基于分布式爬虫的高性能 Tor 网络内容监控系统

郑献春[1], 王 瑞[1], 闫皓楠[1], 赵兴文[1], 李 晖[1,2], 李凤华[3,4]

[1]网络与信息安全学院 西安电子科技大学 西安 中国 710126
[2]综合业务网理论及关键技术国家重点实验室 西安电子科技大学 西安 中国 710071
[3]信息安全国家重点实验室 中国科学院信息工程研究所 北京 中国 100093
[4]网络空间安全学院 中国科学院大学 北京 中国 100049

摘要 随着网络的发展和普及, 人们对于安全性、匿名性、反审查等信息安全的需求快速增强, 越来越多的人开始关注和研究 Tor 匿名通信网络。目前针对 Tor 网络内容监控的研究工作大部分存在功能少、性能弱等劣势, 如缺乏为暗网设计的专用爬虫, 网络连接速度较慢, 本文设计开发了一套综合性的 Tor 网络内容动态感知及情报采集系统, 包含数据采集爬虫以及网页内容分类两个部分。其中爬虫部分使用了分布式架构, 包括了任务管理模块、爬虫调度模块、网页下载模块、页面解析模块、数据存储模块, 同时创新性地优化了 Tor 连接链路以提高爬取速度和稳定性; 网页内容分类部分使用了自然语言处理技术, 建立训练模型并对抓取到的信息进行精准高效分类, 解决分类的准确度和复杂性问题, 最后根据结果分析暗网的内容结构和敏感信息。我们也相应地为保障系统运行设计了容错模块和预警模块, 从而对系统各个组件的当前状态进行实时监控, 并将系统的状态数据进行整合、收集和展示。最后我们将该系统放到了实际 Tor 网络环境中进行了测试, 从系统网页爬取效果、内容分类效果及系统性能等三方面进行了评估和分析, 并与国内外 7 中现有的框架的功能进行了对比, 证明本文提出的方案在暗网域名、网页、数据爬取的量级和速度性能方面均为最佳。

关键词 洋葱路由; 暗网; 爬虫; 自然语言处理
中图法分类号 TP391 **DOI 号** 10.19363/J.cnki.cn10-1380/tn.2023.01.11

A High Performance Tor Web Content Monitoring System Based on Distributed Crawlers

ZHENG Xianchun[1], WANG Rui[1], YAN Haonan[1], ZHAO Xingwen[1], LI Hui[1,2], LI Fenghua[3,4]

[1] School of Cyber Engineering, Xidian University, Xi'an 710126, China
[2]State Key Laboratory of Integrated Services Networks, Xidian University, Xi'an 710071, China
[3] State Key Laboratory of Information Security, Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100093, China
[4] School of Cyber Security, University of Chinese Academy of Sciences, Beijing 100049, China

Abstract With the development and popularization of the network, people's demands for information security such as security, anonymity, and anti-censorship are rapidly increasing, and more and more people begin to pay attention to and study the Tor anonymous communication network. At present, most of the research work on Tor network content monitoring has disadvantages such as few functions and weak performance. For example, there is a lack of a dedicated crawler designed for the dark web, and the network connection speed is slow. This paper designs and develops a comprehensive set of Tor network content dynamic perceptions, and intelligence collection system, including data collection crawler and web content classification. The crawler part uses a distributed architecture, including task management module, crawler scheduling module, web page download module, page parsing module, data storage module, and innovatively optimizes the Tor connection link to improve the crawling speed and stability. The web content classification part uses natural language processing technology to establish a training model and classify the captured information accurately and efficiently, to solve the problem of classification accuracy and complexity, and finally analyze the content structure and sensitive information of the dark web according to the results. We also designed a fault-tolerant module and an early warning module to ensure the operation of the system, to monitor the current status of each component of the system in real-time, and integrate, collect and display the status data of the system. Finally, we put the system into the actual Tor network environment for testing, and evaluated and analyzed from three aspects of system web page crawling effect, content classification effect, and system performance, and compared with the functions of existing frameworks at home and abroad. A comparison is made, and it is proved that the scheme proposed in this paper is the best in terms of the magnitude and speed performance

通讯作者: 李晖, 博士, 教授, Email: lihui@mail.xidian.edu.cn。
本课题得到国家自然科学基金重点项目(No. 61732022), 公安部技术研究计划(No. 2019JSYJA01), 陕西省自然科学基金项目(No. 2019ZDLGY12-02), 陕西省创新团队(No. 2018TD-007)的资助。
收稿日期: 2021-09-24; 修改日期: 2022-01-18; 定稿日期: 2022-11-03

of dark web domain names, web pages, and data crawling.

Key words tor; dark net; crawler; natural language processing

1 引言

匿名网络可以保证公民在阅读网页、银行存取和网上购物时的个人隐私, 也能够为外国间谍、恐怖分子的不法行为提供保护伞。近几年, 暗网中不断有数据泄露事件被爆出, 使得这个犯罪信息汇聚和违法交易横行的“不法之地”逐渐走入大众视野。

暗网整体信息的研究表明, 截至到 2019 年 5 月暗网公开节点总数有近万个[1], 且节点的整体分布主要集中在互联网发达的西方国家。Hurlburt 等人[2]对超过 1000 个隐藏服务样本进行的一项研究中表明, 68%的暗网内容是非法的。Zulkarnine[3]在近 5000 个 Tor 网站中确定了其中的 1547 个含有非法内容。

国内外相关机构以及政府组织近年来都开始把目光高度凝聚在暗网上, 有关的研究内容也有不少, 但是都存在一定的局限性。明网相关的爬虫研究[4-6]由来已久, 但能够应用于暗网网络的并不是很多。许多研究人员研究了明网网站内容的分类以及深网的分类[7-10], 而暗网分类的相关研究[11-12]仍处于早期阶段。相较于传统明网场景下的网页内容分类, 暗网爬虫应针对暗网独特的网络环境进行针对性的设计, 需要将系统网络接入暗网当中, 并且保证连接的快速、稳定运行。此外 Cloudflare 的一项研究[13]表明目前暗网主要是通过 Tor 协议进行网络访问, 约占 94% 以上, 因此本文从监控暗网内容角度出发, 研究的主要对象是 Tor 网络。

