**信息隐藏第二次大作业**

**1410658 杨旭东**

1. **Word文档简介**

Microsoft Word是当今全世界使用最普遍的文字处理软件。Word给人们的日常办公和文字编辑带来了极大的便捷。随着网络和通信的发展，Word作为电子文档的标准格式之一，在各种基于Web的系统中应用非常广泛。

Microsoft Word的存储格式随着它的新版本的变化而变化，而Microsoft公司并不对外公开Word文档的存储格式；Word文档的文字、图片、表格等显示信息通过自己的CSS来描述，而且有自己自定义的命名空间；Word文档是一种以字符信息和格式信息为主要内容的具有多种信息的文件；Word文档是一种OLE的复杂文档，Word文档分简单文档和复杂文档；但是由于Word 97以后，都有Piece table，可以不用区别这两种文档；为了能在各国文字之间通用，Word文档采取了Unicode；同时考虑到有的文字是全字符的，Word采用压缩方式并且用ASCII码来表示字符，以尽可能的减少存储量；文字格式和段落格式的存储也具有自己的特点，在它的文件中有一份默认的格式，而文字和段落格式的保存都采用的是“差额保存”，而且将格式一样的连续文字作为run统一保存，此做法可以节省大量的空间，因此较大的word文档的存储空间并不会太大。

1. **Word文档格式的发展**

Microsoft Office Word 2007提供了一种新的默认文件格式，叫MicrosoftOffice Word XML格式，该种格式基于开放打包约定(Open PackagingConventions)，XML Paper Specification(XPS)也是基于这个约定；Microsoft Office97到Microsoft Office 2003中使用的二进制文件格式仍然可以作为一种保存格式来使用，但是它不是保存新文档时的默认文档。

在1999年，Microsoft通过Microsoft Office Excel 2002中的Spreadsheet ML，将XML引入了Microsoft Office XP当中。Spreadsheet ML是一个良好的开始，但是它没有提供完全真实的功能。在下一个版本的Microsoft Office产品中，Microsoft Office Word 2003引入了Word processing ML。Word processing ML是非常重要的一步，因为它是Microsoft Office提供的第一个完全真实的XML文件格式。通过Microsoft Office 2003，您可以解析Word processing ML文件，向其中添加、更新或处理数据。但是，仍然存在着一些限制。例如，您必须将二进制文件(例如图片)编码为XML文件中的文本，如果文件中包含大量图片的话，这必然会增加文件的尺寸。另外，Word 2003直接将所有自定义XML数据嵌入到描述文档的Word processing ML当中。这样使得自定义XML难于访问和处理所有的外部过程。

Word 2007中新的XML文件格式将文件分成一些文档部件，文件的部分内容通过每个这样的部件定义，这样就可以方便的实现：当您希望查看或者更改文件中的某些内容的时候，您可以快捷地查找您想得到的文档部件，然后编辑它。例如文字格式，而其它的基于XML的文档部件不会受到任何影响。类似的，所有的自定义XML数据都有它自己的部件。现在，处理自定义的XML就更加简单的。这允许您使用很少的代码来生成文档。新的文件格式更加的强大，并且可以更为简单地处理自定义XML，而且它比二进制文件格式要更小。新的文件格式使用开放打包约定，利用ZIP技术。

1. **XML格式的开放打包约定**

开放打包约定指定了Word 2007文档中新文件格式的结构。

为了理解一个Word 2007文档的结构，您必须理解这种新文件格式中的三个主要组成部分：

(1)部件项。包解压后的文件与部件项形成一一对应的关系。例如，如果您右击一个Microsoft Office Excel工作簿，然后选择提取它，您将会看到一个workbook．xml文件，一些sheetn．xml文件，和其它的一些文件。这些文件都是包中的文档部件。

(2)内容类型项。内容类型项是对一个文档部件存储的文件类型的描述。

例如，image／jpeg表示了一个JPEG图片。这种描述允许一些工具，例如，MicrosoftOffice和其他工具，来修改它的内容，进而确定它其中的内容。

(3)关系项。关系项描述：几个部件项是如何组成和链接成一个文档的。

关系在文档包中是以XML部件存储的，例如／rels／．rels。

1. **Word 2007文档包**

Word 2007提供了一种新的默认文件格式，叫做Microsoft Office WordXML(Word XML格式)，其默认的保存格式为"．docx”，文档占用空间在改变格式后有一定程度的缩减。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；Documentxml．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。

