失函数来训练模型^[13], 因此目标函数如下。

$$L = \sum_{(h, r, d) \in S} \sum_{(h', r', d') \in S'} [f_r(h, d) + \gamma - f_{r'}(h', d')]_+ \quad (4)$$

其中, $[x]_+ = \max[0, x]$, 表示求正的最大值, 其中 γ 为大于 0 的先验参数, 表示正负三元组之间的距离。在求最小化损失时, 应满足以下约束条件。

$$\forall e \in E, \|e\|_2 \leq 1 \quad (5)$$

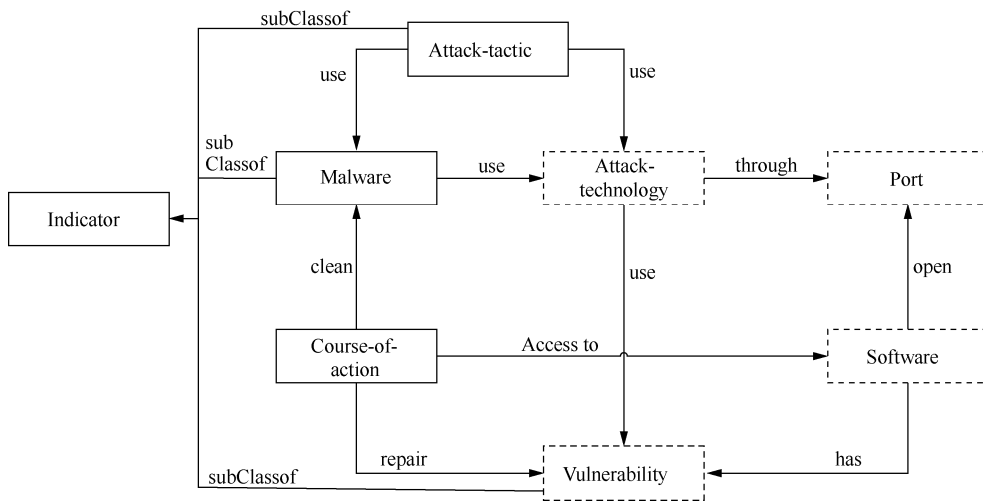


图 3 网络安全本体模型
Figure 3 Network security ontology model

$$\forall r \in R, \frac{|\mathbf{w}_r^T|}{\|\mathbf{r}\|_2} \leq \varepsilon \quad (6)$$

$$\forall r \in R, \|\mathbf{w}_r\|_2 = 1 \quad (7)$$

式(5)保证所有实体向量的归一化, 式(6)保证关系向量 $\mathbf{r}$ 在超平面上, 并采用随机梯度下降法实现最小化损失函数; 式(7)保证法向量的模为 1。该算法将网络威胁知识图谱中三元组作为正三元组进行随机遍历, 在进行遍历的同时构建对应的负三元组, 通过不断遍历迭代计算梯度进行最优化解。

随后利用 TransH 算法对知识图谱的实体进行训练获取实体向量。由于 ATT&CK 是一个分层模型, 不同的层次对应不同实体, 具有相同层次的数据实体紧凑而不同层次间数据实体稀疏的特性, 因此数据间距离并不是均匀分布的, 需要计算不同层次下实体数据相似度。本文将威胁本体定义的 8 类实体 13 类关系(每类关系表示一个网络安全动作, 作为网络安全事件主题), 主题向量 $\mathbf{V} = \{V_1, V_2, V_3, \dots, V_{13}\}$, 利用余弦函数计算不同主题下数据实体 $\mathbf{v} = \{v_1, v_2, v_3, \dots, v_j\}$ 的相似度。余弦相似度计算式如下。

$$\cos(V_i, v_j) = \frac{\sum_{n=1}^n (\mathbf{v}_{in} \times \mathbf{v}_{jn})}{\left(\sqrt{\sum_{n=1}^n (\mathbf{v}_{in})^2} \right) + \left(\sqrt{\sum_{n=1}^n (\mathbf{v}_{jn})^2} \right)} \quad (8)$$

其中, $\mathbf{v}_{in}$ 表示主题向量 $\mathbf{V}_i$ 的第 n 维分向量, 同理 $\mathbf{v}_{jn}$ 表示数据实体 $\mathbf{v}_j$ 的第 n 维分向量。

