北 京 邮 电 大 学

本科毕业设计（论文）开题报告

学院	计算机学院	专业	计算机科学与技术	班级	2018211303
学生姓名	马嘉骥	学号	2018211149	班内序号	17
指导教师姓名	方维	所在单位	北邮计算机学院	职称	副教授
设计（论文）题目	（中文）基于漏洞知识图谱的可视化系统的设计与实现				
	（英文）Design and Implementation of Visualization System Based on Vulnerability Knowledge Graph				

一、选题的背景和意义

1.1 选题背景

近年来，随着互联网产业迅速发展，互联网安全漏洞问题显著性也急剧增加。根据公共漏洞和暴露(Common Vulnerabilities and Exposures，CVE) 等权威漏洞数据档案的数据，自 1999 年安全漏洞首次披露以来，互联网安全漏洞数量呈现增长趋势，2019 年全年，新增漏洞两万余个；2020 年全年粗略估计，新增漏洞三万五千余个。对攻击者而言，不仅漏洞攻击的学习成本和难度下降，其可以利用的漏洞数量也明显增多；对企业与开发者而言，随着开源化逐渐成为一种趋势，各类互联网产品对开源系统与组件的依赖性逐步升高，正如近期 Java Log4j 日志组件漏洞造成大规模安全隐患，互联网产业蓬勃发展的同时也面临与日俱增的安全挑战。

1.2 项目意义

本课题针对上述问题，提出一种漏洞知识图谱可视化系统。基于知识图谱、图数据库等技术，对互联网漏洞数据包括受影响产品、可利用代码及补丁等信息进行收集与分析，形成知识结构；对多个数据源抽取的知识进行融合、抽取漏洞的实体及关系、建立漏洞的图数据库、形成漏洞知识图谱；基于漏洞知识图谱搭建基于 B/S 架构的可视化系统，提供易于使用的交互接口进行展示、知识筛选等操作。

本系统采用自动化的方式，实现对漏洞信息的持续收集与整理，极大节省了人力资源的消耗。结合抽取关键信息，对漏洞间关联性进行分析、构建漏洞知识图谱，将离散的漏洞信息转化为相互联系的图结构，为开发者提供项目依赖安全性参考、为互联网安全研究人员提供数据与服

务支撑，促进构建更高效安全的互联网环境。

二、研究的基本内容和拟解决的主要问题

2.1 研究的基本内容

1、调研互联网安全漏洞数据信息的来源与收集渠道

2、调研 Scrapy、Selenium、Puppeteer 等数据获取方案

3、调研知识图谱的思想、基本原理、构建技术

4、调研 Neo4j 等图数据库存储方案

5、调研针对图数据的 Web 技术方案，进行后端 Django/Flask/Springboot 框架技术选型、进行 Vue/React/Angular 等前端框架技术选型

6、设计实现基于数据爬取、规则或机器学习方法的漏洞知识图谱构建系统

7、设计实现基于图数据的漏洞知识图谱存储系统

8、设计实现基于漏洞知识图谱的可视化系统

9、对上述系统进行系统测试、排错、功能扩展、优化，编写文档记录

2.2 拟解决的主要问题

本课题系统包含数据采集、数据分析、图谱构建、数据库存储、前后端搭建等内容，经前期调研，提出主要问题如下：

1、安全漏洞信息数据来源：数据采集是整个系统的源头。如何获得可靠、准确、高质量的漏洞信息，对数据分析过程实施、最终系统可用性具有决定性作用。本课题需调研互联网安全漏洞数据源，并提出具备可行性的信息采集方案。

2、基于多源异构漏洞数据的安全漏洞信息抽取：互联网相关漏洞、威胁情报、漏洞利用代码等信息分布在各大平台，杂乱且不够全面。本课题需提出一种采用多源异构数据融合的方案，从原始语料提取实体、关系、属性等知识要素，再经知识融合、消除歧义，得到一系列基本的事实表达，实现漏洞信息的完整收集与格式化。

3、漏洞知识图谱的构建方法：知识图谱是本项目核心所在。根据采集整理的格式化漏洞信息，进行本体构建、知识推理、质量评估的知识加工过程，最终获得结构化、网络化的知识体系。同时，还应考虑知识图谱的迭代更新过程实现。

4、漏洞知识图谱的可视化平台设计与实现：构建知识图谱、实现基于图数据的漏洞知识图

谱存储后，本系统应具有易于访问与交互的界面供用户使用。本课题需调研通过基于 B/S 架构的 WebApp 与前述图数据库的交互式访问与可视化图结构呈现方案，实现漏洞知识图谱的可视化系统。

