第 34 卷 第 11 期 计 算 机 工 程 2008 年 6 月
Vol.34 No.11 Computer Engineering June 2008
·开发研究与设计技术· 文章编号：1000—3428(2008)11—0263—03 文献标识码：A 中图分类号：N945

基于关键词匹配的打印数据获取系统

余扬武，吴顺祥
(厦门大学系统与控制研究中心，厦门 361005)

摘 要：针对计算机取证中的电子证据问题，设计并实现基于关键词匹配的打印数据获取系统。通过对硬盘中假脱机文件关键词的搜索，定位打印数据所在的物理扇区并读出相关内容。性能测试显示，该系统能快速有效地获取打印内容，适用于信息保护、电子取证等环境，具有较高的实用价值。

关键词：电子证据；数据获取；关键词匹配

Printing Data Acquisition System Based on Keywords Matching

YU Yang-wu, WU Shun-xiang
(Center of System and Control Research, Xiamen University, Xiamen 361005)

【Abstract】 In order to solve the problems of the electronic evidence about computer forensics. This paper designs and implements a printing data acquisition system. based on keyword matching. The physical sector of printing data can be located by searching keywords of spooling file in hard disk. Then the printing data can be acquired efficiently. The testing results prove that this system has a good performance and is useful in information protection and computer forensics.

【Key words】 electronic evidence; data acquisition; keywords matching

电子证据综合了文本、图形、音频及视频等多种媒体信息，其来源主要有系统日志、系统的审计记录、网络监控流量和 E-mail 等。另外，打印机打印的内容也可以作为电子证据，如犯罪嫌疑人打印过的报表、文档、图片等。目前针对此类电子证据的取证办法不多，因此，本文提出了基于关键词匹配的打印数据获取系统。

1 相关技术简介

1.1 计算机取证

目前国内许多法学界学者对计算机证据定义如下：在计算机或计算机系统运行过程中产生的以其记录内容来证明案件事实的电磁记录物，也称为电子证据[1]。电子证据的来源很多，主要有操作系统日志、IDS和防火墙等安全设备的日志、网络上采集的数据流、传输的数据等[2]。

“计算机取证”首先由International Association of Computer Specialists(IACIS)在 1991 年举行的第一次年会中正式提出。计算机取证专家JuddRobbins对计算机取证定义如下：将计算机调查和分析技术应用于确定并获取潜在的、有法律效力的证据。计算机紧急事件响应和取证咨询公司New Technologies进一步拓展了该定义：计算机取证包括对以磁介质编码信息方式存储的计算机证据的保护、确认、提取、归档[3]。

综上所述，计算机取证是对能被法庭接受的、足够可靠且有说服力的、存在于计算机和相关外设中的电子证据的确认、保护、提取和归档的过程，也称为计算机法医学[4]。

1.2 Windows 打印系统原理

Window 打印系统主要由 2 个部件组成：图形设备接口(Graphics Device Interface, GDI)及其支撑模块，打印假脱机(Simultaneous peripheral operation on-line, Spooling)系统。这 2 个部件互相协作，共同完成 Windows 操作系统的打印工作。

(1)GDI 及其支撑模块

支撑模块主要是设备无关位图(Device Independent Bitmap, DIB)引擎和打印机驱动程序，它们将与具体打印设备无关的应用程序输出信息转换为与具体打印设备相关的输出信息。GDI 是 Windows 系统的一个核心部件，它是 Windows 图形功能的心脏，所有图形图像处理及字体处理、颜色管理等功能都在 GDI 中实现，同时其功能以应用程序编程接口(Application Programming Interface, API)的形式供应用程序调用。在 Windows 系统中所有与图形设备(如显示器、打印机、扫描仪等)相关的处理都与 GDI 有密切关系。对于图形输出设备而言，GDI 用设备描述(Device Context, DC)来维护任何输出设备的信息。应用程序不能直接对设备进行输出，它必须用 DC 设备描述或其他逻辑对象来调用 GDI 的功能。GDI 通过调用指定设备的设备驱动程序，将与设备无关的输出信息转换成与指定设备相关的输出信息，发送到指定的输出设备。这种结构安排消除了应用程序的设备依赖，使 Windows 系统能适应各种输出设备。