定义 3 $D = \{d_1, d_2, d_3, \dots, d_n\}$ 为 13 个主题下推荐的数据集合, d_n 代表主题下推荐的第 n 个数据。在此主题中计算各数据余弦相似度, 根据相似度构建主题矩阵 $\mathbf{sim}_{d_r}(d_j, V_i)$ 。

2.4 基于处置行为的相似度计算

ATT&CK 分层特性造成网络安全数据不同主题间的数据相似度较低, 但在实际网络威胁处置的过程中, 用户需要分析不同主题下的数据。因此在知识图谱的基础上获取网络威胁数据实体的相似度之后, 需要对网络安全事件处置文献进行知识提取和分析, 构建处置行为矩阵, 挖掘处置行为与网络威胁知识间的关联性。

分析网络安全事件需要对网络安全事件处置资料进行知识抽取, 获得事件处置过程中关注的主题、摘要及相关安全知识。在抽取过程中使用 jieba 分词工具对文献题目进行分词处理和词性标注, 筛选网络安全事件处置文献中包含的动词和名词, 抽取的三元组为(名词, 动词, 名词)。为了防止在题目分词过程将部分网络安全专业词汇分割, 根据需要自定义网络安全字典^[14], 包含 4 564 个网络安全词组, 提高分词准确率, 如文献题目“基于 P2P 的僵尸网络检测及防御”在进行分词和词性标注过程中加入网络安全字典前后效果对比如图 4 所示。

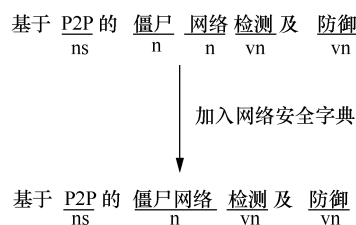


图4 加入网络安全词典前后效果对比
Figure 4 Effect of adding network security dictionary

在加入网络安全字典前,“僵尸网络”被分割为僵尸和网络两个词汇, 加入字典后准确提取词性, 在提取过程中, 若题目中只有两个名词和一个动词则正常构建三元组, 但实际题目中包含多个名词和动词甚至有一些冗余信息, 如果直接将名词作为实体, 动词作为关系将会产生多个三元组交叉。针对这一问题, 使用 TextRank 算法提取文献中的关键词, 并对关键词进行优化, 提升提取三元组的准确率。

TextRank 算法是一种基于图的排序算法^[15], 其中图 $G = (V, E)$, V 为节点集合, E 为边集合, 根据文本中心内容提取关键字, 计算如下。

$$S(v_i) = (1 - d) + d \cdot \sum_{v_j \in \ln(v_i)} \frac{\omega_{ji}}{\sum_{v_k \in \text{outout}(v_i)} \omega_{jk}} S(v_j) \quad (9)$$

其中, $S(v_i)$ 表示节点 v_i 的关键度, d 为阻尼系数, 表示图中一点指向其他点的概率, 一般设为 0.85, ω_{ji} 表示节点 v_i 和节点 v_j 间边的权重, 用余弦相似度表示, 其他点指向给定点 v_i 的集合为 $\ln(v_i)$, $\text{out}(v_j)$ 表示节点 v_j 指向其他点的集合, 每个节点

初始化关键度一致。使用 jieba 分词工具对标题进行划分和词性标注,利用 TextRank 算法实现文献关键字排序,提取关键词数量为 N 。若文献中有两个以上动词,则在关系提取的过程中选取重要性最高的动词表示实体间关系。若文献中有 3 个以上名词时,利用 TextRank 算法在题目中选取与摘要相关度最高的前两个词作为头实体和尾实体,以文献“基于 P2P 的僵尸网络检测及防御”为例,利用关键词提取三元组如图 5 所示。一篇网络安全事件文献往往涉及威胁发现、威胁影响范围、威胁处置等多个步骤,因此分别提取主题、摘要和副标题的三元组作为处置行为数据。

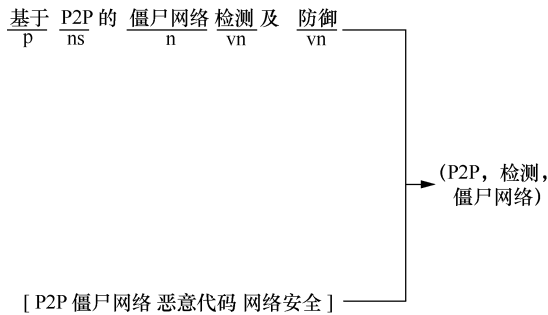


图 5 利用关键词提取三元组
Figure 5 Extracting triples by using keywords

定义 4 利用 TransH 算法将生成文献中的三元组进行向量化,利用式(8)计算文献中的三元组与主题向量相似度,取相似度最高的主题作为处置行为数据 $U = \{u_1, u_2, \dots, u_m\}$,处置行为数据与主题构成处置行为矩阵,示例如表 2 所示。