三、研究方法及措施

1、安全漏洞信息数据来源

目前国际上有公共漏洞和暴露(Common Vulnerabilities and Exposures，CVE) 数据库、中国国家信息安全漏洞共享平台(China National Vulnerability Database，CNVD) 、中国国家信息安全漏洞库(China National Vulnerability Database of Information Security，CNNVD) 、美国国家漏洞库(National Vulnerability Database，NVD) 、中国关键基础设施安全应急响应中心、美国工业控制系统网络应急响应小组等。以上公开漏洞信息发布来源都是信息安全领域内最具权威性的平台。

本课题计划以上述公开漏洞信息源作为漏洞知识的主要数据来源。原始数据收集拟由本基于漏洞知识图谱的可视化系统的数据采集模块完成，拟采用定时增量方式从多个数据源进行信息收集。①对于漏洞相关数据，拟从上述漏洞信息平台、CPE（公共平台枚举）、CWE（常见缺陷列表）等数据源收集；②对于 POC/EXP 漏洞验证/利用代码，拟从 packet storm、exploit-db、github 上进行收集。

拟采用 Scrapy、Selenium 或 Puppeteer 框架进行漏洞数据爬取。例如对于提供 CVE 数据下载的 https://www.cve.org/Downloads，使用爬虫框架或 wget 命令直接访问该页面 https://cve.mitre.org/data/downloads/allitems.csv 链接进行数据下载，再使用 diff 等命令进行差异比对从而获取漏洞增量数据。对于需要从网页内容获取信息的数据源，使用 Scrapy 配合 XPath 获取页面信息，或采用 Selenium、Puppeteer 框架对爬取网页的 DOM 树进行操作获取需要的信息。根据数据更新频度需求，可使用 crontab 命令定时进行数据增量爬取。

2、基于多源异构漏洞数据的安全漏洞信息抽取

当前大多数的漏洞信息平台中涵盖了各种类别的漏洞信息，尽管有些漏洞信息平台会有行业分类，但其中的漏洞信息不够全面，存在交叉，不利于安全研究人员对行业内安全漏洞及相关威胁情报信息的获取和分析，亦不利于开发者增强漏洞防范本领。

信息抽取是知识图谱构建的第一步，其中关键问题是从异构数据源中自动抽取信息得到候选只是但愿。信息抽取是一种自动化地从半结构化和无结构数据中抽取实体、关系、实体属性等结

构化信息的技术，包括实体抽取、关系抽取、属性抽取等。本课题拟基于从多个数据源采集的数据，进行去重、融合、关键信息抽取，整理漏洞标题、对策、CVSS、CWE、CPE 等信息，从而实现漏洞信息的格式化。

去重：对于多源异构的数据，大多数的漏洞都有 CVE-ID，保证了其唯一性，对于没有 CVE-ID 的漏洞数据，可以通过相关链接，受影响平台及 CVSS 等信息结合人工分析比对现有的漏洞数据，判断是否为重复数据。

融合：结合各大漏洞平台的特点，针对漏洞的不同属性，参照其规范程度从不同源头选取，例如：漏洞标题及对策从 JVNDB 中选取，对于一些标准信息如 CVSS、CWE-ID 等从 NVD 中选取，相关链接则直接合并并去除重复链接。同时，利用 CVSS 及 CWE 的详细信息进一步扩充漏洞本体。对于威胁情报这种非结构化数据，通过 CVE-ID 和受影响软件和相关链接信息关联相应的漏洞。

关键信息抽取：通过规则匹配的方式从已有的漏洞数据中获取训练数据，针对漏洞描述信息，可以进行翻译、去除非文本、词形还原、转化为小写和删除停用词等数据清洗后训练 NER（命名实体识别）模型。使用训练好的模型抽取信息后，为保证漏洞信息的准确性和完整性，系统会对使用模型扩充的漏洞信息进行标注，随后进行人工审核和纠正。

3、漏洞知识图谱的构建方法

知识图谱的定义：知识图谱是结构化的语义知识库，用于以符号形式描述物理世界中的概念及其相互关系。它本身是一个具有属性的实体通过关系链接而成的网状知识库，知识图谱将互联网中积累的信息组织起来，成为可以被利用的知识。知识图谱的应用价值在于它能①通过推理实现概念检索、②以图形化方式向用户展示经过分类整理的结构化知识，从而使人从人工筛选过滤网页寻找答案的模式中解放出来。

