

Office XML 文档信息隐藏方法

郝 宇,施 勇,薛 质

(上海交通大学 电子信息与电气工程学院,上海 200240)

摘 要:随着互联网和通信技术的飞速发展,计算机网络信息化等方面的发展对信息安全技术提出了越来越高的要求。电子文档已成为储存及传送信息的最常用载体,计算机泄密问题随之产生,且更具隐蔽性、潜伏性和危害性,同时也增加了泄密问题预防难度。Office 2007 文档采用了一种新的默认文件格式,即 Office Open XML 格式,为在 Office 文档中隐藏信息提供了新的思路。为此,结合 Office 文档的自身特点和 XML 格式规范,提出了一种符合 Office XML 格式规范的信息隐藏方法。该方法针对 DOCX、PPTX 和 XLSX 三种不同类型格式文档,匹配与格式相对应的特定属性,选取或构建包含该属性的 XML 段落,通过替换或构造特有的标识属性值将隐秘信息嵌入目标文本载体中,以实现信息隐藏的目的。实验结果表明,该方法使用文本容量大,安全性高,较好地解决了现有方法鲁棒性不足的问题。

关键词:Office XML;信息隐藏;Excel XML;Word XML;PowerPoint XML

中图分类号:TP301

文献标识码:A

文章编号:1673-629X(2017)10-0096-05

doi:10.3969/j.issn.1673-629X.2017.10.021

Office XML Document Information Hiding Method

HAO Yu,SHI Yong,XUE Zhi

(School of Electronic Information and Electrical Engineering,Shanghai Jiaotong University,
Shanghai 200240,China)

Abstract:With the rapid development of Internet and communication technology,the computer network pays more and more attention in information,which also emphasizes the increasing importance of information security technology. Electronic documents have become the most commonly used carrier in information storage and transmission and thus the problem of computer file leakage has generated more hidden,latent and harmful properties than other classical methods,increasing the difficulty in prevention of leakage. Since a new default file format is introduced in Microsoft Office 2007 document,which is called Office Open XML format and provides a new ideas for information hiding in Excel documents. An information hiding method based on Office Open XML file format is proposed in investigation of features of Office documents and XML specifications. With different kinds of files,such as DOCX,PPTX and XLSX,it adopts different identifier attribute,which finds or creates a segment with this attribute to hide information by changing the attribute value. Experiment results show that it can hide large capacity information and thus is safety with high capacity and security,which has solved poor robustness of the existing methods.

Key words:Office Open XML;information hiding;Excel XML;Word XML;PowerPoint XML

0 引 言

随着互联网电子商务的迅速发展以及电子文档的广泛使用,对于涉及私密信息的电子文档的保护显得尤为重要。信息隐藏技术^[1-3]是实现电子文档保护的重要手段,目前已成为多媒体信息安全领域的一个重要部分。信息隐藏涉及的方面有很多,如图像、音频、视频等,而且取得了较多研究成果。对于图像、音频、

视频,主要利用的是其载体的较大冗余性。与这些载体相比,文本信息的冗余空间非常有限,所以兼顾文本信息隐藏的安全性和鲁棒性难度较高,导致文本信息隐藏技术的研究相对滞后。

Office 文档是目前在电子商务、电子政务中使用最为广泛的文档之一,因此利用 Office 文档进行信息隐藏并实现追踪文件的目的具有重大意义。针对不同

收稿日期:2016-06-02

修回日期:2016-10-10

网络出版时间:2017-07-19

基金项目:公安部信息安全重点课题支持(C14612)

作者简介:郝 宇(1991-),男,硕士,研究方向为网络安全、大数据分析;施 勇,博士,讲师,研究方向为网络安全、网络攻防;薛 质,博士,教授,研究方向为网络安全、网络攻防。

网络出版地址:cnki.net/kcms/detail/61.1450.TP.20170719.1107.002.html

用途,Office 有三种主要的文档格式,即 Word,Excel 和 PowerPoint。目前已有学者提出通过修改 Word 中文本的字符大小或文本颜色来隐藏信息^[4-10]。这些方法主要是针对 Word 文档提出的。自 Office 2007 版起,Microsoft Office 采用基于 Office Open XML 的文档格式^[11],因其可以通过多种方式访问、降低文件损坏的风险等特点,已被越来越多的用户使用。目前,已有一些学者针对此版本提出隐藏信息的方案^[12-19]。

