

是否参加“青年红色筑梦之旅”项目	<input type="radio"/> 是 <input checked="" type="radio"/> 否
是否国家重点领域项目	<input checked="" type="radio"/> 是 <input type="radio"/> 否

# 大学生创新创业训练计划 项目申报表

推 荐 学 校	西安财经大学
项 目 名 称	互联网安全卫士——病毒知识图谱构建及检测
项 目 类 型	创新训练项目
项 目 负 责 人	武晨明
申 报 日 期	2024.5

陕西省教育厅 制  
二〇二四年四月

项目名称		互联网安全卫士——病毒知识图谱构建及检测					
项目类型		(√) 创新训练项目    ( ) 创业训练项目    ( ) 创业实践项目					
		是否为“青年红色筑梦之旅”项目(否) 是否为国家重点领域项目(是)					
项目实施时间		起始时间：2024年6月                          完成时间：2025年6月					
申请人或申请团队		姓名	年级	学校	所在院系/专业	联系电话	E-mail
	主持人	武晨明	2023级	西安财经大学	信息学院软件工程	18792846399	2331051525@xaufe.efu.cn
		成员	张森郁	2023级	西安财经大学	信息学院软件工程	17729208986
	武禹辰		2023级	西安财经大学	信息学院软件工程	16754057786	3285527348@qq.com
	左倩倩		2023级	西安财经大学	信息学院软件工程	13571440638	13571440638@163.com
	尹伊涵		2023级	西安财经大学	信息学院软件工程	17791306260	3485269070@qq.com
指导教师	姓名	罗养霞		研究方向		系统安全、软件安全	
	年龄	50		行政职务/专业技术职务			教授
	手机	13991269478		电子邮箱		Yxluo8836@163.com	
	主要成果	<p>个人简介：罗养霞，计算机软件与理论工学博士，西安财经大学信息学院软件工程专业负责人，省级“双带头人”党支部书记工作室负责人，中国计算机学会 CCF 成员，国家软件学院示范联盟主要参与者。省级创新创业学院主要成员，省级创新基地与实验室主要成员。2018 年于美国密西根大学计算机与软件学院留学访问一年。主要研究软件项目管理与体系结构、软件安全等。主持或参与国家级、省厅级项目 20 多项；发表科研及教改论文 30 多篇；出版专著 1 部，专利 3 项；获得科研或教学奖励 5 次。长期从事“双创”教育与“赛教结合”、“产学研训赛”等教学实践研究。指导学生学科竞赛或大创项目，获得省级以上奖励有 20 多项（国家级 8 项）。多次被评为“创新创业优秀指导教师”。</p> <p>一、科研项目</p> <p>1. 陕西省社会科学基金项目：基于高阶社会网络的重大危机事件舆情传播及干预机制研究项目(2022M005)，2022.8- 2024. 8，3/5；</p> <p>2. 陕西省科技厅 2021JQ-765 ,基于优化特征约束的杨官寨遗址的虚拟复原与展示研究, 2021.6-2023.6，2/3；</p>					

	<ol style="list-style-type: none"> <li>3. 陕西省教育厅项目：“基于空间聚类区域经济数据挖掘与建模评价研究”，项目编号：18JK0318，主持；</li> <li>4. 西安市社会科学规划基金项目：“大数据驱动的区域经济评价与决策支持研究”，项目编号：17J08，主持；</li> <li>5. 陕西省教育厅专项科研项目：“带克隆优化的粒子群聚类高维数据分析研究”，项目编号：15JK1274，主持；</li> <li>6. 陕西省科技厅项目：“基于约束聚类和信息度量的软件胎记特征选择研究”，项目编号：2014JM2-6100，主持；</li> <li>7. 陕西省教育厅项目：“基于 n-gram 分割和聚类分析的软件零水印算法研究”，项目编号：11JK0987，主持；</li> <li>8. 陕西省科技厅项目：“基于模糊聚类特征化（胎记）软件零水印算法研究”，项目编号：2011JM8016，主持；</li> <li>9. 陕西省教育厅项目：“基于门限方案的动态图软件水印”，项目编号：09JK441，2/6；</li> <li>10. 陕西省科技厅项目：“基于 Asmuth-Bloom 体系水印算法研究”，项目编号：2009JM8007，2/6。</li> </ol> <p>二、教学研究与建设项目</p> <ol style="list-style-type: none"> <li>1. 陕西省课程思政示范课、示范团队 1 项，陕教高〔2021〕64 号,1/8。</li> <li>2. 陕西省一流本科课程团队 1 项，陕教〔2021〕107 号,1/5。</li> <li>3. 陕西省重点教改研究项目“+数智”赋能电子商务类应用型创新人才培养探索与实践, 21BZ056, 3/5；</li> <li>4. 陕西省高等教育教学改革研究项目：“以“双创竞赛”为牵引的财经类信息技术专业创新能力培养与赛教结合实践探索”,项目编号：17BY070，2017.10-2019.9,1/5；</li> <li>5. 全国高校计算机基础教育研究会教学改革项目：“以计算思维为导向结合“财经类”专业需求的计算机基础教学研究”,项目编号：2016002，2016.7.1-2017.7.1,1/5；</li> <li>6. 陕西省教育厅新工科研究与实践项目，“财经类”院校新工科人才创新能力培养模式与赛教结合实践探索”,陕教〔2017〕497 号 2018.1-2019.12, 1/5；</li> <li>7. 西安财经大学 2016 年度教学研究重点项目：“以多元化“双创大赛”为导向的信息专业大学生创新意识激励与创新能力培养研究” 项目编号：16xcj02，2016.6.1-2017.6.1, 1/5；</li> <li>8. 参与省级创新创业示范学院建设单位 1 项，2019。</li> <li>9. 参与省级高等学校创新创业教育研究与培训基地 1 项，2021。</li> <li>10. 参与教育部产学校外实践基地---达内时代科技基地，2018.7.</li> <li>11. 参与省级实验教学示范中心 1 项——信息技术与电子商务示范中心，2017.5。</li> </ol> <p>三、科研及教学论文代表</p> <ol style="list-style-type: none"> <li>1. 罗养霞,马迪，常言说. PID 参数调节的谱多流形聚类算法研究,计算机科学与探索 2019.8</li> </ol>
--	---

	<ol style="list-style-type: none"> <li>罗养霞, 郭晔. 基于数据依赖特征的软件识别, 《吉林大学学报(工学版)》2017.vol47(11) P1894-1902</li> <li>YangXia LUO. Statistics and Recognition for Software Birthmark Based on Clustering Analysis[J]. Journal of Applied Statistics. 2017, Vol44(2), P308-324. SCI: EL4BX</li> <li>罗养霞. 选择和提升多属性特征检测恶意代码[J]. 小型微型计算机系统, 2016.vol37(6) P1268-1275.</li> <li>YangXia LUO. Malicious Detection Based on ReliefF and Boosting Multidimensional Features[J]. Journals of Communications, 2015.10.</li> <li>罗养霞 房鼎益. 基于聚类分析的软件胎记特征选择[J]. 电子学报, 2013, 41(12): 2334-2338.</li> <li>罗养霞 房鼎益. 基于多属性特征胎记的软件盗版检测[J]. 吉林大学学报(工学版), 2013 Vol 43(05): 1359-1366. Ei (JA) : 20134416914192;</li> <li>罗养霞, 王浩鸣. 财经院校新工科人才创新能力培养研究. 大学教育, 2019.8.</li> <li>罗养霞, 王浩鸣, 杜延庆. 创新能力培养课程改革与体系优化实践探索, 课程教育研究, 2018.11.</li> <li>罗养霞, 李答民. 财经院校软件工程专业理论与实践教学保障体系探析, 教育教学论坛, 2018.11.</li> <li>罗养霞, 王浩鸣, 杜延庆. 创新“文化驱动”与大学生创新意识培养研究. 高教学刊, 2017, 3 (51), 59-61.</li> <li>罗养霞, 冯居易. 以多元化“双创大赛”为导向的大学生创新能力培养研究. 教师, 2017.2. (5), 94-96.</li> </ol> <p>四、专利与专著</p> <ol style="list-style-type: none"> <li>罗养霞, 国家发明专利: 约束限定聚类和信息度量软件胎记特征选择方法、计算机, 专利号 ZL2017.1.1253690.1, 2021.4.13.</li> <li>罗养霞, 实用新型专利 一种基于大数据的信息处理设备, 专利号 ZL2020.2.1334727.0 2021.3.2.</li> <li>罗养霞、张示泽(本科生), 实用新型专利: 一种便携式软件工程用信号接收装置, 专利号 NO.202121829552.5, CN 215488638U, 2022.1.</li> <li>张示泽(本科生)、罗养霞, 一种计算机大数据机箱散热装置专利号 ZL2021.2. 1829552.5, 2022.1.</li> <li>专著: 基于水印和特征的软件保护技术研究, 科学出版社 2015.09, ISBN: 978-7-03-045593-2, 368 千字。</li> </ol> <p>五、获奖</p> <ol style="list-style-type: none"> <li>2021 年校巾帼建功先进个人。</li> <li>2021 年校教学成果特等奖, 财经院校计算机赋能“经管+数智”应用型创新人才培养探索与实践, 2021.5, 3/6。</li> <li>2021 年校优秀共产党员。</li> <li>2020 年校教学成果二等奖, 基于 OBE 理念《数据库原理与应用》1/5;</li> <li>2019 年获校师德先进、优秀教师称号。</li> <li>2019 年校教学成果一等奖, 财经类工程技术专业创新能力培养与赛教改</li> </ol>
--	--