知识图谱的构建过程如图所示，从原始数据出发，采用自动或半自动手段，从原始数据中提取知识要素，将其存入知识库的数据层和模式层，不断迭代更新。每轮迭代包含三个阶段：信息抽取、知识融合、知识加工。

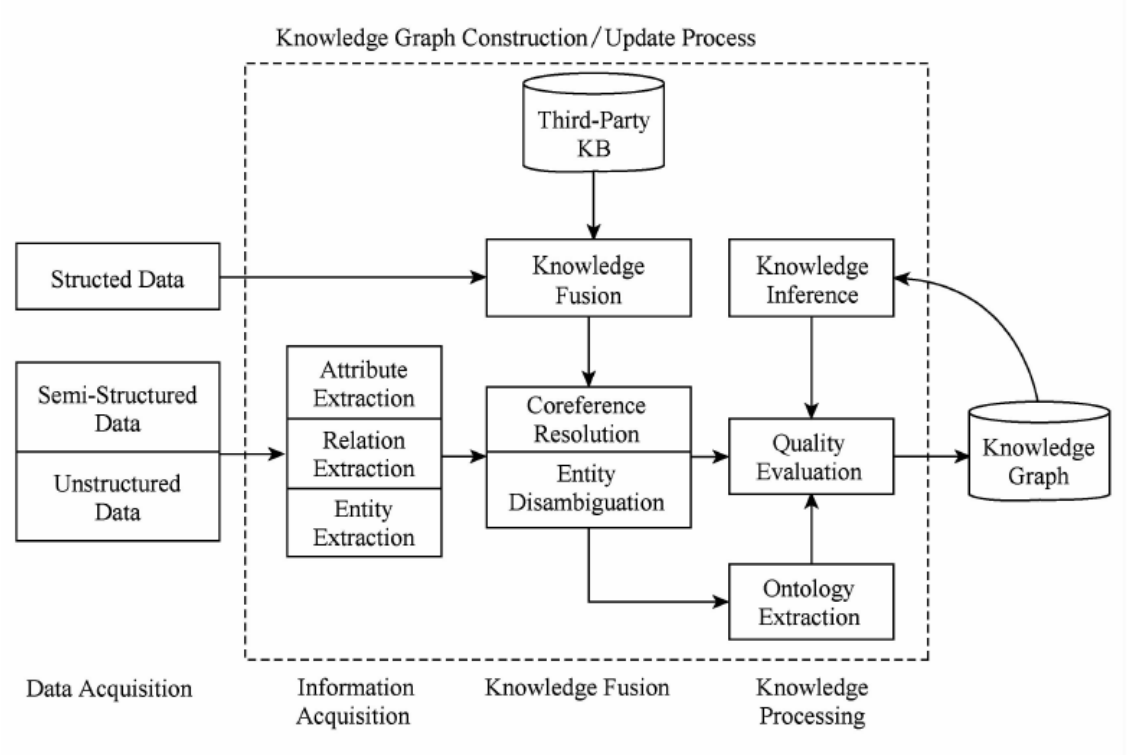

Fig. 1 Technical architecture of knowledge graph.

图 1 知识图谱的技术架构

知识融合包括实体链接与知识合并。知识抽取过程实现了从非结构化和半结构化数据中获取实体、关系、实体属性信息的目标，但这些信息的质量往往不高，数据之间的关系也缺乏层次性与逻辑性。通过知识融合，消除概念的歧义、剔除冗余和错误概念，从而确保了知识图谱中知识的质量。其中，实体链接（entity linking）过程通过获得实体指称、实体消歧与共指消解、链接操作，将从文本中抽取得到的实体对象链接到知识库中对应的正确实体对象。知识合并过程可遵循 W3C 制定的 RDF 和 R2RML 标准将第三方知识库产品或已有结构化数据输入知识图谱。

知识加工主要包括本体构建、知识推理、质量评估。本体构建以形式化方式对概念及其之间的联系给出明确定义，常用数据驱动的自动化构建方法，利用统计机器学习算法抽取概念之间的关系。知识推理从知识库中已有的实体关系数据出发，经计算机推理建立实体间的新关联，从而扩展和丰富知识网络，是知识图谱构建的重要手段和关键环节。

知识图谱有自顶向下与自底向上两种构建方法。自顶向下构建是借助公开数据库等结构化数据源，从高质量数据中提取本体和模式信息，加入到知识库中；自底向上构建是借助一定的技术手段，从公开采集的数据中提取资源模式，加入到知识库中。

