软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2014,25(12):2877−2892 [doi: 10.13328/j.cnki.jos.004727]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

社交网络取证初探*

吴信东[1,2], 李亚东[1], 胡东辉[1]

[1](合肥工业大学 计算机与信息学院,安徽 合肥 230009)

[2](Department of Computer Science, University of Vermont, Burlington, VT 05405, USA)

通讯作者: 吴信东, E-mail: xwu@hfut.edu.cn

摘 要: 社交网络是人类借用计算技术和信息技术进行信息交流、建立人际交互关系等社会活动的一种新型工具,已成为社会计算中研究社会软件的核心课题之一.社交网页取证旨在对用户信息进行证据获取、固定、分析和展示,提供直接、有效、客观、公正的第三方依据.在互联网飞速发展的背景下,社交网页取证面临着用户信息多样、内容动态(实时)变化、海量、交互和图片内容是否可信的挑战,已成为社交网络和社会计算中舆情分析、情感计算、社交网络关系的内容分析以及个人、群体和社会性行为分析的一个重要难题.针对社交网页取证问题,以新浪微博为例,设计了一套取证解决方案,对用户发表的信息、人脸图片、位置信息进行固定,依靠网页取证方法来认证信息的可信性.同时,利用信息可视化展示手段和辅助分析来应对在海量社交网页数据背景下的计算机取证工作.

关键词: 计算机取证;社交网络;社会计算;可信取证

中图法分类号: TP311

中文引用格式: 吴信东,李亚东,胡东辉.社交网络取证初探.软件学报,2014,25(12):2877−2892. http://www.jos.org.cn/1000-9825/4727.htm

英文引用格式: Wu XD, Li YD, Hu DH. Study on social network forensics. Ruan Jian Xue Bao/Journal of Software, 2014,25(12):2877−2892 (in Chinese). http://www.jos.org.cn/1000-9825/4727.htm

Study on Social Network Forensics

WU Xin-Dong[1,2], LI Ya-Dong[1], HU Dong-Hui[1]

[1](College of Computer Science and Information Engineering, Hefei University of Technology, Hefei 230009, China)

[2](Department of Computer Science, University of Vermont, Burlington, VT 05405, USA)

Corresponding author: WU Xin-Dong, E-mail: xwu@hfut.edu.cn

Abstract: Based on advances in computing technology and information technology, social networks have emerged as a new tool for people to exchange information and build interaction networks, and have become a key topic for social software studies in social computing. Social network forensics seeks to acquire, organize, analyze and visualize user information as direct, objective and fair evidence from a third-party perspective. Along with the rapid development of the Internet, social network forensics faces new challenges in dealing with user information being diverse, real-time and dynamic, huge in volume, and interactive, and also photo trustworthiness. It therefore has become a hot issue for opinion analysis, affective computing, content analysis in social networking relations, as well as individual, group and social behaviors in social networks and social computing. This paper designs a forensic model for social network forensics, and implements it on Sina microblogging. This model provides user information analysis, facial image recognition, and location presentation for trustworthiness analysis of digital evidence, and applies visualization to help reduce the difficulty of analysis and forensics on massive data from social networks.

Key words: computer forensics; social networks; social computing; trustworthy forensics

* 基金项目: 国家自然科学基金(61229301, 61272540); 国家高技术研究发展计划(863)(2012AA011005); 国家重点基础研究发展计划(973)(2013CB329604); 教育部创新团队基金(IRT13059)

收稿时间: 2014-05-05; 修改时间: 2014-08-21; 定稿时间: 2014-09-30

计算机技术的飞速发展与互联网的快速传播,极大地改变了人们的日常生活,互联网已经成为政治、经济和社会活动的重要场所.社交网络是人类借用计算技术和信息技术进行信息交流、建立人际交互关系等社会活动的一种新型工具,已成为社会计算中研究社会软件的一个核心课题.与此同时,犯罪分子利用计算机和互联网技术进行各种犯罪活动,出现了计算机和网络犯罪;而传统的犯罪也在计算机和互联网技术的辅助下,变得更加狡诈和隐蔽[1].

针对这些计算机相关犯罪活动,需要利用取证技术对相关证据进行固定,并确保这些证据是可信的.传统计算机取证的证据可信基础依赖于对原始数据存储介质的获取与固定,但是服务器托管技术使得网络服务提供商可以将服务器放置在国外,甚至一个庞大的服务器云中,使得直接获取这些服务器存储介质的难度非常大;同时,这些数据又能被非常方便地远程销毁.因此,在无法直接获得存储介质的情况下及时取证并确保取证得到的证据的可信性,是一个值得研究的问题.特别是随着近几年来社交网站的兴起,人们可以更加迅速地进行远程信息交流、建立人际交互关系,用户也越来越主动地把自身的信息分享到社交网站上.社交网站提供的个性化服务,使得用户能够上传大量的文字、图片、音频、位置等信息.这些社交网络信息具备多样、动态(实时)、海量和交互等特性,极大地增加了取证的困难性,从而导致传统的计算机取证方法已经不能适应,需要发展面向社交网站的有效取证方法.此外,社交网站是社交网络中人与信息系统交互(并通过信息系统与他人进行交互)的重要场所,可以通过对社交网页中相关证据的获取、固定、分析和展示形成相对完整和可信的证据(链).社交网页取证旨在对用户发布在社交网页上的信息进行证据固定,提供直接、有效、客观、公正的第三方依据,已成为社交网络和社会计算中舆情分析、情感计算以及个人、群体和社会性行为分析的一个重要研究课题.

本文针对网页取证,特别是社交网页取证中所面临的问题,主动对网页内容进行取证,探索针对社交网页取证的若干关键技术和方法,并以新浪微博为例,设计了一个社交网站取证方案,固定新浪微博上的文本、位置和人脸图片等信息,确保整个取证过程中的数据可信.同时,对取证得到的相关数据结合社会计算相关分析手段,分析并提取出有用的数据,探索海量复杂社交网络环境下计算机取证的新方法.社交网络的用户关系的挖掘和分析是社交网络取证的重要部分,通过利用社交网络关系的挖掘和分析,可以为网络犯罪侦查提供有价值的线索.由于该部分内容属于相对独立的研究范畴,本文不特别研究该部分内容.

1 研究背景与挑战

1.1 研究背景

1.1.1 传统计算机取证

传统计算机取证对象的主要目标是嫌疑犯的计算机或者存储介质的数据,并按照严格的操作规范从中提取证据并固定,一般情况下会使用专业的数据获取工具,精确地复制存储介质中的每一个字节,生成镜像证据文件.针对固定下来的镜像证据文件,构造独立的系统环境,对原始数据进行分析,并找到关键犯罪证据[2,3].而存储介质上的是最原始的不易被人理解的二进制数据,通过对这些数据的进一步解码、提取、分析,获得相关的文档、音频、视频等文件,最终取得能够被多数人理解、能够用于法庭作证的证据.最终的证据一方面要保证其可信性和不被篡改,另一方面也要保证其可理解性[4].目前,计算机取证的主要有两个研究方向是分析总结现有的相关取证模型[5,6]和针对不同取证场景设计相关取证方案[7−9].