		<p>结合实践探索研究, 2019.5,16, 1/4;</p> <p>7. 2018 年陕西省高等学校科研技术成果三等奖, 基于水印和特征的软件保护技术研究, 2018.3, 1/4;</p> <p>8. 2016 年校第六届科学技术成果一等奖, 基于水印和特征的软件保护技术研究, 2016.1, 1/2;</p> <p>9. 2016 年获得西安财经大学“创新创业优秀指导教师” 西财教发[2016]11 号文件。</p> <p>10. 2015 年校第五届科学技术成果一等奖, 基于聚类分析的软件胎记特征选择, 2015.5 ;</p> <p>11. 2013 年校第三届科学技术成果二等奖, 水印技术在计算机领域中的研究与应用, 2013.5;</p> <p>六、指导学生情况</p> <p>1. 指导国家级大学生创新创业项目 1 项 S202111560018, 张示泽组。</p> <p>2. 指导学生获批省级大创项目 S202011560053, 崔泽豪组 2020.4。</p> <p>3. 指导 2020 年美国数学建模竞赛获得国际一等奖 1 项, 刘卓文组。</p> <p>4. 指导 2020 年国家数学建模, 省二等奖, 宋雪瑶组。</p> <p>5. 指导省级大创项目 S202011560070X, 李向华组。</p> <p>6. 指导 2019 年计算机设计大赛全国三等奖 (指导赵达伟组)。</p> <p>7. 指导 2019 年省级大学生创新创业项目 2 项王丹组、孙浩组。</p> <p>8. 指导 2019 年数学建模竞赛获得省级一等奖 1 项杨朝组。</p> <p>9. 指导 2018 年国家级大学生创新创业项目 1 项, 指导杨森组。</p> <p>10. 指导本科生获得西北赛区计算机设计开发大赛二等奖 4 项; 2019 年计算机设计大赛国家级三等奖 1 项;</p> <p>11. 指导本科生获得 2016 国际数学建模一等奖 1 项;</p> <p>12. 指导本科生获得 2014 和 2015 全国数学建模省级一等奖 2 项;</p> <p>13. 指导研究生获得 2015 年国家数学建模二等奖, 基于多个聚类模型对数据多流行结构分析;</p> <p>14. 指导本科生获得国家级大学生创新训练项目 (201311560008) “提升聚类多属性特征识别迷惑恶意代码”;</p> <p>15. 指导学生获得国家级大学生创新训练项目 (201211560009), “软件特征分析及代码特征分割与识别”;</p> <p>16. 指导本科生 2015 第六届蓝桥杯全国软件大赛省一等奖 1 名, 二等奖 2 名, 三等奖 3 名。</p> <p>17. <b>学生论文:</b> 张雨筱, 罗养霞, 供应链网络建立与最短路径优化问题研究, 现代工业经济与信息化, 2021.12.</p> <p>18. <b>学生论文:</b> 崔泽豪等, 基于数据挖掘的经济销售行为分析, 数字技术与应用, 2020.38(6), 219-220;</p> <p>19. <b>学生论文:</b> 张示泽, 基于灰色预测模型的食品价格变动研究, 现代食品, 2023.7.</p> <p>20. <b>学生论文:</b> 张翰冰等, 太空采矿对全球公平评估模型影响研究, 中国科技人才, 202, 5。</p>
--	--	--

## 一、项目简介（200 字以内）

随着互联网的普及，多态病毒混淆盛行，使病毒数量和复杂性攀升，对网络安全、信息系统造成严重威胁。

项目研究分析病毒多样性、演化路径、威胁方式与分类关联等，构建病毒知识图谱构建方法。项目拟通过模式层指导数据层的方法将知识抽取、融合与存储，构建知识图谱，研究利用 Neo4j 图数据库对知识图谱进行可视化再现，使图谱构建结合人工智能和机器学习，提高智能化的病毒检测，进而提高网络的安全性。项目总体构建流程如图 1 所示。