表 2 处置行为矩阵
Table 2 Disposition behavior matrix

主题	数据				
	u_1	u_2	u_3	...	u_m
V_1	1	0	0	...	1
V_2	0	1	1	...	1
V_3	0	0	0	...	0
...	...	...	...	...	...
V_{13}	1	0	0	...	0

由表 2 中处置行为矩阵数据构造网络安全主题基于处置行为的向量,得到每个主题所对应的 m 维向量,同样利用余弦相似度计算各个主题间的相似度,根据处置行为确定在处置过程中主题

间的关联程度,相似度越高,则说明在处置过程中这几类数据关联程度越高,即当运维人员查看该数据时,往往需要结合相似度较高的主题数据进行关联分析。将基于处置行为的主题相似度矩阵定义为 $\text{sim}_u(V_i, V_j)$, 其中 $i, j \leq 13$ 。

2.5 相似度融合

为实现基于知识图谱威胁数据和处置行为中威胁数据二者的关联融合,由上述计算得到的基于网络威胁知识图谱的相似度矩阵 $\text{sim}_d(d_i, d_j)$ 和基于处置行为的相似度矩阵 $\text{sim}_u(V_i, V_j)$ 以不同权重进行融合。利用基于网络威胁知识图谱的相似度矩阵判断实体数据归属主题,确定归属主题后计算主题间相似度,根据权重 α 将网络威胁数据实体的相似度与基于处置行为的网络威胁数据实体的相似度进行相似度融合,获得全局相似度 $\text{sim}(d_i, d_j)$, 融合方式如下。

$$\text{sim}(d_i, d_j) = \alpha \max(\text{sim}_{d_r}(d_i, V_i)) + (1 - 2\alpha) \text{sim}_u(V_i, V_j) + \alpha \max(\text{sim}_{d_r}(d_j, V_j)) \quad (10)$$

通过式(10)得到实体数据间的相似矩阵,利用重排序算法在实体相似度矩阵取出每行的最小值放置在向量数组中,把向量数组中的数据按照从大到小的顺序排列输出,选取前 n 个数据输出为数据推荐列表。

3 实验结果及分析

3.1 实验评估指标

本文利用 KDD99 和 CNNVD 数据集构建网络威胁知识图谱,并以数据集中的威胁标注数据作为威胁事件,网络安全事件文献信息从 AilenVault 威胁情报博客、CVE 安全事件、threat post 社区安全新闻、思科安全威胁类博客、以网络威胁和网络安全应急响应为主题的知网文献中,共获取文献 3 246 篇,随机选取 100 篇作为题目、摘要和内容标题提取三元组,用于训练主题向量间相似度。

通过准确率 (AC) 评估三元组提取算法的效果,将实验文献中正确提取的三元组个数记为 TP,将实验中提取错误或未成功的三元组个数记为 FP,准确率计算如下。

$$\text{AC} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (11)$$

同时在 KDD99 中随机挑选 80 个安全告警作为安全事件，提取 KDD99 和 CNNVD 中安全事件相关数据作为数据推荐集，将推荐集 80% 的相关威胁数据作为训练集，20% 作为测试数据集。推荐准确程度用精确率（Precision）和召回率（Recall）进行衡量。将给运维人员 u 推荐的数据集合记为 $K(u)$ ，测试集上满足安全事件关联分析的数据集合记为 $F(u)$ ，精确率指满足安全事件关联分析要求的数据在推荐数据集合中的占比，召回率指满足安全事件关联分析要求的数据中推荐数据占总数据的百分比，计算分别如下。

$$\text{Precision} = \frac{\sum_{u \in U} |K(u) \cap F(u)|}{\sum_{u \in U} |K(u)|} \quad (12)$$

$$\text{Recall} = \frac{\sum_{u \in U} |K(u) \cap F(u)|}{\sum_{u \in U} |F(u)|} \quad (13)$$

