

基于 COM 技术的 DBX 邮件文件解析

曾春溪, 蔡剑怀, 杨俊彬, 吴顺祥

(厦门大学自动化系, 厦门 361005)

(wxs1009@163.com)

摘要: 电子邮件客户端 Outlook Express 下保存的邮件数据文件, 蕴藏着丰富的个人信息, 挖掘分析其中的有用线索已成为计算机调查取证的重要手段和研究方向, 而首要条件就是需要对这些经过编码的数据文件进行解析, 将所有的邮件信息还原出来。针对此问题, 提出了 1 种利用 Outlook Express 提供的 COM 组件接口对这些邮件数据文件直接进行解析处理的方法, 避免了研究其编码及内部逻辑架构的繁琐。

关键词: 电子邮件; Outlook Express; DBX 文件; COM 技术; 文件解析

中图分类号: TP393.098 **文献标志码:** A

DBX mail file parsing based on COM technology

ZENG Chun-xi CAI Jian-huai YANG Jun-bin WU Shun-xiang

(Department of Automation, Xiamen University, Xiamen Fujian 361005, China)

Abstract The Email data files kept by Outlook Express (OE) contain the rich personal information. So mining and analyzing the useful clues inside has become a significant means and research area for the computer investigation and forensics. The foremost is to parse the encoded data files and then restore all the mail information. Aiming at the problem, a solution that we adopt the COM component interfaces of OE to parse these mail data files was presented, avoiding the troubles of research on their encoding format and complex internal logical structure.

Key words E-mail; Outlook Express; DBX file; COM technology; file parsing

0 引言

随着信息技术和 Internet 的迅速发展, 电子邮件作为一种快捷便利的通信手段, 已经普及到人们的日常工作与生活中, 其间蕴含着丰富的个人信息, 是进行计算机调查取证的重要途径^[1]。电子邮件客户端 Outlook Express (OE) 作为网民常用的收发邮件的工具, 其保存的邮件数据文件是进行邮件调查取证的重要对象, 挖掘分析其中的有用线索能为案件侦破提供有力依据。实现这些目标的首要工作就是进行邮件数据文件的解析, 将经过编码的二进制数据文件还原成通常的邮件信息, 即从中提取相关邮件的收发件人地址、邮件主题、发送时间、邮件内容和附件等信息。

然而, OE 很多技术细节保密, 直接对其 DBX 数据文件进行格式解析, 研究其编码及内部逻辑架构, 国内外相关资料较少, 难度较大, 且解析效率和软件升级后的兼容性难保证。但考虑到 OE 和 Visual C++ 同为微软公司的产品, 微软提供了相关的 API 函数, 本文提出了一种基于 COM 技术的 DBX 邮件数据文件的解析方法, 即利用 OE 内置的 COM 组件接口实现对 DBX 数据文件的访问读取, 获得相应的邮件信息。经实例验证表明, 该方法稳定可行, 达到了解析的目的。

1 相关知识简介

1.1 OE 及其 DBX 数据文件

OE 是一种用于日常电子邮件和订阅新闻组管理的实用

软件, 通常集成在 Microsoft 的各种操作系统中, 灵活地使用这些功能, 可以更好地为工作和生活提供方便。

OE 所有的邮件和它的一些系统设置都存放在后缀名为 .dbx 的数据文件里。如果使用 Windows 2000/XP 操作系统, DBX 数据文件默认存放目录为: G:\Documents and Settings\ (User Name)\Local Settings\Application Data\Identities\ {User ID}\Microsoft\Outlook Express^[2], 其中: User Name 是操作系统的登录用户名, User ID 为 OE 的用户 ID, 对于不同的计算机和用户, 其值都不一样。该目录下默认的文件有: Folders.dbx, Pop3uid1.dbx, Offline.dbx, 收件箱.dbx, 发件箱.dbx, 已删除邮件.dbx, 已发送邮件.dbx, 草稿.dbx。通常, OE 的相关邮件信息存储在后 5 个 DBX 格式数据文件中。

1.2 COM 技术

组件对象模型 (Component Object Model, COM), 是组件对象之间相互接口的规范, 其目的是为了 提高软件的可复用度, 解决不同程序间的通信、互操作性问题, 以及软件的跨平台、跨网络应用问题。凡是遵循 COM 接口规范的对象彼此之间就能相互通信和交互, 即使这些对象是由不同的厂商、用不同的语言、在不同的 Windows 版本甚至不同的机器上编写和建立的^[3]。一个 COM 组件可以实现多个 COM 对象, 每一个 COM 对象能为客户提供多种服务。COM 对象对客户是透明的, 要使用 COM 对象提供的服务, 必须通过接口实现。每一个 COM 对象都可能有多个接口, 使用某接口之前, 必须先得到这个接口的指针, 然后才能调用接口中的相关函数^[4]。