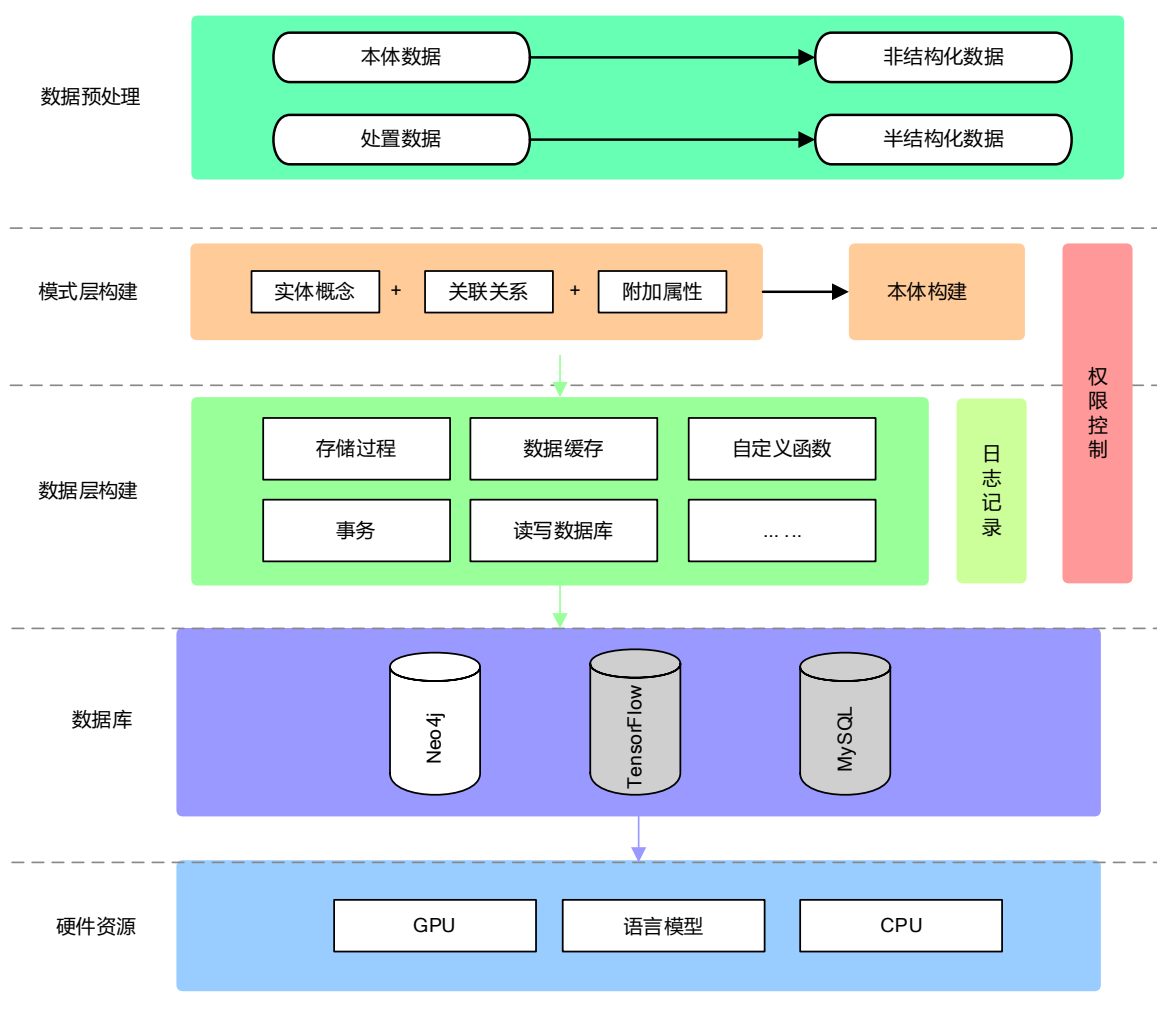


图 1 病毒知识图谱构建流程

## 二、项目相关研究现状及发展动态（不少于 200 字）

### 研究现状：

随着互联网的普及，病毒的数量和复杂性持续攀升。根据 AV-TEST 的最新数据<sup>[1]</sup>显示，到 2023 年病毒数量已增加至 12.3 亿个，不仅数量上呈指数型上涨，形式多以及变种的能力也越来越强。同时，使得识别和揭示病毒分类与多样性、病毒演化路径与关联、特征与家族变种等识别更为困难。随着多态病毒与代码混淆等技术的发展，病毒复杂性增强、构建病毒家族关联，对保障互联网安全、信息安全具有重要意义。

知识图谱<sup>[2]</sup>是一种多边关系图，它由不同实体之间的连接（边）以及这些实体的节点组成。这种图谱依赖于各类相关数据库，并利用 NLP 等相关技术对数据资源进行实体关联和抽取。其通过 Neo4j 图数据库提供可视化和查询功能。Neo4j 将结构化数据存储在网络上，是高性能的 NoSQL（Notonly SQL）图形数据库，是目前知识图谱领域内应用最为广泛的图数据库。

### 发展动态：

随着知识图谱的兴起，网络安全厂商和研究团队也在逐步探索安全知识图谱的构建和应用，面临的挑战与问题主要集中在构建过程的前几个阶段，主要在本体构建、数据收集、信息抽取和知识存储四个部分。

- **本体模型：**定义完备和准确的本体模型，上一章本体建模中提到在确定知识图谱的应用场景、范围和目标后，首先需要安全专家根据规范和约束划分概念，层次结构和关系。对于复杂的应用场景，预定义的本体模型可能不够完备，需要随着信息挖掘的过程逐步完善。
- **数据收集：**从互联网开源数据中自动化获取准确和高质量信息，公开的网络安全数据质量参差不齐，甚至存在由黑客发布的虚假威胁情报数据。
- **知识抽取：**高效准确地从非结构化数据中抽取出实体和关系，主流的机器学习或深度学习模型难以高效抽取。一方面当前没有通用的网络安全语料库，仅靠人工标注的数据量较少，另一方面安全领域专业性强，安全实体和关系类型多样，结构复杂。
- **知识存储：**合理存储图数据需要结合多种存储方法以支持快速或模糊检索，高效知识推理等知识应用。

可以预见，安全知识图谱技术的发展，将全面提升网络安全关键应用场景下的知识推理技术水平，推动安全智能从感知智能，向认知智能和决策智能阶段演进。安全知识图谱目前仍处于蓬勃发展阶段，技术演进仍存在诸多问题需要解决。在此从关键问题着手，如图 2 所示从四个方面展望安全知识图谱技术发展的关键趋势。



图 2 知识图谱技术发展的关键趋势

未来研究将朝向自动化、智能化、实时性与动态性、隐私保护与安全性、跨领域合作以及标准化与互操作性方向发展。自动化和智能化将减少人工参与，提高知识图谱的准确性和智能水平；实时性与动态性将使知识图谱适应快速变化的网络环境；跨领域合作将促进数据共享和技术交流；标准化与互操作性将使知识图谱在不同系统和平台间实现数据交换和协同工作。项目研究病毒知识图谱的构建,进而便于在知识图谱上开展下游任务的延伸，结合人工智能和机器学习，构建相关联的防御体系和提高智能化的病毒检测。

目前在知识图谱的构建和优化中已有不少研究。史慧洋<sup>[3]</sup>等人提出由情报搜索、信息抽取、本体构建和知识推理构建威胁情报的知识图谱构建框架，用于发现攻击者的威胁情报；刘善玲<sup>[4]</sup>提出了基于知识图谱的深层次域名检测方法，从域名、IP 地址、地理位置、解析记录等多个维度抽提构建知识图谱所需的实体信息。Dutta<sup>[5]</sup>等人通过手工注释的 RDF 三元组建立了威胁情报知识图谱，收集了 83 份威胁情报非结构化的数据，构建的知识图谱适用范围较小。当前，已经出现 Cyc、Dbpedia 等依赖专家系统构建的知识图谱以及搜狗知立方等中文知识图谱<sup>[6]</sup>。除此之外 Duoyuan Ma<sup>[7]</sup>、马铎原<sup>[8]</sup>等人构建了 Android API 知识图谱，Lianqiu Xu<sup>[9]</sup>、朱朝阳<sup>[10]</sup>等人构建了病毒行为知识图谱。Aritran Piplai<sup>[11]</sup>等人利用公开网络安全知识图谱进行了病毒分析。



总结以上研究，目前大多是针对恶意代码、API、病毒行为等信息进行病毒分析，并未对病毒族群及病毒分类、安全防御等方面进行探索。

项目研究的特点与创新性: 1.利用病毒知识库进行恶意软件知识图谱的构建，不同其他研究，拟构建的实体为恶意软件名称、发源地、发现时间、相关依赖等信息作为实体，并且将病毒特征、删除指令等详细描述病毒信息文本作为实体属性，具有易提取和识别高效性。2.针对病毒知识库，项目研究 NER 模型：BERT-BILSTM-CRF，通过机器学习训练模型进行恶意软件实体抽取任务，具有智能化检测特点；3.项目拟通过对知识图谱存储更新可以方便安全相关人员进行病毒分析以及对其进行归类，提高利用率和可操作性。

#### 参考文献

- [1] Malware Statistics & Trends Report | AV-TEST[EB/OL]. [2023-12-26]. <https://www.av-test.org/en/statistics/malware/>.
- [2] Singhal A. Introducing the Knowledge graph: Things, not strings[EB/OL]. [2024-01-19]. <https://blog.google/products/search/introducing-knowledge-graph-things-not/>.
- [3] SHI huiyang, WEI jinghuan, CAI xingye, et al. Research on Threat Intelligence Extraction and Knowledge Graph Construction Technology [J]. Journal of Xidian University, 2023, 50 (04) : 65-75.
- [4] 刘善玲. 大数据背景下基于知识图谱的恶意域名检测 [D]. 江苏: 南京邮电大学, 2022.
- [5] Sharmishtha Dutta, Nidhi Rastogi, Destin Yee, Chuqiao Gu, Qicheng Ma. Malware Knowledge Graph Generation[J]. Rensselaer Polytechnic Institute. 2021. 56, 66-72.
- [6] 赵晔辉, 柳林, 王海龙等. 知识图谱推荐系统研究综述[J]. 计算机科学与探索. 2023, 17(4): 771-791.
- [7] Ma, Duoyuan, et al. "A knowledge graph-based sensitive feature selection for android malware classification." 2020 27th Asia-Pacific Software Engineering Conference (APSEC). IEEE, 2020.
- [8] 马铎原. 基于知识图谱的 Android 恶意软件家族分类研究[D]. 天津大学, 2023.
- [9] Xu L, Zhang C, Tang K. A malware analysis method based on behavioral knowledge graph[C]//International Conference on Electronic Information Engineering and Computer Science (EIECS 2022). SPIE, 2023, 12602: 571-579.
- [10] 朱朝阳, et al. "基于行为图谱的恶意代码可视化分类算法." 信息网络安全

全 21.10(2021):54-62.