(2)Spooling 系统

Spooling 主要指当主机处理器给外部设备传送数据时，为了减少占用主机处理器的时间(因为端口速度通常远低于处理器速度)而采用的“把辅助存储器(通常为硬盘)作为端口的缓冲存储器来使用，具体的发送工作由后台来处理”的一

基金项目：福建省教委科技基金资助项目(JA05290)；厦门大学“985”二期信息创新平台基金资助项目

作者简介：余扬武(1979－)，男，硕士研究生，主研方向：智能信息系统，网络安全；吴顺祥，教授、博士

收稿日期：2007-10-08 **E-mail：**yuyangwu@hotmail.com

种方法。在 Windows 多任务环境下，其打印假脱机系统具有重要意义，体现为以下 3 点：

1)后台发送避免了长时间占用 CPU，使控制权迅速返回给用户以便进行其他操作。

2)便于多路由输出，可以将打印作业调度到本地或网络打印机，或者写到磁盘文件以便随后打印。

3)将底层有关端口的操作通过独立部件实现，简化了系统上层模块的实现。

打印假脱机系统由打印请求路由器、本地打印提供者、网络打印提供者、打印处理器、打印作业语言监视器和端口监视器 6 个子部件组成。

在 Windows 中有 3 种典型的打印流程：利用原始假脱机文件的打印流程，利用增强型图元文件(EMF)的打印流程和直接打印流程。下面将给出典型的利用原始假脱机文件的打印过程流程图，来明确 GDI、打印驱动程序及打印假脱机系统各子部件在一次打印过程中的作用和相互联系。

图 1 给出了从一个应用程序通过调用 GDI 发出一个打印请求开始，到本地打印提供者将假脱机文件写到硬盘并启动一个后台线程为止的过程。这个后台线程最终用于触发假脱机文件的解析过程，用图 2 来描述。

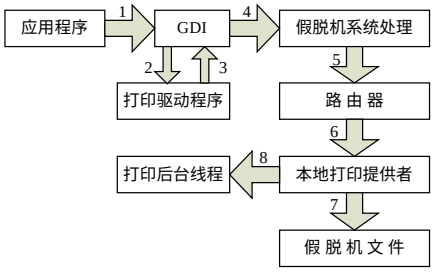

图 1　利用原始假脱机文件的打印流程

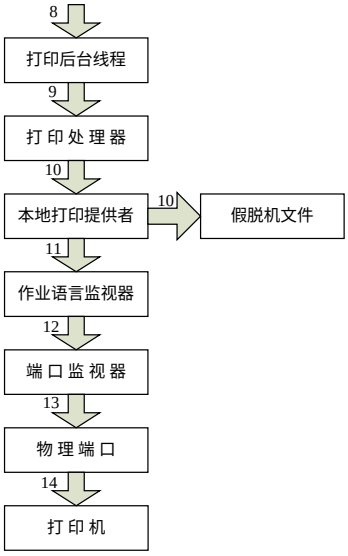

图 2　假脱机文件的解析过程

图 1 中调用过程的步骤如下：

1)应用程序创建一个 DC 并在其上绘制一个对象，然后调用 GDI 发出相对于使用此 DC 的打印机的一个打印请求。

2)GDI 调用相应的打印驱动程序来处理具体打印请求。

3)打印驱动程序创建一个打印作业，并调用 GDI 功能将处理结果送出。

4)GDI 对假脱机系统的处理器进行进程间调用。

5)假脱机系统处理器调用打印请求路由器将打印作业发送到应用程序指定的打印机。

6)路由器将打印作业发送到本地打印提供者(它还可以将打印作业发送给网络打印提供者)。

7)本地打印提供者将打印作业以原始假脱机文件的格式存在硬盘上。

以上第 1)步到第 7)步可能重复多次以产生一个完整的假脱机文件。

8)本地打印提供者启动一个后台线程，它将选定一个最佳时刻触发假脱机文件的解析过程。

图 2 是打印假脱机文件的解析过程。它从后台线程调用打印处理器开始，到本地端口监视器将具体指令和数据通过它所控制的端口发送给与之相连的打印机结束。

9)主线程基于对打印假脱机子系统资源的监视，选定一个最佳时刻触发假脱机文件的解析过程。在这个最佳时刻，主线程通过调用 StartDoc 函数启动打印处理器中的一个新线程开始解析工作。