Kent 等人针对计算机取证过程提出了一种基础的取证模型[10],如图 1 所示.模型将整个取证过程分为 4 个步骤:

(1) 搜集:主要目标是搜集存储有犯罪信息的计算机存储设备,通过这个步骤能够获取原始的存储介质,例如硬盘、光盘等.对于搜集到的存储介质,会立即进行逐字节的只读镜像操作,保证原始的存储设备不被任何修改操作,而之后所有的数据提取与分析均会从镜像数据上读取;

(2) 提取:从镜像数据中提取有用的文本、图形、图像、动画、音频及视频等多媒体信息,这些信息已经是可以理解的有意义的文件数据;

(3) 分析:主要针对提取出的多媒体信息进行详细的分析,寻找相关的犯罪内容.在这一步骤中,需要过滤

掉大量的无关内容;

(4) 报告:将相关犯罪信息通过归纳总结形成报告,作为最终的证据向法庭提交.这些报告需要易于被法庭控辩双方理解,客观地表述事实.

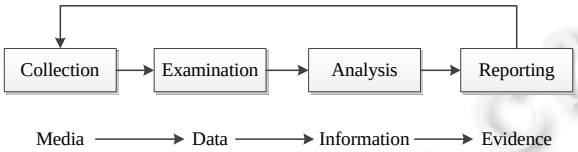


Fig.1 Basic model of computer forensics

图 1 计算机取证基础模型

Kent 等人提出了计算机取证的基本要求,定义了 4 大步骤,适用于普遍的计算机取证过程.但是此模型的主要问题是过于粗略,没有具体定义取证过程中的要求和步骤.Ankit 等人对 4 大步骤进行了细分,论述了各个部分所包含的详细步骤,同时,强调了建立独立的取证环境和严格记录整个取证过程的时间轴来保证取证过程不受干扰与取证结果真实可信[5].

1.1.2 网络取证技术

网络取证技术主要针对网络数据流、网络设备日志的实时监控和分析,发现网络系统的入侵行为,自动记录犯罪证据,对网络的动态信息进行收集和对网络攻击的主动防御[6].通常,这些网络数据流是一次性内容,易丢失,不会直接被存储于存储介质上或者存储于有限空间的存储介质中,通过利用嗅探技术对这些流式网络数据包进行记录固定和取证分析[7,8].但是,网络数据流往往含有大量的用于保证传输可靠性的冗余数据,因此,相比直接分析存储设备上的数据,需要进行更高强度的提取过滤与分析操作.为了能够有效应对这一挑战,Robert 等人进一步提出了网络取证的可视化模型,利用可视化数据分析技术对固定的内容进分析和取证[11].

1.1.3 网页取证技术

网页取证的目标是主要获取用户浏览到的网页数据,可以从以下 3 个方面进行相关取证:

(1) 服务器端网页取证

主要利用传统的计算机取证方法获取网页服务器的相关存储设备,然后,针对网页服务器所使用的相关架构,获取网页内容和数据库内容,并进行固定取证.文献[12]中设计了一种针对 ASP 动态网站的取证方法.但是随着互联网的快速发展与云计算技术的广泛应用,给服务器端取证带来的很大的困难.通常,犯罪分子会将相关犯罪信息放到网络服务提供商的服务器中,这些网络服务提供商的服务器可能托管在国外,直接增加了获取存储设备的难度.而对于犯罪分子,依靠远程访问与服务技术,可以随时随地快速销毁相关证据.而在云计算环境下,通常会将数据存储于一个云存储集群中,并且被分片存储于多台计算机中,即使可以封存这些服务器,然后进行固定取证,也需要耗费大量的人力物力进行多设备数据的提取与固定,代价极高[13].因此,传统的计算机取证技术对服务器端的网页取证也有一定的局限性.

(2) 客户端网页取证

对于客户端的网页取证,同样采用了传统的取证手段,通过对个人用户电脑中的浏览器等软件浏览网页记录进行固定和分析,获取到用户浏览网页的内容.主要难点在于需要详细分析每一种相关的软件,寻找软件的日志并提取信息[9];此外,客户端的网页浏览工具通常不会完整记录网页的信息,会存在一部分的信息丢失情况.本文作者参考客户端网页取证方法,提出了利用网络爬虫模拟用户浏览网页再保存相关数据的方法,通过网络爬虫主动产生访问请求数据对网页的内容进行取证[14].但是该方法的主要问题是保存下来的证据多为文本文件,比较容易受到篡改,证据的可信度较低.

(3) 数据流网页取证

主要利用了网络取证中的相关取证技术,对用户浏览网页过程中产生的网络数据流进行取证固定.这种方法能够保存较为完整的数据,篡改数据的难度也比较大.但是此方法是一种比较被动的取证方式,需要用户去主

动触发浏览网页的行为,大量的数据也提高了分析的难度.

1.2 社交网络取证问题

社交网络由于其信息多样、动态(实时)、海量、交互和可信性质疑等特点,相比一般的网络取证和网页取证具更大的困难性.主要表现在:

(1) 动态实时数据获取问题

不同于其他类型的网站,社交网站普遍采用服务器集群和云计算技术,使得服务器存储设备更加难以直接获取取证.而社交网站的交互要求,使得大多数社交网页采用了 AJAX 等技术,网页内容动态变化实时刷新,更新速度快.社交网站含有大量的个人信息,因此,一般的社交网站均要进行用户登录后才提供完整详细的信息,甚至需要通过对方用户的权限审核才能获取对应用户的信息.在 Huber 等人[15]设计的社交网站取证方案中,主要通过社交网站的开放 API 和网页会话来通过网站验证获取相关信息,但是这只解决了数据的获取问题,无法满足计算机取证过程中对取证内容(证据)的可信要求,这两种取证方法获得的证据难免降低了可信度.

(2) 海量数据内容分析问题

普通网页中含有大量的多媒体信息,多样的内容给具体的分析带来了很大的困难.而社交网站中,用户是发布信息的主体,用户能够每天发布大量的数据.加上新浪微博中的文本长度限制,使得信息的浓缩度高,但是同时也加大了发布频率.此外,社交网站的信息分享属性,使得用户发布的多媒体信息均会与用户自身高度相关.例如在用户新浪微博中,可以发布大量的关于某个话题的内容、用户的个人照片、位置等信息,这些与用户自身密切相关的数据具有很高的取证价值.如何从这些海量多样的原始信息中获得与案件相关的证据,是社交网络取证面临的一个主要挑战.

(3) 复杂交互数据展示问题

目前的社交网站大量采用了动态网页技术,包含大量的交互操作.简单地将网页内容保存到本地,会丢失大量交互信息.同时,社交网络中用户的交互活动还涉及行为、位置、情感等因素,如何有效获取并展示这些证据,也是社交网站取证需要解决的问题.