[11] Piplai, Aritran, et al. "Creating cybersecurity knowledge graphs from malware after action reports." IEEE Access 8 (2020): 211691-211703.

### 三、项目实施的目的、意义（不少于 200 字）

#### 项目的目的：

该项目实施的目的是构建一个病毒知识图谱，以分析病毒的名称、发现地、特征等文本信息，揭示其多样性、演化路径、威胁方式和分类关联。通过知识图谱，可以更有效地理解和追踪病毒的行为，支持网络安全的智能化检测和防御体系构建。

#### 项目的意义在于：

1. 提供了一个结构化的病毒信息存储和分析平台，有助于安全专家快速定位和理解病毒的属性和威胁。
  2. 利用深度学习模型进行命名实体识别，提高了病毒实体识别的准确率，提高了知识抽取的效率。
  3. 知识图谱的构建有助于发现病毒之间的关联，为多态病毒引擎和代码混淆技术带来的识别难题提供解决方案。
  4. 通过 Neo4j 图数据库实现知识图谱的高度可视化，便于直观查询和分析。
- 通过实验验证，该研究对网络安全防御具有实际应用价值。

### 四、项目研究内容和拟解决的关键问题（不少于 300 字）

#### 项目研究内容

##### 4.1 知识图谱的构建内容

项目研究知识图谱的构建过程主要包括知识抽取、知识融合、知识存储、知识更新等主要步骤。在知识抽取环节，实体、关系、属性等要素按需从各类结构化、半结构化、非结构化数据中提取出来。在知识融合阶段，研究各类实体的对齐、关系语义的消歧和知识的映射等工作，将提供满足知识图谱质量要求、设计

规范的数据资料。知识存储阶段，主要是将结构化语义网络数据存储到数据库中，一般的存储介质是各种类型的图数据库。在知识更新阶段，将根据数据层信息的实时性、置信度、语义明确性等维度和更新策略，剔除失效数据，更新最新状态，保证知识图谱信息的高价值属性。

恶意软件知识图谱构建方法包括 3 种：自顶向下、自底向上和两者相结合的方式<sup>[9]</sup>。项目拟采取自顶向下的方法构建知识图谱，首先构建模式层，然后再基于模式层从文本数据中抽取实体构建相应的数据层，如图 3 所示。

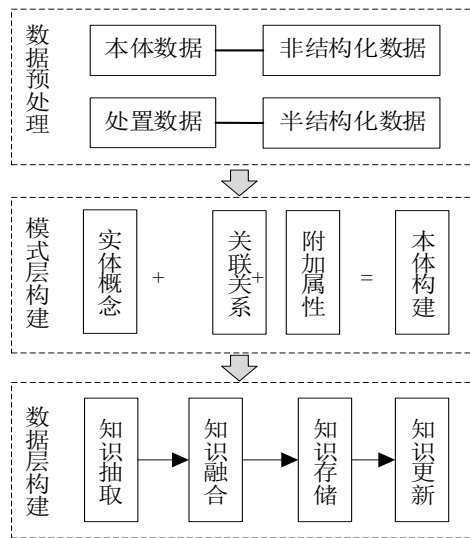


图 3 恶意软件知识图谱构建流程

项目拟利用本体数据和特性描述数据，构建本体知识图谱和处置知识图谱，攻击发生后，安全人员可通过本体属性查询，锁定攻击的方式和病毒、恶意软件的名称，实现快速定位。然后通过本体知识图谱和知识图谱之间的联系，可以快速地找到处置方案以及了解病毒威胁等级及关联。

4.2 构建具有感知、认知、决策智能的安全应用

构建具有感知、认知、决策智能的安全应用，需要研究解决数据的统一建模、实体抽取与关系构建、复杂语义的推理分析和场景化的应用适配等不同层次关键问题。对应这些主要问题，对应本体建模、图谱构建、知识表示和图谱推理等网络安全知识图谱关键技术。如图 4 所示知识图谱的核心框架，对不同来源数据通过关键技术构建知识图谱。

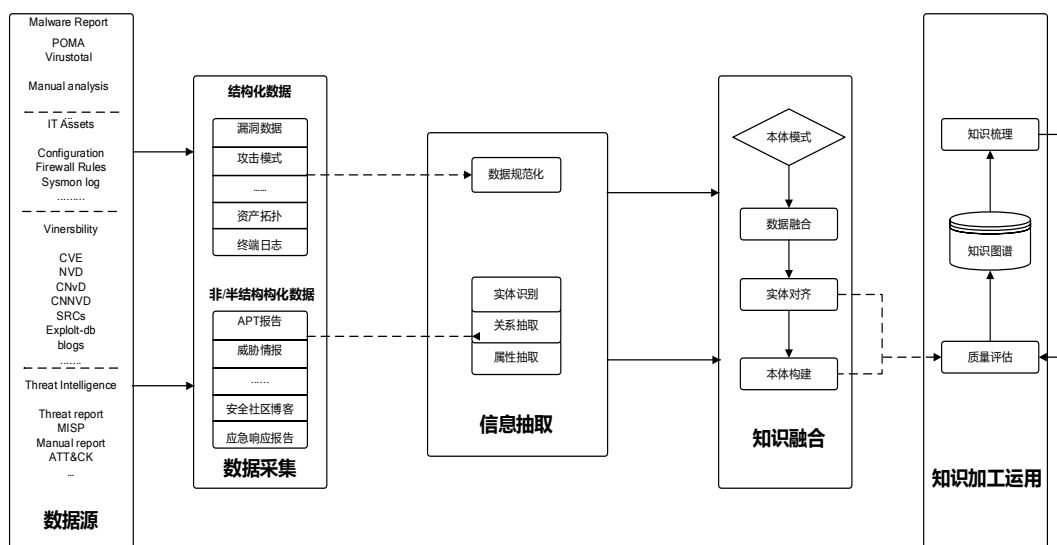


图 4 知识图谱构建关键技术

当前研究主要集中在整合多源数据、表示与存储知识、抽取实体与关系以及应用知识图谱于网络安全任务。这包括从网络流量、病毒样本等数据源提取信息，使用图数据库存储知识，应用规则或机器学习方法抽取实体和关系，并将知识图谱应用于病毒检测、分类等任务。

主要包括以下几个步骤：

### 1.数据预处理

对病毒的结构化和半结构化数据进行实体标识和规范化，包括病毒的名称、别名、发现时间、特征等信息。

恶意软件相关数据可以分为经过安全中心处理后的半结构化数据和公开的未经处理的非结构化数据。针对恶意软件名称、编号、特征等本体信息，因为其数量较多且多为非结构化数据，项目研究病毒领域 NER 模型，对识别实体进行知识加工并整理为三元组。针对发生攻击后采取措施和解决办法等结构化数据，项目拟采取“机器+人工”方式进行人工标注，编写规则采用“关键词+文本”形式。数据分类样式如表 1。

表 1 二源数据

数据名称	数据类型	识别内容
本体数据	非结构化数据	名称、别名、时间、地点等
特性描述数据	半结构化数据	解决办法、危险等级等

2.模式层构建

模式层构建的目标是定义恶意软件领域内三元组内部直接实体以及属性的关系。通过对数据的实体概念、关联关系和属性的表示抽取，形成其对应的本体规则，用于构建知识图谱的模式层。

2.1 本体数据模式层

通过对获取的数据进行分析，得到的实体包括：恶意软件名称、恶意软件别名、特征、首次出现时间、首次出现地点等，根据以上信息建立拓扑关系，抽象出部分实体之间的关联关系，如表 2 所示样式。