这种格式有一个压缩的ZIP包组成，包中包括文档所有的内容，如图所示：



1. **Word 2007文档部件**

Word包中的内容由一些逻辑部件组成，每个部件的内容是一个特定的文档部分，这些可以使第三方方便的完成Word 2007文档的组建，例如：

(1)格式定义

(2)注释

(3)列表定义

(4)文档内容

(5)页眉

(6)图表

(7)图片

(8)关系图

在包中，Word文档中的每一部分都由一个单独的部件来表示。这些部件可以由XML文件组成，例如包含标记的Word XML格式的文档部件，和一些附加的内容(例如二进制的图片或嵌入的OLE文件)。所有的这些都包含在包中，但是，更为重要的是，实际的文件目录结构是任意的，通过开放打包约定定义的一些除外。

文件的有效性不是由文件结构决定的，而是由包中文件的关系决定的。您可以重新排布和命名ZIP容器当中的任何部件，当保证了正确无误的更新关系，和保证文档部件之间正确的关联关系的情况下。更新容器中的部件后，如果这种关联关系是正确的，那么您就可以打开文档。Word 2007文档的初始文件结构只是一种Word创建的默认结构，它使得您可以轻松的定位文件的组成。只要您保证文档部件之间正确的关联关系，那么您可以修改这种文件结构。

例如，在Word 2007中，容器文件代表了一个文档。在容器文件中，文档

由按序排布的部件组成。例如，一个Word 2007文件可能会包含一些下列的目

录和文件：

(1)[Content\_Types]．xml： 描述文件中包含的所有的内容类型。

(2)rels folder： 存储所有指定部件的相关联的关系部件。

(3)rels file：称为关系部件，描述了文档结构中的起始关联关系。

(4)datastore folder：包含文档中的自定义XML数据部件(自定义XML数据部件是一个XML文件，您可以将节点绑定到文档的内容控件中)。

(5)iteml．xml file：包含了一些文档中出现的数据(例如一些自定义XML数

据部件)。

(6)docProps folder：包含了一些应用程序的属性部件。

(7)App．xml file：包含了一些应用程序特定的属性。

(8)Core．xml file：包含了所有基于开放打包约定的文档格式的通用文件属性。

为了理解基于Microsoft Office Open XML格式(Office XML格式)的组成，您可以将它的所有部件提取出来。要打开这个文件，您需要在您的计算机中安装一个ZIP工具，例如WinRAR。您可以通过以下步骤在Word 2007中打开一个WordXML格式：

(1)创建一个临时目录：存储文件及其部件。

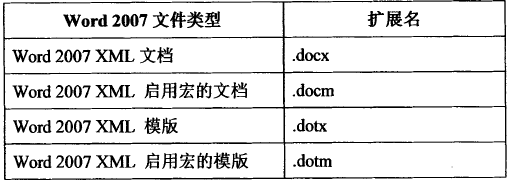
(2)创建一个Word 2007文档，包括文本，图片，以及其它信息，保存为一个．docx文件。