根据 Office XML 的特点,在分析 Office Open XML 文件的构造方法的基础上,提出了一种适用于 Word XML,Excel XML 与 PowerPoint XML 三种文档格式的信息隐藏方法,给出了相应的实现算法并对其进行了实验验证。

1 Office Open XML 文件格式

自 2007 Microsoft Office 系统开始,Microsoft Office 使用基于 XML 的文件格式,例如.docx、.xlsx 和.pptx。这些格式和文件扩展名适用于 Microsoft Word、Microsoft Excel 和 Microsoft PowerPoint。Office Open XML (Open XML)是一种国际认可的文件格式标准,Office 软件套件实施这种标准来保存和交换信息。Open XML 遵循 ECMA-376 及 ISO/IEC 29500 标准,这意味着创建、编辑和保存 Open XML 文件等操作均需符合标准。Office Open XML 格式使用 Zip 压缩技术来存储文档,这种新的文件格式采用开放打包协议,整个文档由一个压缩的 Zip 包组成,同时减少了文档的大小。而文件结构则以模块形式进行组织,从而使文件中的不同数据组件彼此独立。

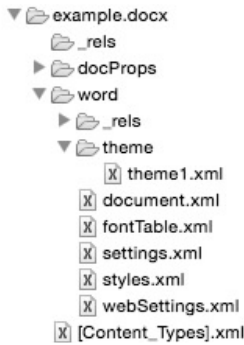


图 1 Word XML 文档的层次化文件结构

图 1 是一个只包含纯文本的 Word XML 文档。在该文档包中,_rels 文件夹存储所有指定部件的关系部件主文档 document.xml;docProps 文件夹包含应用程序的属性部件;而 word 文件夹存储着 example.docx 文件的核心数据。其中 document.xml 文件则是文件包的主文档,其记载了 Word XML 文件的文字内容及其他相关属性;[Content_Types].xml 文件描述了出现在文件中的每个内容类型。PowerPoint XML 和 Excel

XML 的文档包的层次化文件结构与图 1 所示相似,其主文档 slide1.xml 等文件和 sheet1.xml 等文件,分别保存在 ppt/slides 文件夹和 xl/worksheets 文件夹中。

2 基于“标识属性”的信息隐藏方案

2.1 标识属性

基于 XML 的 Office Open XML 中,最基本的单位是元素,元素可以带有若干个属性及属性值作为附加信息。无论是.docx、.pptx 还是.xlsx 格式的文件,元素的命名及其属性值的定义都应遵循 Open XML 规范,同时每个元素也有其特定的意义及作用。

图 2 是某 Word XML 文档包中的 document.xml 文件,其根元素为元素 w:document。

```
<? xml version="1.0" encoding="UTF-8" standalone="yes"? >
<w:document xmlns:wpc="http://schemas.microsoft.com/office/word/2010/wordprocessingCanvas" xmlns:mo="http://schemas.microsoft.com/office/mac/office/2008/main" xmlns:mc="http://schemas.openxmlformats.org/markup-compatibility/2006">
  <w:body>
    <w:p w14:paraId="7F2394E7" w14:textId="77777777" w:rsidR="00F04602" w:rsidRDefault="004A08E6">
      <w:pPr>
        <w:rPr>
          <w:rFonts w:hint="eastAsia"/>
        </w:rPr>
      </w:pPr>
      <w:proofErr w:type="spellStart"/>
      <w:r>
        <w:t>A</w:t>
      </w:r>
      <w:bookmarkStart w:id="0" w:name="_GoBack"/>
      <w:bookmarkEnd w:id="0"/>
      <w:proofErr w:type="spellEnd"/>
    </w:p>
    <w:sectPr w:rsidR="00F04602" w:rsidSect="006452E4">
      <w:pgSz w:w="11900" w:h="16840"/>
      <w:pgMar w:top="1440" w:right="1800" w:bottom="1440" w:left="1800" w:header="851" w:footer="992" w:gutter="0"/>
      <w:cols w:space="425"/>
      <w:docGrid w:type="lines" w:linePitch="312"/>
    </w:sectPr>
  </w:body>
</w:document>
```