表 2 本体及关系

实体 1	关联关系	实体 2
名称	直连	别名
名称	直连	特征
特征	包含	时间
特征	包含	地点
环境	显示	设备类型

2.2 特性描述数据模式层

对恶意软件特性描述数据进行分析，得到的实体有：社会影响、威胁等级、攻击特性、运行环境、传播方式、防御措施等等。再对各实体进行关联关系的分析，抽象出实体间的关联关系，部分安全处置实体和实体间的关联关系，如表 3 所示。

表 3 特性描述数据本体及关系

实体 1	关联关系	实体 3
运行环境	下连	解决办法
威胁等级	下联	波动指标
攻击特性	下联	防御措施
防御措施	包含	解决办法
传播方式	影响	防御措施

3.数据层构建

3.1 知识抽取：拟使用模型对非结构化数据进行实体识别，将病毒的非结构化

信息转化为结构化数据。

(1) 基于 BERT-BiLSTM-CRF 模型的知识抽取

项目拟研究的恶意软件本体数据为非结构化数据，构建 BERT-BiLSTM-CRF 实体识别模型，对其进行知识抽取。通过分析数据的结构特征可知，设定的实体类型有：恶意软件/病毒名称、别名、设备类型、时间、地点、环境，如表 4 所示，抽取模型和算法过程见 3 部分。

表 4 恶意软件实体示例

实体类别	实体示例
名称	火焰病毒
别名	Flamer、Skywiper
时间	2012 年 5 月
地点	中东
设备	Windows 操作系统
环境	以 Lua 和 C++语言写成

(2) 基于多类协同标注的知识抽取

项目拟研究恶意软件特性描述数据多为半结构化数据，利用标注平台+人工标注的方式进行知识抽取，识别的实体为：威胁等级、攻击特性、防御措施、传播方式、社会影响，如表 5 所示。

表 5 火焰病毒实体示例

实体类型	实体示例
威胁等级	★★★★
攻击特性	此类病毒能自主监测并解析自身的网络传输模式，同步实现自动录音功能，记录用户的密码输入行为和键盘按键节奏，并将收集到的数据连同其他关键文件一道，秘密传输至远程控制病毒的服务器端。在完成数据收集使命之后，这类恶意软件还能自我销毁，确保行动痕迹彻底清除。
防御措施	及时更新杀毒软件、不点击下载未了解的邮件链接。
传播方式	通过钓鱼邮件诱骗其点击链接然后进行秘密安装。
社会影响	被认为以军事为目的的攻击，对中东计算机造成大量危害。

### 3.2 知识融合

解决不同实体名表示同一实体的问题,拟通过实体对齐和实体合并实现数据和信息的融合。知识融合指实体链接和知识合并,解决不同实体名表示同一实体的问题。

项目研究在一个统一结构下对源自不同恶意软件类别的异构、多样的知识进行梳理、整合与对应映射。以此达到数据、信息等不同角度的融合,主要分本体融合和数据融合两类。

#### (1) 本体融合

现有研究针对本体间的异质性问题,提出了多元化的解决方案,主要可归纳为两个方向:本体整合与本体映射策略。本体整合旨在将源自多个异构数据源的独立本体体系汇聚为一个统一的整体,实现知识结构的标准化与一致性构建<sup>[16]</sup>。而本体映射则是通过设立一套规则体系,以促进不同本体之间的信息对接和相互转化。

在本研究中,拟采用本体整合方法来融合各个分散的本体资源,该过程涵盖了若干关键步骤,包括但不限于选择合适的整合策略、识别与确认潜在的待整合本体对象、筛选作为整合起点的源本体,以及实际执行整合操作。

另外,在数据层面的知识融合中,实体对齐与实体合并是至关重要的环节。实体对齐作为多源知识融合的核心组成部分,其目的在于有效解决因来源各异导致的实体指向混乱和语义冲突现象,通过精确匹配与对应关系的确立,实现数据源中实体的一致化表述。

#### (2) 数据融合

在数据融合的过程中,知识整合的关键层面涉及实体配对和实体合并两大技术手段。实体配对是多源知识深度融合不可或缺的组成部分,其主要作用在于解决由于数据来源多样化引起的实体指向不协调与冲突矛盾问题,旨在通过精准的实体匹配,确保各源头实体信息的一致性和连贯性。项目对抽取的实体采用 Cosine 相似度算法在词典中进行匹配,得到与抽取实体相似度最高的目标实体,避免在知识图谱构建中重复节点造成不同实体对应相同属性问题。计算流程图如图 2。

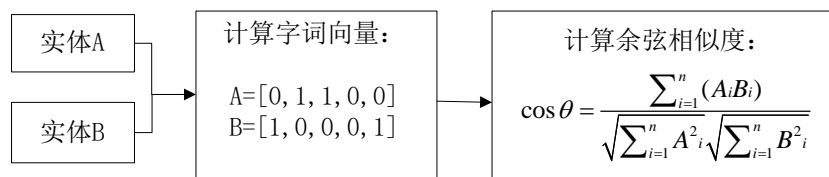


图 5 文本相似度计算流程

### 3.3 知识存储

使用 Neo4j 图数据库存储构建的知识图谱，实现数据的可视化和查询。在知识图谱内，信息以高度结构化的形式储存于庞大的知识库存中。这些知识按照类别划分，遵循规范标准，分别储存在知识库存内各自独立的模块单元，有利于深化知识发掘。项目研究运用了 Neo4j 图数据库技术以承载这些知识，利用图论为基础的搜索算法，并借助直观的 Cypher 查询语言，使得用户无需手动编写复杂的遍历图结构代码，即可轻松实现对恶意软件相关图形化数据的高效存储和便捷查询服务。图 6 Neo4j 图数据库构建两个实体范例图。

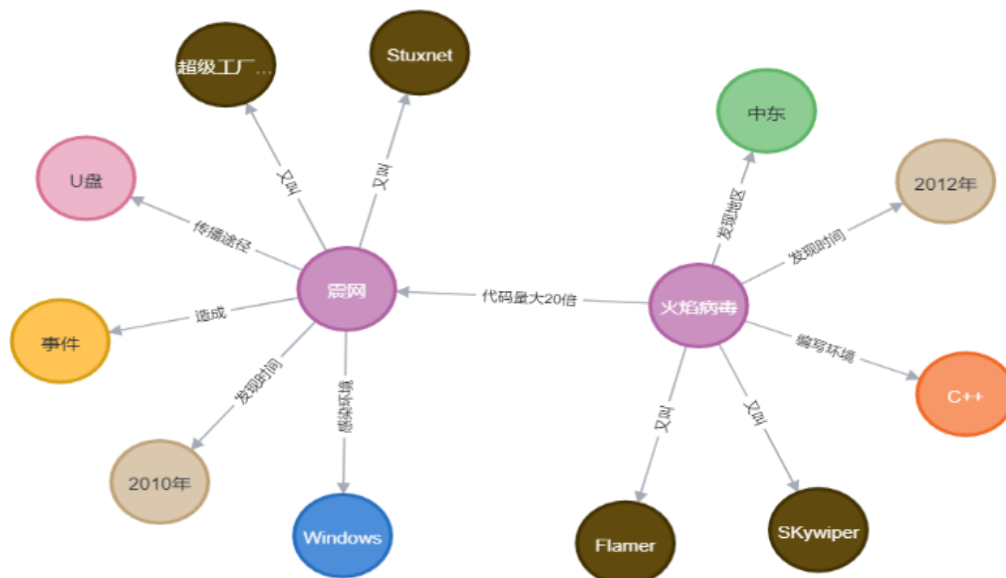


图 6 两个恶意软件节点 Neo4j 范例图

### 3.4 知识更新

随着病毒的增加，对知识图谱进行必要的更新和维护。在知识图谱构建完成后，随着恶意软件和病毒不断地增加，对已构建好的知识图谱实施必要的更新，主要包括模式结构更新与数据内容更新两大部分。模式层的更新是通过增加实体和