(3)在文件名的末端添加一个．zip扩展名。

(4)双击文件。这时将会在ZIP应用程序当中打开该文件。您可以查看组成文件的每个部件。

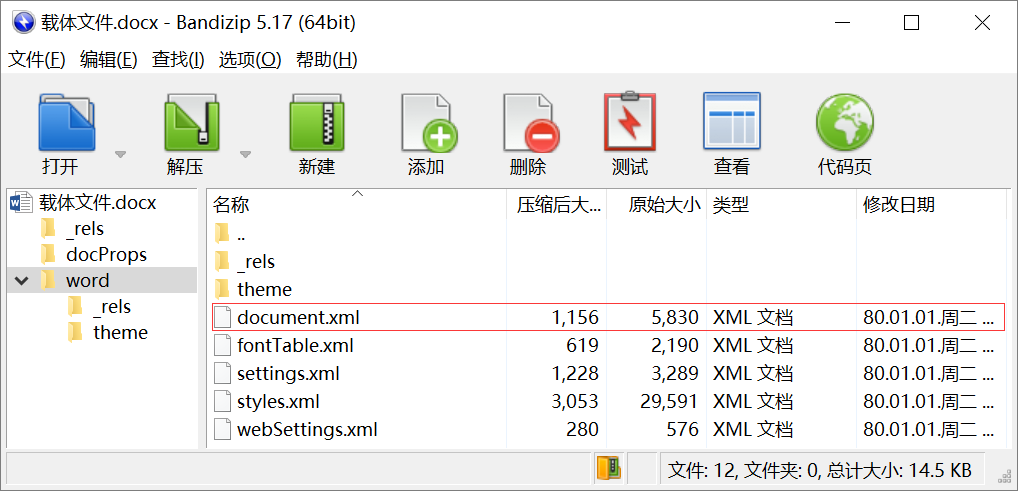
(5)将这些部件提取到刚才创建的临时目录当中。

(6)集成的ZIP压缩可以将文件的尺寸减少75％。文件最终被分成一些模块化的文件结构，它使得数据的恢复更为可能，并且增强了安全性。新的格式将文件分成了不同的组件，从而可以独立的管理和修复。以新格式创建的文件也可以根据不同的文件类型，拥有一些特殊的扩展名，如下表所示。



1. **Word 2007中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。



如上图所示为用Bandizip（解压软件）打开一个后缀为．docx的文件，docx格式的文件实际上是一个ZIP文件；将一个docx文件的后缀改为ZIP后，是能够用解压工具例如WinRAR、Bandizip等打开或着解压的；实际上，Word 2007的基本文件可以看做是docx文件的容器。docx格式文件的主要内容是以XML格式保存的，文件是直接保存在一个ZIP文件中，而不是直接保存于磁盘，然后将文件取扩展名为docx，如果将其后缀改为ZIP后解压，可以发现解压出来的文件夹中有word这样一个文件夹，其中包含了Word文档的大部分内容。而其中的document．xml文件则包含了文档的主要文本内容。



上图是一个Word 2007文档包中的document．xml的截图，元素w：document中的主文档所引起的文档部件的关系是一系列名字空间；w：body是指正文元素，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)；文字的属性，字体、颜色等子属性被包含在w：rPr元素中；文字的RGB值是由w：eolor的属性w：val的属性值描述的；元素w：p的属性w：rsidR、w：rsidRr、w：rsidRDefault、w：rsidP以及w：r的属性w：midRr就是“修改标志”。

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，如果出现差异，那么表示这个段落在后面的编辑中被修改，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、下划线时候也会用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。

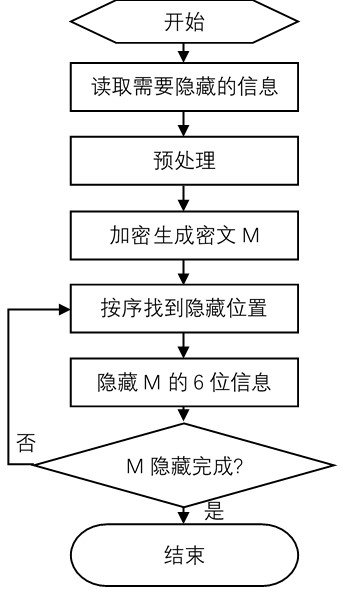
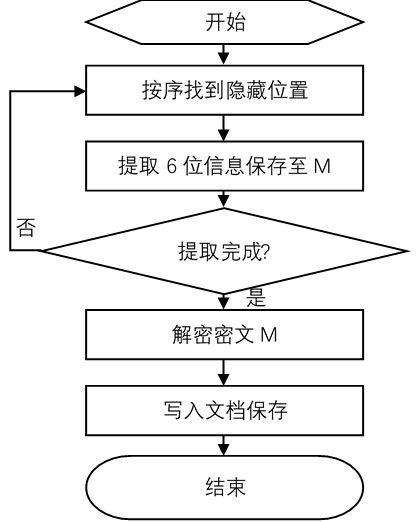
每当Word文档被打开编辑时，都会生成1个独一无二的ID号；这个ID号被保存在“修改标识”的属性值里面，而且它是随机产生的一个字符串，和时间等没有关系。

大量的“修改标识”存在于document．xml文件中。如上图中第4行w:rsidP="00D1EED0"、w:rsidR="00B15B20"、w:rsidRDefault="00B15B20"、w:rsidRPr="00913DB4"和第24行的w:rsidP="00F1B6AB"、w:rsidR="00B15B20"、w:rsidRDefault="00B15B20"、w:rsidRPr="00913DB4"，且“修改标识”属性值共8位。

