资料提交

提交作品说明如下：

作品名称：个人隐私信息泄露风险评估系统

英文名称：Personal privacy information leakage risk assessment system

作品介绍：

随着大数据技术和人工智能的迅速发展，数据的创建量呈指数级增长。数据的应用场景也愈加广泛，这使得用户不再享有对数据的绝对控制权，数据的安全性受到威胁，从而频发个人隐私泄露事件。数据安全是数字经济发展的压舱石，是国家安全的重要组成部分，数据安全已成为事关国家安全和经济社会发展的重大议题。为了能够对数据内容的潜在隐私风险做出评估，设计了个人隐私信息泄露风险评估系统，针对多源数据进行个人信息层面的泄露风险评估，用于保证数据安全，此前提下充分发掘数据要素价值及推动生产力发展。

去年国家发布实施的《中华人民共和国网络安全法》《中华人民共和国数据安全法》《中华人民共和国个人信息保护法》彰显了我国高度重视数据安全保护工作的相关法规，以法为基准指定风险泄露标准，帮助发现数据存在的风险隐患，能够有效预防信息泄露情况的发生。此系统面向个人隐私保护领域，符合A-ST专项指南范围。针对现有的检测工具（例如Firefox monitor）依据多样但不明确，且检测的数据部分来自存储着已泄露数据的数据库，并不能反映个人隐私信息的泄漏情况的问题，系统收集指定目标的多源数据，可以根据不同的数据隐私需要，匹配不同事物数据项，并调整不同数据项的风险等级。系统相较于其他检测方案更加符合国家法律法规，例如在新冠肺炎流调数据个人隐私泄露检测的应用场景中，系统可以在数据发布前做出风险评估，数据发布方可以根据评估结果对数据做出适当修正，从而减少隐私泄露情况的发生。

系统针对重要待检测目标，从数据源中利用深度学习、文件解析、分布式爬虫等技术获取数据，提取可能伴随隐私泄露风险的数据并归类。随后通过检测模型对收集的数据进行泄露风险评估，模型对多条数据类型进行建模，结合基于人工经验的固定规则和基于机器学习的关联规则算法推断，实现对风险等级的科学评估。风险评估模型以现有的法律法规为依据，根据不同数据之间的关联程度最终给出评估结果，也可以根据用户需求进行调整，满足不同源数据存在的差异化风险需求。并根据实际网络部署情况，分布式部署爬虫和检测程序，在一定程度上提高了系统运行的效率和稳定性，帮助解决了数据孤岛带来了检测问题，加强了数据检测过程中的安全性。

提交文档内容如下：

项目介绍

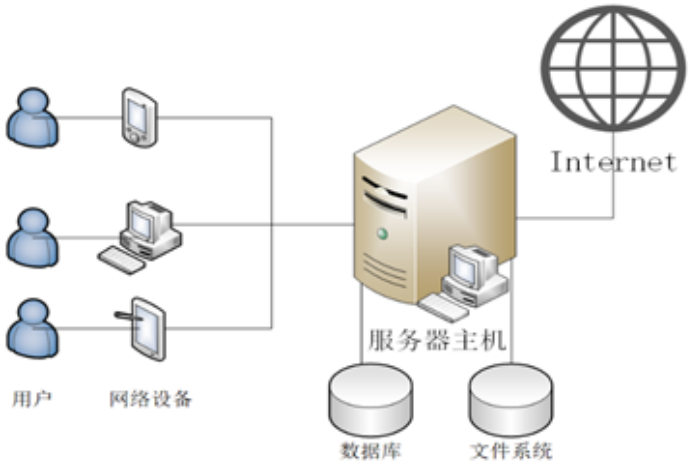
随着大数据技术和人工智能的迅速发展，尤其是5G技术的普及， 数据的创建量呈指数级增长。信息技术的发展带动人类社会发生重大变革，人们的衣食住行、健康医疗等信息都被数据化，信息数字化的应用场景也愈加广泛。这就使得用户不再享有对数据的绝对控制权，数据的安全性受到威胁，从而频发个人隐私泄露事件。2020 年国务院颁布的《关于构建更加完善的要素市场化配置体制机制的意见》进一步提出，加快培育数字要素市场，充分挖掘数据要素价值，数据要素已成为数字经济深化发展的核心引擎，数据安全也成为事关国家安全和经济社会发展的重大议题。党中央、国务院高度重视数据安全工作，数据安全是数字经济发展的压舱石，是国家安全的重要组成部分，已成为世界主要国家战略布局重点，去年发布实施的《数据安全法》《个人信息保护法》与《网络安全法》彰显了我国高度重视数据安全保护工作的重视程度。为此，能够对数据内容的潜在隐私风险做出评估，是在保证数据安全的前提下，充分发掘数据要素价值以及生产力的重要前提。

