Software Engineering and Applications 软件工程与应用, 2023, 12(6), 844-850
Published Online December 2023 in Hans. https://www.hanspub.org/journal/sea
https://doi.org/10.12677/sea.2023.126082

Hans汉斯

知识图谱嵌入的安全性问题分析

纪　雷，张志敏，王心怡，赵国生

哈尔滨师范大学计算机科学与信息工程学院，黑龙江　哈尔滨

收稿日期：2023年10月9日；录用日期：2023年12月5日；发布日期：2023年12月14日

摘　要

知识图谱作为新型数据语义分析模型得到广泛应用，人们对提交数据安全的需求也日益提高。为帮助目前面临的数据孤岛和关联网络构建问题，知识图谱得到了现有研究中的广泛关注和研究。虽然知识图谱作为一种极具影响的机器学习技术，能够对离散数据进行关联关系挖掘分析，但是也存在很多安全性问题。因此本文对知识图谱构建过程中面临的隐私安全隐患进行总结和分析，这对知识图谱的发展及应用具有重要意义，本文首先对知识图谱的基本概念和知识图谱嵌入过程进行详细阐述；接着，深入分析知识图谱建模过程中遇到的隐私泄露问题，包括模型逆向攻击、模型萃取攻击、投毒共计。然后，归纳总结了不同的知识图谱嵌入过程中的攻击防御方法。最后，总结与展望了知识图谱嵌入的应用前景及未来重要研究方向。

关键词

知识图谱，数据安全，隐私保护

Analysis of Security Issues in Embedding Knowledge Graph

Lei Ji, Zhimin Zhang, Xinyi Wang, Guosheng Zhao

School of Computer Science and Information Engineering, Harbin Normal University, Harbin Heilongjiang

Received: Oct. 9th, 2023; accepted: Dec. 5th, 2023; published: Dec. 14th, 2023

Abstract

Crowd sensing, as a new data collection mode, has been widely applied, and people's demand for submitting data security is also increasing. In the multi-participant joint construction of graph models, there are many privacy and security risks. Therefore, this article summarizes and analyzes the privacy and security risks faced in the process of constructing knowledge graphs, which is of great significance for the development and application of knowledge graphs. Firstly, this article

文章引用: 纪雷, 张志敏, 王心怡, 赵国生. 知识图谱嵌入的安全性问题分析[J]. 软件工程与应用, 2023, 12(6): 844-850. DOI: 10.12677/sea.2023.126082

elaborates on the basic concepts of knowledge graphs and the embedding process of knowledge graphs in detail; Next, an in-depth analysis will be conducted on the privacy leakage issues encountered in the process of knowledge graph modeling, including. Then, the privacy protection methods in different knowledge graph embedding processes were summarized and summarized. Finally, the application prospects and important future research directions of knowledge graph embedding were summarized and prospected.

Keywords

Knowledge Graph, Data Security, Privacy Protection

1. 引言

随着人工智能技术的发展，知识图谱(knowledge graph)作为一种图模型语义网络在语义分析[1]、命名实体消歧[2]、信息提取[3]、问答系统[4]等方面得到广泛应用。知识图谱嵌入学习方法主要是对图谱中的实体及关系学习低秩且密集的向量，通过图嵌入向量表达实体和关系间语义网络信息同时度量实体之间、实体和关系之间、关系之间的关联关系。然而图数据蕴含信息丰富，实体间关联关系复杂，给知识图谱的隐私保护带来了很大的挑战。首先，图数据上信息的多样性增大了隐私定义的难度。图数据中结点所代表的实体身份、语义属性、结点所在的子图结构、结点本身在图中的存在性，以及图中边上的语义属性、边的存在性，都可能是需要保护的敏感信息。如何选择并综合各类敏感信息进行合理的隐私定义，是图数据隐私保护上的一个难点。其次，图数据中结点之间复杂的关联关系增大了隐私保护技术设计与应用的难度。同一个结点可能与大量其它结点存在各种不同的链接关系，并且结点上的语义信息与结点所在子图的结构特征也存在一定的关联，对图中任何一个结点、一条边或一条语义信息稍做更改，都可能牵一发而动全身，大大降低图数据整体的可用性。