1. **基于Word 2007的信息隐藏算法**
2. **算法思想**

在Word 2007文档中嵌入秘密信息后的载密文档需要满足：1)载密对象(嵌入秘密信息后的载密文档)必须满足Word XML格式的要求。2)载密对象必须能够正常显示和使用。

经研究发现，修改Word 2007文件中document．xml的某些“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Word2007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，本算法选取修改标识w：rsidP属性值的后六位用于信息隐藏。信息隐藏、提取的流程图如下图所示：



开始

解密密文M

写入文档保存

按序找到隐藏位置

提取6位信息保存至M

提取完成?

结束

否

是

开始

读取需要隐藏的信息

预处理

加密生成密文M

按序找到隐藏位置

隐藏M的6位信息

M隐藏完成?

结束

否

是

1. **基于Word XML算法的实现**

**(1)信息隐藏**

i)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中修改标识个数的随机二进制串S2(由0、1组成)；

ii)将与S1异或后的秘密信息M转换为十六进制串M\_h；

iii)按顺序查找T中待修改标识，并判断S2中对应位置的值是否为1，若为1，转iv)，否则，转iii)；

iv)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v)，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iii)，否则转v)；

v)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(如6个0)；

vi)生成载密文档，并替换T，结束。

**(2)信息提取**

i)读取Word 2007文档中的document．xml(载体文档T)，输入随机二进制串S2，秘密信息M赋值为空串；

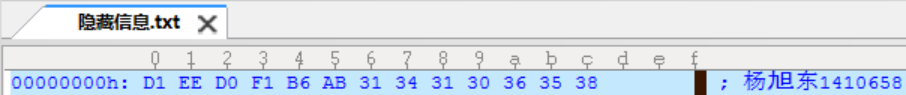
ii)按顺序查找T中待修改标识，若此标识为特定标识，转iii)。否则，判断S2中对应位置的值是否为1，若为1，转iii)；否则，转ii)；

iii)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ii)；

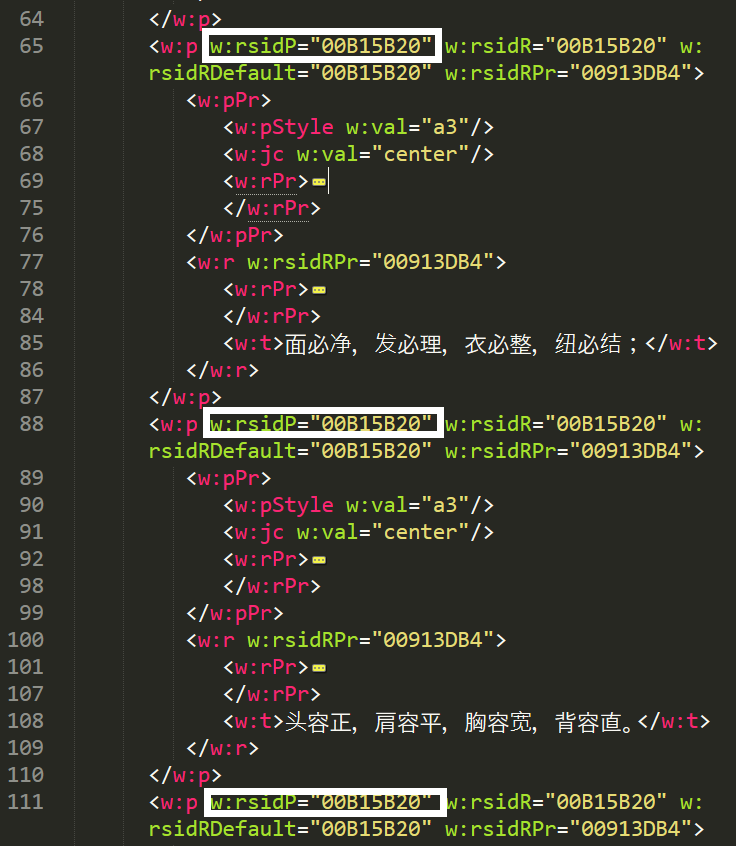
iv)把M转换为二进制字符串与S1异或，得到隐藏的明文信息。

1. **基于Word XML算法的具体实现**

待隐藏信息“杨旭东1410658”，其16进制编码为“D1EED0F1B6AB31343130363538”，



在后面补充“0”成为6的倍数，并且添加结束标记“000000”，新的串为“D1EED0F1B6AB313431303635380000000000”，将其按6个一组划分，之后存入w：rsid属性值之中。在修改之前，原始w：rsid属性值均为w:rsidP="00B15B20"，如下图：