收稿日期: 2008-03-17; 修回日期: 2008-05-26。 基金项目: 国家“十一五”科技支撑计划项目 (2007BAK34B04); 国家自然科学基金资助项目 (60704042); 厦门大学 985 二期信息创新平台项目 (2004-2007)。

作者简介: 曾春溪 (1979-), 男, 湖南邵阳人, 硕士研究生, 主要研究方向: 网络安全与取证; 蔡剑怀 (1975-), 男, 福建厦门人, 博士研究生, 主要研究方向: 智能信息系统、网络安全与取证; 杨俊彬 (1984-), 男, 福建漳州人, 硕士研究生, 主要研究方向: 智能信息系统、网络安全与取证; 吴顺祥 (1967-), 男, 湖南邵阳人, 教授, 博士, 主要研究方向: 智能信息系统、数据挖掘与知识发现、网络安全与取证。

1.3 操作 OE 的 COM 接口

在操作 Outlook Express 6 的头文件“msoeapi.h”中,定义了对 OE6 进行各种操作的函数,接口 IStoreNamespace 和 IStoreFolder 的定义就在其中。另外,在头文件“mimeole.h”中,定义了对 OE6 中邮件进行各种操作的函数,其中本文主要用到了接口 IMimeMessage 和 IMimeBody。

1) IStoreNamespace Interface。OE 的数据存储空间接口,利用此接口可以创建、遍历和修改邮件夹,同时可以拷贝及移动邮件^[5]。

2) IStoreFolder Interface。OE 的邮件夹接口,利用此接口可以创建、遍历和修改邮件^[6]。

3) IMimeMessage Interface。用于创建及操作邮件,包括获取和修改邮件的具体信息^[7]。

4) IMimeBody Interface。用于操作邮件体的标题及内容等^[8]。

2 基于 COM 技术的 DBX 邮件文件解析

Microsoft 公司的很多软件均深度支持 COM 技术,OE 也不例外,COM 提供了访问软件服务的一致性,不管要访问的服务存在于链接库还是另一个进程或系统软件中,均可将它们当成 COM 对象^[9],开发者能够利用面向对象的方法设计和开发程序去访问这些服务,不过前提是必须在运行系统上安装这些软件。

本文参考了 MSDN 的相关例程^[10-11],利用 OE 提供的 COM 组件,调用其接口函数直接按顺序找邮件,获得相应的邮件信息,从而间接实现了对 DBX 数据文件的解析,其解析流程如图 1 所示。

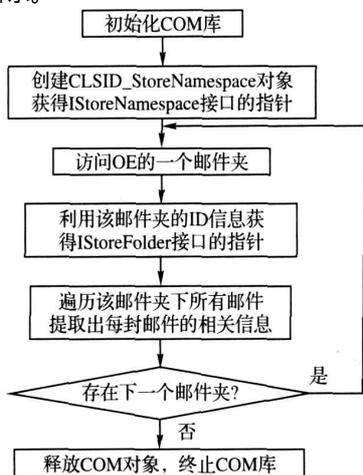


图 1 基于 COM 技术的 DBX 邮件文件解析流程

本文采用 Visual C++ 编程实现上述方法的详细步骤如下:

1) 实现任意路径下的 DBX 文件解析。利用 OE 提供的 COM 接口对 DBX 数据文件进行解析处理时,是通过读取注册表来获得 DBX 数据文件夹的路径信息。OE 默认的 DBX 数据文件夹路径信息存放在注册表 HKEY_CURRENT_USER \ Identities \ {User ID} \ Software \ Microsoft \ Outlook Express \ 5.0 目录下的 Store Root 子键中,其默认值为: % UserProfile% \ Local Settings \ Application Data \ Identities \ {User ID} \ Microsoft \ Outlook Express

因此,在创建 COM 组件实例之前,通过程序首先修改注册表对应的 Store Root 键值,把默认路径替换为所要解析的 DBX 数据文件夹所在的路径,便可实现任意路径下 DBX 数据

文件夹的解析。

在修改注册表前,需要将注册表的默认值保存起来,以便在解析结束时能还原注册表。

2) 创建 COM 组件实例。用 CoInitialize 函数初始化 COM 库。初始化成功后,调用 CoCreateInstance 函数创建标识符为 CLSID_StoreNamespace 的 COM 组件实例,请求标识符为 IID_IStoreNamespace 的接口,并将其指针赋给 m_pStoreNamespace,该接口类型为 IStoreNamespace,其指向 OE 的 Dbx 数据文件夹的存储空间。主要实现程序如下:

```

HRESULT hr = CoInitialize(0);
hr = CoCreateInstance(CLSID_StoreNamespace,
  NULL, CLSCTX_SERVER, IID_IStoreNamespace,
  (LPVOID*)&m_pStoreNamespace);
hr = m_pStoreNamespace->Initialize(NULL, NULL);
  