10)收到 StartDoc 调用后，在第 9)步中启动的打印处理器线程将用 ReadPrinter 调用来激活本地打印提供者以便从硬盘读取之前生成的打印假脱机文件。

11)在收到 StartDoc 调用后，上述打印处理器线程还调用了一个 WritePrinter 函数来激活打印作业语言监视器(通过本地打印提供者)以便将数据通过物理端口发送到所连接的打印机上。

12)打印作业语言监视器调用端口监视器的功能来给打印机发数据。

13)端口监视器监测物理端口，通过物理端口给打印机发送数据。

14)物理端口与打印机的通信。

第 10)步到第 14)步将重复多次直到遇到假脱机文件的结尾或打印作业被取消，最后解析线程中止。

利用增强型图元文件(EMF)的打印过程以 EMF 为打印假脱机文件，而对于直接打印，打印提供者直接将打印作业发往目标打印机，没有形成或解析打印假脱机文件的过程。

2　打印数据的获取

2.1　基本思路

根据 Windows 操作系统打印任务的工作流程，为了获取打印内容，打印过程中产生的假脱机文件是首先需要研究的对象。每个打印任务产生后缀名为 SHD 和 SPL 的 2 个假脱机文件，它们在经过图 1 的流程后生成，并存放在系统后台打印文件夹，在打印任务完成之后由系统自动删除。后缀名为 SHD 的假脱机文件包含了一些和打印任务相关的重要信息。后缀名为 SPL 的假脱机文件包含了具体的打印内容，它有 2 种类型：和具体的打印机及其驱动程序相关的原始(RAW)打印内容，统一的 EMF 的打印内容。因此，在了解假脱机文件的基础上，可以采用以下步骤获取尽可能多的打印内容：

(1)根据关键词在硬盘中搜索得到所有 SHD 打印假脱机文件。

(2)从每个打印假脱机文件中获取有关打印该文件行为的相关信息，如文件名、打印机型号、打印的用户和计算机名称。

(3)根据文件名信息的字符串在硬盘上搜索对应 SHD 假脱机文件的 SPL 假脱机文件，并尽可能完整地恢复搜索到的 SHD 和 SPL 假脱机文件。

(4)如果得到的 SPL 假脱机文件类型是 EMF 格式的，不管它是否完整，都可以根据 EMF 的文件格式从中得到完整的或部分的打印内容。

(5)如果得到的 SPL 假脱机文件类型是原始 RAW 格式的，并且是完整的，就可以根据 SHD 假脱机文件里的打印机型号信息建立一个相似的软硬件仿真环境(操作系统版本和打印机型号一样即可)，最后把恢复的 SHD 和 SPL 假脱机文件拷贝到后台打印文件夹，重新启动计算机后打印机就可以把假脱机文件的内容打印出来。

(6)可根据文件名使用计算机取证工具 ENCASE 软件在硬盘上搜索并恢复相应文件，如果成功，就能获取打印内容。

打印数据获取系统的结构如图 3 所示。

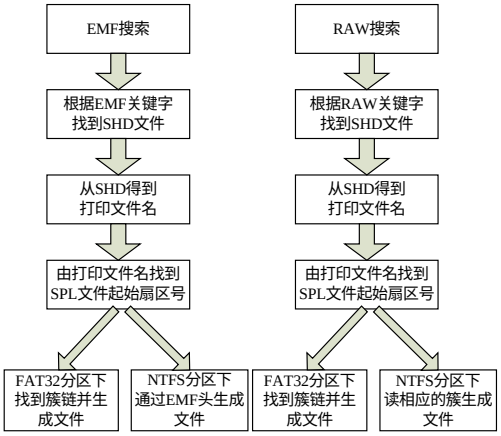

图 3　打印数据获取系统框架

2.2　系统的具体实现

系统首先对硬盘实行第 1 次遍历，将关键词与硬盘所有扇区的内容进行匹配，在命中扇区中解析出用户名称、打印机型号、端口号、文件名等信息。此后根据文件名信息的字符串对硬盘实行第 2 次遍历，找到目的扇区，提取扇区内容，生成 SPL 假脱机文件，最终从该假脱机文件得到完整的或部分的打印内容。

