基于 Unicode 编码文本水印的分密级公文系统①

郭福亮, 周 钢

(海军工程大学 电子工程学院计算机工程系, 武汉 430033)

摘 要: 介绍公文系统基本安全管理方法, 对传统的文本数字水印技术进行分析, 提出一种基于 Unicode 编码的文本数字水印算法, 采用密钥对水印信息加密后通过异或运算嵌入到文本 Unicode 编码中, 同时附加密级编码、奇偶校验码和纠错码构建水印文本编码, 实现公文分密级和全过程的安全管理, 在定性分析中发现算法具有广泛适用性和较强安全性.

关键词: 文本水印; Unicode 编码; 分密级; 公文系统;特征编码

Classification Document System Based on Unicode Text Watermarking

GUO Fu-Liang, ZHOU Gang

(Department of Computer Technology, College of Electronics Eng. Naval Univ. of Engineening[1)], Wuhan 430033, China)

Abstract: The method of document system security management is introduced and traditional text fragile watermark technology is analyzed firstly, and then proposes a new text fragile watermarking algorithm based on Unicode. And using security classification code, parity check code and error correction code to build the watermark text coding for achieving document security classification and the entire process of safety management. At last the qualitative analysis shows the algorithm has the good applicability and safety.

Key words: text watermarking; Unicode; Classification; document system; feature encoding

1 引言

随着网络技术、信息技术和计算机技术为基础的自动化办公技术快速发展, 以及政府、军队信息化建设的不断推进, 电子公文系统在政府、军队等要害战略部门得到了广泛应用. 这些部门通过内外网隔离等技术建立内部物理专网体系阻断外界破坏和攻击, 但是在公文在内部网中的生成、传输、浏览、存档和销毁等环节确保保密性、完整性、不可否认性仍是重点研究工作.

对于公文系统的安全性研究, 国内外学者主要从下面几个方面已经进行很多工作, 文献[1]研究了 GAP 技术阻断内部网的虚拟隔离实现内部电子公文的安全保密, 文献[2]设计一种基于 PKI 技术和 CA 系统建立的安全体系保护公文系统, 文献[3]建立电子印章和 CA 系统实现公文安全管理. 这些方法均能提高公文信息安全性, 但在有效进行分密级进行安全保护, 涉密公文追踪等方面研究较少.

本文通过研究设计一种通过文本编码技术和加密算法嵌入密级等信息的公文数字水印, 设计分密级一个分密级公文系统完成公文的生成、传输、阅览和打印中的水印嵌入和读取, 实现系统分密级公文安全管理, 并定性分析了系统优点.

2 数字水印技术

数字水印技术是一种将版权、用户和产品信息等秘密信息隐藏到数字载体中的信息隐藏方法[4]. 数字水印具有安全性、不可见性、鲁棒性、可检测性和水印容量等特征. 根据嵌入数字载体不同可分为图像水印、音频水印、视频水印、文本水印和数据库水印等, 随着数字技术发展, 会出现针对新数字载体的水印技术.

文本是公文系统中最常用的数字载体, 在文本中嵌入数字水印进行版权保护、内容认证、操作追踪等

① 收稿时间:2013-07-02;收到修改稿时间:2013-07-19

具有十分重要意义. 文本数字水印是将带有标识文本创建者或版权拥有者的水印信息嵌入文本中, 并采用合适技术提取出来的技术[5]. 基本的文本数字水印算法设计包括文档结构微调法, 基于语法水印算法, 基于语义水印算法, 基于汉字特点的水印算法和基于变换域的水印算法, 其中基于文档结构特点算法通过引入人类视觉系统模型(HVS, Human Vision System), 具有较好的水印容量、鲁棒性和适用性, 包括行移编码、字移编码、特征编码和空格编码等方式.

文本数字水印还可以利用文本 Unicode 编码进行水印信息的嵌入, 文献[6]采用字符 Unicode 编码末位的奇偶性嵌入“0”或“1”, 通过嵌入位构建数字水印序列, 但水印容量较小; 文献[7]通过在字符 Unicode 编码的空字符处隐藏信息嵌入水印, 但适用性较差.

