

一种基于 Word XML 的信息隐藏新方法

董 艳,徐江峰

(郑州大学 信息工程学院,河南 郑州 450000)

摘 要:在应用广泛的 Word 文档中隐藏秘密信息意义巨大,在对 Word 2007 中 document.xml 文件的“修改标识”特性分析的基础上,提出了一种新的信息隐藏方法。该方法首先根据 logistic 方程伪随机序列生成器产生的伪随机二进制序列,去抽取载体文件 document.xml 中特定的“修改标识”属性值,而后用秘密信息的十六进制码替换抽取的属性值后六位,从而达到隐藏秘密信息的目的。实验结果表明,基于 document.xml 文件中“修改标识”属性值的信息隐藏新方法,与既有的基于 Word 的信息隐藏方法相比,具有安全性更高、隐蔽性更强等优点。

关键词:信息隐藏;Word XML 格式;修改标识;伪随机序列

中图分类号:TP301

文献标识码:A

文章编号:1673-629X(2015)02-0122-04

doi:10.3969/j.issn.1673-629X.2015.02.028

A New Information Hiding Way Based on Word XML

DONG Yan, XU Jiang-feng

(School of Information Engineering, Zhengzhou University, Zhengzhou 450000, China)

Abstract: It is of great significance to hide some secret information in a Word document. Based on analyzing the "revision identifiers" in document.xml of Word 2007 document, propose a new information hiding method. According to a pseudo-random binary sequence which is produced by logistic pseudo-random sequence generator, extract certain attributes of "revision identifiers" in document.xml, and change the last six attributes values with the hexadecimal secret information, achieving the purpose of hiding secret information. The experimental result show that compared with the previous hiding ways based on Word, the new hiding way based on the "revision identifiers" is better in robustness and safety and so on.

Key words: information hiding; Word XML; revision identifiers; pseudo-random sequence

0 引 言

随着网络技术的迅速发展,通过网络传输和获取信息变得普及,随之也产生了网络信息安全问题。如何保护信息在传输过程中的安全问题已经成为信息技术研究领域的重要内容,随之出现的信息隐藏技术成为解决信息传输安全问题的解决方案之一。信息隐藏^[1](information hiding)是指在图像、视频、音频、文本、网页等载体中嵌入一些秘密信息,让第三方难以察觉秘密信息的存在。

基于文本格式文档的信息隐藏技术是近年来发展细化出来的一个新分支。其中的 Word 文档支持文字、图形图像等的多格式文件,是目前使用最广泛的文本处理软件,因此基于 Word 文档文本格式的信息隐藏得到了广泛的研究,已经取得了一系列成果。刘显德等提出了一种根据字符间距编码的方法实现在

Word 文档中隐藏秘密信息的方法^[2];刘玉玲、孙星明根据特征编码的方法,针对 Word 文档格式的特点,通过改变文档中某些字符的大小以嵌入和检测水印^[3];王海春等提出了修改 Word 文档汉字的西文字体来隐藏信息^[4];莫佳提出通过微调 Word 文本的字符大小而隐藏信息^[5];付兵,肖小玲提出修改 Word 字体的 RGB 颜色低位以及下划线的值来隐藏秘密信息^[6];王智,周洪玉提出改变 Word 文档文字字体以实现图片的隐藏^[7]。

1 Word 2007 文档分析

1.1 Microsoft Office Word 2007 文档包

Word 2007 提供了一种新的默认文件格式,叫做 Microsoft Office Word XML (Word XML 格式)。它的默认保存格式为“.docx”,改变格式后文档占用空间将有

收稿日期:2014-03-24

修回日期:2014-06-25

网络出版时间:2014-12-27

基金项目:国家自然科学基金资助项目(61071211)

作者简介:董 艳(1984-),女,硕士研究生,研究方向为信息隐藏;徐江峰,博士,教授,研究方向为数字水印、信息隐藏、混沌等。

网络出版地址:<http://www.cnki.net/kcms/detail/61.1450.TP.20141227.1348.046.html>

一定程度的缩小。Word 2007 文档包除了主文档 document.xml 外,还包括 styles.xml,它定义了文档的样式;themel.xml 定义文档的模板;document.xml.rels 用于重新将这些组件组合成一个完整文档用的指示文件等。事实上,Word 2007 的基本文件是 ZIP 格式的^[8]。