(4) 证据的可信性问题

正是由于社交网络中信息的多样、动态(实时)、海量和交互等特点,造成了证据的获取难、分析难、展示难,通过一般计算机取证方法所获取的证据在证据的完整性、真实性和可理解性方面有所欠缺,不能形成法律上可以接受的可信证据.

本文针对这些问题和挑战,提出一种多层次社交网页可信取证模型系统,探索社交网页复杂海量数据的证据分析、展示方法.

2 一种多层次社交网页可信取证模型系统

为了能够有效地对社交网页进行取证,本节设计了一种多层次取证模型系统,利用不同层次的证据获取与固定,实现对社交网页的可信固定,并能够进行有效展示.

本文所适用的取证场景主要有两种情况:

(1) 针对某个指定网站或者某个指定网页内容,进行一次性完整的快照取证;

(2) 针对某个指定网站或者某个指定网页,进行连续的多次取证对比.此时为了减少固定的工作量,需要自动识别已经被固定过并且未被修改的网页内容,避免重复固定.

2.1 模型系统的整体框架

图 2 给出了一种多层次社交网页取证系统的整体框架,从底层到顶层分别为:

(1) 网络数据层

该层主要记录取证过程中的网络数据通信,保存的是二进制网络数据包,是最原始的数据,数据量大,最可信也最不易被篡改.但由于是大量的二进制流,分析困难,无法给人直观的感受,可理解性差,也不能直接用于法

庭作证.

(2) 内容爬取层

该层主要将网页相关的文件保存下来,大多是可读的文本数据和图像,相比网络数据层,容易提取其中的信息,但因为大多是文本形式,所以容易遭到篡改,也并不是直观显示内容,需要经过一定的转换才能够重新恢复到可理解的证据.

(3) 截图取证层

利用 Webkit 对爬取的网页文件进行渲染,并保存为图片.所获得的网页截图是网页内容最直观的展示,与人浏览同样网页的效果一致,但是同时会忽略掉原始网页文件中肉眼不可见(例如白色文字)的信息,也无法直接和用户进行互动,因此含有的信息量比前两者少,无法直接从网页截图中提取文本等数据.但是这些图片却最为直观.

(4) 数据分析层

对内容爬取层获得的社交网页相关内容进行分析,提取和展示与案件相关的有用的信息,并形成法庭可以理解的证据报告.

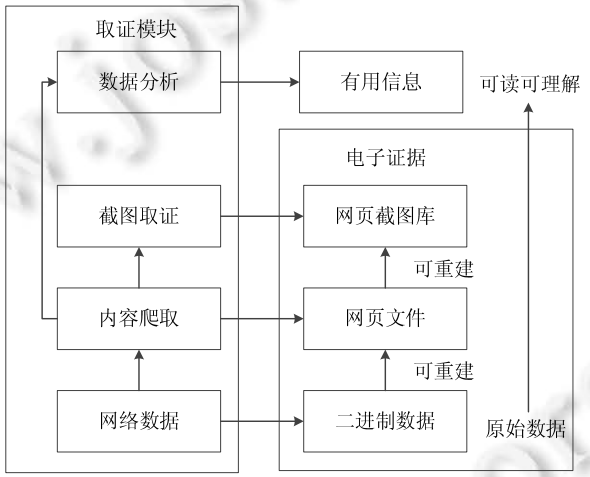


Fig.2 A multi-layer forensic model of social network Websites

图 2 一种多层次社交网页取证模型

对于上述所获取的数据,由最无原始的二进制数值(网络数据包)到含有可读信息的文本文件,再到可视化的网页截图与经过分析提取的信息,依次提高了人们对证据理解和直观感受的程度.其中,前三者能够作为原始证据,包含有完整的网页内容数据.但是这三者包含信息的冗余和完整程度不一样,造成篡改的难度也依次降低,证据的可信性有所降低.因此,只有结合这三者数据,才能形成既具备可信性又能直接提交法庭供直观审阅的证据.第 4 层数据分析层则将这些取证下来的可信信息,进行进一步分析以便用于犯罪侦查.

在具体取证过程中,取证需求多变,需要实时针对取证对象(不同的网页)进行灵活调整取证策略,修改信息的获取和过滤、模拟登录规则.因此,本系统主要采用脚本语言 Python 进行开发.Python 语言本身非常灵活,语法简洁,能够在应对不同需求时快速的调整策略.

此外,网页取证对象数量庞大,采用传统单一取证节点易造成取证时间缓慢、内容分析困难.本文利用现有的开源海量数据处理与存储方案,在未来爬虫库 Scrapy 的基础上,利用分布式内存数据库 Redis 实现分布式爬虫队列,进行分布式爬取,提高爬取速度和降低负载;利用分布式数据库 MongoDB 存储相关爬取内容,并基于 MongoDB 实现的 Gridfs 分布式文件系统对海量小文件数据进行动态存储.通过利用这些现有开源分布式平台,来提高取证系统应对海量数据取证需求的能力(如图 3 所示).

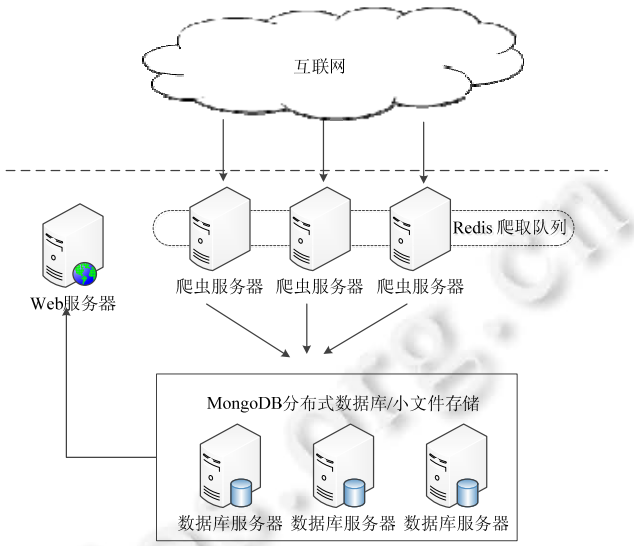


Fig.3 Framework of distributed forensics

图 3 分布式取证框架

2.2 网络数据流

网络数据层的主要功能是:在系统进行取证操作时,利用软件从操作系统底层获取网络报文数据.这些数据的特点是均为原始的二进制数据流,各个数据包只有简单的校验码,含有大量非结构化的数据,难以进行直接的修改,篡改的难度大;但同时,这些数据包必须经过解析才能读取,不可以直接转换为直观的文本数据,难以直接应用于法庭.

网页取证过程中,主要利用 HTTP 协议进行数据的获取,因此,嗅探的主要对象是 HTTP 协议;此外,为了防止 DNS 欺骗攻击,还需要抓取 DNS 协议数据.对于其他的大量网络数据包可以进行选择性地忽略,以减少抓取数据量,降低分析复杂度.