本文在总结文本编码水印技术基础上, 根据应用需求采用 Unicode 编码方法嵌入数字水印, 通过对文本字符根据公文系统密钥将要嵌入信息加密后转化为二进制, 按照一定序列嵌入到公文的文本 Unicode 编码中, 同时通过字符特征编码嵌入水印完成公文打印输出的追踪.

3 基于Unicode编码的数字水印算法

本文设计基于 Unicode 编码的数字水印, 采用文本普遍采用的 Unicode 编码技术对水印包含的用户信息等进行 Unicode 编码加密, 同时附加密级、奇偶校验和纠错码, 最后嵌入到文本载体中实现水印算法. 本文水印信息算法基本步骤为:

(1) 编码转换, 将水印中用户信息转换为 Unicode 编码;

(2) 编码加密, 用户信息 Unicode 编码结合密钥进行加密;

(3) 附加信息, 将密级信息、奇偶校验码和汉明纠错码附加在用户信息编码中组成数字水印;

(4) 水印嵌入, 将水印信息编码嵌入到文本载体的分组编码中.

3.1 Unicode 编码

Unicode 编码是一种多字节等长编码, 通过双字节或多字节表示一个字符, 从而在更大范围内将数字代码映射到多种语言的字符集. Unicode 标准提供三种编码格式, UTF-8, UTF-16 和 UTF-32, 其中 UTF-16 得到最广泛应用, 能够包含书写系统中绝大多数字符. Unicode 编码字符集如表 1.

表 1　Unicode 编码字符集

Unicode 编码	字符集
0000-1FFF	字母表
2000-2FFF	符号和标点
3000-4DFF	CJK 辅助符号
4E00-9FFF	CJK 统一表意字符
A000-DFFF	保留部分
E000-FFFD	限制使用

3.2 水印信息加密和编码

为了提高系统水印安全性, 在嵌入水印前对水印信息进行加密处理, 提取时再检测解密. 系统中水印信息 W 包括密级信息 WS 和用户信息 WU 两部分, 其中 WU 信息采用密钥 K 循环取模方法进行加密, 密级 WS 信息根据 Unicode 编码保留部分直接嵌入到文本载体 T 的开头部分. 具体嵌入编码和对应密级如表 2.

表 2　密级信息嵌入 Unicode 编码表

Unicode 编码	密级
A000	公开
A001	内文
A002	秘密
A003	机密
A004	绝密

将水印用户信息 W_u 的字符 Unicode 编码转换为二进制序列 $W_u=w_1w_2\cdots w_n$, 密钥 K 的字符 Unicode 编码转换为二进制序列 $K=k_1k_2\cdots k_m$, 将 W_u 和 K 进行一对一循环取模加密得到加密后新水印序列 $M=m_1m_2\cdots m_n$, 其中 $m_i=w_i \oplus k_{(i \bmod m)}$, $1\leqslant i\leqslant n$. 那么水印信息序列变为 $W=W_s+M$.

为了提高水印安全性在进行加密基础上, 同时嵌入奇偶校验码和纠错码:

(1) 奇偶校验信息 W_p 嵌入

在嵌入水印信息后, 如果仍有可填充位, 那么结合数据信息的奇偶数, 对填充进行填充. 如果校验的数据中有奇数个“0”, 那么填充“1”, 如果有偶数个“0”, 那么填充“0”, 嵌入奇偶校验码的水印信息序列 $W_1=W+W_p$.

(2) 纠错码 W_h 嵌入

汉明码是一种能自动检测并纠正一重错的线性纠错码[8], 设数据位数为 m, 校验位数为 k, 总编码数为 n=m+k, 那么根据汉明不等式:

$$2^k-1\geq n, 2^k\geq m+k+1$$

采用(7, 4)分组码最小码距 d=3, 能够纠 1 个错或检 2 个错, 假设 $A=a_1a_2a_3a_4$, 那么添加校验位 $a_5=a_1+a_2+a_3$, $a_6=a_2+a_3+a_4$, $a_7=a_1+a_3+a_4$, 其中“+”表示位的“与”运算, 最后构造含有纠错位的汉明码为 $A=a_1a_2a_3a_4a_5a_6a_7$. 根据汉明纠错编码要求, 按照属性数值数据进行汉明编码添加在填充位中. 最终得到水印信息加密和进行奇偶校验和纠错码嵌入的二进制序列 $W'=W_s+M+W_p+W_h$.