图 2 Word XML 主文档 document.xml 示例

以 w:sectPr 元素为例,它定义了文档最后一部分的属性,并拥有两个属性:w:rsidR (Section Addition Revision ID) 和 w:rsidSect (Section Properties Reversion ID)。在微软官方公开文档中表明:所有拥有相同值的 rsid * 属性的区域均指向同一编辑会话期间。如图 2 中 w:p 与 w:sectPr 两元素,其属性 w:rsidR 的属性值均为“00F04602”,这意味着这两个元素的修改是在同

一编辑会话内完成的。类似于 w:rsidR 这样的属性,称之为“标识属性”(Identifier Attribute)。“标识属性”一般用于区分文本、表格等数据或属性,其特点是拥有独一无二且由系统随机生成的属性值,并且该属性值与用户及修改时间等无任何关系。经研究发现,Office XML 各文档格式中均含有“标识属性”,研究结果举例由表 1 所示。

表 1 Office XML 各文档中“标识属性”举例

文档类型	标识属性举例	标识属性值位数	标识属性值进制	标识属性值举例
Word XML	w:rsidR	8	16	00F04602
Excel XML	x14:id	32	16	E45B70CA-B6A0-4E88-BB95-1D31570318B9
PowerPoint XML	a:tableStyleId	32	16	5C22544A-7EE6-4342-B048-85BDC9FD1C3A

2.2 信息隐藏方案

研究表明,对于标识属性值的修改不会对文本内容造成影响。因此通过将待隐藏信息写入其属性值中,可以实现在 Office XML 文档中隐藏信息的目的。为了保证安全性,可以先将待隐藏信息转化为十六进制的 Unicode 码,然后选择加密算法将其进行加密,并添加校验位以便于之后的隐藏信息提取工作。

对于 Word XML 中的 w:rsidR 属性而言,其出现的次数及频率较多,因此上述基于替换原有标识属性属性值的隐藏方案在 Word XML 中较易实现。但对于 Excel XML 和 PowerPoint XML 格式文档,在一个简单的文本文档中,如 x14:id 和 a:tableStyleId 这样的标识属性可能不存在,这时则需要通过在特定位置构造 Office Open XML 元素嵌入隐藏信息。

在 PowerPoint XML 文件中,a:tableStyleId 用于标识表格样式。当某页幻灯片中存在一个表格样式的数据,则相应的 slide.xml 部件中则会记录其 a:tableStyleId 属性。同时在该文档的 tableStyles.xml 部件中也会记录下该表格样式信息。tableStyles.xml 部件用于记录 PowerPoint XML 文档在整个编辑过程中曾经使用过的表格样式,无论这些表格样式目前是否仍在使用。所以若利用替换原有属性值的方法,则需要同时更新 slide.xml 与 tableStyles.xml 部件中的属性值。若各个 slide.xml 部件中不存在 a:tableStyleId 属性,可以在 tableStyles.xml 部件中进行添加以达到信息隐藏的目的。例如:待隐藏的已加密信息为“AAAAAAAA-1111-1111-1111-AAAAAAAAAAAA”,则可通过图 3 的方式添加元素。图 3(a)中,a:tblStyleLst 元素用于记录表格样式列表。因此,该方法通过对 a:tblStyleLst 元素添加子元素 a:tblStyle 即新的表格样式来嵌入隐藏信息。值得注意的是,a:tblStyle 本身拥有两个属性:styleId 和 styleName,分别用于记录样式标识和样式名称。而 styleName 不可省略,否则 Office 软件系统会判