关系结构进行完善；数据层的更新，是随着实体和关系数据的不断增多，定期增加新出现的新型实体。

**拟解决的关键问题包括：**

1. 如何有效识别和规范化病毒的非结构化数据，构建实体和属性。
2. 如何设计和训练深度学习模型以提高病毒实体的识别率。
3. 如何构建和维护一个动态更新的病毒知识图谱，以反映病毒的多样性和演化。
4. 如何通过知识图谱实现病毒的快速查询、分析和防御策略的制定。

**五、项目研究与实施的基础条件（不少于 300 字）**

**项目成员的基本条件：**

- **技术能力：**具备扎实的计算机科学基础，熟悉自然语言处理（NLP）、深度学习、知识图谱构建和图数据库技术。
- **编程技能：**熟悉 Python 编程，熟练使用深度学习框架，能够进行模型的搭建、训练和调优。此外，需要了解 Neo4j 等图数据库的操作，用于知识图谱的存储和查询。
- **数据处理能力：**具备数据预处理和分析技能，包括数据清洗、标注（如 BIO 标注）、分割数据集等，以及使用统计学方法评估模型性能。
- **创新与解决问题能力：**能够针对项目中出现的问题提出创新解决方案，比如优化模型结构、改进训练效率或提升知识图谱的质量。
- **团队协作与沟通：**具有良好的团队合作精神，能够与不同背景的成员有效沟通，共同推动项目进展。
- **持续学习：**愿意不断学习新技术和方法，跟进行业动态，因为 AI 和 NLP 领域发展迅速，新技术层出不穷。

**实验所需的基本条件：**

- **硬件资源：**高性能计算设备，包括 GPU（如 NVIDIA V100）和 CPU（如 Intel Xeon），保证模型训练的高效运行。

- **软件环境：**Python 3.7, TensorFlow 1.5 或更高版本, Neo4j 图数据库用于知识图谱的存储, 以及其他必要的库和工具。
- **数据集：**大量病毒相关文本数据, 包括病毒名称、特征等, 用于模型的训练和验证。
- **实验设计：**合理的数据集划分 (训练、验证、测试), 以及明确的评价指标 (如精度、召回率、F1 分数等), 用于衡量模型性能。
- **知识图谱构建与更新：**掌握知识图谱的构建流程, 包括实体识别、关系抽取、知识融合, 以及图谱的更新策略以适应新的病毒信息。

**文档与报告：**撰写项目文档, 包括模型设计、实验结果、数据分析和结论, 以及可能的改进方案, 促进知识分享和项目透明度。

## 六、项目实施方案 (不少于 300 字)

项目计划构建一个病毒知识图谱, 以分析病毒的名称、发现地、特征等文本信息, 揭示其多样性、演化路径、威胁方式和分类关联。试图通过构建知识图谱, 更有效地理解和追踪病毒的行为, 支持网络安全的智能化检测和防御体系构建。

针对目前病毒命名复杂、命名实体识别准确率较低, 以及受到恶意攻击后无法系统性地解决提示等问题, 项目计划提出一种基于 BERT-BiLSTM-CRF 模型构建病毒知识图谱的方法。

其过程概述是: 首先从公开数据集获取病毒相关的非结构化文本资料, 通过 BERT-BiLSTM-CRF 模型处理为结构化标注文本; 其次, 按照自顶向下-模式层指导数据层的方法构建病毒知识图谱和病毒处理方法知识图谱; 随后, 所构建的两个知识图谱被存储于 Neo4j 图数据库中, 并进行了可视化的展示。最后将模型与同类方法进行评价分析。

实施方案流程如图五个部分实施技术路线图, 如图 7 所示。

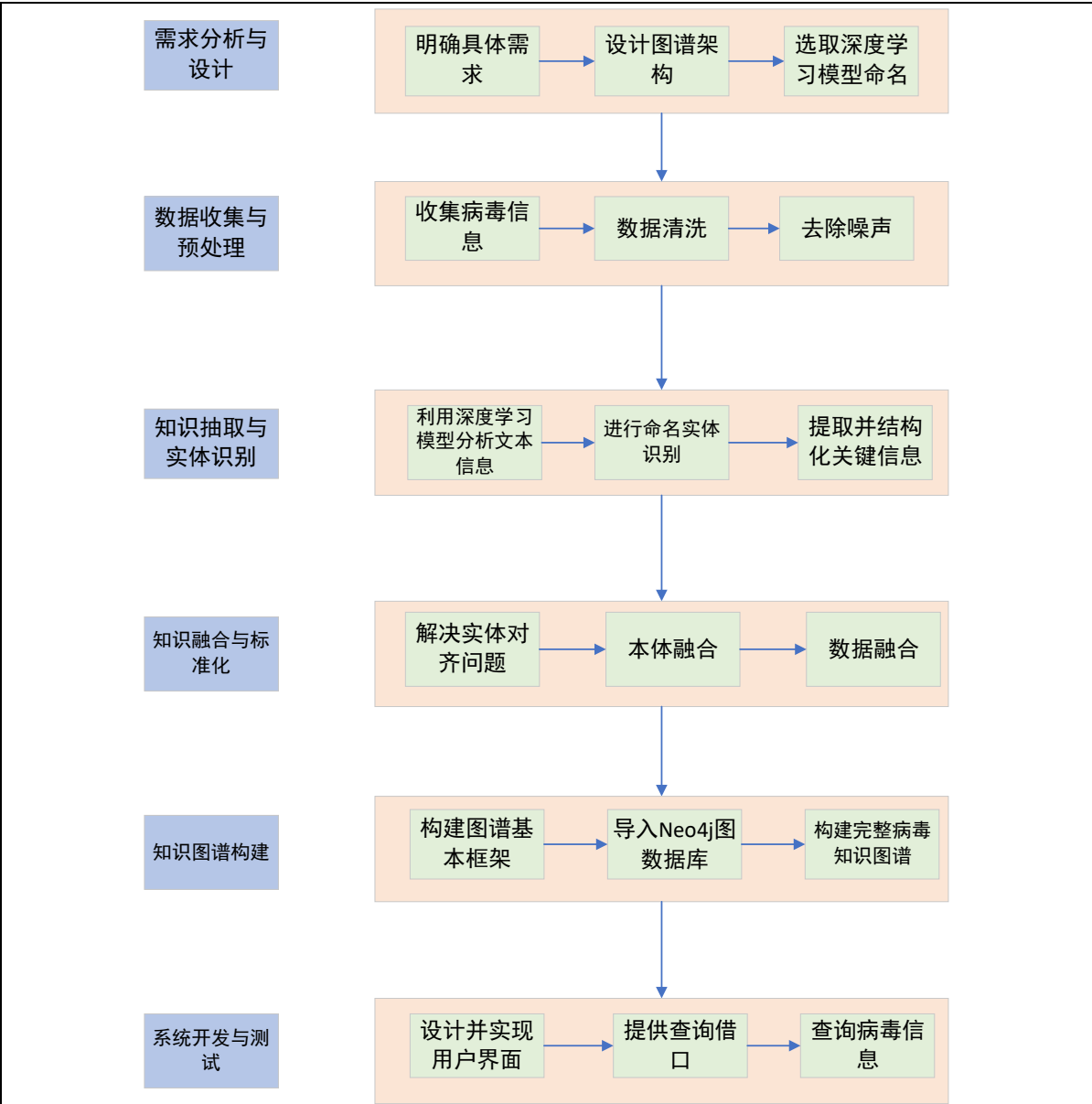


图 7 项目实施技术路线图

具体实施方案如下：

1. 需求分析与设计：明确知识图谱构建的具体需求，包括数据来源、所需抽取的实体类型、属性定义等。设计知识图谱的架构，确定节点类型、边的含义及其关系。选定适合的深度学习模型用于命名实体识别和实体关系抽取。
2. 数据收集与预处理：从公开的网络安全报告、蜜罐系统、反病毒数据库等渠道收集病毒相关信息。对收集到的数据进行清洗，去除噪声，如重复记录、错误条目等，确保数据质量。
3. 知识抽取与实体识别：拟利用预训练的或定制的深度学习模型，如 BERT、CNN 等，对文本信息进行分析，进行命名实体识别（NER），提取病毒名称、