知识图谱的隐私保护方法需要考虑边、点之间的关系，图数据内部存在关联性，其中的关于节点、链路、子图等信息都可能为攻击者提供背景知识，极大增加图数据的隐私泄露问题，因此，现有针对图像处理、自然语言处理相关方法的攻击和防御手段无法直接适用于知识图谱模型中，知识图谱的隐私保护具有其自身的独特性。尽管知识图谱作为一种关系语义网络模型得到了大量关注与研究，但是在隐私保护方面仍存在一定问题。本文主要对知识图谱嵌入的定义和方法进行了介绍，并对知识图谱嵌入过程中的安全性问题和现有的防御方法进行分析，最后对展望了知识图谱未来的应用前景和发展方向。

2. 知识图谱嵌入概述

2.1. 知识图谱嵌入的定义

知识图谱作为一种异构性数据语义网络，常用于表示多实体间复杂关系，图结构中节点代表实体，节点间的连接边表示实体间存在关系。知识图谱 G 可形式化表示为事实三元组 $G=\left(e_i,e_j,r\right)$ 或 $G=\left(E,R,T\right)$，其中 e_i 表示头实体，r 表示实体间的关系，e_j 表示尾实体，E 表示实体集合，R 表示关系集合，T 表示事实三元组集合。面向知识图谱的图嵌入学习，目的在于为每个实体 $e\in E$ 学习相应的低维特征向量 $v_e\in R^{d_e}$，以及为每个关系 $r\in R$ 学习相应的低维表示向量 $v_r\in R^{d_r}$。其中 d_e 表示实体嵌入的维度，

d_r 表示关系嵌入的维度。上述任务目的在于通过学习实体及关系之间的关联矩阵，有效建模实体关系间的语义网络，有利于对数据进行挖掘和应用。图 1 为群智感知领域的部分知识图谱。

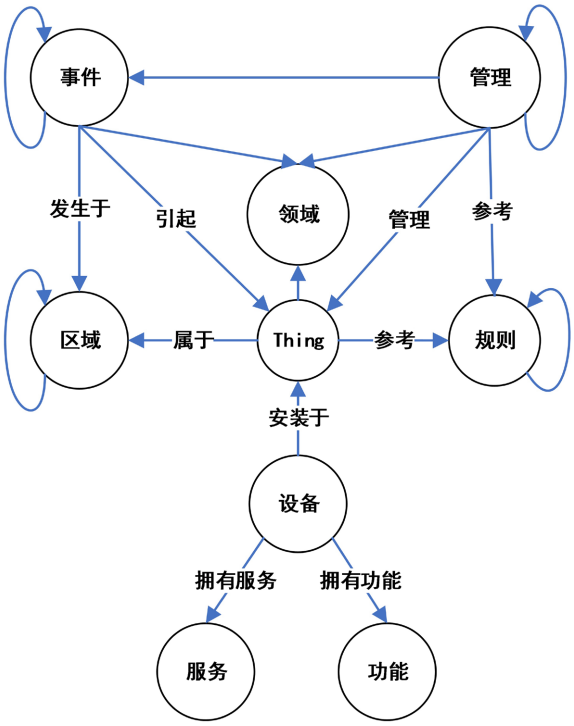


Figure 1. Internet of things knowledge graph
图 1. 群智感知领域知识图谱

2.2. 知识图谱嵌入模型

基于转移思想(translation-based)的图嵌入学习方法最早由 Bordes [5]提出，提出首个基于转移的表示学习模型 TransE，由词向量空间的平移不变现象启发。给定事实三元组 (e_i, e_j, r)，关系 r 的向量 v_r 被定义为在头实体向量 v_{ei} 和尾实体向量 v_{ej} 之间的平移。所以，嵌入头尾实体 e_i 和 e_j 可经过平移向量 v_r 以低偏差链接，从而满足 $e_i + r \approx e_j$，综上所述，对于给定事实三元组 (e_i, e_j, r)，TransE 设计的优化目标函数为：