设计一套能够评估特定内容中是否存在信息泄露风险的系统，帮助发现数据存在的风险隐患，能够有效预防信息泄露情况的发生。现有的检测工具依据多样但不明确，且检测的数据部分来自存储着已泄露数据的数据库，并不能反映个人隐私信息的泄漏情况，针对此问题，系统收集指定目标的多源数据，可以根据不同的数据隐私需要，匹配不同事物数据项，并调整不同数据项的风险等级。系统基于最近的《中华人民共和国网络安全法》《中华人民共和国数据安全法》《中华人民共和国个人信息保护法》等法规对个人隐私数据做出的定义和判定，相较于其他检测方案也更加符合国家法律法规。例如在新冠肺炎流调数据个人隐私泄露检测的应用场景中，系统可以在数据发布前做出风险评估，数据发布方可以根据评估结果对数据做出适当修正，从而减少隐私泄露情况的发生。

系统针对重要待检测目标，从数据源中利用深度学习、文件解析、分布式爬虫等技术获取数据，提取可能伴随隐私泄露风险的数据并归类。随后通过检测模型对收集的数据进行泄露风险评估，模型对多条数据类型进行剑建模，结合基于人工经验的固定规则和基于机器学习的关联规则算法推断，实现对风险等级的科学评估。风险评估模型以现有的法律法规为依据，根据不同数据之间的关联程度最终给出评估结果，也可以根据用户需求进行调整，满足不同源数据存在的差异化风险需求。

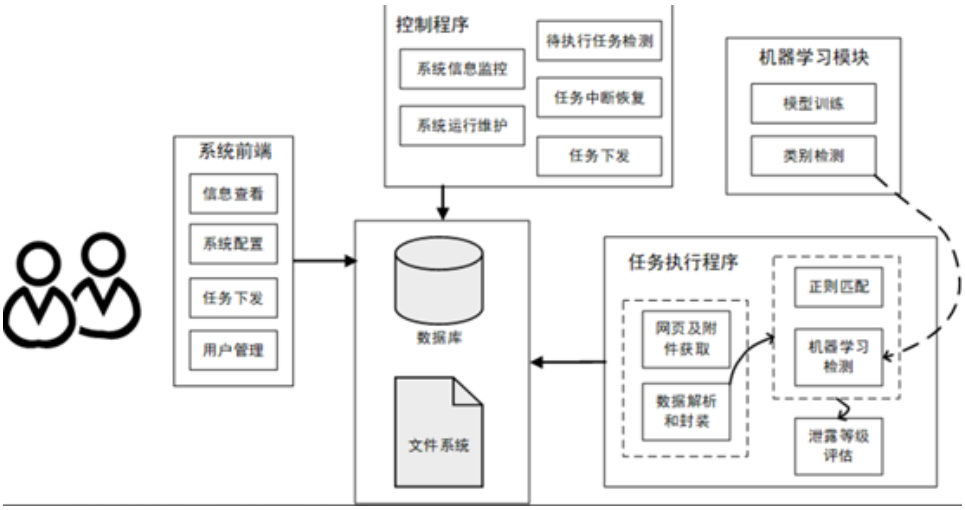
开发环境的配置

需求名称	详细要求
系统	Centos7
数据库	MySQL 5.6MySQL Community Server
Web服务器	Gunicorn
前台语言环境	HTML、JavaScript
后台语言环境	Python 3.8
磁盘	40GB
内存	4GB



网络拓扑图

系统运行于服务器上，为用户提供网页作为系统外部接口，用户需要拥有可访问服务器的网络设备才能使用系统，通过浏览器访问系统提供的页面，用户可以使用系统功能。因为系统涉及到爬虫对网页的扫描，所以服务器主机需要连接到互联网。提供网页的前台程序和负责业务逻辑的后台程序，以及系统的数据库、文件存储系统部署在同一台服务器主机上。