这种格式由一个压缩的 ZIP 包组成,包中包含了文档所有内容^[9],如图 1 所示。

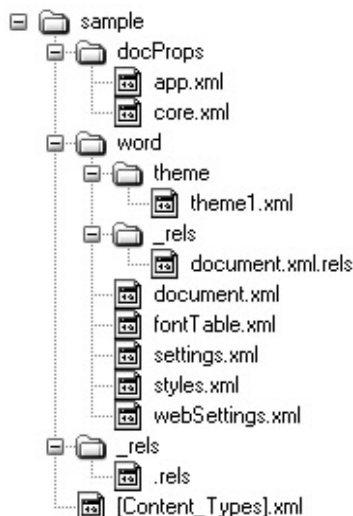


图 1 一个 Word 2007 文件包含的内容

Word 2007 文档包括:主文档 document.xml;app.xml file 包含了应用程序特定的属性;core.xml file 包含了所有基于开放打包约定文档格式的通用文件属性;docProps folder 包含了应用程序的属性部件;rels file 描述了文档结构中的起始关系,它称为关系部件;rels folder 存储所有指定部件的关系部件;[Content_Types].xml 描述出现在文件中的每个内容类型。

1.2 Word 文件夹中主文档 document.xml 格式分析

document.xml 是 Word 2007 的主 XML 文档,一些信息隐藏方案都是在这个 xml 文档中进行的。图 2 是一个 Word 2007 文档包中的 document.xml 文件截图,元素 w:document 中的一系列名字空间是主文档所引起的文档部件的关系;w:body 是正文元素,包含了文档的各个段(w:p),每个段包含一个或多个文本域(w:r),每个文本域又包含一个或多个文本块(w:t);在 w:rPr 元素中描述了文字的属性、字体、颜色等子元素;w:color 的属性 w:val 的属性值即为文字的 RGB 值;元素 w:p 的属性 w:rsidR、w:rsidRr、w:rsidRDefault、w:rsidP 以及 w:r 的属性 w:rsidRr 就是“修改标志”^[10]。

在 w:p (一个自然段落)中,w:rsidR (Revision Identifier for Paragraph)指定唯一一个标识符,用来跟踪编辑在修订时表行标识,作用:在文档修改中记录修改的信息,尤其是合并等情况。特性:只是属性,没有特殊意义,可以关闭,信任中心“存储随机数以改善合并准确性”。w:rsidP (Revision Identifier for Paragraph

Properties)是段落属性修改标识符。w:rsidRPr (Revision Identifier for Table Row Glyph Formatting)是段落字形修改标识符,用来跟踪编辑在修改时字符或字形发生的改变,所有段落都应该拥有相同的属性值,如果出现差异,那么表示这个段落在后面的编辑中被修改。w:rsidRDefault (Revision Identifier for Runs)是默认的版本标识符,“w:rsidR”属性变化之前的默认属性,也就是版本属性。

在 w:r (一个样式串,即文本的显示样式)中,(1)段落中以相连续的中文或英文字符串,作为开始和结束。目的就是要把一个段落中的中英文字符区分开来。(2)当中文字符有属性时,比如粗、斜、下划线时,也会用 w:r 进行分割和标识,并且会含有一个修改标识的属性 w:rsidRPr。

每一次 Word 文档被打开编辑时,都会产生 1 个独一无二的 ID 号。这个 ID 号被保存在“修改标识”的属性值里面,而且它是随机产生的一个字符串,和时间没有关系^[11]。document.xml 的内容如图 2 所示。

```

xmlns:w="http://schemas.openxmlformats.org/wordprocessingml"
xmlns:wne="http://schemas.microsoft.com/office/word/2001"
- <w:body>
- <w:p w:rsidR="00457A87" w:rsidRDefault="000A512F"
  w:rsidP="000A512F">
- <w:r>
  <w:t>采莲南塘秋,莲花过人头;低头弄莲子,莲子清如水。</w:t>
</w:r>
</w:p>
- <w:p w:rsidR="00287396" w:rsidRPr="000A512F"
  w:rsidRDefault="000A512F" w:rsidP="000A512F">
- <w:pPr>
~