随着目前暗网技术的更迭, 现有暗网爬虫存在爬取速度慢、单机运行性能弱、获取数据量少的缺点, 为了更好地维护国家网络环境, 发现暗网中的非法内容, 本文从暗网情报的采集与动态感知等角度出发, 设计了一个暗网情报系统, 主要解决了以下几个问题: 暗网信息的爬取、数据的存储、对爬取信息的分析及分类、保持系统稳定持久运行。系统主要功能是对暗网中的内容进行爬取和索引, 并对爬取的内容进行分类, 达到精确定位非法信息的效果。然后我们利用该系统在实验中收集了一万多条暗网网站分类的数据, 并实现了对暗网非法内容的精准分类。

本文的主要贡献有: 1)建立了一套基于分布式爬虫的高性能 Tor 网络内容动态监控系统, 能够对大量的暗网信息进行爬取, 配套开发了系统保障模块, 可监控各个组件的运行状态并提供自动化的解决方案; 2)基于自然语言处理研究并实现了一套暗网网站内容处理算法, 用来去除网站分类中的干扰信息, 实现了对暗网网站内容的分类; 3)对链路进行了优化, 减少每次建立 Tor 链路的中继节点数, 提高了网络连接的速度及稳定性。

本文的文章结构做如下安排: 首先介绍了 Tor 网络和暗网爬虫的相关工作, 接着分别从整体设计和模块化实现对我们的内容监控系统进行了介绍, 然后分析评估了实际 Tor 网络工作环境中爬虫的性能和内容分类的效果, 最后展望了未来的研究方向并对全文进行了总结。

2 相关工作

2.1 Tor 基本原理

Tor(The Onion Router, 洋葱路由)是一个自由软件, 致力于保护其用户的隐私。Tor 的网络流量通过一些志愿操作的服务器(也称为“节点”)进行引导, 网络的每个节点都对其盲目传递的信息进行加密, 不登记流量的来源和流向。Tor 建立的目的之一就是帮助用户隐藏自己, 通过其志愿节点请求网页或其他数据, 使用 Tor 路由将请求内容发送给用户, 从而增加了流量跟踪和审查的难度[14]。

2.2 暗网爬虫

网络爬虫是一种软件或编程脚本, 能够系统化、自动化地浏览网页内容。常规爬虫通常从一个受欢迎的网站开始, 索引其网页上的文字并跟踪网站中找到的每个链接, 而暗网爬虫不同于常规的爬虫, 需要根据暗网独特的网络环境进行针对性的设计。暗网爬虫的工作流程为: 先搜集一定数量的 URL 作为初始爬取的基础, 这些初始的 URL 主要来自于明网中公开的 Onion 地址列表、暗网搜索引擎等线索, 初始数据规模为 139 个; 将待爬取的 URL 发送至爬取队列, 爬虫解析 URL 后调用网页下载模块下载网页, 使用页面解析模块提取网页文档中的 URL, 将解析的结果传输给内容存储模块进行存储。整体流程如图 1 所示。

Moore 等人[11]通过暗网网站爬虫收集并分析了 5000 个洋葱域名, 发现 Tor 的隐藏服务最常见的用途是犯罪和非法活动, 包含毒品、武器和各种色情内容。随着 Tor 网络的不断更新换代以及大众对暗网越来越多的了解, 会有更多的人接触使用 Tor 网络。

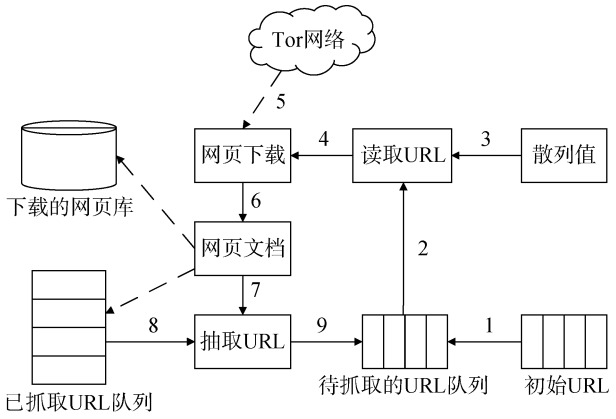

图 1 Tor 爬虫基本流程

Figure 1 Tor crawler basic process

Portsmouth University 的研究人员[15]在 Tor 网络中运行了 40 个中继节点以达到探测 Tor 网络中流量流向以及获取部分 Tor 网络数据的目的, 这项研究探测到了超过 4.5 万个暗网节点。2006 年, Overlier 等人[16]的研究描述了一种通过运行具有 HSDir 标志的中继节点来探测暗网隐藏服务的方法, 这项研究至今仍然有效, 殷帅[17]在此基础上进行了基于 PageRank 和描述符查询行为的流行性计算。Zulkarnine[3]在研究中设计并开发了一种暗网爬虫, 它基于预定义的关键词和图像散列值进行网站分析, 以当时两个流行的暗网搜索引擎 Ahmia 和 Onion City 为出发点, 获取了大约 5205 个以.onion 结尾的 Tor 网站, 其研究开发的爬虫在每个站点上最多深入 5 层, 最多不超过 100 个网址, 经过一段时间的爬取之后, 研究组在 Tor 网络中访问了大约 300000 个地址, 以 205000 个唯一页面的形式产生了高度多样化和重要的数据语料库。于浩佳等人[18]在 2017 年有过对于暗网信息爬取的相关研究, 但系统只是单机运行, 且使用的 klassify 文本分类器需要大量的人工操作, 性能存在明显的上限。

2.3 暗网页面内容分析

Tor 网络中隐藏着大量的非法信息, 针对 Tor 网络内容进行分析, 可精确识别网站的类型和主要内容, 有助于提高 Tor 网络的违法内容监控能力, 达到精准定位非法信息的效果。

关于页面内容分析的研究, Kaur 等人[10]在 2014 年引入了几种算法来对 Web 内容进行分类, Kan 等人通过解析和分割特征来提取特征探讨了在 Web 分类中使用统一资源定位器(URL)[7-8], 但是上述技术并不能应用到对 Tor 网络隐藏服务的分类中。Sun 等人[19]在 2002 年采用支持向量机(SVM)实现了利用上下文特征对 web 内容进行分类的功能。关于暗网, Moore 等人[11]在 2016 年提出了一项基于 Tor 隐藏服务对暗网进行分析和分类的新研究。