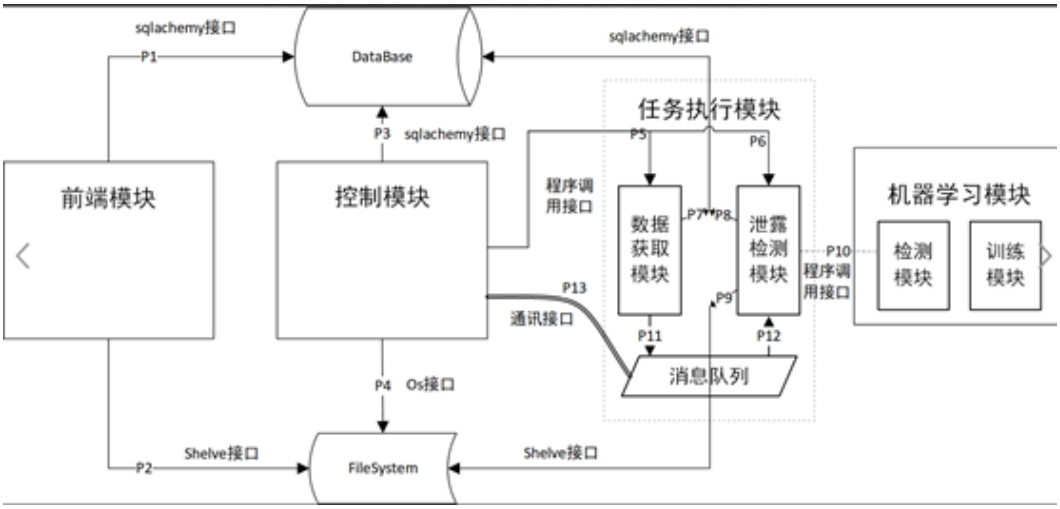


系统实现框架图

系统基于B/S框架，由系统前端、后台控制程序、任务执行程序、机器学习模块以及相应的数据存储，包括数据库和文件系统共同组成。

系统的**前端**部分只起到记录用户操作和向用户展示信息的功能。对用户操作的处理逻辑都由系统的**控制程序**完成，后台控制程序是独立于系统前端的管理程序，控制程序负责检测系统中是否有待执行任务并持续更新和维护系统信息。如果控制程序检测到待执行任务，就会调用任务执行程序完成任务。**任务执行程序**的工作是完成系统的扫描任务，并将扫描结果保存到数据库，任务执行分为**数据爬取**和**泄露检测**两部分，泄漏检测需要进行数据信息的归类和是否构成写泄露的判断，数据爬取由爬虫完成，在爬虫获取到网页数据后，分别由**正则匹配**和**机器学习检测**敏感信息内容以及类别，再由泄露判断部分完成对是否构成泄露的分析。机器学习模型分为训练和检测两部分，训练部分用于模型检测自身的构建，由人工利用数据集进行训练，不对系统检测程序可见，检测程序只调用机器学习模块的检测功能，用于判断符合的敏感信息种类

系统的核心功能是完成一次检测任务，网页监控管理功能可以视为周期性完成同一件扫描任务，而临时扫描任务则是仅完成一次扫描，系统辅助测试则是完成一次简单的测试扫描，不同任务的区别只在任务添加方式和使用爬虫上有区别，在执行过程上都相同。

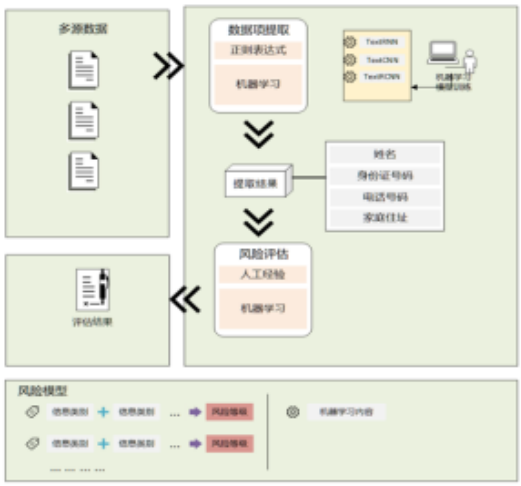


系统模块结构图

模块划分	模块描述
前端模块	提供用户界面和系统对应用户功能的外部接口
控制模块	调度任务和系统信息维护
任务执行模块	爬取网页和对内容进行正则匹配，并将结果数据保存到数据库和文件
机器学习模块	对机器学习模型进行训练，并为任务执行模块提供机器学习检测