获取网络数据包可以通过网络嗅探的方式,在取证服务器或者网关上进行网络数据嗅探.在取证服务器上,能够及时地获取相关数据;而在网关上进行数据嗅探,则能在一定程度上加强取证的可信性,但是可能无法及时地采集相关数据,同时也容易受到其他非取证通信数据的干扰.

2.3 网页内容爬取与重现

2.3.1 普通浏览网页过程

用户浏览网页时的主要过程如下:

(1) 用户在浏览器中输入网址,浏览器自动向该网址发送请求.此后,浏览器会在后台执行余下的步骤(2)~步骤(6);

(2) DNS 解析.通过查询域名得到对应的 IP 地址.在这一步骤中,通常由操作系统自动进行解析,并且现在的多数浏览器均对 DNS 解析结果进行了缓存,防止多次无用请求,提高浏览速度.但是目前广泛存在着 DNS 欺骗攻击,攻击者将 DNS 查询结果篡改为钓鱼网站的 IP 地址,导致用户上当受骗,这会直接影响到取证结果;

(3) 请求主页面信息.当域名解析成功后,会根据解析结果和网址向对应 IP 的网页服务器发送请求数据,网页服务器随后返回一个主页面内容;

(4) 解析主页面内容.主页面内容是一个 HTML 或者 XML 页面,但是该页面中会包含多条需要获取的其他资源,将这些页面资源组合之后,才会组成一个完整的浏览页面.因此,浏览器会自动解析主页面,并

从中提取出相关的资源链接,并进一步对这些资源发起请求;

(5) 请求相关资源.过程与步骤(2)~步骤(4)类似,根据请求获得对应的资源文件,但是此时获取的资源文件不仅仅是 HTML 页面,也包含 CSS 网页样式文件、图片、JS 脚本乃至音视频等各种类别的资源;

(6) 合成渲染网页.浏览器汇总所有获得的网页和资源,然后渲染网页,绑定事件,并最终向用户展示出来;

(7) 浏览网页.用户浏览最终的网页,并在网页上进行信息的获取与界面的交互;

整个过程中,用户只需要输入网址,主要的操作都是由浏览器完成.但是浏览器通过多步较为复杂的过程,从服务器获取相关数据并展示成网页.因此,针对网页进行主动取证,主要的手段是模仿用户浏览网页过程,获取网页内容,并保存相关文件.本文利用网络爬虫实现数据的获取、分析和存储,将用户浏览的页面保存下来,作为证据.

2.3.2 取证过程

为了能够远程对服务器上的网页内容进行取证,本文利用网络爬虫模仿用户浏览网页的过程,将服务器提供的信息爬取下来并固定座位证据.网络爬虫是按照一定规则自动抓取网络信息的脚本或程序.利用网络爬虫,能够进行定向有选择的数据抓取.

具体的取证过程如图 4 所示,主要过程如下:

Step 1. 发送请求.针对需要取证的 URL,发送访问请求.

Step 2. 记录 DNS 解析结果.提取 URL 中的域名,记录其解析结果.

Step 3. 针对一个取证 URL,发送获取主页面的请求.

Step 4. 获取主页面内容,并进行内容解析,获得该页面包含的相关静态资源.

Step 5. 针对页面内的静态资源继续发送请求并进行固定.

Step 6. 将所有获取的资源固定为证据.

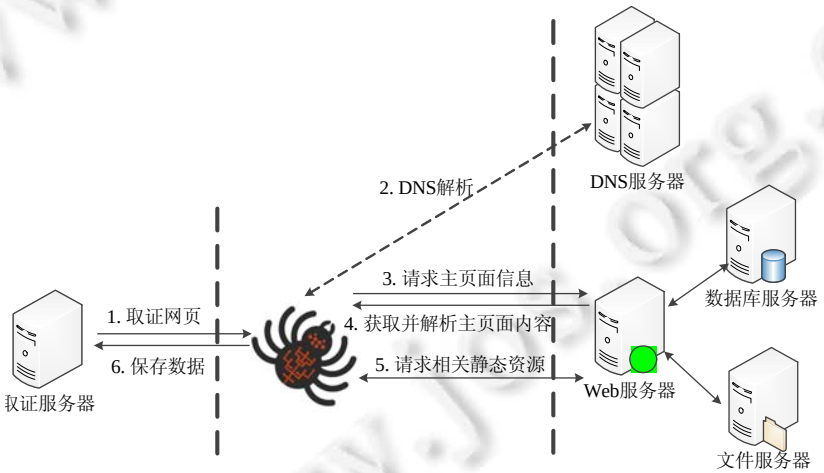


Fig.4 Procedure of web crawling

Fig.4 网络爬虫取证过程

在整个网页取证系统如图 5 所示,包括以下几部分:

(1) HTML 和 CSS 解析器解析

此模块的主要功能是解析 HTML 和 CSS 中所存在的网页资源链接.解析得到的链接会放入 URL 队列中,供爬虫进一步爬取.

(2) 过滤及数据提取

在处理爬取数据时,可以经过过滤和数据提取,以便提高实际的爬取效率.在爬取过程中,使用正则表达式

和关键字来预先定义要忽略的网页链接,用户可以根据需求忽略某些链接,加快爬取速度,减少无用数据.用户也可以定义需要提取的内容,例如文章标题等,辅助对网页内容的分析.

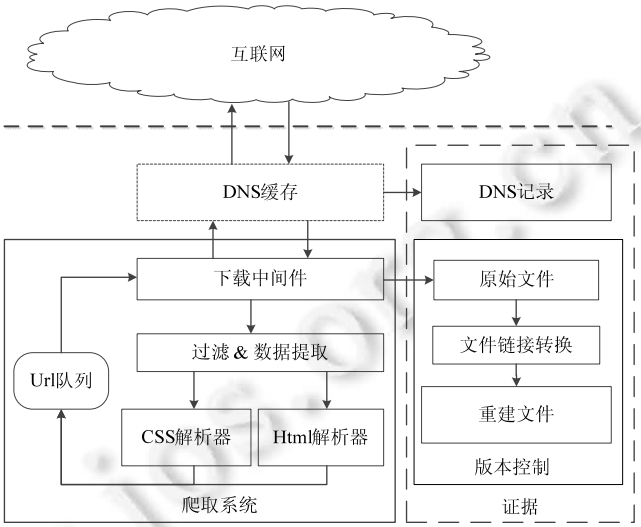


Fig.5 Structure of the crawling system

图 5 爬取系统结构图

(3) 版本控制与文件存储

网页会实时快速变化,因此需要对网页每个版本根据不同时间来保存.通过版本控制,系统可以在用户指定的时间间隔内检测网页是否发生变化,并对每次保存的不同文件均做记录.通过设置 HTTP Header 中的 Last-Modified 字段来实现文件的缓存技术,对于返回为“304”,则表示当前网页文件未发生变化,无需再次保存.利用版本控制功能,在节省数据存储空间的同时,也能随时记录服务器上文件的变化情况.保存下的文件,大多是不会改变的文本形式的小文件,可以存储于基于 MongoDB 的分布式文件存储系统中,具备横向扩展功能,提高实际分析时的读取速度.