TF-IDF(Term Frequency–Inverse Document Frequency)是一种信息加权方案, 可根据文档的术语频率(TF)和反向文档频率(IDF)为文档中的每个术语分配权重, 具有较高权重分数的术语被认为更重要, 因而 TF-IDF 常被用来表示信息检索和文本挖掘的权重。Graczyk 等人[12]提出了一种使用 TF-IDF[20]进行文本特征提取的管道架构, 该架构将暗网上著名的黑市 Agora 的产品分为 12 类, 准确率能够达到 79%。Noor 等人[21]讨论了用于从暗网数据源中提取内容的常用技术, 称为“查询探测”, 这种技术基于监督学习算法以及“可见表格特征”[22]。Barbosa 等人[23]提出了一种无监督的机器学习聚类管道, 其中文本频率逆文档频率(TF-IDF)用于文本表示, 余弦相似性用于 k-means 的距离测量。

3 系统设计

本文提出了一个分布式暗网内容监控系统, 可针对 Tor 网络进行高性能的信息爬取监控工作, 且系统对链路连接速度和连接稳定性进行了优化, 结合文本分析技术对爬取信息进行了精确分类, 解决了已有研究的功能和性能问题。

大数据技术和工具的选用支撑了海量暗网网站内容数据的存储、分析和监控等功能, 因此本文选用了目前性能表现最先进的技术栈, 并在此基础上设计了系统。建立的整套动态感知系统具备爬取 Tor 网络中的信息并对其进行分析及内容分类的功能, 且能够为用户提供监控服务的能力, 因此系统完成了以下功能需求: 1)能够对暗网中的信息进行高效抓取; 2)建立训练模型并对抓取到的信息进行精准分类, 解决分类的准确度和复杂性问题; 3)运行保障部分可监测整套系统的运行状态, 对各部分的状态进行严密的把控, 能够及时发现系统异常, 并拥有一定的容错能力。

系统整体设计如图2所示, 该系统由高性能暗网爬虫、内容处理分析、系统运行保障三部分组成。

3.1 高性能分布式暗网爬虫的设计

要满足系统的整体信息爬取、数据分析等需求, 暗网爬虫系统应具备以下几个功能模块: 任务管理模块、爬虫调度模块、网页下载模块、页面解析模块、数据存储模块。其中, 我们创新性地改进了 Tor 原有连接架构, 将三跳连接变为一跳, 并集成到了网页下载模块中, 提高了爬虫的爬取速度和稳定性。通过各模块之间的相互配合, 爬虫系统运行逻辑的

时序图如图 3 所示:

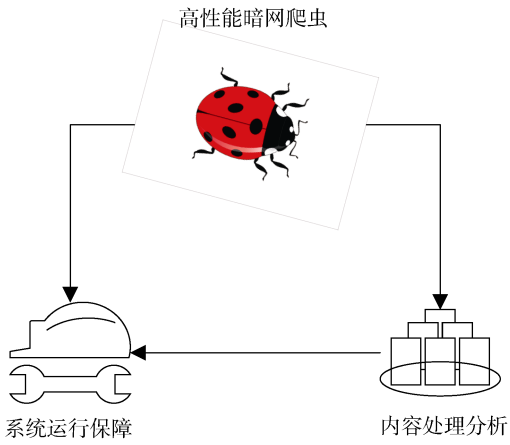

图 2　系统整体结构图

Figure 2　Overall system structure

爬虫系统各个模块的功能和设计如下:

任务管理模块主要实现任务管理的功能, 即对任务添加和任务下发的总体控制, 该模块需要解决任务重复出现的问题。面对大量任务需要去重的应用场景, 任务管理模块选用 Bloom Filter 算法[24]来进行网页的去重, Bloom Filter 是一种概率数据结构, 能够快速、高效地判断元素是否存在于集合中, 相较于其他的数据结构, Bloom Filter 在时间和空间上均具有一定优势因而可以用于网页的去重操作。对于 Bloom Filter 的实现, 模块使用了 Redis 进行存储, 同时开发了两种 Redis 版本的 Bloom Filter: ①基于 Redis 的 STRING 类型的 setbit 和 getbit 方法; ②基于 Redis 原生的外部拓展模块。这两种实现方式分别具有兼容性强和效率高的特点, 适合不同应用场景。

爬虫调度模块负责启动、停止、监视爬虫的情况, 是整个爬虫过程的入口。该模块负责接收从总控制器下发的任务, 并把任务分配给相应的爬虫进行爬取, 可以接收控制器的各种命令, 对爬虫的参数进行微调或者整体停止和重启, 同时进行性能统计和回馈。

网页下载模块的主要功能是从暗网中下载资源。该模块使用开源组件 Privoxy 将 Tor 的代理从 Socks5 转为 HTTP, 然后使用 HTTP 请求下载网页中的资源, 同时为了增加 Tor 的匿名级别, Privoxy 使用 Socks5 代理以确保 DNS 请求通过 Tor 完成, 且可获得更精确的错误消息。为了提高爬虫的爬取效率, 爬虫系统对 Tor 的源码进行了修改, 具体修改代码的作用为可让 Tor 每次建立链路只使用一个中继, 如图 4 所示。然后修改配置文件 torrc 中的 AllowSingleHopCircuits 为 1。

上述修改后的代码使得 Tor 每次建立链路只使用一个中继, 优化前后连接建立方式的对比如表 1 所述。

同时, 关闭和调整大量的匿名性选项如流量填充等参数以降低无用流量, 以上做法虽然降低了匿名性, 但是提高了连接的速度及稳定性。Tor 客户端每 60 s 更换一组守卫节点(Guard Entry), 可避免守卫节点被过多的爬虫请求阻塞链路的问题, 修改的选项如表 2 所述。