3.3 水印信息的嵌入与提取

待嵌入水印信息 W’的数字载体文本 T, 水印信息嵌入基本思想是将文本 T 按照水印信息 W’长度划分为若干段 r, 对每一段 T 按文献[9]方法转化为 Unicode 编码, 将 W’中 W_s 信息直接附加在数字载体文本编码文首, W’其余部分同文本 T 的 r 段分组进行异或运算得到隐藏信息后的二进制编码.

水印信息嵌入的基本模型如图 1.

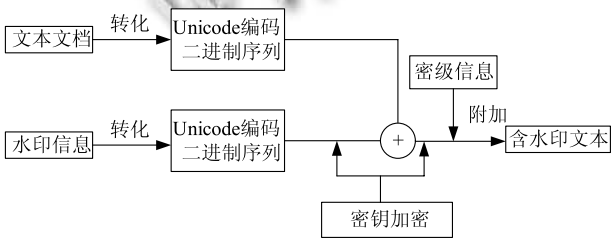

图 1 水印信息嵌入模型

水印信息嵌入基本步骤为:

(1) 编码扩展, 将水印信息的 Unicode 编码按照 0→01, 1→10 的方法进行编码拓展;

(2) 编码分隔, 将扩展后的 Unicode 编码按照 16 位一组进行分隔;

(3) 文本编码, 从载体文本中顺序读取一个字符并得到其转化的 Unicode 编码;

(4) 文本分组, 将载体文本编码按照水印信息扩展编码长度分组;

(5) 水印嵌入, 对每一组文本编码同水印编码采用异或运算得到嵌入水印后文本.

水印信息的提取是水印信息嵌入文本载体的反过程, 通过对文本编码文首字符编码提取若为保留字符则检测文本含有水印并提取文本密级, 然后按照文本水印嵌入的反过程利用密钥进行解密和编码分解提取水印信息和文本信息. 水印信息提取基本流程如图 2.

图 2 水印信息提取流程

文本文档打印中的水印嵌入和提取采用文本特征编码方法, 结合行移编码和字移编码, 采用文献[10]设计的模型完成不可见水印文本的嵌入打印和扫描提取实现涉密文档管理.

4 分密级公文系统设计

4.1 系统设计

为了实现公文系统的分密级管理, 控制涉密文本文档生成、修改、浏览和打印的流转全过程的安全性, 设计了基于 Unicode 编码分密级水印算法以及采用特征编码实现纸制密级文档跟踪. 系统采用 C/S 结构, 信息安全管理部门为服务器端, 系统内用户为客户端, 系统基本流程图如图 3.

系统完成对文本文档的分密级、全流程安全管理, 主要实现以下功能:

(1) Unicode 编码水印嵌入和提取功能. 通过对 WORD、PDF 等文本文档的 API 接口编程实现在文档保存时嵌入 Unicode 编码水印, 在文本打开浏览时进行 Unicode 编码水印检测和提取.

(2) 文字特征编码水印嵌入和提取功能. 该功能主要针对文本打印环节实现涉密文本追踪, 在文本打

印时嵌入文字特征编码水印，同时能够利用精度扫描实现水印检测和提取.