定该构造元素(a:tblStyle)为无用元素并删除,导致隐藏信息失败。

<? xml version="1.0" encoding="UTF-8" standalone="yes"? >
 <a:tblStyleLstxmlns:a="http://schemas.openxmlformats.org/drawingml/2006/main" def="{5C22544A-7EE6-4342-B048-85BDC9FD1C3A}"/>
 (a) tableStyles.xml 部件原始表格样式列表
 <? xml version="1.0" encoding="UTF-8" standalone="yes"? >
 <a:tblStyleLstxmlns:a="http://schemas.openxmlformats.org/drawingml/2006/main" def="{5C22544A-7EE6-4342-B048-85BDC9FD1C3A}">
 <a:tblStylestyleId="{AAAAAAAA-1111-1111-1111-AAAAAAAAAAAA}" styleName="aaa"/>
 </a:tblStyleLst>
 (b) tableStyles.xml 部件隐藏信息后表格样式列表

图 3 PowerPoint XML 文档 tableStyles.xml 部件元素节点添加前后对比

在 Excel XML 文件中,x14:id 可用于标识条件格式规则(Conditional Formatting Rule)。如图 4 所示,条件格式元素 x14:cfRule 由属性值为“C5A286DA-8583-446B-B1AC-FC4211EE1663”的标识属性 id 标识。图 4 中 extLst 元素用于记录扩展列表(extension list),且位于 worksheet 元素节点中,其子元素为 ext 用于标记扩展(extension)。因此,如果某 Excel XML 文档中不存在标识属性 x14:id,同样可通过在 slide.xml 部件中对其进行添加,从而实现信息隐藏的目的。

例如:待隐藏的已加密信息为“AAAAAAAA-1111-1111-1111-AAAAAAAAAAAA”,则可通过图 5 的方式添加元素。与 PowerPoint XML 中的方法相比,在 Excel XML 中使用的方法略有不同,此方法将标识属性 x14:id 构造为元素。这是因为在研究的测试中发现,若构造如图 4 中所示的含有 x14:cfRule 元素的 ex-

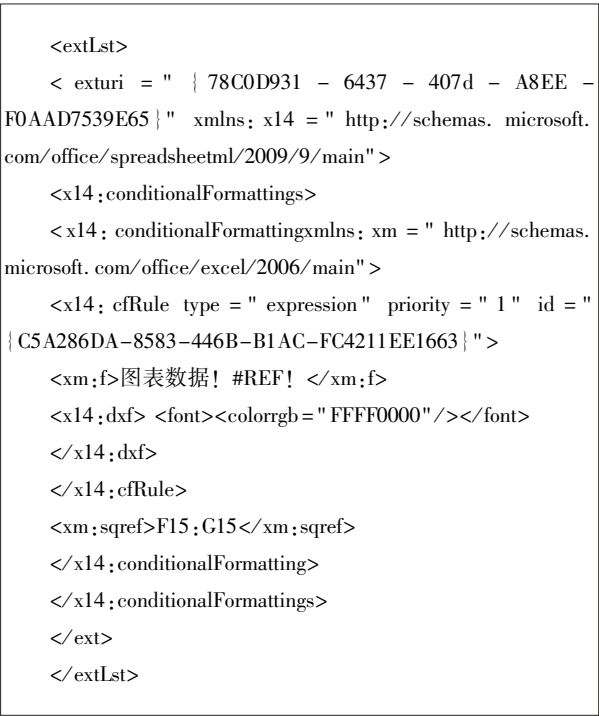


图 4 Excel XML 文档 slide1.xml 部件中
的条件格式元素 x14:cfRule
tLst 元素,则加密信息的鲁棒性不能得到保障,加密文档可以正常打开,但是 Office 软件系统无法找到实际

的条件规则与构造的 x14:cfRule 元素相关联,因此使用文件过程中的任何修改都会导致加密信息的丢失。经研究发现,每个 ext 元素将 uri 属性作为标识符来指示扩展的信息,同时其对于子元素则没有具体要求。因此,可以通过构造符合已申明的 XML 命名空间 (xmlns:x14) 的子元素 x14:id 来隐藏加密信息,如图 5 所示。