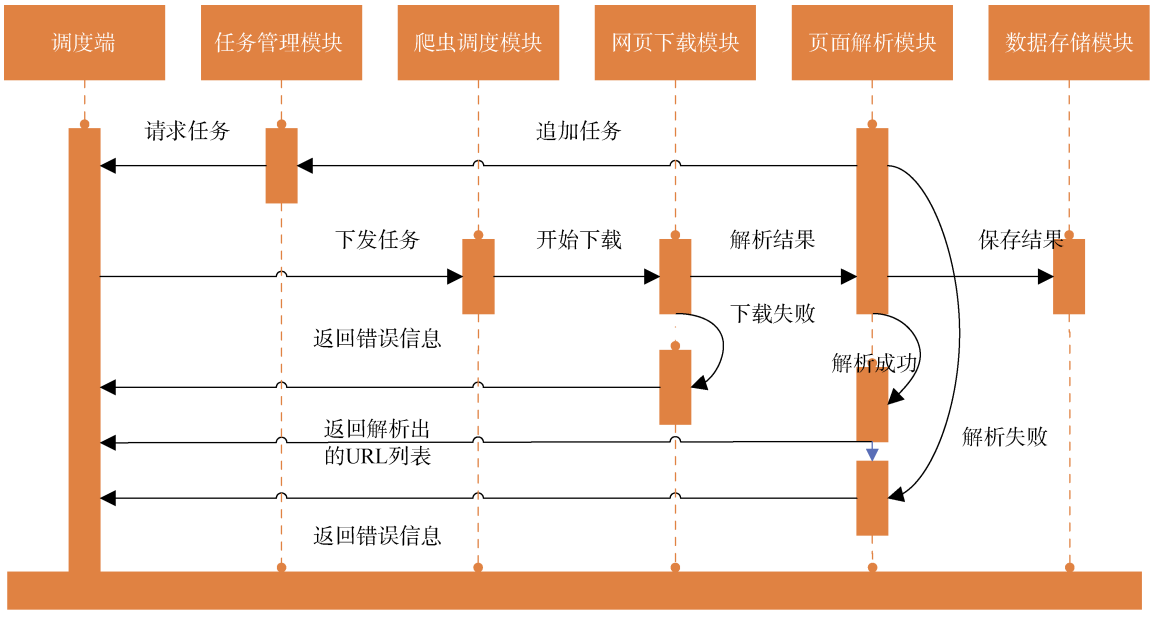

图 3　爬虫运行逻辑图

Figure 3　Crawler operation logic

```
--- tor/src/or/control.c        2007-06-08 10:55:51.000000000 +0200
+++ tor-mine/src/or/control.c   2007-06-08 11:03:14.000000000 +0200
@@ -2014,11 +2014,6 @@
                       conn);
     return 0;
   }
-  if (circ && (circuit_get_cpath_len(circ)<2 || hop==1)) {
-    connection_write_str_to_buf(
-                       "551 Can't attach stream to one-hop circuit.\r\n", conn);
-    return 0;
-  }
   if (circ && hop>0) {
     /* find this hop in the circuit, and set cpath */
     cpath = circuit_get_cpath_hop(circ, hop);
```

图 4　修改 Tor 源码

Figure 4　Modify Tor source code

表 1　优化前后连接建立方式对比

Table 1　Comparison of connection establishment methods before and after optimization

连接建立方式	节点数	连接建立过程
Tor 默认方式	3	客户端选择第一跳节点，第一跳节点选择第二跳，第二跳节点选择第三跳，第三跳节点作为出口节点进行通信
优化后的方式	1	客户端选择第一跳节点，第一跳节点作为出口节点进行通信

表 2　修改 Tor 客户端的配置项

Table 2　Modify configuration item of Tor client

配置项	值
Circuit Build Timeout	10
Circuits Available Timeout	30
Max Client Circuits Pending	256
Learn Circuit Build Timeout	0
Connection Padding	0
Reduced Connection Padding	1
Num Entry Guards	40
Num Directory Guards	10

页面解析模块的主要任务是从下载模块获取的 HTML 中提取 URL 并回传给 URL 调度模块。该模块将同时使用基于正则表达式、基于 CSS 和基于 HTML 解析三种 URL 提取方法以尽可能的减少漏报，URL 在经过任务去重之后就可以加入到爬取队列中继续运行。URL 提取解析的整体流程如图 5 所示。

图 5　URL 提取解析的整体流程

Figure 5　The overall process of URL extraction and parsing

数据存储模块的主要任务是将经过 URL 提取解析后的网页数据储存起来，本系统主要使用的是分布式存储，鉴于使用目的不同，数据有以下两种存储位置：Elasticsearch 集群和 HBase 集群。其中 Elasticsearch 作为搜索和分析引擎，数据输入来源于 Logstash；HBase 作为爬取内容原始数据的存储数据库，同时为爬虫提供灾备机制，系统使用 Thrift Server 作为 HBase 集群的代理，即作为数据查询和存储的接口。综合多种技术，数据存储部分的整体结构如图 6 所示。

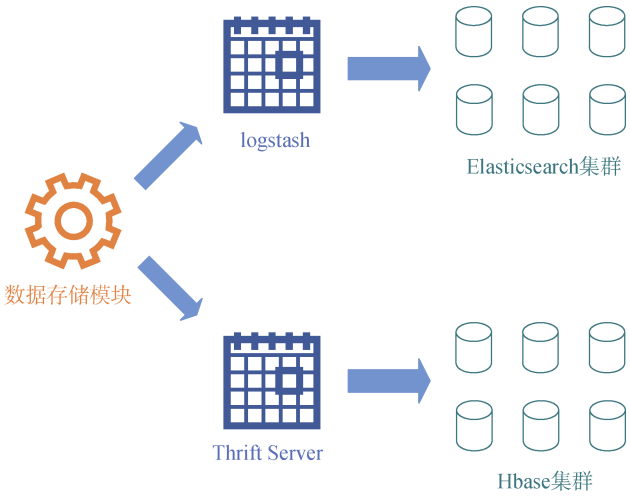

图 6　数据存储模块架构图

Figure 6　Data storage module architecture

3.2　内容处理分析的设计

内容处理分析部分的功能主要是对爬取的信息进行处理和分析，具体的任务是对爬取的网页内容进行分类，而后根据结果分析暗网的内容结构和敏感信息。经过分析和研究，系统共划分了毒品、黑客、色情、暴力、伪造、数字货币等 9 种网页类型。其中论坛、市场、空白等内容归入了其他类型，“市场”类型的网页会提供市场交易的平台，“空白”类型是指文字很短或只有图像没有文字等情况。