发现地、特征等关键信息，并将其结构化。

知识抽取主要算法如下：

### 1.基于 BERT-BiLSTM-CRF 模型的知识抽取

在知识图谱数据获取阶段，大量的文本数据为非结构化文本。故项目提出安全领域的 NER 模型，利用此模型将非结构数据转换为结构化数据。将文本信息转换为相应的词向量嵌入到模型中进行训练，最后输出文本信息中需要的实体名称。项目所构建的模型如图 8。

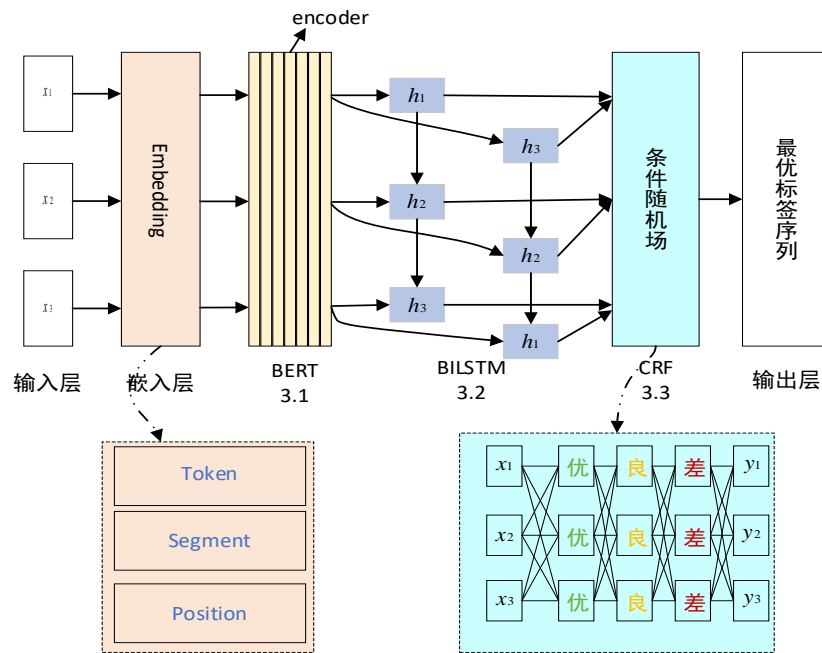


图 8 模型框架图

#### 1.1 BERT 模型

BERT 模型架构基于双向 Transformer 编码器构建，具备同时参考输入序列前后的信息能力，这有助于模型高效捕获文本的上下文信息。在预训练阶段，BERT 采用了“遮蔽语言模型”策略，随机隐藏约 15% 的句子词汇，然后要求模型依据周围环境推测被遮蔽的词语。由此，BERT 模型不仅展现出强大的语义理解与抽取技能，还在实体关系识别上表现出色，尤其擅长解决词汇的多义性问题。BERT 模型生成的词嵌入涵盖三个方面：字符级词嵌入（Token Embeddings）、段落标识嵌入（Segment Embeddings）以及位置嵌入（Position Embeddings）。在这一过程中，首先运用字向量技术，将病毒相关的字符序列映射至一个 768 维的词空间，从而实现对病毒字符的向量化表达。对于句向量

部分，其核心功能在于捕捉并编码句子层面的语义特性，特别是在 **Segment Embeddings** 层面上，系统采用二元标识法进行处理，即首个句子的所有字符对应的词向量均使用全零向量表示，而第二个句子则以全 1 向量来标识其内部字符的词向量，以此区分不同句子间的语义边界。至于位置向量，则着重于凸显同一句子内不同位置字符的特定语义上下文信息。借助 **BERT** 的预训练模型架构，能够有效地从丰富的语言环境中抽取出具有深度语义含义的特征，并最终将病毒相关文本转化为富含语义信息的序列向量表示，效果图如图 9。

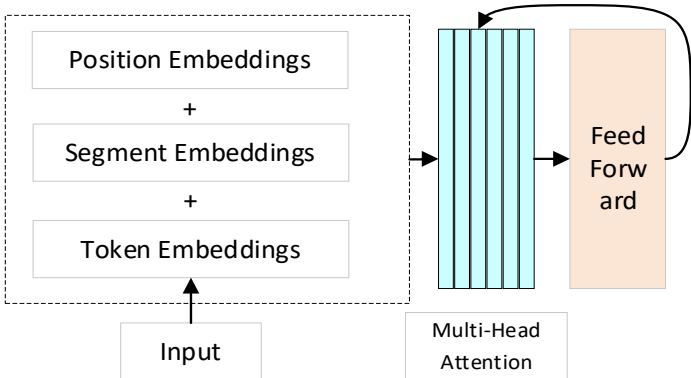


图 9 BERT 模型图

### 1.2 双向长短时间模型（BiLSTM）

为应对循环神经网络训练期间出现的梯度消失和梯度爆炸问题，S. Hochreiter 及合作者提出了长短期记忆（LSTM）结构作为解决方案。该结构创新性地采用了门控设计（Gating Mechanism），通过三个关键的调控组件——输入门、遗忘门以及输出门来智能管理信息流的流动与保留。

在 LSTM 中，输入门负责决定在当前时间步应该让多少新的候选状态信息得以保留并整合到单元状态中。另一方面，遗忘门承担着筛选任务，它根据需要进行选择丢弃上一时间步单元状态中的哪些历史信息。最后，输出门扮演着过滤器的角色，确定当前时间步单元状态中有多少信息应当被输出至网络的下一阶段作为有用信号。

这三个门控组件均采用“软”阈值形式运作，而非硬性的开关，这样能够在连续处理时序数据的过程中更加平滑和精细地调节信息的存储与传播，从而有效解决传统循环神经网络中存在的长期依赖学习难题。即按照一定的概率允许信息通过，计算公式为：

$$i_t = \sigma(W_i X_t + U_i h_{t-1} + b_i) \tag{1}$$

$$f_t = \sigma(W_f X_t + U_f h_{t-1} + b_f) \quad (2)$$

$$o_t = \sigma(W_o X_t + U_o h_{t-1} + b_o) \quad (3)$$

其中  $\sigma(\bullet)$  为 Logistic 函数，其输出区间  $(0, 1)$ ， $x_t$  为当前的环节的输入， $h_{t-1}$  为上一环节的外部状态。

LSTM 引入新的内部状态进行线性循环信息传递，其内部状态  $c_t$  公式为：

$$c_t = f_t \boxtimes c_{t-1} + i_t \boxtimes \tilde{c}_t \quad (4)$$

$$h_t = o_t \boxtimes \tanh(c_t) \quad (5)$$

其中  $\boxtimes$  为向量元素乘积， $\tilde{c}_t$  为候选状态， $c_{t-1}$  为上一时刻的记忆单元。

记忆单元  $c$  中的保存信息的周期长于循环神经网络中隐状态存储的历史信息，然而，相较于神经网络中普遍存在的长期记忆能力，其存储容量有限，故此被称为长短时记忆网络。

项目拟采用双向长短期记忆网络，拼接两个独立的 LSTM，一个用于处理病毒数据正序输入，另一个用于处理病毒数据逆序输入，分别提取正序和逆序的特征，拼接后形成的词向量作为该次的最终病毒图谱特征表达。

## 系统开发与评价

### 1.系统开发与界面

拟设计并实现用户界面，提供查询接口，使用户能够便捷地查询病毒信息、分析关联网络。

### 2.性能评估与优化

拟通过模拟查询、实际应用案例，评估知识图谱查询效率、信息准确度。拟根据反馈调整模型参数，优化算法，提升知识图谱的实用性和效率。评价指标如下：

项目拟选用命名实体识别领域内通用评估方法对项目构建的模型进行评估，评估指标为精确率  $P_{recision}$ 、召回率  $R_{ecall}$  以及  $F_{1\_score}$ 。

$$P_{recision} = \frac{T_p}{T_p + F_p} \quad (6)$$

$$R_{recall} = \frac{T_p}{T_p + F_N} \quad (7)$$

$$F_{1\_score} = \frac{2P_{recision}R_{recall}}{P_{recision} + R_{recall}} \quad (8)$$