综合评价指标 eve 是精确率和召回率的调和平均值，公式如下。

$$eve = \frac{2 \cdot \text{Recall} \cdot \text{Precision}}{\text{Recall} + \text{Precision}} \quad (14)$$

3.2 实验结果

3.2.1 知识图谱构建

根据构建的本体模型对网络威胁数据进行抽取，得到概念化实例实体，在 Neo4j 中创建实体

并通过关系进行连接，知识图谱部分节点可视化结果如图 6 所示。

3.2.2 基于文本的处置行为实验分析

本文实体三元组的提取直接影响网络安全事件处置行为的向量化表示，从而影响网络威胁数据主题间的关联性判断。在判断过程中，文献的主题、摘要或标题中包含多个动词和名词时，关键词的提取数量 N 直接影响三元组构造过程中实体关系的选取。因此需要选取一个合适的 N 值，以提高处置行为判断准确率。设置 $N \in [0, 20]$ 间隔为 2，实体准确率与提取数量 N 的关系如图 7 所示。

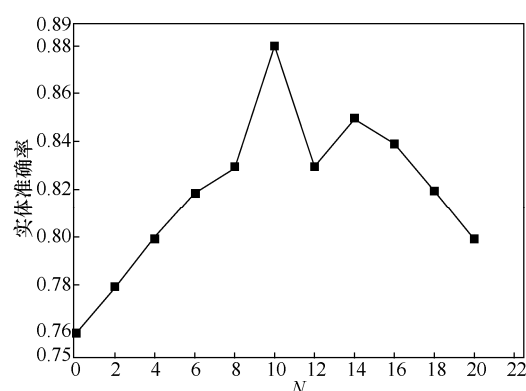


图7 实体准确率与提取数量 N 的关系
Figure 7 Relationship between entity accuracy and extraction quantity N

根据图 7 可知，当 $N=0$ 时，即未加入文献内容对三元组提取进行优化时，获取三元组的准

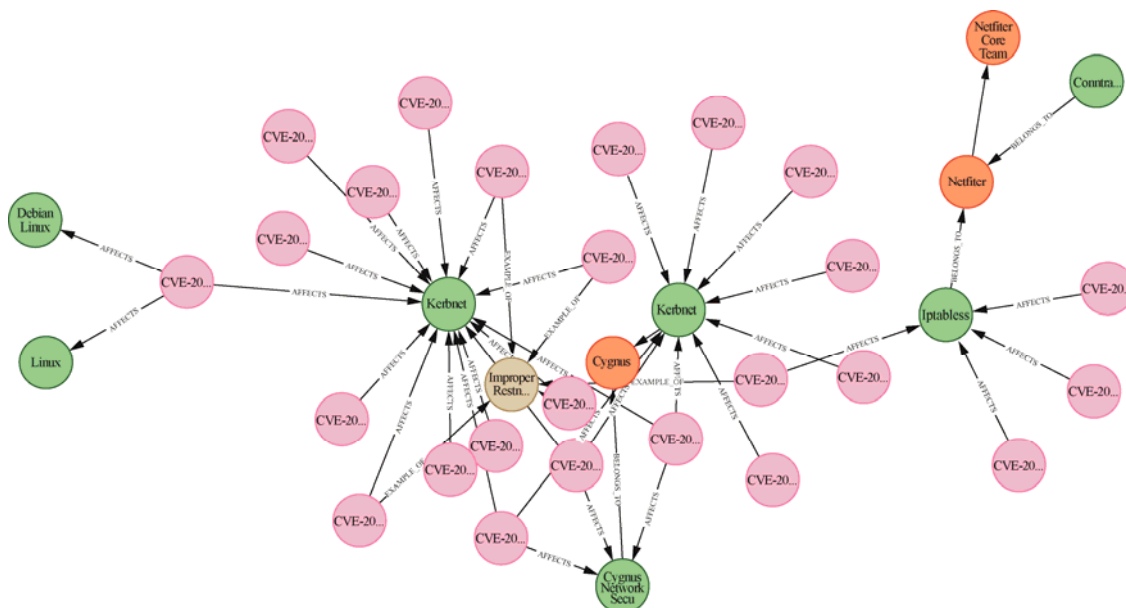


图6 知识图谱部分节点可视化
Figure 6 Visualization of knowledge graph section nodes

准确率最低。三元组提取的准确率随着关键词提取数量的增加而提高,当 $N=10$ 时,数量最大,但是三元组准确率不是一直随着关键词提取数量的增加而提高的,当 $N>10$ 时,三元组准确率出现下降,当提取词过多时会出现词性分散,出现多个主题反而难以提取三元组。因此,本文将关键词的提取数量设为10,在此条件下提取三元组作为网络安全事件处置行为。

