第 34 卷 第 6 期 计 算 机 工 程 2008 年 3 月
Vol.34 No.6 Computer Engineering March 2008
·安全技术· 文章编号：1000—3428(2008)06—0155—03 文献标识码：A 中图分类号：TP309

基于 XML 的信息隐藏方法

丁光华，刘嘉勇，孙克强
（四川大学信息安全研究所，成都 610064）

摘 要：分析 XML 的文件结构，提出一种基于 XML 的信息隐藏方法。对待隐藏信息进行预处理后，按照文件中元素的结构，通过对文件中的元素名称进行同义词替换来实现信息隐藏。实验表明，该方法实现简单，有较好的可扩展性和较高的安全性。
关键词：信息隐藏；XML 文件；同义词；元素名称

Information Hiding Algorithm Based on XML

DING Guang-hua, LIU Jia-yong, SUN Ke-qiang
（Information Security Institute, Sichuan University, Chengdu 610064）

【Abstract】 The paper proposes an information-hiding method based on the structure of XML, based on analysis of the data structure of XML file. The secret data is preprocessed, then according to the structure of the elements, the secret data is hided by replacing the element name with thesaurus. Experimental results prove the method is easy to implement, and has high expansibility and high security.
【Key words】 information hiding; XML file; thesaurus; element name

1 概述

传统基于密码学的信息隐藏方式，通过特定的数学算法，把信息变换为不被其他人所理解的形式，但是同时也暴露出信息的重要性。信息隐藏把信息以不可见的方式隐藏在普通的通信载体中，隐蔽了通信过程的存在性，为安全通信提供了另一重保障，成为信息安全领域的一个新的研究方向[1-2]。

文本信息具有信息量大、传递快捷等特点，是人们广泛使用的信息交换载体。目前，Web应用的主要载体是HTML，随着人们需求的不断提高，Web应用空前复杂，HTML的局限性也逐渐显现出来。W3C开发了一种超越HTML的新语言XML语言[3]，由于其具有使用灵活、可扩展性强等特点，一经提出就受到业界的广泛支持，各种基于XML的技术和应用被相继开发出来，并逐渐取代HTML语言。

文献[4]提出的基于 XML 的句法规则和文档逻辑结构信息隐藏算法，通过改变 XML 文档中标记的属性的表示方法和命名，变换标记的字体，改变同名元素的排列顺序来进行信息隐藏。本文在研究 XML 文件结构的基础上，利用 XML 使用灵活、数据和显示相分离的特点，通过对 XML 元素名称进行同义词替换，使用不同的元素名称表示不同的信息，实现了信息隐藏。

2 XML 文件结构特点和分析

2.1 XML 文件结构特点

XML 即扩展标记语言，它是 SGML 的优化子集，本质上是一种元语言。XML 的主要特点是可扩展性、灵活性、自描述性，它定义了一种文件格式，其基本思想是数据按照其固有的结构以层次化的格式存储，从而形成一种基于内容的格式。它与 HTML 相似，但是各方面的性能确优于 HTML。XML 所使用的标记由用户根据自己的需要自己定义，极大地增强了文件的可读性。XML 的优势还在以其内容和显示格式相分离，通过使用样式表语言(如:CSS,XSL)，可以很方便地实现同一内容的不同显示。

就组成而言，一个 XML 文档由内容(文档中的数据部分)和标记组成。标记说明了文档数据的意义和组织结构，元素是标记的基本类型之一，同时也是 XML 文件的基本组成部分，文件的内容在各个元素中描述，元素使用成对出现的标记将数据内容划分为具有不同意义的组成部分，如：<元素名称>内容</元素名称>。由于 XML 的可扩展性，元素的名称和属性值都可以由用户自己定义。作为文件的基本组成部分，各元素的数据内容相互独立，为了表达复杂的数据结构，元素之间可以嵌套，即一个元素的内容可以包含另一个元素，具有父子关系。

就结构而言，一个完整的 XML 文档可以映射为一幅文档树图，根节点是文档节点，它代表了整个 XML 文档，文档节点有序言节点和文档元素节点 2 个子节点。序言节点是对 XML 版本声明，字符编码和文档类型定义等说明信息。XML 的标准字符编码集是 UNICODE，同时它也支持 GB2132 和 BIG5。文档元素节点主要描述文件内容，一个 XML 文档只能有唯一的一个文档元素节点，但它可以包括元素子节点，元素子节点又有自己的子节点，节点之间通过嵌套表达复杂的文件结构。

2.2 XML 文件结构特点分析

元素是 XML 文档的基本组成部分，元素的名称作为标识用来描述元素的内容，可以由用户根据需要自己定义，使得对元素名称的替换成为可能。其次 XML 中各元素以树状结构进行存储，元素之间有明确的父子关系，为信息隐藏及检测提供了有利条件，而且一个实际的 XML 文件通常都由大量的元素组成，为信息隐藏提供了大量的载体，保证了信息隐藏容量。由于 XML 同时支持中文和西文，这种信息隐藏方法对中文和英文同样适用。但是根据 XML 的特点，在