修改之后，变为要隐藏的内容，分别是

w:rsidP="00D1EED0"（隐藏信息开始）

w:rsidP="00F1B6AB"

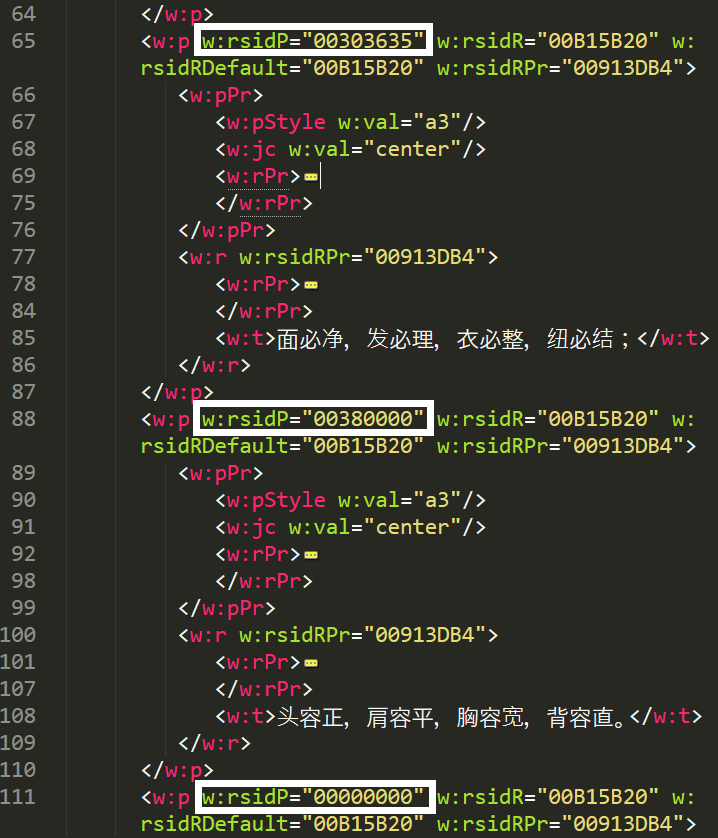
w:rsidP="00313431"

w:rsidP="00303635"