3.2.3 融合权重选择

融合权重 α 的选择对推荐算法的准确率起到重要作用,不同权重会影响数据融合矩阵的主题数据关联重要度,从而影响不同主题数据间的关联程度。本文在不同融合权重 α 下,通过对比网络安全事件数据推荐算法效果评价指标召回率以及准确率,找寻融合权重 α 与推荐数据数量 K 的关系。选取推荐列表数据数量范围为[4,14],间隔为2,不同融合权重 α 下,数据推荐算法的召回率和精确率如图8和图9所示。

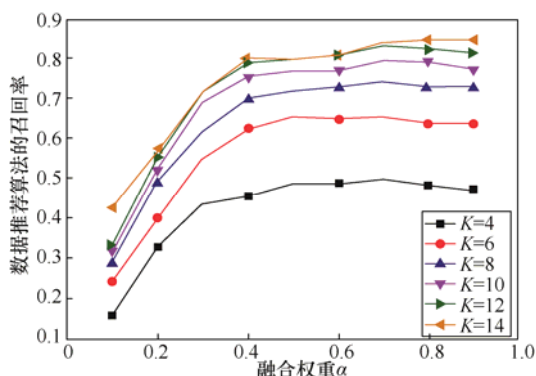


图8 不同融合权重 α 下,数据推荐算法的召回率
Figure 8 Recall of data recommended algorithm under different fusion weights α

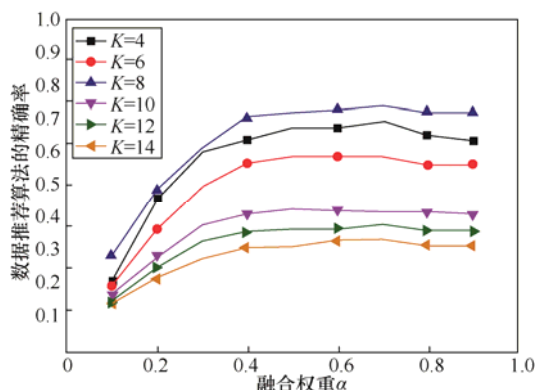


图9 不同融合权重 α 下,数据推荐算法的精确率
Figure 9 Precision of data recommended algorithm under different fusion weights α

从图8可知,数据推荐算法的召回率与融合权重 α 呈正相关,即随着基于知识图谱的威胁数据相似度所占的比重变大,数据推荐算法的召回率随之增加,且 K 越大召回率越高。类似地,在图9中,数据推荐算法的精确率随着融合权重 α 增加而提高,当 $\alpha=0.7$ 时精确率最高,之后随着 α 增加准确率出现下降的趋势,因此将 $\alpha=0.7$ 作为最优融合权重。从图8和图9中可知,推荐数据数量对推荐数据质量也有影响。

因此,需要在最佳权重 $\alpha=0.7$ 的条件下,通过改变推荐数据数量 K ,查看召回率和准确率的变化情况,找寻最优推荐数据数量。其中 $K \in [2,15]$,推荐数据数量与召回率和精确率的关系如图10所示。

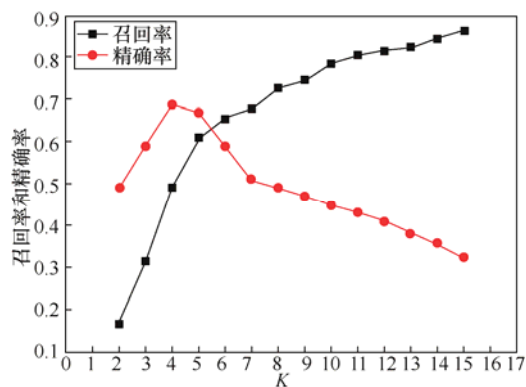


图10 推荐数据数量与召回率和准确率的关系
Figure 10 Relationship between the quantity of recommended data, recall and precision