(4) DNS 缓存记录

记录 DNS 查询结果,使得取证服务器如果遭受了恶意 DNS 欺骗攻击,也能够正确区分错误结果,避免提供错误的证据,确保整个取证过程的可信.

(5) 网页重建

取证的社交网页内容实时变化,为了能够重现对取证时刻的网页,需要在保留原始网页文件副本的前提下,修改和重定向网页中的链接内容,使其指向本地取证资源,随后,利用 WEB 服务器重新提供对这些网页的访问功能.

在开始取证时,用户需要设定起始爬取网址、爬取层次数、爬取方式以及过滤链接.设置完毕后,网络爬虫会自动根据相关设置执行主动取证任务;同时,网络嗅探程序会开始嗅探 DNS 和 HTTP 协议数据包,并记录保存到本地.

2.3.3 证据固定

根据网络爬虫的爬取结果,每次访问一个链接,便保存为一个文件.保存的内容不经过任何修改,对保存下来的文件计算哈希值,并保存到数据库中,作为证据,并防止遭受后期的篡改:

$$hash_1=f_{hash}(file) \tag{1}$$

此外,抓取时间、抓取的 URL 地址都是证据的重要部分,这些数据也是需要保存到数据库中防止被篡改(如图 6 所示):

$$hash_2=f_{hash}(file+time+url) \quad (2)$$

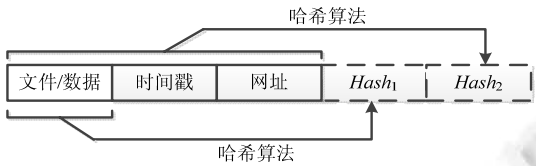


Fig.6 Components of an evidence item

图 6 证据组成

完整的一组证据内容包括网页文件、时间戳、网址,表 1 列出了对 www.qq.com 取证所得到的部分文件与对应的哈希值.

Table 1 Some crawling data from the wetsite of http://www.qq.com and their hash values

表 1 从 http://www.qq.com 下取证的部分内容及其哈希值

序号	抓取时间	网址	SHA-256
0	2013-06-13 23:51:11	http://www.qq.com/index.html	3526b1e8db7cf746f459ea035096609c 527d8f6d53d3d5d8a3a8d243fc9fb5d1
1	2013-06-13 23:51:11	http://mat1.gtimg.com/www/icon/favicon2.ico	a139a76e213346733427ad54e330f06 c23a4ca66576b11d6ed3b8439e739c6
2	2013-06-13 23:51:11	http://mat1.gtimg.com/www/images/qq2012/floatContract.png	b8c5b90a3043c7c572245d06c199569d 6b31c6a3dd85a7a476d5c50e949fcdd4
3	2013-06-13 23:51:11	http://mat1.gtimg.com/www/images/qq2012/alphabg50.png	2c83be886dde64561a202b2bc1f0caa1 b07ed792c902c176252659274ed3f1e2
4	2013-06-13 23:51:11	http://pingjs.qq.com/ping.js	e374e7a91aa4bfe28bfa4b5733ad99bc d84a503e6dddb7fb25deec0287364a9a
5	2013-06-13 23:51:15	http://adsrich.qq.com/web/crystal/v1.9Beta07Build073/crystal-min.js	8b62476d5905b8e8d8b200327eb902 a3f55171af2b45557639516fa4853c3b
6	2013-06-13 23:51:15	http://mat1.gtimg.com/www/js/Koala/Koala.min.1.5.js	c3f97969241998541a193dea7f4edd05d fbca5ed2890ee40ba0bfe8d550e9e2d
…	…	…	…

2.3.4 再现过程

保存下的内容只是文本文件,虽然可以直接进行浏览,但是其中包含了大量的 HTML 标记语言,难以直接理解.为了有效地在法庭上直观展示取证结果,需要将这些文本内容中重现在浏览器中.

默认取证得到的是原始网页文件,可以直接从浏览器中打开,但是其加载的资源仍然为 Web 服务器上所提供的远程资源,一旦远程资源在取证之后有所变化,会导致显示错误的内容.为了能够正确地显示取证内容,需要对保存的内容进行修改,将指向远程内容的链接修改为指向本地取证的内容,使其调用取证时所保存的相关资源.此时的网页内容上已经发生了变化,仅仅只能作为辅助展示的内容,而不是原始证据.

对于网页中的链接,需要使用 HTML 解析器从中解析出可用链接,并根据数据库中取证时该链接对应的本地文件,替换为本地链接.除了网页文件外,CSS 样式文件中也包含一定的链接地址,需要进行解析和重定向.

2.4 网页截图

内容爬取层所固定的文件可以提供大量的证据,但是因为其主要的内容都是文本信息,因此在抵抗篡改能力上相对较差.虽然能够将网页内容再现到浏览器中,但是保存的文件可以提供大量信息,但是这些信息多为带有大量冗余标记的文本语言,并不是可视直观的内容,不适合在法庭上直接出示给无相关专业知识的控辩双方.此外,固定的网页内容也仍然存在一定的被篡改的可能性.为了提高可信度,针对取证的网页,在取证过程中实时进行网页截图.

网页截图部分主要利用 Webkit 模块,对爬取的内容进行渲染和截图.本部分的具体实现利用 Webkit 的 Python 封装接口实现,利用了 Python 语言具有的快速开发、功能强大等特点.Ghost.py 是 Webkit 在 Python 环境下的一个 Web 客户端,它对 Webkit 的功能进行集成后封装,以方便用户快速、高效地进行网页固定开发.本文

主要用到 Ghost.py 的网页解析模块、Waiters 模块以及截图模块.

具体过程是:

Step 1. 使用 Ghost.py 提供的网页打开模块把 HTTP 文件以及所有相关的 CSS 文件、Javascript 文件和图片文件等加载到本地 HttpRessource 对象中.

Step 2. 使用提供的 HTTP 解析模块中 URL 提取模块提取所有子页面 URL,经过筛选后加入等待队列.

Step 3. 使用截图模块把本地缓存的 HttpRessource 对象保存为经过解析的网页图片并保存在本地.

Step 4. 对等待队列中待处理的 URL 重复 Step 1~Step 3,直至等待处理队列为空.

2.5 证据交叉验证

通过 3 层内容固定,同时能够得到 3 份可以交叉验证的证据:

(1) 二进制数据:嗅探得到的网络数据流;

(2) 保存文件内容:主要是以文本形式存在的 HTML 文件、CSS 文件、JavaScript 文件等;

(3) 网页截图:是一份直接可视化的内容.

对于这 3 份证据副本,均计算了哈希值,保证不被篡改,能够交叉印证,确保证据的可信度,即使其中一份被篡改时,也仍然可以依靠三者之间的交叉校验,能够确保相关数据的原始性.多副本的证据也能在一定程度上提高篡改的难度.