在常规句子中，噪声数据可以定义为文本文件页眉、页脚、HTML、XML、标记数据，由于这些类型的数据没有意义且不提供任何信息，因此必须删除这些类型的噪声数据。当检测到图像标记时，保留了图像名称并删除了扩展名，同时把 URL 单独提取出来。该模块使用 langdetect python 库[25]对训练集数据进行了处理，并使用 NLTK[26]中的 PorterStemmer 提取词干，用 WordNetLemmatizer 进行词形还原，利用词库删除了特殊字符和停用词，并且将所有电子邮件、网址和货币转化为常规的记号。最终训练集采用的数据预处理方式必须和真实应用时一致。

将自然语言文本转换为数字的过程在机器学习

中被称为矢量化。在本系统中, 使用实现较为简单的 BoW[27]和 TF-IDF 进行特征提取。计算每一个词项 ω_i 的 TF-IDF 值, 通常采用如下公式[28]:

$$\begin{aligned}\mathrm{TF-IDF}(\omega_i) &= tf(\omega_i)\times idf(\omega_i)\\ &= tf_i(\omega_i)\times\log(N/df(\omega_i))\end{aligned}$$

其中 $tf_i(\omega_i)$ 表示当前词项 ω_i 在文本 j 中出现的频率, N 表示文本集合中所有文本的总数, $df(\omega_i)$ 表示文本集合中有多少篇文本出现了当前词项 ω_i。

BoW 是一种从文本中提取特征以用于机器学习算法的方法, 这种方法将成篇的文本看作一堆文字, 在忽略语法、语义以及文字出现顺序的基础上计算每个单词的频率, 将频率向量作为结果输出给机器学习算法。在大文本语料中经常会出现一些出现频率高但所包含信息量极少的词语, 例如常见的冠词、连词等, 为避免这些词语的影响, 文本分类模块引入 TF-IDF 模型, 该种模型可根据文档的单词频率(TF)和反向文档频率(IDF)为文档中的每个单词分配权重, 对文本特征的频数做进一步权重划分, 最终具有较高权重分数的单词将被认为代表着更重要的意义。

对于每种特征表示方法, 分别使用了如下三种分类器: 支持向量机(SVM)、逻辑回归(LR)和朴素贝叶斯(NB)。处理过后的数据将被用于机器学习模型的训练, 且是一个有监督的文本分类机器学习问题。

3.3 系统运行保障的设计

运行保障部分主要负责对系统各个组件的当前状态进行实时监控, 以及提供一定的自动化问题解决能力, 模块主要对以下部分的状态进行监控: 1)当爬虫出现速度降为零、URL 队列数据不足等情况时需要及时进行预警和提醒; 2)持续监控爬虫系统服务器的 CPU 和内存占用、Redis 服务器的内存占用、存储节点服务器的存储空间以及内存空间占用进行。该模块使用 Prometheus 提供系统运行状况的数据信息采集, 通过开源 Exporter 和自行编写的 Exporter 进行系统和软件性能监控, 并使用 Grafana 对系统的状态数据进行整合、收集和展示。此外系统还设计了一套的容错模块和预警模块, 主要完成对爬虫异常状态的统计、预警和监控。

4 实验评估

4.1 网页爬取效果

爬虫同时在 8 台服务器上并发运行, 每个爬虫有 15 个并行工作线程, 可以完成对隐藏服务的 HTML 代码的下载和解析, 每个线程尽可能多地收集文本, 同时减少对同一个隐藏服务器的爬取, 在完成下载之后, 将内容保存以供后续分析。爬虫在 7 天时间里, 共爬取不重复网页 4080725 个, 域名 6550 个, 平均每 30s 爬取 400 个网页, 整体性能比较稳定。

在爬取的网页内容中, 文本占比达 77.14%, 其次是图片 5.25%, 内容类型为 HTML 的占比高达 93.8%, 而之后是图片占比 2.92%, 由此可看出暗网中的文字内容占主要部分, 我们认为这是由于暗网网络速度较慢, 过多的图片会导致网站的可用性下降, 影响用户的使用体验。

服务端占比中 Nginx 的使用比例最大, 其次是 Lighttpd、Apache, 甚至还有 Cloudflare 做 CDN 的隐藏服务, 针对不同的服务器情况可以对他们进行进一步的分析和研究。

表 3 状态信息数量表

Table 3 Quantity table of status information

状态信息	数量
OK	3 691 060
Service Unavailable	150 095
Not Found	111 874
Forbidden	55 567
Moved	39 020
Other	32 837

从表 3 中可以看出, 由于暗网内容更新不及时, 所以存在很多的错误状态, 尤其是 Service Unavailable, 平时在表层网络出现较少, 但在暗网中随处可见, 说明暗网的网络情况比较不稳定, 且大多数隐藏服务也只会上线不多的时间。

表 4 网站链接数量表

Table 4 List of website links

网站域名	网页数量
http://saufca42reinzasa.onion	466688
http://xfmro77i3lixucja.onion	207779
http://center222xdihudu.onion	202650
http://clivl6rf3vft7ihw.onion	134605
http://zqktlwi4fecvo6ri.onion	121581
http://www.flibustahezeous3.onion	110591
http://flibustahezeous3.onion	80438
http://archivecaslytosk.onion	69233
http://nnmclub5toro7u65.onion	59150
http://32pbf32xi6ccm63z.onion	54326

表 4 中链接数量靠前的几个网站大部分是一些内容存储类的网站, 因此网页数量较多。

4.2 内容分类效果

需要进行的数据收集并不是原始数据, 而是经过标记的的分类数据, 为了对数据进行准确的分类, 在这个阶段我们基于公开的相关研究[29]和手动标记的方法, 经过一段时间工作之后, 共收集到各类型网页分类数据 10367 条, 其中“主机”和“空白”内容约占一半比重。经过分类整合并进行大量数据处理之后, 系统得到的网站内容分类表 5 所示。