从图10可知,召回率与 K 呈正相关,随着 K 的增加而增加,精确率则表现为先增后减,当 $K=4$ 时取得最大值,且随着 K 的增加召回率和精确率也在提高,无法直接获取最优的 K 值,因此在此基础上计算召回率和精确率的调和平均值,根据综合评价给出 K 的最优取值,推荐数据数量与调和平均值的关系如图11所示。

由图11可知, K 从2到5时,调和平均值随着推荐数据数量的增加而增加,当 K 大于5之后,调和平均值随着推荐数据数量的增加而减小,因此在 $K=5$ 时,调和平均值取得最大值68.4%,在 $K=5$ 时,召回率和精确率分别为62.37%和68.23%。

3.3 算法对比分析

为了验证所提算法的有效性,将所提算法分

别与未进行相似度融合算法（V）和类似推荐算法（BPR、CKE、FMG）进行召回率和精确率对比，不同算法召回率和精确率与 K 值的关系如图 12 和图 13 所示。

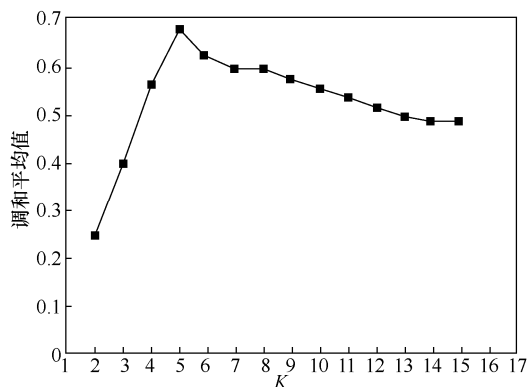


图 11 推荐数据数量与调和平均值的关系

Figure 11 Relationship between the number of recommended data and the harmonic mean

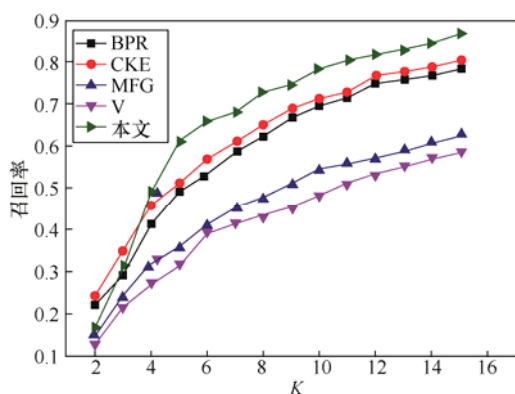


图 12 不同算法召回率与 K 值的关系

Figure 12 Relationship between recall of different algorithms and K -values

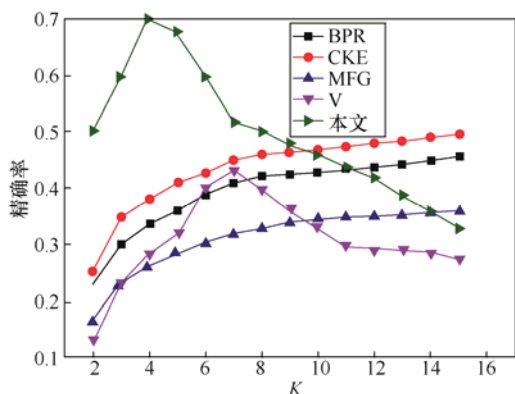


图 13 不同算法精确率与 K 值的关系

Figure 13 Relationship between precision of different algorithms and K -value

从图 12 和图 13 可知，不同算法的召回率随着 K 值增加而提高，由于本文算法在数据相似度的基础上加入了处置行为相似度，更接近事实处置行为，所以召回率优于其他算法。但是由于余弦函数只考虑数据向量方向，并没有考虑数据大小，因此在数据数量较小时能够获得较高精确率，但是随着数据数量增加会出现下降趋势，当 $K > 9$ 时，本文算法不再具有优势，《信息安全技术 网络安全事件分类分级指南》中网络安全事件从 8 个方面进行分类分级评判^[16]，因此该算法精确率在网络安全领域具有一定优势。

4 结束语

本文提出一种基于知识图谱的网络安全事件数据推荐算法，构建基于网络威胁框架 ATT&CK 的知识图谱，将攻击技术、漏洞、防御措施等离散的安全数据提炼为互相关联的安全知识，在此基础上引入基于网络安全事件处置行为的实体相似度，根据处置行为发掘不同主题数据间的内在关联性，对网络威胁数据实体的相似度和基于处置行为的网络威胁数据实体的相似度进行权重融合，形成推荐列表，通过与类似算法进行比较，该算法的召回率具有明显优势，精确率在一定范围内具有一定优势。