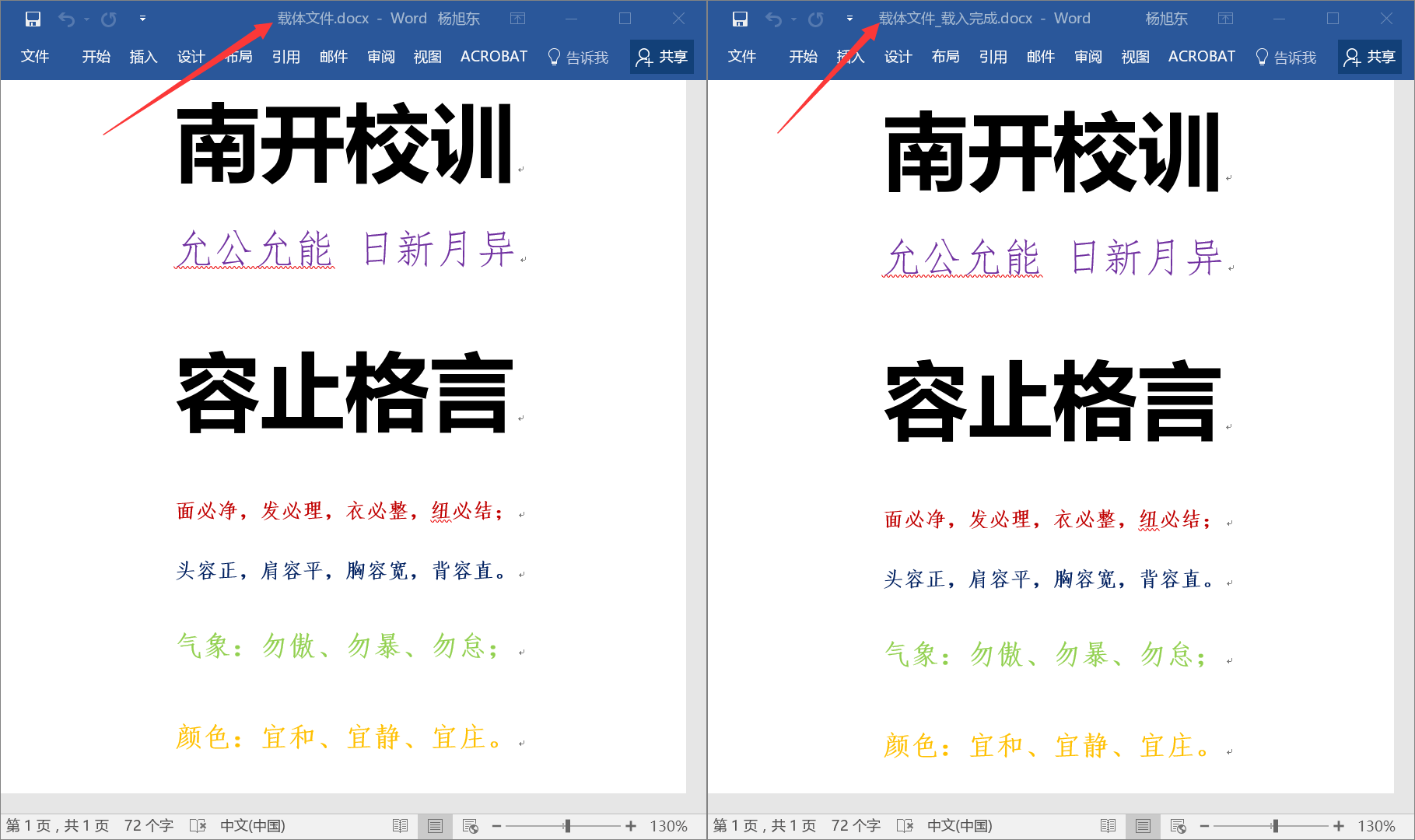
w:rsidP="00380000"（隐藏信息结束）

w:rsidP="00000000"（结束标记）

篇幅有限只截取了其中三个修改位置，如下图：



隐藏后对载体文件无影响，对比如下：



提取就是直接将w：rsid属性值后六位提取并拼接，解码后得到隐藏信息，对比可知，成功提取：



详细代码（用MATLAB实现）在“代码与结果”文件夹中，其中“encode.m”是嵌入，“decode.m”是提取。

1. **算法分析**

该算法的基础是Word 2007文档采用的新的Word XML格式，这种格式使得该算法容易实现。只用抽取document．xml单个文档，就可以实现该算法。

(1)透明性：隐藏秘密信息后的载密对象与原始载体对象相比，在正常情况下，主观上是很难察觉到载体对象的变化。隐藏后的载体文件与原始载体文件只存在细小差别，例如大小的细小差别，但是视觉上完全一样。它的透明性，也称不可感知性或者隐蔽性非常好。很好的满足信息隐藏基于隐蔽性的基本要求。

(2)鲁棒性：对文档的“修改标识”的属性值进行修改，正常情况下，该隐藏算法容易实现，而且容易提取。经实验验证，人为的对Word 2007文档实施一些攻击行为，例如：改变字体、改变文字大小、删除文字等等，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，本算法选取修改标识w：rsidP属性值的后六位用于信息隐藏。

(3)嵌入容量：document．xml文件中的“修改标识”的数量非常可观，可隐藏的信息量相对较大。但由于不是所有的“修改标识”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小，但隐秘性增加，安全性更高。

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位二进制数的秘密信息。也就是说：假如隐藏5000位二进制秘密信息，那么需要“修改标识”的属性值的个数为5000／6\*4个，隐藏秘密信息的容量很可观，但是与既有算法的隐藏容量相比，该算法隐藏秘密信息的容量有所下降。

(4)不可检测性：隐藏后载体中“修改标识”的属性值，仍然是以“00”开头，且隐藏位置是由伪随机序列中“1”的位置确定，在无法获知密钥的情况下，非法拦截者几乎不能确定被修改的“修改标识”的属性值的位置，因此也不可能获得隐藏的秘密信息。