对于这 3 种同时存在的取证固定方式,可以根据不同层次形式同时记录来保证取证数据的正确性以及不可被篡改性.一旦其中一份被篡改,可以使用其他两份通过各自方法检验其正确性,甚至恢复被篡改的证据.

由于 HTTP 数据包中所包含的具体数据内容可以通过各种解析得到第 2 层文件证据内容,并且第 1 层向第 2 层的提取过程是不可逆的;而 HTTP 文件、CSS 样式、JavaScript 脚本等文件与网页截图是多对一的关系,而由于代码编写习惯等因素存在,不同的 HTTP 文件、CSS 样式、JavaScript 脚本通过 Webkit 的解析和页面渲染后存在得到同一个页面效果的可能性较低,并且当初的网页渲染截图仅仅包含图像信息,因此,第 2 层向第 3 层的数据提取过程也是不可逆的.由于这两层信息提取不可逆关系的存在,导致只有低层次证据能恢复高层次的证据,而无法从高层次的信息恢复得到更加丰富的低层次数据.

- 假设只有第 3 层的数据被篡改,可以使用第 1 层的数据通过模拟数据包的解析与爬虫框架爬取,重新得到第 2 层证据,再与第 2 层数据作对比.如果两者内容完全一样,则证明证据依旧有效.也可以经过被还原的第 2 层证据或原始的第 2 层证据来恢复被篡改的第 3 层数据;
- 假设只有第 2 层的证据被篡改,可以使用第 1 层的证据先恢复出第 2 层的数据,再由恢复出的证据提取出第 3 层的证据,再与原始的第 3 层数据作对比.如果两者内容一致,则证明证据有效;
- 假设只有第 1 层的证据被篡改,由于第 1 层证据提取的不可逆性,导致不能完全恢复第 1 层的证据.只能使用第 2 层的证据提取后与第 3 层的证据做对比,如果两者内容一致,则证明证据有效.

3 社交网页取证数据分析

社交网页相比一般的网页有着特定的访问权限与隐私设置,这意味着用户并不是简单地将自己的信息公开在网上,可以选择性地针对不同的人群进行公开.社交网页的限制主要有两种情况:

(1) 网站本身的限制

需要用户登录后,才能(完整)看到用户公开的内容,例如新浪微博等.对于这种限制,网络爬虫可以模拟用户登录动作,获得登录后的会话,最终获取用户内容的访问权限.

(2) 用户指定隐私限制

用户只对指定的浏览者开放浏览权限,普通浏览者无法获取这些内容,例如人人网、QQ 空间等,只有成为好友用户才能访问主页内容.这时候,如果需要固定这些内容,则需要取证网站提供相应的授权和权限.

3.1 文本内容分析

微博中的文本是用户发布最多的信息形式,同时,新浪微博也以多种形式扩展了文本的内容,使得单一的文本能够含有更多的含义,这些扩展形式如下:

(1) @符号表示文本中提到某人;

(2) #符号表示涉及到某个话题;

(3) //符号表示其后内容为他人发表,可以是对另一用户内容的回复,或者是直接转发.

通过对这些文本内容的分析,可以了解到微博内容中的相关话题、参与者等信息.这一部分主要涉及到对文本内容的分割与提取,利用微博中含有特定含义的字符对文本内容进行分割,提取出评论、转发内容、话题与提到的人等信息.

3.2 文本感情色彩分析

随着互联网的迅速发展,社交网络也呈现出爆炸性的增长.大量微博数据的出现,已经严重地影响了人工取证速度,增加了大量的时间成本和人工费用,也严重影响了取证工作的效率.为此,开发出一种能够自动甄别微博文本感情色彩并快速分析和分类的功能迫在眉睫.

基于文本的情感分析是一个复杂的交叉研究,它涉及多项研究,例如数据挖掘、自然语言处理、机器学习等等.基于文本的情感分析主要包括机器学习方法和语义分析的方法:基于语义的情感分析一般由现有字典或者词库拓展生成不同语义倾向的语义库,利用语义分析等方法实现对文本的情感识别;基于机器学习的情感分析一般方法有根据句子句义分词提取特征和计算权值,再通过训练库对分类器进行训练后由分类器进行分类.短文本的独特性有:长度很短,微博的长度一般被限定到 140 字内;不规范的语句多,交互性强,语义分析艰难,存在大量的新词和变形词,并且特征词数量少.由于这些特征,微博短文本中所包含的感情色彩分析一直是难题.因为短文本语义分析艰难,本文采用机器学习的方法进行情感分析.

本节使用常见的正负词词汇组成的传统情感词典,并且在词典中添加最近新出现的网络词汇,组成适用于微博情感分析的情感字典.由分词系统提取出文本主要词语,利用情感词典提取文本的情感特征值,再用朴素贝叶斯分类器,利用提取出的特征值对文本进行分类.

本文使用了基于统计的分词算法:首先,基于 Trie 树结构实现高效的词图扫描,生成句子中所有可能成词情况所构成的有向无环图(DAG);采用动态规划,从有向无环图中找到最大概率路径,找到基于词频的最大组合;对于未找到的词组,使用 HMM 模型进行分词[16].

本文使用的特征提取方法为基于词典的正负词词频.使用经过扩展的微博情感字典与文本分词得出数据进行比对,即得到一个 *dict*,键为字典中的词,值为该词在句子中出现的次数,并使用该 *dict* 结构作为特征值.

通过微博文本情感色彩分析,可以在社交网络的海量数据中查找到潜在的犯罪证据或嫌疑,缩小取证范围,加速取证过程和提高准确性.有些与犯罪行为关联度高的,也可以作为法庭上使用的相关网络犯罪的佐证.

3.3 位置信息提取

当用户使用移动设备发布微博时,可以选择将由手机获得的位置信息附带发送到微博上.这些位置信息体现了用户的移动轨迹与出现场所,是重要的取证内容.

手机等移动设备的定位,主要是基于手机通信基站位置确定具体位置的,部分设备也会根据内置的 GPS 和 WIFI 连接到的无线网络进行定位.对于新浪微博等大公司,通过对用户位置信息的搜集,可以对 3 种定位方式的定位结果进行交叉比对,最终达到较高的定位精度.虽然定位结果并不能非常精确,但是仍然可以作为用户活动地点的一个实际证据.

新浪微博主要利用了谷歌和高德地图提供的地图服务.每当用户发送位置时,新浪微博会直接在微博消息中附带上当前地点的附近地点的一个名称,然后将详细的地址映射为一个短连接,如 http://t.cn/xxxxxxx 的形式.还原此链接之后,能够得到一个显示该地点详细位置的网页,该网页包含经纬坐标.通过解析这个网页,可以获取比较精确的位置.

位置链接包含两种形式:

(1) 明文经纬度

解析后的位置链接形如 http://weibo.com/p/100101*AAA.aaa_BB.bbb*,其中,*AA.aaa* 和 *BB.bbb* 分别是经度、纬度的坐标.通过这两个数值可以直接地图定位.