作者简介：丁光华(1981－)，男，硕士研究生，主研方向：信息隐藏，网络安全；刘嘉勇，教授；孙克强，硕士研究生
收稿日期：2007-04-12 **E-mail**：acedgh2004@yahoo.com.cn

信息隐藏时须用同义词进行元素名称的替换，以保证 XML 文档的自描述性的特点，并增强隐蔽性，同时为了保证 XML 的可用性，所使用的同义词需要在文档类型定义中声明，但对同一类型的 XML 文件，文档类型只需要定义一次，之后可重复使用。

3 基于 XML 文件结构的信息隐藏和提取算法

3.1 信息隐藏的基本思想

根据以上对 XML 文档的结构的特性，本文的信息隐藏的基本思想如下：首先获得 XML 文件中文档元素节点的所有子元素节点的元素名称(由于文档元素节点的唯一性，文档元素的名称不能作为隐藏载体)，为每个元素名称构造一组同义词，形成元素名称的同义词库，然后在文档类型定义中进行声明，最后按序遍历文档中的各元素，根据待隐藏信息的值按序对各元素的名称进行替换实现信息隐藏。

3.2 构造同义词库

在进行信息的隐藏之前，必须先构造同义词库，此词库由信息隐藏方通过安全信道传送给接收方。同义词库包括隐藏所使用的同义词组和各个词所对应的信息值。针对中文和英文，可使用不同的方法构造同义词组。对于中文，可采用两种方法：(1)直接寻找同义词，如“厂商”与“厂家”，“标题”，“题目”和“主题”；(2)通过增词或减词的方式，如：“厂商”与“生产厂商”，“标题”，“本文标题”与“文章标题”。对于西文，可寻找同义词，如：“title”和“headline”，由于 XML 对大小写敏感，不同的大小写代表不同的意义，因此还可通过不同的大小写方式来构造同义词库，如：“title”，“Title”和“TITLE”。在每个同义词组中，除了代表不同信息值的词外，还有一个词用来标识是否使用此词组来进行信息隐藏，如：“title”，“Title”和“TITLE”中“title”表示“0”，“Title”表示“1”，“TITLE”表示无隐藏信息。

3.3 创建文档类型定义

为了保证 XML 文档的可用性，信息隐藏所使用的同义词库必须在文档类型定义中声明，对于元素声明的内容不仅是包括其存在性声明，而且要说明它们与其他元素的依赖关系和位置关系。为了简化声明过程，可使用参数实体，实现具有相同属性的元素的定义的共享。

3.4 信息隐藏算法

假设所构建的同义词库中每组同义词有$(N+1)$个同义词，它们对应的数值为 0，1，…，N，其中，0，1，…，$(N-1)$依次代表 N 进制信息的值；N 表示不含隐藏信息。

信息隐藏时，首先预处理待隐藏的信息，即将待隐藏信息加密并进行进制转换，然后遍历原始 XML 文档对应的逻辑树，按照转换后的 N 进制信息通过查同义词库的方式依次变换元素的名称，框图如图 1。

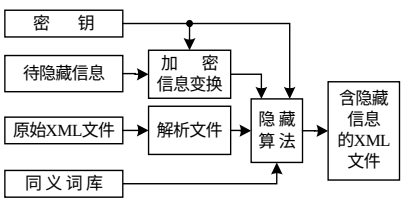

图 1　信息隐藏框图

信息隐藏过程：

Step1 加密和信息变换。首先在密钥的控制下，将待隐藏信息进行加密，然后将加密后的信息进行变换，将加密后的信息转化为 N 进制信息，同时获得 XML 文档中除文档元素外所含有的元素个数的总数，若元素的个数大于转换后的 N 进制信息个数，则在 N 进制待隐藏信息后补上若干数值 N，直到其数量和元素的个数相等。若元素的个数小于转换后的 N 进制信息个数，则说明隐藏容量受限，不能进行信息隐藏。

Step2 解析原始 XML 文件，获得文件内容根节点文档元素节点，根据 XML 文档元素的树状结构，在密钥的控制下，读取 XML 文档中的元素节点的元素名称，并同步读取待隐藏信息值，查找同义词库，找到元素名称对应的同义词组，再在同义词组中找到与待隐藏的 N 进制信息值所对应的词，并用此词替换原来的元素名称。直到所有的元素名称替换完毕。