```

3) 遍历 OE 中的邮件夹。通过调用 IStoreNamespace 接口的 GetFirstSubFolder 函数获得第一个邮件夹的信息,将其赋给一个 FOLDERPROPS 的结构体,可获得该邮件夹的名称和 dwFolderId 值(邮件夹的 ID 值),然后调用 GetNextSubFolder 函数获得下一个邮件夹的信息,反复直至遍历所有的邮件夹。程序流程如图 2 所示。

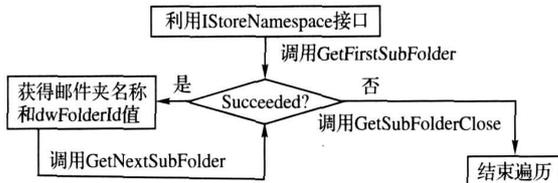


图 2 遍历 OE 邮件夹程序流程

4) 遍历 OE 某邮件夹下的所有邮件。遍历邮件夹下的所有邮件之前需要调用 IStoreNamespace 接口的 OpenFolder 函数,通过传入参数 dwFolderId 值,返回一个 IStoreFolder 接口的指针,其指向一个对应的邮件夹空间。

通过调用 IStoreFolder 接口的 GetFirstMessage 函数获得邮件夹下的第一封邮件的信息,将其赋给一个 MESSAGEPROPS 的结构体,可获得该邮件的基本信息(邮件收发人名字、主题、发送时间等)和 dwMessageId 值(邮件的 ID 值),然后调用 GetNextMessage 函数获得下一封邮件的信息,反复直至遍历该邮件夹下的所有邮件。程序流程如图 3 所示。

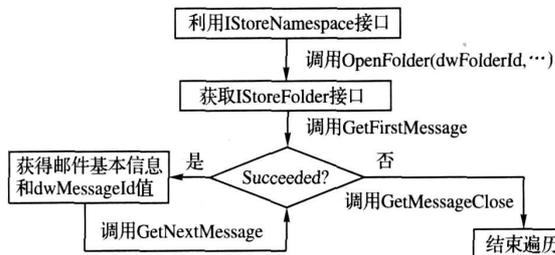


图 3 遍历 OE 某邮件夹下所有邮件程序流程

5) 提取 OE 每封邮件的信息。在提取一封邮件的信息之前需要调用 IStoreFolder 接口的 OpenMessage 函数,通过传入参数 dwMessageId 值,返回一个 IMimeMessage 接口的指针,其指向一个对应的邮件空间。

通过调用 IMimeMessage 接口的 GetBody 函数和 GetAttachments 函数,可分别获得对应邮件的邮件体标识和附件体标识,以这些标识作为传入参数,调用该接口下的相应函数,便可提取该邮件的详细信息(邮件收发人地址、内容和附件等)。程序流程如图 4 所示。