(2) 转码网址

解析后的位置链接形如 http://weibo.com/p/100101*FFFFFFFFFFFFFFFFFFFF*,包含一串 16 进制.当定位到的位置是一个商业性地点时,会采用这种链接形式,并附带有一定的商业宣传广告.通过对这个网页进行进一步的解析,可以从中提取出地图的经纬度.

本节以微博用户 Tinyfool 为例,对其进行了位置取证实验.Tinyfool 是国内知名的 IOS 开发者,在微博上非常活跃,近期,他从上海到杭州进行了一次游玩,并在相关微博中发布了相关的地点信息.通过对其近期的微博数据的爬取,总共获得了 13 条与此次旅行相关的含有位置信息的微博,见表 2 和如图 7 所示.

Table 2 Detailed information of Tinyfool's locations

表 2 Tinyfool 近期活动地点的详细信息

时间	地点名称	精确坐标	
		经度	纬度
2014-04-10 21:33	上海市 虹桥火车站	121.32152	31.19268
2014-04-10 21:46	上海市 动物园/虹桥机场	121.346159	31.193901
2014-04-10 22:24	上海市 宁路/七浦路	121.486748	31.25086
2014-04-11 13:02	上海市 五角场	121.505699	31.297501
2014-04-11 16:26	上海市 96 广场	121.52503	31.22758
2014-04-12 11:37	杭州市 南庄兜枢纽	120.119003	30.386999
2014-04-12 16:11	杭州市 迪臣南路	120.081001	30.302999
2014-04-12 16:22	杭州市 迪臣南路	120.081001	30.302
2014-04-12 16:43	杭州市 花蒋路	120.078003	30.299999
2014-04-12 18:06	杭州市 留祥路	120.081001	30.313
2014-04-13 20:26	杭州市 新井路	120.208	30.292999
2014-04-13 21:15	嘉兴市 长安互通	120.433998	30.476999
2014-04-14 18:30	上海市 机场南路	121.498001	31.312

Fig.7 Tinyfool's location path

图 7 Tinyfool 位置轨迹图

通过社交网络中用户关键位置信息的提取,结合相关案例及其时间信息,可以形成有效的证据链,可以为相关案件提供法律上的佐证.

3.4 人脸图像识别与标注

随着手机等移动设备的普及,人们能够随时随地在社交网站上发布照片,而发表自己或者与他人的合影俨然已经成为一种时尚.对于计算机取证而言,图像是重要的取证对象,而含有人脸的图像则能够表明人的行为和人与人之间的密切关系,是取证过程中需要重点关注的内容.但是社交网页中的图片数据量巨大,仅仅依靠人力去进行低效的人脸分类、辨别是一件几乎不可能完成的任务.因此,需要依靠技术手段实现对人脸图片的自动发掘、标注和识别.本文利用 AdaBoost 人脸识别方法和 SVM 训练器[17]实现了在社交网页下自动发现识别人脸,依靠已有的手工标注内容,自动对为标注人脸进行匹配标注,在一定程度上降低了社交网络取证中人脸发现和识别的难度,提高了取证效率.

(1) 人脸提取

人脸识别的第 1 步需要检测到图像中的人脸,然后将人脸部分的图像区域提取出来.本系统主要利用 AdaBoost 级联分类器,该分类器是目前比较成熟的人脸检测算法,在 OpenCV 库中有较为成熟的级联分类器代码实现与较全面的人脸模型,利用现有的算法库,可以方便地从图像中识别出人脸.

(2) 人脸矫正

经过初步的人脸检测,可以提取出 AdaBoost 级联分类器检测出的人脸.但是提取出的脸都不一定是正脸,经常会出现倾斜甚至翻转的人脸,这时会对以后的人脸标注产生不利的影响.为解决这个问题,可以利用人脸中的眼睛来做人脸姿态判定,通过合成平均滤波器(ASEF)来识别并判断双眼的位置,并最终确定人脸的中间位置,再加以辅助调整人脸姿态,能够有效提高后续标记效果.

(3) 提取人脸特征

将调整好姿态的人脸切割为 120×160 像素图片,然后使用高通滤波器滤波以减少灯光对图片的影响,再经过主成分分析(PCA)提取特征值后,放入 SVM 分类器中进行分类.

(4) 人脸手动与被动标注

对于最初几次出现的人脸,需要手动地添加标签,添加标签后,根据分类结果决定是否继续添加或更正标签.一般性情况下,每张人脸前几次的标签均需要手动进行更正与重新标注,但当标注次数达到一定程度,分类正确率到达阈值,SVM 分类器可以自动对人脸进行较高准确率的标注,对后续标注的内容,仍然会自动提取特征并添加到相应训练库中,达到训练库自动训练、自动识别、自动标注的效果.

本次实验选取的图像来自社交网络图片(林志颖、王诗龄、李湘和何炅这 4 位名人的微博).各取 25 张图片进行实验.对 100 张微博原始图片使用 AdaBoost 算法进行人脸检测,并切割成固定大小.检测效果可见表 3:王诗龄的识别率是最低的,通过重新观察所选取的照片发现,原因是儿童拍照时并不如其他成人那样始终保持正脸,会出现大量的侧脸与歪脸,正是这些侧脸与歪脸导致了人脸的识别成功率较低;李湘的识别率最高,全部识别成功,因为相比何炅和林志颖,李湘的微博照片中人脸都是正脸居多.

Table 3 Detection rates of face recognition

表 3 人脸检测识别率

	林志颖	何炅	李湘	王诗龄
总样本数	25	25	25	25
检测成功数	21	22	25	17
检测成功率 (%)	84	88	100	68

针对已提取到的人脸图像进行裁剪和调整,整理后见表 4.从识别成功的图像中各抽出 5 张作为训练集,将剩下的作为测试集.例如:王诗龄识别成功的 17 张照片中,5 张作为训练集,剩下的 12 张作为测试集进行测试.测试结果见表 5,可以发现:李湘照片的标签识别率最低,主要原因是因为李湘经常化妆,而其他 3 人浓妆照片的数量与李湘相比明显较少,因而识别成功率比李湘高;王诗龄识别率依然较低,原因仍然是人脸姿势非常多样化.