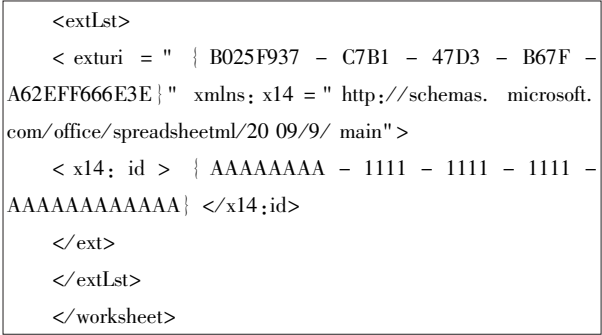


图 5 Excel XML 文档 slide1.xml 部件添加 extLst 元素
针对 Office Open XML 的三种不同格式文档,采用如图 6 所示的信息隐藏流程图。通过在相应 xml 部件中替换或构造含有标记属性特征的元素节点来达到信息隐藏的目的。

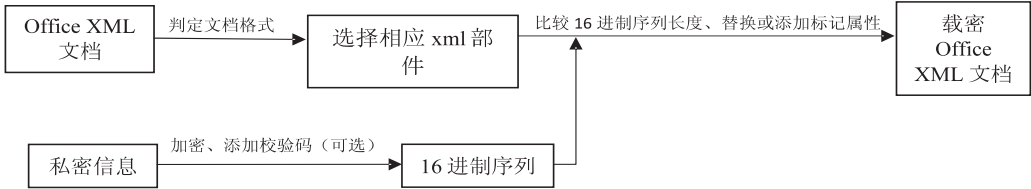


图 6 Office XML 文档信息隐藏流程

3 基于 Office XML 的信息隐藏算法

3.1 信息嵌入算法

输入: Office XML 文档 T , 待隐藏信息 M , 密钥 K 。

输出: 载密文档 T' 。

- (1) 判定文档 T 的格式类型 x ($x \in \{.docx, .pptx, .xlsx\}$);
- (2) 将待隐藏信息 M 转为 16 进制序列 (组) S , 长度由文档格式 x 决定;
- (3) 对 16 进制序列 (组) S 通过密钥 K 进行加密并添加校验码得到新的 16 进制序列 (组) S' ;
- (4) 通过其文档类型 x , 遍历文档中相对应的部件, 寻找可隐藏 S' 的标记属性值。结果与 S' 进行比较, 替换或添加对应 x 格式的标记属性元素;
- (5) 保存新文档 T' , 信息嵌入完毕。

3.2 信息提取算法

输入: 载密 Office XML 文档 T' , 密钥 K 。

输出: 载密信息 M 。

- (1) 判定载密文档 T' 的格式类型 x ($x \in \{.docx, .pptx, .xlsx\}$);
- (2) 通过其文档类型 x , 遍历文档中相对应的部件, 遍历标记属性值, 得到 16 进制序列组 S' ;
- (3) 针对 S' 中每项序列进行校验, 判断并通过密钥 K 解密, 得到序列组 S ;
- (4) 拼接序列组 S , 并转换为字符, 得到载密信息 M , 信息提取完毕。

3.3 算法分析

实验环境是 Windows 8, OS X Yosemite, Microsoft Office 2016 以及 Eclipse 4.5。所使用的实验文件是从 Baidu 上搜索并下载的一些 Office XML 文档。下面从鲁棒性、信息隐藏容量和隐蔽性方面进行分析。

- (1) 鲁棒性: 在各式文档中, 隐秘信息嵌入到标记属性值中。经试验测试, 在 Office 系统软件中, 对加密文档进行常用的各种格式设置和内容的添加删减均不会造成隐秘信息的遗失, 因此该算法鲁棒性较强。但

是若将含有隐密信息的 Word 文本段落整段删除,或将含有隐秘信息的 Excel 整页表格完全删除,隐秘信息均会丢失。

(2)信息隐藏容量:针对 PowerPoint XML 及 Excel XML 格式文档,一个标记属性值允许隐藏长度为 16 字节即 128 比特的加密信息。而对于 Word XML 文档,一个标记属性值允许隐藏长度为 4 字节即 32 比特的加密信息。而对于一个大小 3 MB 左右的.docx 格式文档,其存在着大约 300 个 w:rsidR 属性,即约 1 200 字节。由此可见,该算法拥有较大的信息隐藏容量。