接下来，在持续优化网络威胁知识图谱的基础上，将优化数据相似度算法或引入调整余弦相似度算法，使其能够适应更加复杂和数据量更大的场景。

参考文献：

- [1] 张淑英. 网络安全事件关联分析与态势评测技术研究[D]. 长春: 吉林大学, 2012.
ZHANG S Y. Research on correlation analysis and situation evaluation technology of network security events[D]. Changchun: Jilin University, 2012.
- [2] CUPPENS F, MIEGE A. Alert correlation in a cooperative intrusion detection framework[C]//Proceedings of 2002 IEEE symposium on Security and privacy, 2002: 202-215.
- [3] ZANG Y C, ZHOU T Y, GE X Y, et al. An improved attack path discovery algorithm through compact graph planning[J]. IEEE Access, 2019, 99: 1
- [4] LI K, ZHOU H, TU Z, et al. Cskb: a cyber security knowledge base based on knowledge graph[C]//Proceedings of First International Conference on Security and Privacy in Digital Economy (SPDE 2020). 2020: 100-113.

- [5] SYED Z, PADIA A, FININ T, et al. UCO: a unified cybersecurity ontology[C]//Proceedings of AAAI Workshop on Artificial Intelligence for Cyber Security. 2016: 14-22.
- [6] 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(1): 151-166.
SUN C, HU H, YANG Y J, et al. Prediction method of 0day attack path based on cyber defense knowledge graph[J]. Chinese Journal of Network and Information Security, 2022, 8(1): 151-166.
- [7] WANG H, ZHANG F, WANG J, 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. 2018: 417-426.
- [8] WANG X, WANG D, XU C, et al. Explainable reasoning over knowledge graphs for recommendation[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2019: 5329-5336.
- [9] RENDLE S, FREUDENTHALER C, GANTNER Z, et al. BPR: Bayesian personalized ranking from implicit feedback[C]//Proceedings of UAI 2009. 2009: 452-461.
- [10] ZHANG F, YUAN N J, LIAN D, et al. Collaborative knowledge base embedding for recommender systems[C]//Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016: 353-362.
- [11] ZHAO H, YAO Q, LI J, et al. Meta-graph based recommendation fusion over heterogeneous information networks[C]//Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2017: 635-644.
- [12] 徐增林, 盛泳潘, 贺丽荣, 等. 知识图谱技术综述[J]. 电子科技大学学报, 2016, 45(4).
XU Z L, SHENG Y P, HE L R, et al. Overview of knowledge graph technology[J]. Journal of University of Electronic Science and Technology, 2016, 45(4).
- [13] ZHOU X, ZHU Q, LIU P, et al. Learning knowledge embeddings by combining limit-based scoring loss[C]//Proceedings of the 2017 ACM on Conference on Information and Knowledge Management. 2017: 1009-1018.
- [14] 李忠伟, 高东, 刘昕, 等. 基于知识图谱的海洋数值预报数据推荐算法[J]. 计算机工程与设计, 2023, 44(5).
LI Z W, GAO D, LIU X, et al. Recommendation algorithm of marine numerical forecast data based on knowledge graph[J]. Computer Engineering and Design, 2023, 44(5).
- [15] RANI U, BIDHAN K. Comparative assessment of extractive summarization: TextRank, TF-IDF and LDA[J]. Journal of Scientific Research, 2021(1).
- [16] GB/T 20986-2023. 信息安全技术 网络安全事件分类分级指南[S].
GB/T 20986-2023. Information security technology-classification of cyber security incidents[S].

[作者简介]



祝现威 (1991-), 男, 河南虞城人, 中国人民解放军 61660 部队工程师, 主要研究方向为 SDN 安全、知识图谱、云计算安全。

刘伟 (1982-), 男, 河北邯郸人, 中国人民解放军 61660 部队高级工程师, 主要研究方向为网络安全、知识图谱、指挥控制。

刘自豪 (1993-), 男, 湖北十堰人, 中国人民解放军 61660 部队助理工程师, 主要研究方向为网络安全、知识图谱。

顾泽宇 (1993-), 男, 山东滨州人, 中国人民解放军 61660 部队助理工程师, 主要研究方向为网络安全、态势感知。