获取邮件信息后, 调用 Release 函数释放 COM 组件各个接口的使用。

6) 解析结束后, 释放 COM 对象, 终止 COM 库, 并将注册表中“Store Root”键值还原成原来的默认值。

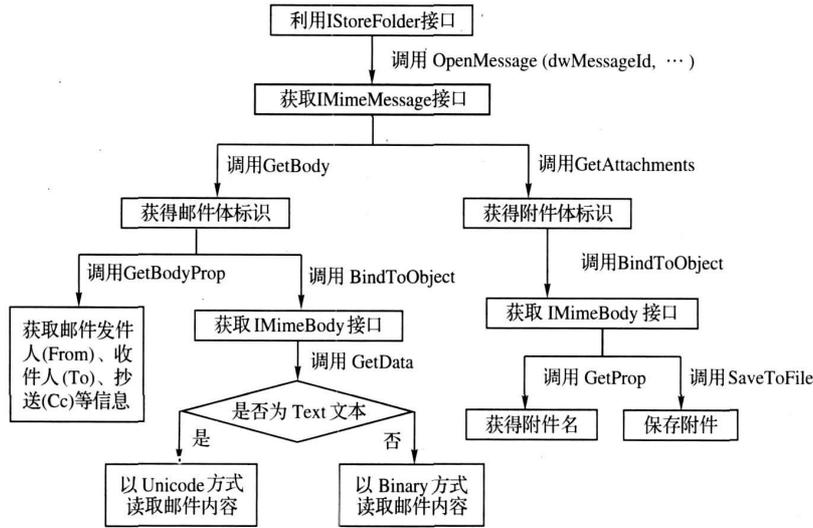


图 4 获取 OE 邮件信息程序流程

3 实例验证

本文研究中, 采用 Windows XP+ SP2 操作系统, 在 VC++ 6.0 开发环境下编程实现了上述方法。在默认安装 OE 的前提下, 通过在 Windows 2000/XP/2003 和 Vista 四个不同版本的操作系统下, 更换多组数据反复进行验证, 均能得到正确的解析结果, 运行界面如图 5 所示, 结果表明, 基于 COM 技术的 DBX 邮件文件解析方法是可行的, 具有较高的稳定性。

参考文献:

- [1] 杨泽明, 刘宝旭, 许榕生. 电子邮件取证技术 [J]. 信息安全, 2002(6): 33- 34.
- [2] 刘浩阳. 电子邮件的调查与取证 [J]. 辽宁警专学报, 2007(5): 27- 31
- [3] 潘爱民. COM 原理与应用 [M]. 北京: 清华大学出版社, 2001
- [4] 梁忠杰, 思敏, 李婷. COM 技术和动态链接库技术的应用研究 [J]. 微计算机应用, 2006 27(6): 702- 705.
- [5] Microsoft Corporation. IStoreNamespace interface [EB/OL]. [2007- 10- 23]. <http://msdn2.microsoft.com/en-us/library/ms710214.aspx>
- [6] Microsoft Corporation. IStoreFolder interface [EB/OL]. [2007- 10- 24]. <http://msdn2.microsoft.com/en-us/library/ms710250.aspx>
- [7] Microsoft Corporation. IMessage interface [EB/OL]. [2007- 10- 25]. <http://msdn2.microsoft.com/en-us/library/ms711861.aspx>
- [8] Microsoft Corporation. IMBody interface [EB/OL]. [2007- 10- 27]. <http://msdn2.microsoft.com/en-us/library/ms712525.aspx>
- [9] 李美满, 夏汉铸. 基于 COM 技术的通用题库系统的研究与实现 [J]. 计算机工程与设计, 2006 27(15): 2770- 2773
- [10] Microsoft Corporation. Windows Mail programming examples [EB/OL]. [2007- 10- 30]. <http://msdn2.microsoft.com/en-us/library/ms715241.aspx>
- [11] YABO P. Reading and writing messages in outlook express [EB/OL]. [2007- 11- 13]. http://www.codeproject.com/Outlook_Express_Messages.asp



图 5 DBX 邮件数据文件的解析界面

4 结语

本文针对直接解析 DBX 邮件数据文件难度较大的问题, 提出了基于 COM 技术的解析方法, 即利用 OE 提供的 COM 组件接口, 调用相关 API 函数直接获取文件中的邮件信息, 避免了研究分析复杂的 DBX 邮件文件格式, 为该数据文件的解析找到一种新的解决方法, 对电子邮件调查分析软件的研究开发具有较好的参考价值。

(上接第 2162 页)

- [4] HU JUN, ZHANG RONG, HE JIN-LIANG. Novel method of corrosion diagnosis for grounding grid [C] // International Conference on Power System Technology. New York: IEEE, 2000 3 1365- 1370.
- [5] 张晓玲, 黄青阳. 电力系统接地网故障诊断 [J]. 电力系统及其自动化学报, 2002, 14(1): 48- 51
- [6] 刘健, 王建新, 王森. 一种改进的接地网故障诊断算法及测试方案评价 [J]. 中国电机工程学报, 2005 25(3): 71- 77
- [7] 刘健, 王树奇, 李志忠, 等. 接地网故障诊断的可测性研究 [J]. 高电压技术, 2008 34(1): 64- 69.

- [8] 刘健, 王树奇, 李志忠, 等. 基于网络拓扑分层约简的接地网腐蚀故障诊断 [J]. 中国电机工程学报, 2008 28(3): 60- 65
- [9] 黄文武, 文习山, 朱正国. 接地网腐蚀与断点诊断软件系统的开发 [J]. 高电压技术, 2005 31(7): 42- 44.
- [10] 王湘中, 黎晓兰. 基于关联矩阵的电网拓扑辨识 [J]. 电网技术, 2001, 25(2): 10- 13
- [11] 陈竟成, 张学松, 汪峰, 等. 接地网络建模与网络结线分析 [J]. 电网技术, 1999, 23(5): 52- 54
- [12] 宋久旭, 刘健, 刘玟权. 一种接地网络拓扑快速增量建模方法 [J]. 继电器, 2005 33(5): 21- 26