

网络空间安全命名实体识别综述

高宸¹, 张璇^{1,2,3}, 韩梦婷¹, 刘会¹

¹云南大学软件学院, 中国昆明市, 650091

²云南省软件工程重点实验室, 中国昆明市, 650091

³网络空间工程研究中心, 中国昆明市, 650091

摘要: 随着互联网技术飞速发展和大数据时代到来, 越来越多网络空间安全文本出现在互联网上。这些文本不仅包括安全概念、事件、工具、指南和政策, 还包括风险管理方法、最佳实践、保证和技术等。整合大规模、异构和非结构化的网络空间安全信息, 对网络空间安全实体进行识别和分类, 有助于处理和解决网络空间安全问题。由于网络空间安全领域文本的复杂性和多样性, 使用传统的命名实体识别 (NER) 方法难以识别该领域中的安全实体。本文介绍该领域NER的各种方法和技术, 包括基于规则的方法、基于字典的方法和基于机器学习的方法, 并讨论该领域NER研究面临的问题, 如实体词组的结合与分离、非标准化的命名约定、缩写和大量嵌套等。最后, 提出NER在网络空间安全方面的3个研究方向: (1) 应用无监督或半监督技术; (2) 开发更全面的网络空间安全本体; (3) 应用更加有效的深度学习模型。

关键词: 命名实体识别 (NER); 信息抽取; 网络空间安全; 机器学习; 深度学习
<https://doi.org/10.1631/FITEE.2000286>