```

图 2 document.xml 中的内容

在 document.xml 文件中存在大量的“修改标识”。如图 2 中,w:rsidR = “00457A87”、w:rsidRDefault = “000A512F”、w:rsidP = “000A512F”和最后两行的 4 个“修改标志”。

2 基于 Word 2007 的信息隐藏算法

2.1 算法思想

在 Word 2007 文档中隐藏信息需要满足下列条件:

(1)载体文档(隐藏信息后的文档)必须满足 Word XML 格式的要求。

(2)载体文档必须能正常显示。

经实验发现,修改 Word 2007 文件中 document.xml 中“修改标识”的属性值并不会影响文档的正常显示和使用。“修改标识”属性值共 8 位,前两位一般为“00”,可以选择“修改标识”属性值的后六位为修改位。经实验验证,对 Word 2007 文档的修改,w:p 中以下的三个修改标识的属性值不会发生变化,因此,隐藏信息的“修改标识”最好选取 w:p 元素中的 w:rsidR、

w:rsidP、w:rsidRr。文中算法选取修改标识 w:rsidP 属性值的后六位用于信息隐藏。

文中用到了混沌理论中的 logistic 方程,当方程的初始条件取一定值时,产生伪随机序列,再转化成伪随机二进制序列^[12] S_1 ,初始条件的值作为算法的密钥。秘密信息隐藏、提取的流程图如图 3、图 4 所示。

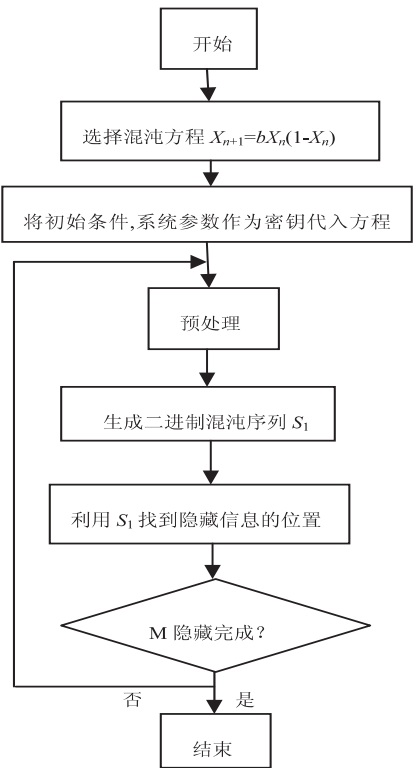


图 3 秘密信息 M 的隐藏过程

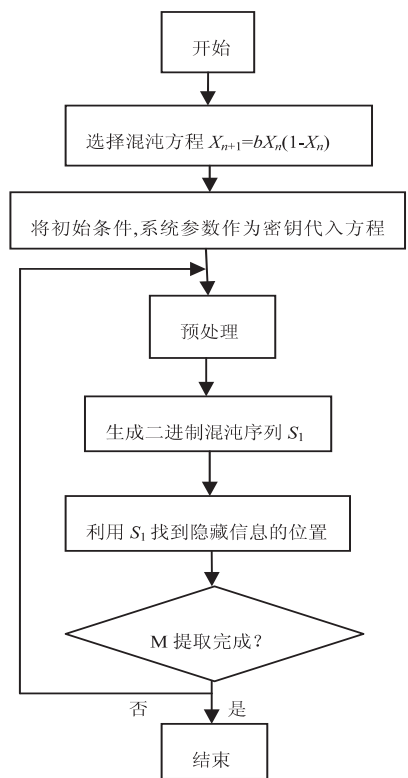


图 4 秘密信息 M 的提取过程

2.1.1 算法描述

信息隐藏过程如下:

- (1)读取 Word 2007 文档中的 document.xml (载体文档 T),输入秘密信息 M,以及一个长度不小于 T 中修改标识个数的伪随机二进制串 S_1 (由 0、1 组成)。
- (2)将秘密信息 M 转换为十六进制串 M_h。
- (3)按顺序查找 T 中待修改标识,并判断 S_1 中对应位置的值是否为 1,若为 1,转(4);否则,转(3)。
- (4)判断 M_h 中剩余元素是否还有 6 位,若不够则在剩余元素后填充 0,使其长度达到 6 位,并用此串替换对应的修改标识,转(5);否则,取出 M_h 的连续 6 位并替换对应的修改标识,判断 M_h 中是否还有元素,若有则转(3),否则转(5)。
- (5)T 中继续查找待修改标识,并把此标识替换为一个特定的标识。
- (6)生成载密文档,并替换 T,结束。

信息提取过程如下:

- (1)读取 Word 2007 文档中的 document.xml (载体文档 T),输入伪随机二进制串 S_1 ,秘密信息 M 赋值为空串。
- (2)按顺序查找 T 中待修改标识,若此标识为特定标识,转(4);否则,判断 S_1 中对应位置的值是否为 1;若为 1,转(3);否则,转(2)。
- (3)取出修改标识中对应的后 6 位,并把其连接到已得到的秘密信息 M 后,转(2)。
- (4)把 M 转换为二进制字符串,得到秘密信息。

2.1.2 隐藏方法举例

假如随机选取的修改标识是图 2 中第一列的 w:rsidP = “000A512F” 和最后一列的 w:rsidP = “000A512F”,假设待隐藏信息为:郑州大学,其十六进制为:90D15DDE59275B66。待隐藏的信息有 16 位,那么就选取修改标识属性值的后六位 0A512F 为替换位。信息隐藏过程如图 5 所示。

原属性值	0	0	0	A	5	1	2	F
修改后属性值	0	0	9	0	D	1	5	D
原属性值	0	0	0	A	5	1	2	F
修改后属性值	0	0	D	E	5	9	2	7
原属性值	0	0	0	A	5	1	2	F
修改后属性值	0	0	5	B	6	6	0	0

图 5 信息隐藏过程

(隐藏前后“修改标识”的属性值的变化)

对应于图 2 信息隐藏前的 document.xml 文件,图 6 为信息隐藏后的 document.xml 文件。

用于隐藏信息的“修改标识”属性值被修改后,新的修改标识分为 w:rsidP = “0090D15D”、w:rsidP =

“00DE5927”、w:rsidP=“005B6600”,如图6所示。

```

xmins:w=http://schemas.openxmlformats.org/wordprocess
xmlns:wne=http://schemas.microsoft.com/office/word/200
<w:body>
- <w:p w:rsidR="00457A87" w:rsidRDefault="000A512F"
  w:rsidP="0090D15D">
- <w:r>
  <w:t>采莲南塘秋,莲花过人头;低头弄莲子,莲子清如水。</w
</w:r>
</w:p>
- <w:p w:rsidR="00287396" w:rsidRPr="000A512F"
  w:rsidRDefault="000A512F" w:rsidP="00DE5927">
- <w:pPr>
  ....

```

图6 隐藏信息后 document.xml 中的内容

现选定待隐藏秘密信息为“郑州大学”,按照上述流程,将其转化为十六进制串并隐藏在 document.xml 中。实验选取的载体文本包含 436 字节,解压包大小为 36 671 字节。使用的软件工具:WinRAR、Word 2007、Matlab 等。

嵌入秘密信息后 Word 文档的显示效果与嵌入秘密信息前相比,没有发生任何改变。嵌入秘密信息后的 Word 文档显示效果如图7所示,并且实验证明,提取的秘密信息也是正确的。



曲曲折折的荷塘上面,弥望的是田田的叶子。叶子出水很高,像亭亭的舞女的裙。层层的叶子中间,零星地点缀着些白花,有袅娜地开着,有羞涩地打着朵儿的;正如一粒粒的明珠,又如碧天里的星星,又如那远处高楼上渺茫的歌声似的。这时候叶子与花也有一丝的颤动,像闪电般,霎时传过荷塘的那边去了。叶子底下是脉脉的流水,遮住了,不能见一些颜色;但叶子却更见风致了,且因为遮住了,所以更见其风致了。