以 Windows 2003 操作系统为例，本文做了如下研究：

(1)对 EMF 类型 SHD 假脱机文件包含的相关打印信息进行研究，结果如图 4 所示。其中包括打印该文件的计算机系统用户名称、打印文件名称、打印机型号、EMF 类型 SHD 假脱机文件关键词(如 WinPrint NT EMF)、计算机名称。

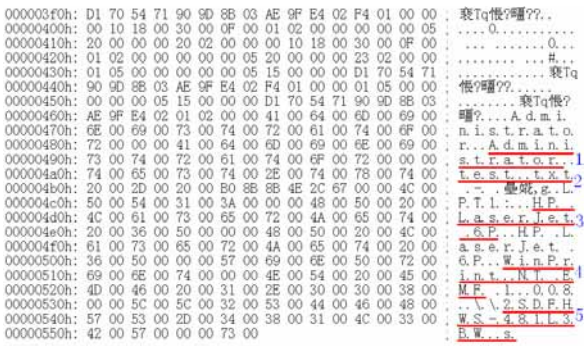

图 4　EMF 类型 SHD 假脱机文件扇区部分内容

该 EMF 类型的 SHD 假脱机文件相应的 SPL 假脱机文件包含的相关打印信息如图 5 所示。

图 5　EMF 类型 SPL 假脱机文件扇区部分内容

由图 5 可以看出，从 SHD 文件中得到的“ 打印文件名称 ”在相应的 SPL 文件中再次出现。

(2)对 RAW 类型 SHD 假脱机文件包含的相关打印信息进行研究，主要是假脱机文件关键词不同，结果与图 4 类似。其中包括打印该文件的计算机系统用户名称、打印文件名称、打印机型号、RAW 类型 SHD 假脱机文件关键词(如 IMFPrint RAW)、计算机名称。

该 RAW 类型的 SHD 假脱机文件相应的 SPL 假脱机文件包含的相关打印信息与图 5 类似，从 SHD 文件中得到的“ 打印文件名称 ”在相应的 SPL 文件中也再次出现。

(3)根据对 Windows 98, Windows 2000, Windows XP 各版本操作系统下 EMF, RAW 两种类型假脱机文件信息和打印系统的研究，可得到以下 6 种关键词：WinPrint NT EMF, WinPrint RAW, IMFPrint NT EMF, IMFPrint RAW, ModiPrint NT EMF, ModiPrint RAW。

由以上分析结果可知，分别对上述 EMF, RAW 类型的 SHD 假脱机文件关键词进行搜索，可以得到硬盘上曾经打印过的文件名称，再根据该文件名信息的字符串定位相应的 SPL 假脱机文件在硬盘中的物理位置，即可读出相应扇区的内容。

本系统在 Windows XP 环境中 Visual C++ 6.0 下编程调试通过，效果良好。

3　结束语

本文在深入分析打印系统原理的基础上，实现了基于关键词匹配的打印数据获取系统。此系统适用于对硬盘的计算机取证，为打印监控和打印管理提供了借鉴，对信息保护部门、行政管理部门维护信息安全有重要意义。

参考文献

[1] 钱桂琼, 杨泽民, 许榕生. 计算机取证的研究与设计[J]. 计算机工程, 2002, 28(6): 56-58.

[2] 梁锦华, 蒋建春, 戴飞雁, 等. 计算机取证技术研究[J]. 计算机工程, 2002, 28(8): 12-14.

[3] Bamshad M, Cooley R, Srivastava J. Data Preparation for Mining World Wide Web Browsing Patterns[J]. Journal of Knowledge and Information System, 1999, 1(11): 285.

[4] 赵小敏, 陈庆章. 打击犯罪新课题计算机取证技术[J]. 计算机技术信息网络安全, 2002, (9): 23-25.