表 5 网站内容分类表

Table 5 Website content classification table

类型	数量	类型	数量
毒品	290	伪造信用卡	392
黑客	182	假币	81
色情	226	身份伪造	42
暴力	47	数字货币	847
其他	6462		

我们对 BoW 和 TF-IDF 两种特征提取方法和朴素贝叶斯 Naïve Bayes(NB)、逻辑回归 Logistic Regression(LR)和支持向量机 SVM 三种分类模型进行了评估。这三种模型是目前比较成熟且通用的分类模型, 适合本工作中需求的文本类型分类任务, 训练后得到的模型简单且易于部署。通过 K 折交叉验证和网格参数搜索进行超参调优之后得出了结论: TF-IDF 的整体效果优于 BoW, 在同种预处理和特征提取方式下, 支持向量机的效果优于逻辑回归的效果优于朴素贝叶斯的效果。

表 6 是使用 BoW 法时, 朴素贝叶斯 Naïve Bayes(NB)、逻辑回归 Logistic Regression(LR)和支持向量机 SVM 三种算法计算每个类型的 Precision、Recall 率、F1-Score 和 Support, 可以看出在使用 BoW 法时, SVM>LR>NB。

表 6 Bow 分类结果得分表

Table 6 Score table of Bow classification results

			precision	recall	f1-score
BoW	NB	micro avg	0.793	0.793	0.793
		macro avg	0.489	0.65	0.487
		weighted avg	0.863	0.793	0.817
	LR	micro avg	0.828	0.828	0.828
		macro avg	0.633	0.845	0.712
		weighted avg	0.879	0.828	0.843
	SVM	micro avg	0.901	0.901	0.901
		macro avg	0.754	0.731	0.738
		weighted avg	0.899	0.901	0.899

使用 TF-IDF 时, 朴素贝叶斯的效果仍旧远远低于另外两种算法, 而逻辑回归的 Precision、Recall 和 F1-Score 已经极为接近支持向量机的得分。

如表 7 中所示, 使用 TF-IDF 之后, 三种算法的得分情况都有明显提高, 最明显的是逻辑回归模型 LR, 提升幅度达到 10%, 其次是支持向量机 SVM, 也有 5%左右的明显提升, 朴素贝叶斯模型 NB 由于其特点, 只是结果产生波动。

表 7 TF-IDF 分类得分表

Table 7 TF-IDF classification score table

			precision	recall	f1-score
TF-IDF	NB	micro avg	0.865	0.865	0.865
		macro avg	0.624	0.307	0.344
		weighted avg	0.846	0.865	0.829
	LR	micro avg	0.927	0.927	0.927
		macro avg	0.814	0.873	0.840
		weighted avg	0.934	0.927	0.929
	SVM	micro avg	0.939	0.939	0.939
		macro avg	0.885	0.840	0.861
		weighted avg	0.939	0.939	0.939

针对几种分类模型在训练集递增的情况下, 我们对训练分数与交叉验证分数的拟合程度进行了绘图(基于 Sklearn 的 Learning Curve 学习曲线), 如图 7 所示:

图 7 证明, 对于 LR 和 SVM 模型, 当训练精度曲线开始略微降低时, 通过增加样本数量, 交叉验证得分曲线正在稳步上升并且分类精度得到提高, 模型正在正确学习, 而对于 NB 模型, 准确度偏低, 同时增加数据对效果没有太大帮助, 所以选择放弃 NB 模型。再对比 SVM 和 LR 模型, 当训练数据数量达到最大时, 训练分数仍远大于验证分数, 增加训练数据有助于提升模型的泛化能力, SVM 模型可以受益于更多数量的训练。此时的 SVM 和 LR 模型都存在过拟合的可能, 所以还是需要增加数据量, 进行更多的验证。

4.3 性能对比

我们将本文提出的分布式爬虫与 Zulkarnine[3]、Christin[30]、Branwen[31]、于浩佳[18]、季节[32]、王相军[33]、赵宗飞[34]等人的方案进行了性能对比, 其中 Zulkarnine 的数据与暗网恐怖主义有关, Christin 主要对暗网市场 Silk Road 进行了商品信息的测量, Branwen 实现了匿名市场产品列表爬虫并创建了数