Table 4 Results of face recognition in a social network

表 4 社交网络中人脸检测结果

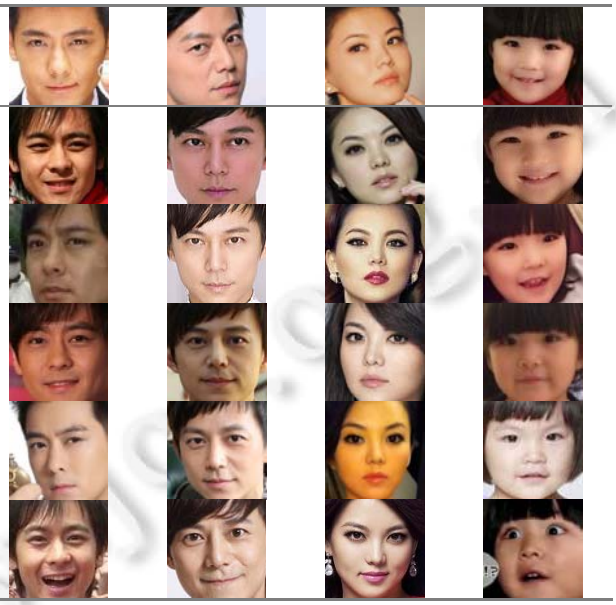
林志颖	何炅	李湘	王诗龄
			

Table 5 Testing rates of face recognition (%)

表 5 人脸识别效果测试(%)

训练图片数量	1	2	3	4	5
林志颖	69	81	93	100	100
何炅	70	76	88	94	100
李湘	60	70	85	90	90
王诗龄	66	75	91	91	100

在社交网页取证系统中,会根据用户发表的微博中的图片自动筛选含有人脸的照片,并对其进行标注,如图8所示.如果系统给出错误的标注,则用户可以直接进行修改,重新标注.结果多次迭代之后,可以达到较高的人脸识别率.

Fig.8 A face recognition interface in Weibo forensics

图 8 微博取证人脸识别界面

4 总结和展望

目前,传统的计算机取证技术由于社交网络取证信息的多样、动态(实时)、海量和交互等特点而不再适合,亟需发展面向社交网络取证的新型取证模型和计算方法.针对社交网页取证证据收集困难问题和证据可信问题,本文提出了一种取证方法——主动通过网络爬虫对网页数据进行取证的方案,设计了 3 层取证模型框架,通过不同层次间的证据验证,确保取证内容的可信度;将证据重现,使其能够被直观有效地展示,提高了证据的可理解性.

针对社交网络中(与人交互的)证据的交互性和复杂性问题,本文针对社交网页中包含的文本、位置、人脸图像等重要信息进行取证.对用户微博位置进行定位,显示到地图上绘制行为路径;自动提取含有人脸的图像,结合人工标记和自动预测的手段标记人脸;对文本的情感色彩进行提取,展示用户的情绪波动.通过上述方法,自动提取信息,提高了在社交网页中取证效率,降低分析难度,提供更加直观和可以被法庭理解和接受的证据.

本文针对社交网络的取证,在网页数据获取、网站内容分析等方面进行了初步探索.如何在海量复杂的社交网络数据中快速而有效地收集、检索、分析和展示客观、真实、完整、丰富、可理解而且有针对性的证据,需要进一步系统而深入地展开研究.社交网站中重要的信息是用户关系,本文还没有对社交网站用户关系进行详细的取证和分析,在我们的后续研究中,需要进一步对用户关系进行取证,并挖掘人群所属社群,以便用于取证侦查.此外,结合具体的社交网络犯罪模式与相关案例,提出针对性的取证模型和方法研究,也是未来的研究方向之一.

References:

[1] Wall D. Cybercrime: The Transformation of Crime in the Information Age. Polity, 2007. 14−16.

[2] Dixon PD. An overview of computer forensics. Potentials, IEEE, 2005,24(5):7−10. [doi: 10.1109/MP.2005.1594001]

[3] Garnkel SL. Digital forensics research: The next 10 years. Digital Investigation, 2010,7:64−73. [doi: 10.1016/j.diin.2010.05.009]

[4] Kozushko H, Digital evidence. 2003. http://infohost.nmt.edu/~sfs/Students/HarleyKozushko/Papers/DigitalEvidencePaper.pdf

[5] Agarwal A, Gupta M, Gupta S, Gupta SC. Systematic digital forensic investigation model. Int'l Journal of Computer Science and Security (IJCSS), 2011,5(1):118−131.

[6] Pilli ES, Joshi RC, Niyogi R. A generic framework for network forensics. Int'l Journal of Computer Applications, 2010,1(11).

[7] Slay J, Simon M, Irwin D. Voice over IP and forensics: A review of recent Australian work. In: Proc. of the 1st Int'l Conf. on Digital Forensics and Investigation (ICDFI). Beijing, 2012.

[8] Meghanathan N, Allam SR, Moore LA. Tools and techniques for network forensics. Int'l Journal of Network Security & Its Applications (IJNSA), 2009,4,1(1):14−25.

[9] Zhong LD. Forensic analysis of Web browser with dual layout engine. In: Proc. of the 1st Int'l Conf. on Digital Forensics and Investigation (ICDFI). Beijing, 2012.

[10] Kent K, Chevalier S, Grance T, Dang H. Guide to integrating forensics into incident response. 2006. http://csrc.nist.gov/publications/nistpubs/800-86/SP800-86.pdf

[11] Erbacher RF, Christensen K, Sundberg A. Visual network forensic techniques and processes. In: Proc. of the 9th Annual NYS Cyber Security Conf. Symp. on Information Assurance. 2006. 72−80.

[12] Wu CS, Qiu M. The method for obtaining electronic evidence from ASP dynamic website. Forensic Science and Technology, 2010,(5):43−45 (in Chinese with English abstract).

[13] Ruan KY, Carthy J, Kechadi T, Crosbie M. Cloud forensics: An overview. In: Proc. of the Advances in Digital Forensics VII. 2012. 15−25.

[14] Li YD, Hu DH, Fan YQ, Wu XD. Web page forensics: A Web spider based approach. In: Proc. of the 2nd Int'l Conf. on Digital Forensics and Investigation (ICDFI). 2013.

[15] Huber M, Mulazzani M, Leithner M, Schrittwieser S, Wondracek G, Weippl E. Social snapshots: Digital forensics for online social networks. In: Proc. of the 27th Annual Computer Security Applications Conf. (ACSAC 2011). 2011. 113−122. [doi: 10.1145/2076732.2076748]

[16] Qian X, Zhang Q, Huang XJ, Wu LD. 2D Trie for fast parsing. In: Proc. of the 23rd Int'l Conf. on Computational Linguistics (COLING 2010). Beijing: Association for Computational Linguistics, 2010. 904–912.

[17] Bolme DS, O'Hara S. PyVision—Computer vision toolkit. 2014. https://github.com/bolme/pyvision

附中文参考文献:

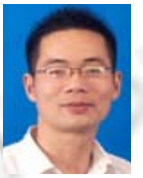
[12] 吴春生,邱敏.ASP 动态网站信息电子取证方法.刑事技术,2010,(5):43–45.

吴信东(1963－),男,博士,教授,博士生导师,IEEE Fellow,AAAS Fellow,主要研究领域为数据挖掘,知识库系统,万维网信息开采.

E-mail: xwu@hfut.edu.cn

李亚东(1989－),男,硕士,主要研究领域为计算机取证.

E-mail: liyadong.hfut@gmail.com



胡东辉(1973－),男,博士,副教授,CCF 会员,主要研究领域为网络信息可信度量,数字取证和隐私保护.

E-mail: hudh@hfut.edu.cn