结合本课题的实际需求，拟采用自顶向下的知识图谱构建方式，使用前述权威漏洞信息发布平台采集抽取的信息，通过知识融合与知识加工，构建漏洞知识图谱。

4、漏洞知识图谱的存储与可视化平台设计与实现

构建知识图谱的工作实现了安全漏洞知识图谱中节点和关系的构建，形成了逻辑上的知识图谱。还需将信息进行存储以实现对知识图谱的实际利用。拟选用 Neo4j 图数据库存储该图谱。Neo4j 数据库具有高效、成熟、稳定、接口丰富、扩展型强等优点。若使用 Django/Flask 作为后端框架可使用 py2neo 库、若使用 SpringBoot 作为后端框架可使用 Neo4j Java Driver 进行数据库的驱动。此外 Neo4j 也支持使用 Neo4j-import 一次性批量导入 csv 数据文件。

设计并实现安全漏洞情报平台，能够将全面的漏洞情报信息及时响应给安全研究人员，并提供多种功能帮助安全研究人员做分析。本课题拟搭建一套基于 B/S 架构的漏洞知识图谱可视化系统，用户通过浏览器访问 Web 应用，通过 RESTful API 与后端进行通信，从图数据库中获取查询结果，并由 Web 应用以可视化的方式展现。

知识图谱的可视化系统首先需要建立多维度检索功能以确定展示范围。使用图结构，通过相似度度量与漏洞体量评估的方式，来展示漏洞之间的关联；展示信息应包含漏洞标题、漏洞发现日期、漏洞 CVE-ID、CWE-ID、CVSS 与 CWE 详细信息、CPE、对策等。此外，拟实现可交互的可视化展示效果，以便用户更清晰地梳理漏洞之间逻辑关联、查看漏洞详细信息。漏洞情报平台可以实现对漏洞和相关情报的及时告警，通过邮件等方式将新增漏洞信息发送给用户。

本系统实现难点在于对于大量漏洞关系数据的高效检索与展示。进一步调研将确定后端 Django/Flask/SpringBoot 框架技术选型、前端 Vue/React/Angular 等框架技术选型，以及 D3.js 等可视化方案的技术选型。

四、研究工作的步骤与进度

2021/12/12 - 2021/12/24（两周）	明确任务，了解课题背景，制定计划，查找相关论文资料，对课题的研究方法形成大体框架并提交开题报告。
2021/12/27 - 2022/01/14（三周）	学习基本 Web 知识和 Scrapy 爬虫框架的使用，爬取公开漏洞、CPE、CWE、POC 等数据。
2022/03/01 - 2022/03/13（两周）	学习 SpringBoot 和 Vue 框架，分析漏洞知识间的关系；分析系统功能，完成系统的需求分析。
2022/03/14 - 2022/03/26（两周）	使用基于规则或机器学习的方法抽取漏洞实体及关系，构建漏洞数据图模型，建立图数据库。完成系统的概要设计。

2022/03/27 - 2022/04/17（三周）	完成系统详细设计以及部分代码实现，完成系统前端漏洞知识图的可视化、知识筛选等基本功能。接受中期检查。
2022/04/19 - 2022/05/02（两周）	完善前端功能，完成前后端全部代码。
2022/05/03 - 2022/05/16（两周）	进行系统测试、排错，及功能扩展、优化，编写文档记录。
2022/05/17 - 2022/05/30（两周）	撰写毕业论文，完成毕业设计。

主要参考文献：

[1]张吉祥,张祥森,武长旭,赵增顺.知识图谱构建技术综述[J/OL].计算机工程:1-16[2021-11-06].https://doi.org/10.19678/j.issn.1000-3428.0061803.

[2]陶耀东,贾新桐,吴云坤.一种基于知识图谱的工业互联网安全漏洞研究方法[J].信息技术与网络安全,2020,39(01):6-13+18.

[3] Han Z, Li X, Liu H, et al. DeepWeak: Reasoning common software weaknesses via knowledge graph embedding[C]// 2018 IEEE 25th International Conference on Software Analysis, Evolution and Reengineering (SANER). IEEE, 2018.

[4] Yan Jia, Yulu Qi, Huaijun Shang, Rong Jiang, Aiping Li. A Practical Approach to Constructing a Knowledge Graph for Cybersecurity[J].Engineering,2018,4(1):53-60.

[5] https://d3js.org/

[6] https://neo4j.com/docs/

[7] https://www.cve.org/

允许进入论文撰写环节：是 ☑ 否 ☐		指导教师	
日期	2022 年 3 月 2 日	签字	方雅

注：可根据开题报告的长度加页。