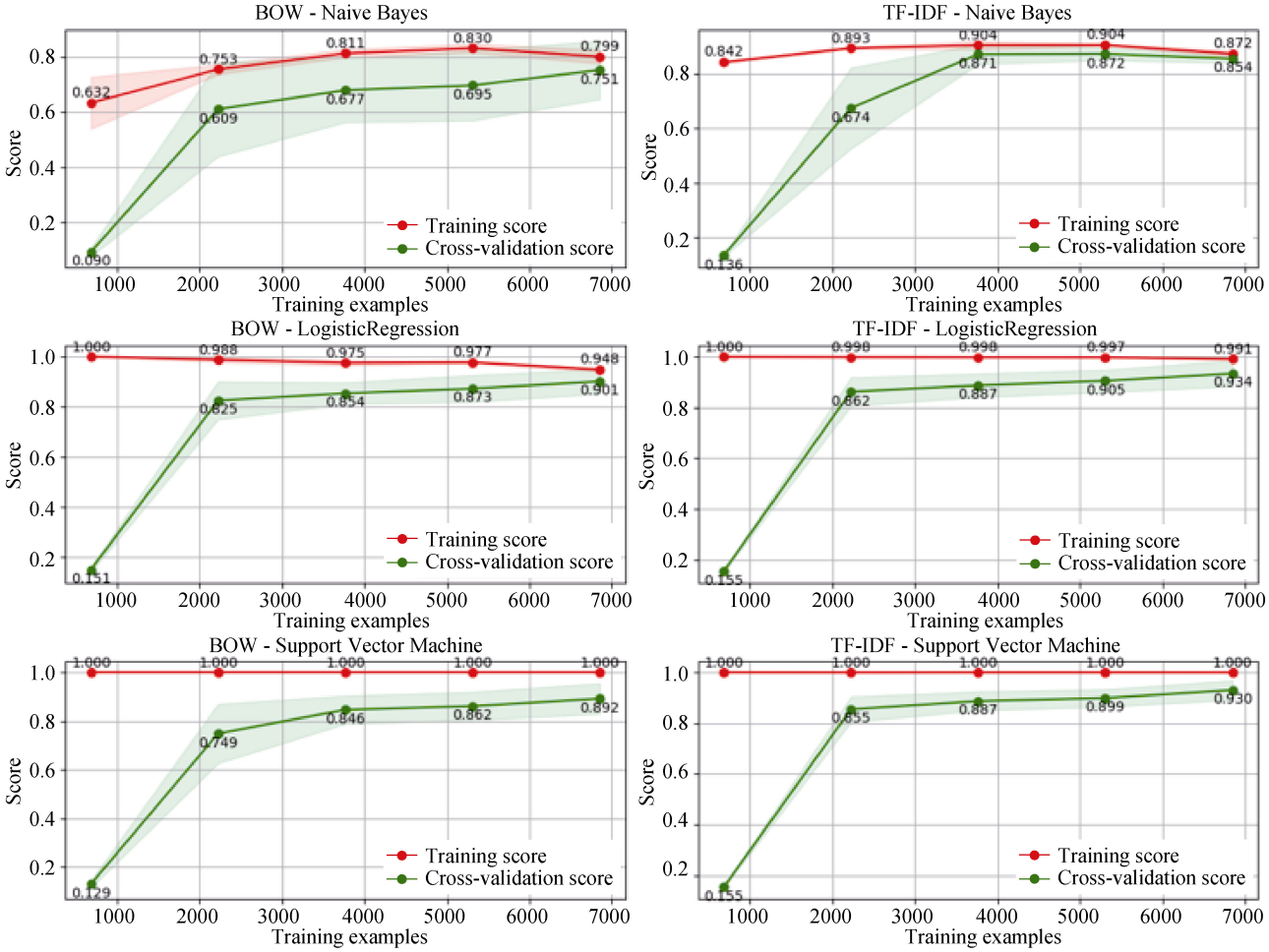

图 7　三种分类器学习曲线拟合图

Figure 7　Learning curve fitting of three classifiers

据集，于浩佳使用 Scrapy 实现了简单的暗网爬虫，季节、王相军、赵宗飞等人依据不同的设计思路设计了明网通用爬虫，功能与性能对比如表 8 所示。

综合对比，我们针对 Tor 网络的分布式爬虫功能最全、性能最好，其中域名数量少于 Zulkarnine 的原因是近几年 Tor 受到了更多的关注，对于暗网的监管打击力度加大使得许多暗网域名被关闭，且伴随着现有暗网网站的纵向发展，网页数量在逐步增多。

5　研究展望

在将来的工作中，本文将针对以下几个方面进行更进一步的研究：1)提高爬虫性能和稳定性：优化

表 8　功能和性能对比

Table 8　Function and performance comparison

		Zulkarnine[3]	Christin[30]	Branwen[31]	于浩佳[18]	季节[32]	王相军[33]	赵宗飞[34]	我们
功能	暗网专用爬虫	√	√	√	√	×	×	×	√
	Tor 快速连接	×	×	×	×	×	×	×	√
	内容分类	√	√	√	√	√	×	√	√
	运行保障	×	×	×	×	×	×	×	√
	分布式	×	×	×	×	×	√	√	√
规模	网页数量	54141	204000	84000	—	—	29112	6740	**4080725**
	域名数量	**10163**	1	—	509	—	—	6	6550
	数据总量	260G	—	—	—	—	—	—	**500G**
	爬取速度	—	—	—	—	3 页/s	0.4 页/s	0.4 页/s	**13 页/s**

爬虫连接 Tor 的速度、优化链路建立过程、优化代码, 全部用异步实现; 2)提高系统健壮性: 引入 Sentry 进行错误监控、加强监控系统的建设; 3)多元化分析: 当前对数据的分析较为单调, 缺乏细化分析, 应继续增加对收集的各种信息的细化分类和分析; 4)收集更多数据: 进行更精准的文本特征提取和筛选, 提高分类算法的性能; 引入半监督方法; 5)优化分类算法: 对现有分类器进行性能调优, 提高性能; 试验更多的分类方法, 增加分类算法的准确性和效率; 引入深度学习算法。

6 结束语

以 Tor 网络为代表的暗网越来越多的引起了关注, 为了解决其中存在的网络内容监控问题, 帮助国家及时发现暗网中存在的犯罪内容, 对犯罪行为进行严格打击, 实现更高效快速的暗网监管的目标, 本文从 Tor 网络的原理和应用层面出发, 设计和实现了一套暗网网络动态信息感知、收集、分析系统。该系统在 Tor 网络中进行了实际的功能测试和性能评估, 结果显示本文提出的分布式爬虫效果优于现有的其他爬虫方案。在最后我们给出了 Tor 网络内容监控系统未来的发展方向与前景。

参考文献

[1] Tor project. https://metrics.torproject.org/networksize.html. Sept. 2019.

[2] Hurlburt G. Shining Light on the Dark Web[J]. *Computer*, 2017, 50(4): 100-105.

[3] Zulkarnine A T, Frank R, Monk B, et al. Surfacing collaborated networks in dark web to find illicit and criminal content[C]. *2016 IEEE Conference on Intelligence and Security Informatics*, 2016: 109-114.

[4] Bai Y L. *The research and implement of data mining algorithms based on hadoop*[D]. Beijing: Beijing University of Posts and Telecommunications, 2011.
(白云龙. 基于 Hadoop 的数据挖掘算法研究与实现[D]. 北京: 北京邮电大学, 2011.)

[5] Cheng J J. *Research and implementation of distributed web crawl based on hadoop architecture*[D]. Beijing: Beijing University of Posts and Telecommunications, 2010.
(程锦佳. 基于 Hadoop 的分布式爬虫及其实现[D]. 北京: 北京邮电大学, 2010.)

[6] Udapure T V, Kale R D, Dharmik R C. Study of Web Crawler and Its Different Types[J]. *IOSR Journal of Computer Engineering*, 2014, 16(1): 1-5.

[7] Dumais S, Chen H. Hierarchical Classification of Web Content[C]. *The 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, 2000: 256-263.

[8] Kan M Y. Web Page Classification without the Web Page[C]. *The 13th international World Wide Web conference on Alternate track papers & posters*, 2004: 262-263.

[9] Kan M Y, Thi H O N. Fast Webpage Classification Using URL Features[C]. *The 14th ACM international conference on Information and knowledge management*, 2005: 325-326.