Step3 保存文件，获得含有隐藏信息的 XML 文档。

信息的检测过程为隐藏过程的逆过程，框图如图 2。

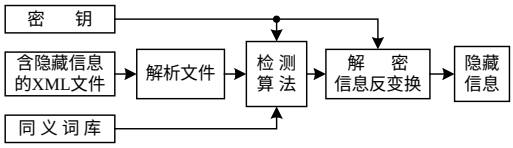

图 2　信息检测框图

信息检测过程：

Step1 解析含有隐藏信息的 XML 文件，获得文件内容根节点文档元素节点，解析原始 XML 文档，获得文件内容根节点文档元素节点，根据 XML 文档元素的树状结构，在密钥的控制下，读取 XML 文档中的各元素节点的元素名称，同步查找同义词库，找到元素名称对应的数值，直到查找完所有的元素。最后保存所得信息。

Step2 解密和信息反变换。删除获得的信息序列中的 N 值，所得 N 进制系列即为隐藏的 N 进制系列，再将 N 进制信息进行进制转换，最后对使用密钥解密转换后的信息，得 XML 中所隐藏得信息。

Step3 关闭文件，检测过程结束。

3.5 实验及分析

根据前文介绍的信息隐藏和检测的方法，进行了实验。下面是一个使用这种方法的实例，隐藏的信息为“12345678910”，同义词库中同义词组的大小为 5。信息隐藏前后文件的截图为图 3 和图 4。图 3 为原始 XML 文件的截图，图 4 为信息隐藏后 XML 文件的截图。

```
<!DOCTYPE LIST SYSTEM "list.dtd">
<DOCUMENT>
 <CUSTOMER>
  <NAME>
   <LASTNAME>Andy</LASTNAME>
   <FIRSTNAME>Fickle</FIRSTNAME>
  </NAME>
  <DATE>October 26,2006</DATE>
  <ORDERS>
   <ITEM>
    <PRODUCT>Bread</PRODUCT>
    <NUMBER>3</NUMBER>
    <PRICE>$3.5</PRICE>
   </ITEM>
   <ITEM>
    <PRODUCT>Cucumber</PRODUCT>
    <NUMBER>6</NUMBER>
    <PRICE>$1.5</PRICE>
   </ITEM>
  </ORDERS>
 </CUSTOMER>
 <CUSTOMER>
  <NAME>
   <LASTNAME>Jim</LASTNAME>
```

图 3　原始 XML 文件

```
<!DOCTYPE LIST SYSTEM "list.dtd">
<DOCUMENT>
 <CONSUMER>
  <NAme>
   <FAMILYNAME>Andy</FAMILYNAME>
   <FIRSTNAMe>Fickle</FIRSTNAMe>
  </NAme>
  <DATe>October 26,2006</DATe>
  <ORDers>
   <LIST>
    <PRODuCT>Bread</PRODuCT>
    <NUMBEr>3</NUMBEr>
    <PRIce>$3.5</PRIce>
   </LIST>
   <LIST>
    <PRODuCT>Cucumber</PRODuCT>
    <SUM>6</SUM>
    <PRIce>$1.5</PRIce>
   </LIST>
  </ORDers>
 </CONSUMER>
 <CONSUMER>
  <NAme>
   <FAMILYNAME>Jim</FAMILYNAME>
```

图 4　含隐藏信息的 XML 文件

3.6 算法特性分析

(1)安全性分析。从上面的讨论和实验可知，本算法具有如下特点：

1)算法只修改 XML 文档的元素名称，并不修改文档的数据内容，且 XML 的内容和显示是分离的，所以在隐藏前后文档的显示没有任何区别，不易被人察觉。

2)算法对元素名称进行的是同义词替换，不影响文档的可阅读性，当通过编辑形式打开文档时也有一定的隐蔽性。

3)算法使用的是同义词替换，替换后文档本身的长度基本不变，若所构建的同义词库中的用作信息隐藏的同义词组的大小为 N，则每个元素名称所能隐藏的信息大小为 lb$N$ bit，当 N 为 4 的时，每个元素所隐藏的信息为 2 bit，因而可以通过提高同义词库的大小来增加信息隐藏容量。

4)算法可扩展性好，可随机选取用做隐藏的元素，或改变遍历元素节点的顺序，来获得较高的安全性。

(2)脆弱性分析及改进。由于 XML 文件所含的冗余信息较少，此种基于 XML 的信息隐藏方法的鲁棒性相对较弱，当文件中的元素名称受到删除或篡改攻击后，隐藏的信息将部分丢失或被篡改，影响秘密信息的检测。但是文件受到删除攻击后，会影响文件的可用性和可读性，因而攻击很容易被发现。

在 XML 文件中常会出现某一元素的兄弟元素与其自身为同一类型元素的情况，此时为了增强隐蔽性，可选择其中的部分元素作为隐藏载体。另外对于英文元素名称若只用是改变标记的大小写的方式来构建同义词库，这样通过类似 ULTRAEDIT 的编辑软件打开时，很容易引起用户的怀疑。为了增强隐蔽性，可同时采用直接寻找元素名称同义词和改变元素名称大小写状态两种方法来构建同义词库，且在改变元素名称大小写状态时只选取大小写书写形式相似度高的字母。