$$f(e_i, e_j, r) = \| v_h + v_r - v_t \|_p \tag{4}$$

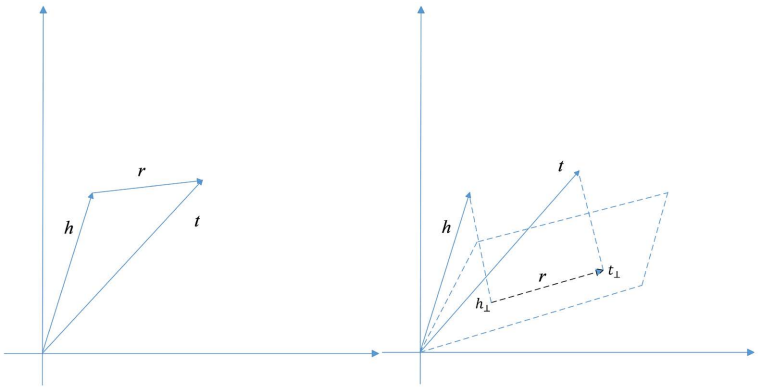
当三元组错误时最大化该目标函数，反之则最小化该函数。图 2(a)为 TransE 的核心思想，而在处理多对多、一对多和一对一的关系时，由于约束 $e_i + r \approx e_j$ 的存在，即使不同实体语义信息不同，仍会在事实空间中将不同实体的分布式聚集在一起，导致实体语义的混乱。为解决上述问题，TransH 引入关系超平面对 TransE 进行改进，平移头实体前，首先将头尾实体分别映射到关系网络超平面。TransH 将实体建模为向量，将每个关系建模为法向量 w_r 的关系超平面上的向量 $r(r \in R^d)$。具体来说，对事实三元组 (e_i, e_j, r)，TransH 将头实体向量 h 与尾实体向量 $t(h, t \in R^d)$ 沿法线 w_r $(w_r \in R^d)$ 投影到关系 r 对应的超平面，投影向量为 $h_\perp$ 和 $t_\perp$，表示如下：

$$\begin{aligned} h_\perp &= h - w_r^T h w_r \\ t_\perp &= t - w_r^T t w_r \end{aligned} \tag{5}$$

对于三元组 (e_i, e_j, r)，TransH 设计的评分函数为：

$$f_r\left(e_i,e_j\right)=-\|h_\perp+r-t_\perp\|_2^2 \tag{6}$$

通过引入关系超平面和投影机制，TransH 能够将不同实体于不同关系中得到不同表现。图 2(b)为 TransH 的核心思想。



(a) TransE (b) TranH

Figure 2. TransE 和 TransH 的核心思想示意图
图 2. Schematic diagram of the core ideas of TransE and TransH

3. 知识图谱嵌入中的安全性问题

3.1. 模型逆向攻击

数据作为构建群智网络系统的重要资源，服务运营商通过收集参与者发布数据、交易记录等方式收集数据，并在此基础上通过提取向量、训练模型、数据融合，对数据进行凝炼形成训练数据集构建知识图谱嵌入模型，因此，对知识图谱的嵌入模型的训练数据应该对外部不可见的，即可用不可见。但是，面对开放环境的知识图谱嵌入模型，攻击者可通过开发服务接口，以模型的输入样本和模型对应训练结果作为攻击行为的启动数据，从而进行“模型逆向攻击(Model Inversion Attack)”，通过知识图谱数据中蕴含的关联关系推理重构匿名的敏感信息，窃取模型训练数据，获取训练数据中相关成员社交网络信息，从而破坏数据的安全性。

3.2. 模型萃取攻击

知识图谱模型是服务运营商基于大量数据资源、能耗资源训练获得，而在开放环境下，现有研究中 Shen [6]等人提出 Inductive GNN 构建萃取攻击模型的新方法，该攻击面向 Graph SAGE，GAT，GIN 等多目标模型根据查询图的节点特征及图结构信息构建背景知识构建统一萃取攻击框架，攻击者通过在知识图谱模型部署后通过多次查询－反馈的结果联系，通过得到信息进行“模型萃取攻击(Model Extraction Attack)”，获得模型架构、参数等指标信息，并重构支撑建模自有模型或对目标系统实施其他攻击，从而在对目标模型不造成影响的前提下实现对知识图谱模型敏感信息的泄露，获取参与知识图谱建模的成员信息、关联网络数据，破坏知识图谱模型的隐私性。