(3)隐蔽性:通过实验证明,在 Office XML 文档中应用该算法,不会引起文档显示的任何改变,也不会影响文档的正常使用。若只采用算法中的替换属性值方案,则对文本大小不会产生任何改变,因此算法的隐蔽性较好。

4 结束语

为提高计算机电子文档的安全性,解决计算机失泄密预防难题,提出了一种基于 Office Open XML 三种不同文档格式的信息隐藏方法。通过对文档格式进行深入的研究分析,在归纳总结替换或添加“标识属性”的信息隐藏思路的基础上,设计并实现了信息嵌入及提取算法。实验结果表明,与以往的在 Office 文档中通过修改文本显示格式以隐藏信息的方法相比,该方法较好地解决了传统方法鲁棒性弱及信息隐藏容量较小等问题。随着 Office Open XML 文档的逐渐普及,今后的工作将主要集中于载密 Office XML 文档格式转换中载密信息保留的研究。

参考文献:

- [1] 徐献灵,崔楠. 信息隐藏技术及其应用[M]. 北京:科学出版社,2007.
- [2] 吴树峰,黄刘生,卢继军,等. 信息隐藏技术及其攻击方法[J]. 计算机科学,2003,30(2):92-96.
- [3] Katzenbeisser S, Petitcolas A P F. 信息隐藏技术:隐写术与数字水印[M]. 北京:人民邮电出版社,2001.
- [4] 刘玉玲,孙星明. 通过改变文字大小在 Word 文档中加载数字水印的设计与实现[J]. 计算机工程与应用,2005,41(12):110-112.

- [5] 莫佳. 基于 Word 文本的信息隐藏系统的设计与实现[J]. 计算机应用与软件,2009,26(12):278-281.
- [6] 付兵,肖小玲. 一种基于 Word 文档的高隐藏率水印算法[J]. 长江大学学报(自科版):理工卷,2007,4(2):55-57.
- [7] Chandramouli R, Kharrazi M, Memon N. Image steganography and steganalysis: concepts and practice [C]//Proceedings of IDWD. [s. l.]:[s. n.], 2015:35-49.
- [8] Khan A, Siddiqua A, Munib S, et al. A recent survey of reversible watermarking techniques[J]. Information Sciences, 2014, 279:251-272.
- [9] Subhedar M S, Mankar V H. Current status and key issues in image steganography: a survey[J]. Computer Science Review, 2014, 13:95-113.
- [10] Murdoch S J, Lewis S. Embedding covert channels into TCP/IP [C]//International workshop on information hiding. Berlin: Springer, 2005:247-261.
- [11] Microsoft. Office (2007) Open XML 文件格式简介[EB/OL]. [2007-07-06]. <http://www.microsoft.com/china/msdn/library/office/office/OfficeOpenXMLFormats.mspx?mfr=true>.
- [12] Park B, Park J, Lee S. Data concealment and detection in Microsoft Office 2007 files [J]. Digital Investigation, 2009, 5(3):104-114.
- [13] Garfinkel S L, Migletz J J. The new XML office document files: implications for forensics [J]. IEEE Security & Privacy, 2009, 7(2):38-44.
- [14] 刘玉玲,万晶,辛国江. Excel2007 文档信息隐藏方法[J]. 计算机工程与应用,2010,46(28):70-72.
- [15] 徐敏,王衍波,李涛. Word2007 文档信息隐藏的新方法[J]. 计算机研究与发展,2009,46:112-116.
- [16] 吴悠,孙星明. 基于正弦波的 Word 文档数字水印[J]. 计算机工程,2005,31(24):175-176.
- [17] What's up with all those rsids? [EB/OL]. [2006-12-11]. https://blogs.msdn.microsoft.com/brian_jones/2006/12/11/whats-up-with-all-those-rsids/.
- [18] 耿建勇. XML 安全技术的应用研究[D]. 北京:中国科学院研究生院(计算技术研究所),2005.
- [19] Liu T Y, Tsai W H. A new steganographic method for data hiding Microsoft word documents by a change tracking technique [J]. IEEE Transactions on Information Forensics and Security, 2007, 2(1):24-30.