由于浏览器在解析 XML 语言时，对大小写不敏感，有些网页作者可能使用大小写标记语言。为了避免含有隐藏信息与不含隐藏信息的文档的混淆，可在 XML 文件中添加标识信息。位置可选择在 XML 的声明中，通过改变声明中的关键字“XML”的书写形式来实现。或者可在待隐藏信息的头部前面添加一段标识码，标识码作为待隐藏信息的一部分隐藏在 XML 文件中，这样做的代价是牺牲了一部分隐藏容量。

为了增强算法的抗攻击能力，可在信息隐藏前，对待隐藏的信息只进行纠错编码，如汉明码。当待隐藏信息量很大时，可先对隐藏信息进行压缩，然后再隐藏。

4 结束语

文本文件主要由一系列的字符编码组成，其所含的冗余信息较少，与基于图像的信息隐藏相比，其实现的难度较大，算法的安全性与基于图像的信息隐藏还有较大的差距，本文提出的基于 XML 的信息隐藏算法，虽然基于特定的文件结构，但是其具有隐蔽性好、实现简单等优点，随着 XML 在互联网及其他领域中的广泛使用，将有较好的应用前景。

参考文献

[1] Izquierdo E, Kim H J, Macq B. Introduction to the Special Issue on Authentication, Copyright Protection, and Information Hiding[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2003, 13(8): 729-730.

[2] 张登银, 邓兰兰. 信息隐藏技术及其性能分析[J]. 南京邮电学院学报, 2004, 24(3): 63-69.

[3] 瞿裕忠, 张剑锋, 陈 峥, 等. XML 语言及相关技术综述[J]. 计算机工程, 2000, 26(12): 4-6, 30.

[4] 吴 晶, 王书文. 基于 XML 语言的信息隐藏方法[J]. 中国安全科学学报, 2005, 15(12): 78-80.

(上接第 154 页)



(a)对图 3(a)的恢复　(b)对图 3(b)的恢复　(c)对图 3(c)的恢复

图 4　图像的恢复结果

本算法的恢复效果能准确且较理想地表达原图信息。算法公开时，在本文的平台穷举破解图 2(b)至少需要 4.173×10^{12} 年。实验表明算法能有效地对抗剪切攻击，是一种有效、快速、安全的置乱方法。

6 结束语

针对彩色图像，本文提出一种快速的置乱算法。对像素的颜色分量进行分存管理，降低了同时受到攻击的概率，从而提高了对抗剪切等攻击的能力。为避免颜色分量存放过于集中，任意 2 个行(列)移位参数之差不宜太小(不小于 10)。图像的恢复依赖于分块内的第 1 行第 1 列像素，因此，n 的取值不宜太大，取值越大分块越少，一旦所依赖的这少数像素被破坏，图像恢复的效果将受到较大影响，实验发现 n 取 6~10 有较好的性能。下一步将研究调色板图像的置乱技术。

参考文献

[1] Zeng Wenjun, Lei Shaw-min. Efficient Frequency Domain Selective Scrambling for Digital Video[J]. IEEE Trans. on Multimedia, 2003, 5(1): 118- 129.

[2] Wyner A D. An Analog Scrambling Scheme Which Does not Expand Bandwidth, Part I: Discrete Time[J]. IEEE Trans. on Inform. Theory, 1979, 25(3): 261-274.

[3] Ville D V D, Philips W, Walle R V, et al. Image Scrambling without Bandwidth Expansion[J]. IEEE Trans. on Circuits Syst. Video Technol., 2004, 14(6): 892-897.

[4] Watanabe O, Nakazaki A, Kiya H. A Fast Image-scramble Method Using Public-key Encryption Allowing Backward Compatibility with JPEG2000[C]//Proc. of IEEE International Conference on Image Processing. Singapore: IEEE Press, 2004.

[5] Tie Xiaoyun, Zou Jiancheng, Ward R K, et al. Some Novel Image Scrambling Methods Based on Affine Modular Matrix Transformation[J]. Journal of Information and Computational Science, 2005, 2(1): 223- 227.

[6] 丁 玮, 闫伟齐, 齐东旭. 基于 Arnold 变换的数字图像置乱技术[J]. 计算机辅助设计与图形学学报, 2001, 13(4): 338-341.

[7] Yang Yali, Cai Na, Ni Guoqiang. Digital Image Scrambling Technology Based on the Symmetry of Arnold Transform[J]. Journal of Beijing Institute of Technology, 2006, 15(2): 216-220.

[8] 李 敏, 费耀平. 基于队列变换的数字图像置乱算法[J]. 计算机工程, 2005, 31(1): 148-149, 152.