3.3. 投毒攻击

在原始数据采集或者图数据提交发布阶段，攻击者通过将模型训练相关参数作为输入，并将精心改造后的训练样本插入到数据集中来操控模型训练集合的内部分布，从而输出投毒之后的数据实现改变模型行为和降低模型性能的目的，即“投毒攻击(Poisoning Attack)”。基于获取到的原始数据，攻击者根据

目标模型的可用信息及具体攻击对象，设计对抗扰动，通过对模型输入样本进行细小修改扰动，产生对抗样本，导致模型错误输出或可控输出，最后影响模型的可用性和隐私性，这种攻击实际上是因知识图谱关联性的特性促成的，因为攻击者可通过知识图谱间构建关联进行投毒数据设计，因此会对知识图谱的应用产生灾难性的后果。

4. 知识图谱嵌入的防御方法

4.1. 知识图谱模型逆向攻击的防御

针对模型逆向攻击，传统机器学习领域通常采用差分隐私、随机失活(Dropout)和模型集成(Model Stacking)等方法实现。He [7]等人在从探索数据和模型角度出发，在模型训练阶段通过随机添加链路，使得模型最终输出结果仅包含预测标签，从而保护目标模型及其训练数据隐私，Zhang [8]等人为防御成员推理攻击，提出在目标模型训练数据集中添加链路、目标模型输出结果仅包含预测标签等两种方法进行针对防御，并且通过实验论证，在多个图神经网络模型中能有效降低模型逆向攻击的推理准确性，但是同时会不可避免的数据失真，降低部分实用性。总体上，当前对知识图谱模型逆向攻击研究工作仍未形成系统研究成果，当前仅有研究只是在传统机器学习领域方法的迁移应用，图模型逆向攻击防御方法研究缺乏充分理论指导。

4.2. 知识图谱模型萃取攻击的防御

为防御模型萃取攻击，传统机器学习领域主要通过限制发布和通过对置信度精度截断实现保护防御。同时，还提出结合查分隐私方法对目标模型输出图数据进行保护，通过降低模型数据的准确性从而实现对萃取攻击的防御，并且，在传统机器学习领域中，还提出基于密码学的模型隐私保护方法、基于模型水印的模型隐私保护方法及基于 Dropout 等策略防止模型过拟合的模型隐私保护方法，但是关于知识图谱模型的模型萃取攻击问题，当前领域研究中还未提出针对性的方法策略，传统的通用模型萃取攻击方法的可用性也有待现实环境验证和分析。

4.3. 知识图谱投毒攻击的防御

4.3.1. 基于预处理的防御方法

预处理是通过在模型训练前对知识图谱数据源中潜在的对抗扰动数据进行查找和剔除，Wu 等人[9]使用相似度度量 Jaccard 攻击对图节点间存在链路的可能性进行识别，发现高度相异特征节点对，通过移除异常链路实现防御对抗攻击。Entezari 等人[10]通过研究 Nettack 攻击，对其攻击模型进行分析，其仅通过影响邻居矩阵奇异值分解的高阶值实现攻击行为，从而提出利用主奇异值低秩近似消除防御对抗攻击。

4.3.2. 基于对抗训练的防御方法

对抗训练通过在训练数据集中加入对抗样本，从而丰富数据源集合的同时提升知识图谱模型鲁棒性，Liu 等人[11]给出基于模型可解释性的对抗训练方案，通过误差平方和局部范围最优近似目标模型 f 建立简单可解释模型，并基于该模型生成对抗数据进行对抗训练。Feng 等人[12]对半监督的节点分类模型提出基于虚拟对抗训练的防御方法 GARV，通过最大－最小框架，在多轮迭代过程中生成对抗样本实现最大化节点关联差异、破坏光滑性，同时最小化目标函数，增加关联节点间的光滑性，最终完成模型训练。Li [13]等人提出一种频谱对抗训练方法，采用基于谱分解的图结构低秩近似，在谱域构造对抗性扰动，通过脱离图结构本体操作，提高 GNN (图神经网络)对抗训练有效性的同时，无需牺牲分类器准确性和模型效率。