[10] Kaur P. Web Content Classification: A Survey[EB/OL]. 2014: arXiv: 1405.0580. https://arxiv.org/abs/1405.0580

[11] Moore D, Rid T. Cryptopolitik and the Darknet[J]. *Survival*, 2016, 58(1): 7-38.

[12] Graczyk M, Kinningham K. Automatic product categorization for anonymous marketplaces[R]. *Technical report, Stanford University*, 2015.

[13] Cloudflare. https://www.pcmag.com/news/cloudflare-94-percent-of-tor-traffic-is-malicious. March. 2016.

[14] Zheng X C, Li H, Wang R, et al. Survey of Anonymous Network Applications and Simulation Platforms[J]. *Journal of Xidian University*, 2021, 48(1): 22-38.
(郑献春, 李晖, 王瑞, 等. 匿名网络应用及仿真平台研究综述[J]. *西安电子科技大学学报*, 2021, 48(1): 22-38.)

[15] Owen, G., and N. Savage. The tor dark net [J]. Paper Series no: Global Commission on Internet Governance. 2015.

[16] Overlier L, Syverson P, Processing C A. Locating hidden servers[C]. *2006 IEEE Symposium on Security and Privacy*, 2006: 15pp.-114.

[17] Yin S. *Design and implements of dark web scanning system based on injection*[D]. Beijing: Beijing University of Posts and Telecommunications, 2018.
(殷帅. 基于节点注入的暗网扫描系统的设计与实现[D]. 北京: 北京邮电大学, 2018.)

[18] Yu H J, Chen B, Liu R. Research on Information Crawling Technology of Anonymous Website[J]. *Journal of Information Security Research*, 2017, 3(10): 922-931.
(于浩佳, 陈波, 刘蓉. 匿名网站信息爬取技术研究[J]. *信息安全研究*, 2017, 3(10): 922-931.)

[19] Sun A X, Lim E P, Ng W K. Web Classification Using Support Vector Machine[C]. *The 4th international workshop on Web information and data management*, 2002: 96-99.

[20] Xiao L, Yao N M. Research on Chinese classification based on TF-IDF[C]. *2021 International Conference on Neural Networks, Information and Communication Engineering*, 2021: 11933: 59-64.

[21] Noor U, Rashid Z, Rauf A. A Survey of Automatic Deep Web Classification Techniques[J]. *International Journal of Computer Applications*, 2011, 19(6): 43-50.

[22] Xian X F, Zhao P P, Fang W, et al. Automatic classification of deep web databases with simple query interface[C]. *2009 International Conference on Industrial Mechatronics and Automation*, 2009: 85-88.

[23] Barbosa L, Freire J, Silva A, et al. Organizing hidden-web databases by clustering visible web documents[C]. *2007 IEEE 23rd International Conference on Data Engineering*, 2007: 326-335.

[24] Frontera S, Lazzeretti R. Bloom Filter Based Collective Remote Attestation for Dynamic Networks[C]. *The 16th International Conference on Availability, Reliability and Security*, 2021: 1-10.

[25] Mimino666. https://github.com/Mimino666/langdetect. Nov. 2021.
[26] Nltk. https://github.com/nltk/nltk. Nov. 2021.
[27] Rupapara V, Rustam F, Amaar A, et al. Deepfake Tweets Classification Using Stacked Bi-LSTM and Words Embedding[J]. *PeerJ Computer Science*, 2021, 7: e745.
[28] Huang C H, Yin J, Hou F. A Text Similarity Measurement Combining Word Semantic Information with TF-IDF Method[J]. *Chinese Journal of Computers*, 2011, 34(5): 856-864.
[29] Al Nabki M W, Fidalgo E, Alegre E, et al. Classifying illegal activities on tor network based on web textual contents[C]. *The 15th Conference of the European Chapter of the Association for Computational Linguistics*: *Volume 1*, *Long Papers*, 2017: 35-43.
[30] Christin N. Traveling the Silk Road: A Measurement Analysis of a Large Anonymous Online Marketplace[C]. *The 22nd international conference on World Wide Web*, 2013: 213-224.
[31] Branwen G, Christin N, Décary-Hétu D, et al. Dark net market archives, 2011–2015[J]. *Retrieved August*, 2015, 28: 2015.
[32] Ji J. *Research and application of crawler algorithm in Internet public opinion system*[D]. Zhenjiang: Jiangsu University of Science and Technology, 2015.
(季节. 爬虫算法在互联网舆情系统的研究与应用[D]. 镇江: 江苏科技大学, 2015.)
[33] Wang X J. *Research on distributed crawler based on crowd-sourcing*[D]. Harbin: Harbin Institute of Technology, 2017.
(王相军. 基于众包协作的分布式爬虫研究[D]. 哈尔滨: 哈尔滨工业大学, 2017.)
[34] Zhao Z F. *Research and application of Internet hot topic detection based on big data background*[D]. Guangzhou: South China University of Technology, 2016.
(赵宗飞. 基于大数据的互联网热点话题挖掘的研究与实现[D]. 广州: 华南理工大学, 2016.)

郑献春 于 2011 年在西安电子科技大学获得工学学位。现在西安电子科技大学网络空间安全专业攻读博士学位。研究领域为匿名网络。Email: 99444696@qq.com

王瑞 于 2019 年在西安电子科技大学信息安全专业获得学士学位。现在西安电子科技大学网络空间安全专业攻读硕士学位。研究领域为匿名网络、网络流量安全。Email: 785340571@qq.com

闫皓楠 于 2018 年在西安电子科技大学信息安全专业获得学士学位。现在西安电子科技大学网络空间安全专业攻读博士学位。研究领域为网络流量安全。Email: yanhaonan.sec@gmail.com

赵兴文 于 2011 年在中山大学计算机软件与理论获得博士学位。现任西安电子科技大学副教授。研究领域为互联网安全应用、隐私保护、数据共享安全。Email: xwzhao@xidian.edu.cn

李晖 于 1998 年在西安电子科技大学获得博士学位。现任西安电子科技大学教授。研究领域为密码信息安全、信息论与编码理论。Email: lihui@mail.xidian. edu.cn

李凤华 于 2009 年在西安电子科技大学获得工学博士学位。现任中国科学院信息工程研究所二级研究员。研究领域为网络与系统安全、信息保护、数据安全。Email: lifenghua@iie.ac.cn