模块划分表

前端模块通过接口P1将用户操作写入数据库，取出数据库中信息展示给用户，通过接口P2取出文件系统中储存的网页解析内容。控制模块通过接口P3操作数据库，包含任务调度、清理定期检测记录、更新系统运行信息等。通过接口P4对本地储存的临时文件进行定期清理。通过接口P7下发任务，调用任务执行模块进行任务的执行。任务执行模块是爬虫脚本，负责具体任务的执行，对爬虫下载的文件和网页文本进行正则匹配，通过接口P5储存检测结果和修改任务调度信息，通过接口P6储存用于正则匹配的文本内容。



系统流程图

系统针对特定的检测目标获取到多源数据后，经正则表达式和机器学习将敏感的目标数据提取，随后将提取结果经过风险评估模块得出评估结果，具有一定的警示作用。

关键问题	主要依赖技术	简述
网页数据获取	分布式爬虫，Redis，Selenium，Chrome-WebDriver	爬虫需要获取到网页内容和解析网页附件。
数据正则匹配	正则表达式	匹配文本中符合指定正则表达式的字符串。
机器学习检测	机器学习（PyTorch框架），词嵌入（Word Embedding），循环神经网络（Recurrent Neural Network，*RNN*）	利用PyTorch和相关数据集训练泄露信息检测模型，利用模型进行泄露信息的检测。

实现效果

系统首页图

序号	任务...	检测范围	任务开始时...	任务结束时间	检测页...	检测文件数...
1	repl...	cstc.hrbeu.edu.cn	2022-04-23 ...	2022-04-23 16:33	965	82
2	计算...	cstc.hrbeu.edu.cn	2022-04-23 ...	2022-04-23 16:12	863	66
3	班主任	http://cstc.hrbeu.ed...	2022-01-03 ...	2022-01-03 09:47	1	
4	演示	cstc.hrbeu.edu.cn	2021-12-21 ...	2021-12-21 22:16	483	33

系统检测任务图

序号	风险级别	网址	操作
1	high	http://cstc.hrbeu.edu.cn/bzrgz/list.htm	查看任务信息 转到原网页
2	low	http://cstc.hrbeu.edu.cn/2021/1203/c3688a2796...	查看任务信息 转到原网页
3	low	http://cstc.hrbeu.edu.cn/2021/1203/c3688a2796...	查看任务信息 转到原网页
4	medium-low	http://cstc.hrbeu.edu.cn/2021/1125/c3688a2791...	查看任务信息 转到原网页
5	medium-low	http://cstc.hrbeu.edu.cn/2021/1111/c11965a278...	查看任务信息 转到原网页
6	medium-low	http://cstc.hrbeu.edu.cn/2021/1116/c11965a278...	查看任务信息 转到原网页
7	medium-low	http://cstc.hrbeu.edu.cn/2021/1116/c11965a278...	查看任务信息 转到原网页
8	medium-low	http://cstc.hrbeu.edu.cn/2021/0419/c11973a267...	查看任务信息 转到原网页
9	medium-low	http://cstc.hrbeu.edu.cn/2021/0417/c6988a2675...	查看任务信息 转到原网页
10	medium-low	http://cstc.hrbeu.edu.cn/2021/0322/c6988a2667...	查看任务信息 转到原网页
11	medium-low	http://cstc.hrbeu.edu.cn/2021/0322/c6989a2667...	查看任务信息 转到原网页
12	low	http://cstc.hrbeu.edu.cn/3687/list2.htm	查看任务信息 转到原网页
13	low	http://cstc.hrbeu.edu.cn/2021/0608/c3688a2715...	查看任务信息 转到原网页
14	low	http://cstc.hrbeu.edu.cn/2021/0413/c11984a267...	查看任务信息 转到原网页
15	medium-low	http://cstc.hrbeu.edu.cn/2019/0916/c11983a264...	查看任务信息 转到原网页

系统评估结果图