4.3.3. 基于认证的防御方法

基于认证的防御方法目前主要通过添加一系列随机噪声依据随机平滑方法测试分类器在当前环境下的鲁棒性，在图学习领域，现阶段 Zhang [14]等人提出一种基于认证的图神经网络面向后门攻击防御算法，主要思想是利用随机子采样方法来建立平滑分类器，并据此对图中标签进行预测，最后通过投票得分形式将得分最多标签作为最终真值标签，以此防止后门攻击。

5. 知识图谱嵌入的应用前景和研究方向

5.1. 知识图谱嵌入的应用前景

知识图谱作为新型数据信息处理网络模型，能将数据和知识整合成一个统一且可读性较强的寓意网络，是实现人工智能和大数据分析的重要工具，在众多领域得到广泛应用，具有广阔应用前景和发展方向。

(1) 智慧金融。金融领域内包含多源数值数据，为了给消费者提供更准确的金融信息和消费体验，往往需要收集消费者的交易记录、投资偏好、经济情况等信息进行训练收集，但通常情况下，多用户收集数据的融合需要消耗大量资源整理，并且数据收集需要不同部门进行，对收集到的数据的关联信息的利用也不够充分，知识图谱能够将多源异构的数据利用机器学习方法进行解决，实现跨数据和跨领域的数据关联知识图谱模型训练。

(2) 智慧医疗。现有的医疗健康服务体系仍在建设阶段，难以满足健康服务诉求，人们更倾向于从网络从获取医疗健康知识，医疗领域知识专业且复杂，在语义网络中非结构化文本往往占大多数，传统数据存储方法对非结构化文本利用率较低，无法充分挖掘非结构化文本的医疗知识，而结合知识图谱的语义解析能力提高相关智慧医疗的检索能力。

5.2. 知识图谱嵌入的研究方向

知识图谱嵌入因能实现非结构化数据的发掘、分析。在智慧城建、智能金融、智慧医疗等领域具有巨大潜力。但同样面临一些安全性分析需要进一步研究。

(1) 知识图谱的隐私保护机制。当前知识图谱模型隐私保护主要以启发式方法为主，往往通过具有目标性问题的结果分析进行，隐私保护水平没有保障，结合图数据特性的理论性、基础研究有待加强，如，基于知识图谱的生成模型对图数据隐私风险有待量化，使用同态加密方法的图数据加密方法和密文数据挖掘方法研究，及其他传统隐私保护方法的图数据相关内容结合框架研究，因此，隐私保护的知识图谱学习框架和模型方法研究是未来研究的重要方向。

(2) 知识图谱数据鲁棒性建模研究。当前攻击和防御模型研究缺少对图数据的内在关联特征的考量，真实数据往往是异构、复杂的，现有对图数据的认识往往基于传统复杂网络等相关成果，对图数据特征的研究仍不成熟，对图数据结构模式与图模型底层原理的关系缺乏分析，在图模型安全方面如，对抗样本生成和攻击检测方法的构建都缺乏对图数据的利用，因此，知识图谱数据鲁棒性的模型设计是知识图谱模型安全性问题未来重要研究方向。

6. 总结

知识图谱作为一种新型的语义网络机器学习方法，有效的发掘数据间的关联语义关系，但是目前研究中仍存在一些潜在的安全问题。本文通过对知识图谱嵌入模型的基本概念进行阐述，并详细介绍了 TransE 和 TransH 这两种经典的知识图谱嵌入框架。同时对知识图谱嵌入过程中的安全性问题进行了分析，主要包含模型逆向攻击、模型萃取攻击和投毒攻击，并在此基础上，对现有的攻击防御方法进行了