图7 嵌入秘密信息后的 Word 2007 文档

2.2 算法分析

Word 2007 文档采用了一种新的 Word XML 格式,只用抽取 document.xml 文档,就可以实现该算法。

(1)透明性:隐藏后的载体文件与原始载体文件相比,在正常情况下,视觉上完全一样,透明性非常好。

(2)鲁棒性:文中选择用于隐藏信息的“修改标识”属性值的位置,是根据二进制伪随机序列中1的位置确定的,而用于产生此二进制串的 logistic 方程的初始值作为密钥,只有加密者和特定的接收秘密信息方才知道。初始值不同,二进制串就不同,因而秘密信息隐藏的位置就不同。与文献[8]相比,该算法几乎不能被蓄意破坏,鲁棒性非常好。

(3)嵌入容量:document.xml 文件中的“修改标识”的数量非常可观。秘密信息采用十六进制编码,替换“修改标识”的属性值后六位,相当于一个属性值可以隐藏 24 位二进制数。也就是说:假如隐藏 5 000 位二进制秘密信息,那么需要“修改标识”的属性值的个数为 $5\,000/6 \times 4$ 个。但由于不是所有的标识都用

来进行隐藏信息,因此隐藏信息的量与传统算法相比有所减少,但隐秘性增加,安全性更高。

3 结束语

文中提出了一种新的基于 Word 2007 文档“修改标识”属性值的信息隐藏方法,通过修改属性值的后六位实现秘密信息的隐藏。XML 格式为基于文本格式的信息隐藏提供了巨大的研究空间。实际上,一些其他用于 XML 文档中的信息隐藏的方法也可以在 Word 2007 文档中的 XML 文件中使用^[13]。下一步的主要研究工作将是在该算法的基础上如何增加隐藏秘密信息的容量和如何在新版本的 Word 中实现秘密信息的隐藏。


参考文献:

- [1] Petitcolas F A P, Anderson R J, Kuhn M G. Information hiding—a survey[J]. Proceedings of the IEEE, 1999, 87(7): 1062–1078.
- [2] 刘显德,唐国维,富宇,等.一种基于 Word 文档的信息隐藏方法[J]. 电子技术应用, 2005, 31(4): 16–17.
- [3] 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.
- [4] 王海春,邱寄帆,邱敦国.一种基于 Word 文档的数字密写设计与实现[J]. 微计算机信息, 2006, 22(30): 47–49.
- [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] 王智,周洪玉.基于 Word 文档的信息隐藏方法的实现[J]. 信息技术, 2008, 32(11): 30–31.
- [8] 李兵兵,王衍波,徐敏.基于 ZIP 文档格式的信息隐藏方法[J]. 计算机工程, 2011, 37(5): 155–157.
- [9] Walk through: Word 2007 XML format[EB/OL]. (2008–08–25)[2008–10–25]. <http://msdn.microsoft.com/en-us/library/bb266220.aspx>.
- [10] What's up with all those rsids? [EB/OL]. [2006–12–11]. <http://blogs.msdn.com/brian-jones/archive/2006/12/11/what-s-up-with-those-rsids.aspx>.
- [11] Andrew R, Juan S, James N. A statistical test suite for random and pseudorandom number generators for cryptographic applications[M]. [s. l.]: NIST Special Publication, 2001.
- [12] 徐敏,王衍波,李涛. Word 2007 文档信息隐藏的新方法[J]. 计算机研究与发展, 2009, 46(z1): 112–116.
- [13] 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.

一种基于Word XML的信息隐藏新方法

作者：[董艳](#)，[徐江峰](#)，[DONG Yan](#)，[XU Jiang-feng](#)

作者单位：[郑州大学 信息工程学院, 河南 郑州, 450000](#)

刊名：[计算机技术与发展](#)

英文刊名：[Computer Technology and Development](#)

年，卷(期)：2015 (2)

引用本文格式：[董艳](#). [徐江峰](#). [DONG Yan](#). [XU Jiang-feng](#) 一种基于Word XML的信息隐藏新方法[期刊论文]-[计算机技术与发展](#) 2015 (2)