其中，正类样本的正确预测记为 $T_p$ ，负类样本的正确预测表示为 $T_N$ ， $F_p$ 指的是实际为负类却被误判为正类的样本， $F_N$ 代表实际情况为正类但不幸被预测为负类的样本。

### 3.持续迭代与反馈循环

拟基于用户反馈和实际应用效果，不断迭代改进知识图谱构建方法、系统功能，形成持续优化的闭环。

## 七、项目创新点及特色（不少于 300 字）

该项目的创新点及特色主要体现在以下几个方面：

1. 病毒知识图谱的构建：项目的核心是构建一个面向互联网安全的病毒知识图谱，它不仅整合了病毒的名称、发现地、特征等基本信息，还深入分析这些元素之间的关联性，以及它们的威胁方式和分类关联。知识图谱的构建可以使病毒行为的模式识别和追踪更为系统化和高效。

2. 多源数据整合与实体关系发现：通过从多种渠道（包括网络安全报告、反病毒数据库等）收集数据，并进行有效清洗与预处理，解决数据异构性问题。知识融合与标准化过程进一步识别并合并不同数据源中的相同实体，减少冗余，保证知识图谱的一致性和完整性。这一过程主要用于揭示病毒之间的隐藏关联。

3. 可视化与智能化分析平台：利用 Neo4j 图数据库，项目可以实现知识图谱的高度可视化，为用户提供一个直观、易用的查询和分析界面。这种可视化不仅帮助用户迅速理解病毒属性和威胁，而且促进信息的快速传播与决策制定，增强应对网络威胁的即时响应能力。

4. 技术综合应用与实践价值：项目综合运用了数据科学、人工智能、图数据库技术，为网络安全领域提供了一种创新的解决方案。通过实验验证，该研究展示了其对实际网络安全防御的直接应用价值，有助于提升智能化检测和防御体系的效能，为对抗日益复杂的网络威胁提供强有力的技术支撑。

综上所述，项目结合了前沿技术和实际需求，通过深度学习驱动的知识图谱构建，为网络安全领域带来了新的分析工具和防御策略，具有显著的创新性和实用性。

## 八、已有基础

### 项目成员的已有基本：

项目成员都是西安财经大学信息学院软件工程专业的学生，对项目中所涉及的Python 编程、Neo4j 图数据库、深度学习基础框架等相关知识都能够熟练运用，喜欢利用所学知识来研究更多的问题，不断探索。我们的每个项目成员都具有独立思考的能力和良好的逻辑思维，懂得相互协作，具有强烈的团队意识。我们对项目充满热情，愿意不断接受挑战，投入到项目的实施中，努力完成项目研究任务。

### 已有的技术基础：

我们所属的学院西安财经学院信息学院专门成立了实践教学中心，目前有电子技术实验室、软件工程技术实验室、电子商务与电子政务实验室、大学物理实验室、校内实习基地等相关实践设施，为学生创新创业工作提供了良好的环境和坚实的基础条件。并且我们团队已经从公开的病毒信息库中获得了病毒的基本信息，如：病毒名称、产生时间、属性等。

### 学校提供的条件：

1. 资金资助和奖励方面：学校对大学生创新创业训练计划项目从资金方面进行配套资助，并奖励创新创业表现优秀的教师和学生，系统的奖励办法已纳入教学制度。
2. 辅助资料来源：学校图书馆可以提供相关资料，同时学校的电子阅览室可以与各大数据库相连接，为资料的查阅提供了便利。
3. 技术交流方面：定期举办校内各项大学生创新创业项目申请，提供和开展交流活动，创建大学生创新创业文化氛围。支持鼓励学生参加校内外学术会议，为学生创新创业提供交流经验、展示成果、共享资源的机会。



## 九、项目研究进度安排及各阶段预期成果

### 1. 项目启动与筹备阶段（2024 年 6 月至 2024 年 7 月）

- 任务：组建项目团队，明确分工，确立研究目标与预期成果。
- 预期成果：项目计划书完成，研究目标明确，预期成果框架设定。

### 2. 需求分析与数据收集（2024 年 7 月至 2024 年 9 月）

- 任务：深入分析病毒领域需求，确定知识图谱所需覆盖的实体与关系类型；收集病毒相关文本数据。
- 预期成果：需求分析报告，数据集初步构建完成，数据预处理方案确定。

### 3. 模型与算法设计（2024 年 9 月至 2024 年 11 月）

- 任务：设计模型结构，定义实体识别与知识抽取策略；选定或设计知识图谱构建框架。
- 预期成果：模型设计文档，算法设计报告，模型初步实现代码。

### 4. 数据预处理与模型训练（2024 年 11 月至 2025 年 2 月）

- 任务：完成数据清洗、标注（BIO 标注）、划分训练/验证/测试集，训练模型。申报软件著作权或实用新型专利 1 项。
- 预期成果：数据集准备完毕，模型训练完成，训练与验证结果分析报告。

### 5. 知识图谱构建（2025 年 2 月至 2025 年 3 月）

- 任务：基于模型输出，进行实体链接、知识融合，构建知识图谱；使用 Neo4j 进行图谱存储与可视化。
- 预期成果：初步知识图谱构建完成，图数据库搭建，可视化展示 1 项。

### 6. 实验验证与评估（2025 年 3 月至 2025 年 5 月）

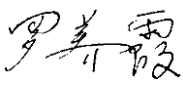
- 任务：对构建的知识图谱进行性能评估，与同类模型对比；调整模型参数优化性能。
- 预期成果：性能评估报告，模型优化方案，与同类模型对比结果。

### 7. 论文撰写与成果总结：（2025 年 5 月至 2025 年 6 月）

- 任务：整理研究成果，撰写论文，准备图表，总结项目经验与教训。
- 预期成果：完成论文，项目总结报告 1 份，成果展示，参加学科竞赛。

### 8. 论文修订与发表：（2025 年 6 月）

- 任务：根据反馈修订论文，提交至期刊/会议，准备演讲或答辩。
- 预期成果：论文发表，会议/期刊接收函，学术交流经验。

十、经费预算		
开支科目	预算经费（元）	主要用途
实验费	8000	构建实验所需的软硬件环境
交流费	5000	调研及外出学习
平台设计费	4000	搭建用户可交互平台
实验外协费	2000	委托外部单位或机构进行服务
其它	1000	
总计		20000 元
十一、导师推荐意见		
<p>1. 项目选题新颖</p> <p>由于互联网的发展，病毒软件形成迅速，对社会影响巨大，且由于各种网络攻击，不利于网络安全。项目研究多学科交叉融合，把网络安全、数据挖掘、知识图谱、深度学习等专业技能相结合，利用技术手段实现对海量的病毒情报信息进行深度挖掘与分析，以快速汇总成知识图谱，以便提高病毒的识别效率，进一步保证网络安全，积极正向思想引导，保持社会稳定，具有重要的意义。</p> <p>2. 项目研究内容充实</p> <p>为实现病毒知识图谱构建及检测，该项目结合“知识图谱+病毒检测”方向，充分利用专业技能，拟设计与实现病毒知识图谱构建及检测系统，通过构建一个病毒知识图谱，以分析病毒的名称、发现地、特征等文本信息，揭示其多样性、演化路径、威胁方式和分类关联，有效地理解和追踪病毒的行为，支持网络安全的智能化检测和防御体系构建。充分结合专业技能，内容详实。</p> <p>3. 项目研究方案及实施的可行性</p> <p>该项目研究针对明确，提出自己的方案技术路线清晰，时间安排合理，符合创新训练项目的要求。而且项目的实施不管是从团队合作、技术支持，还是实验条件都会得到坚实的支援，可行性及可操作性强。</p> <p>因此，同意并支持小组项目申请。</p> <div>签名： </div> <div>2024 年 5 月 20 日</div>		

十二、院系推荐意见	
院系负责人签名：	学院盖章 年 月 日
十三、学校推荐意见	
学校负责人签名：	学院盖章 年 月 日
十四、省教育厅评审意见	
	单位盖章 年 月 日

注：表格栏高不够可自行增加。