总结和介绍，最后对知识图谱的应用前景和未来研究方向进行了探讨总结。

基金项目

黑龙江省高等教育教学改革研究一般研究项目(SJGY20220351)和 2023 年度省规划办重点课题(GJB1423438)。

参考文献

[1] Jain, N. (2020) Domain-Specific Knowledge Graph Construction for Semantic Analysis. In: Harth, A., *et al*., Eds., *The Semantic Web*: *ESWC* 2020 *Satellite Events*, Springer, Cham, 250-260. https://doi.org/10.1007/978-3-030-62327-2_40

[2] Zhang, L., He, Q., Yu, W., *et al*. (2022) Research on Entity Disambiguation Method and Model Construction Based on Knowledge Graph. 2022 4*th International Conference on Machine Learning, Big Data and Business Intelligence* (*MLBDBI*), Shanghai, 28-30 October 2022, 174-177. https://doi.org/10.1109/MLBDBI58171.2022.00041

[3] Wei, L., Zhao, H. and He, Z. (2022) Designing the Topology of Graph Neural Networks: A Novel Feature Fusion Perspective. *Proceedings of the ACM Web Conference* 2022, Nanjing, 25-29 April 2022, 1381-1391. https://doi.org/10.1145/3485447.3512185

[4] Omar, R., Mangukiya, O., Kalnis, P., *et al*. (2023) ChatGPT versus Traditional Question Answering for Knowledge Graphs: Current Status and Future Directions towards Knowledge Graph Chatbots. arXiv: 2302.06466. https://doi.org/10.1145/3588911

[5] Bordes, A., Usunier, N., Garcia-Duran, A., *et al*. (2013) Translating Embeddings for Modeling Multi-Relational Data. *Advances in Neural Information Processing Systems*, **26**, 2787-2795.

[6] He, X., Wen, R., Wu, Y., *et al*. (2021) Node-Level Membership Inference Attacks against Graph Neural Networks. arXiv: 2102.05429.

[7] Wu, B., Yang, X., Pan, S., *et al*. (2022) Model Extraction Attacks on Graph Neural Networks: Taxonomy and Realisation. *Proceedings of the* 2022 *ACM on Asia Conference on Computer and Communications Security*, Nagasaki, 30 May-3 June 2022, 337-350. https://doi.org/10.1145/3488932.3497753

[8] Zhang, Z., Liu, Q., Huang, Z., *et al*. (2021) GraphMi: Extracting Private Graph Data from Graph Neural Networks. *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence* (*IJCAI*-21). Montreal, 19-27 August 2021, 3749-3755. https://doi.org/10.24963/ijcai.2021/516

[9] Wu, H., Wang, C., Tyshetskiy, Y., *et al*. (2019) Adversarial Examples on Graph Data: Deep Insights into Attack and Defense. arXiv: 1903.01610. https://doi.org/10.24963/ijcai.2019/669

[10] Entezari, N., Al-Sayouri, S.A., Darvishzadeh, A., *et al*. (2020) All You Need Is Low (Rank) Defending against Adversarial Attacks on Graphs. *Proceedings of the* 13*th International Conference on Web Search and Data Mining*, Houston, 3-7 February 2020, 169-177. https://doi.org/10.1145/3336191.3371789

[11] Liu, N., Yang, H. and Hu, X. (2018) Adversarial Detection with Model Interpretation. Proceedings of the 24*th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, London, 19-23 August 2018, 1803-1811. https://doi.org/10.1145/3219819.3220027

[12] Feng, F., He, X., Tang, J., *et al*. (2019) Graph Adversarial Training: Dynamically Regularizing Based on Graph Structure. *IEEE Transactions on Knowledge and Data Engineering*, **33**, 2493-2504. https://doi.org/10.1109/TKDE.2019.2957786

[13] Li, J., Peng, J., Chen, L., *et al*. (2022) Spectral Adversarial Training for Robust Graph Neural Network. *IEEE Transactions on Knowledge and Data Engineering*, **35**, 9240-9253. https://doi.org/10.1109/TKDE.2022.3222207

[14] Zhang, Z., Jia, J., Wang, B., *et al*. (2021) Backdoor Attacks to Graph Neural Networks. Proceedings of the 26*th ACM Symposium on Access Control Models and Technologies*, New York, 16-18 June 2021, 15-26. https://doi.org/10.1145/3450569.3463560