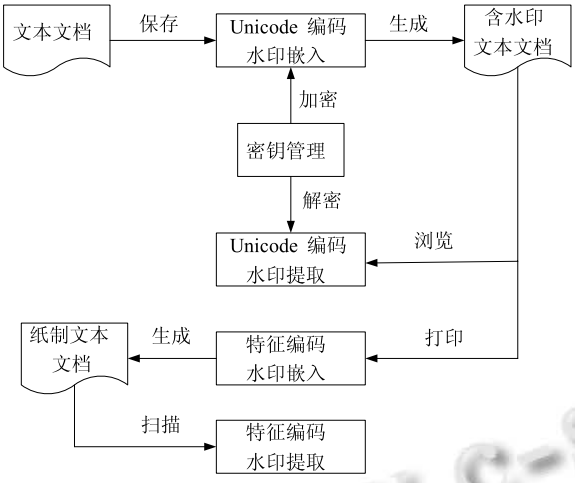

图 3　系统流程图

(3) 用户管理和密钥分配功能. 通过数字水印嵌入时采用的密钥进行分配管理实现涉密文档的知密范围，只有具有同一等别密级用户、同一范围(如同一单位)采用相同密钥，并通过各用户配置密钥表实现不同级别用户上下级隶属密级文档管理. 信息安全管理部门通过定期更新密钥配置表，对客户端程序进行升级实现密钥管理. 系统密钥分配表基本结构如表 2.

表 2　密钥配置表结构

列名	数据类型	标识	说明
ID	Varchar(10)	主键	用户 ID
Grade	Char(4)		涉密等级
Dept_code	Varchar(10)	外键	单位编码
Key	Binary(16)		密钥
IP	Nvarchar(150)		IP 地址范围
Remark	Nvarchar(100)		备注

4.2 系统实现

按照公文系统设计实现分密级公文系统，当文本文档创建、修改时进行保存时调用公文系统嵌入数字水印，用户可以自己定密级，根据用户单位调用密钥配置表使用规定密钥进行加密，加密基本流程如图 4.

系统加密基本界面如图 5.

当文件嵌入水印后，系统通过更改文本文件图标的形式将在文件上添加特殊标签记号，如图 6.

4.3 性能分析

利用本系统对 pdf, word, txt 和 wps 四种常见的公文文本格式文件，按照不同密级和短中长不同文本篇幅进行 Unicode 编码水印信息嵌入和提取，通过对 200 份测试文本文件处理发现:

(1) 有效性: 100%文件能够实现 Unicode 编码的数字水印嵌入，而其中 1.5%的水印文本不能进行有效提取.

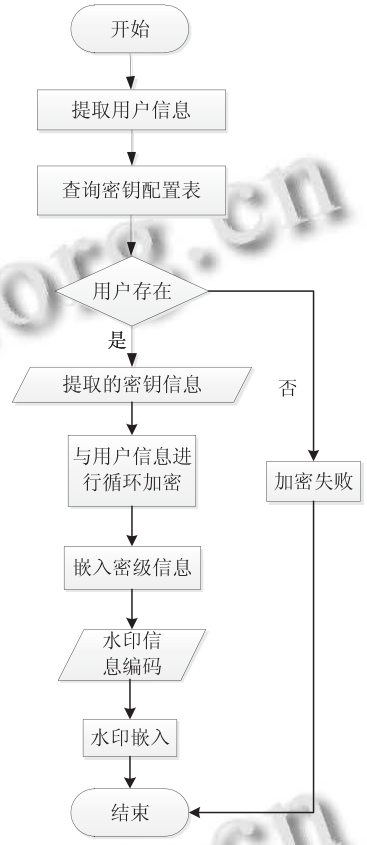

图 4　系统加密基本流程图

图 5　系统实现图

图 6 嵌入水印的文件标记

(2) 实用性: 100%文件能够在不同计算机操作系统, 不能格式中实现水印算法.

(3) 安全性: 100%嵌入水印的文本文件在未安装水印系统中不能打开.

在系统实际应用中, 研究总结近半年来在某单位试运行和反馈情况, 结合软件测评结果, 系统从以下几个标准进行分析:

(1) 安全性: 通过下发密钥配置表完成密钥管理, 通过用密钥对数字水印信息加密和对文本编码循环加密实现文本信息安全保证.

(2) 鲁棒性: 由于加密后水印信息采用循环异或加方法嵌入到文本编码中, 只要有一部分文本编码未被篡改既能提取水印.

(3) 水印容量: 本算法的水印容量主要取决于文本载体编码量, 理论上至多具有与文本等量的数字水印容量.

(4) 适用性: 水印算法主要针对文本 Unicode 编码进行水印嵌入和提取, 能够针对多种格式文本文档, 具有很强的适用性.

5 结论

公文系统安全管理目前主要采用对文件的密钥加密, 在密级控制和文本安全性都有缺陷, 本文提出了基于 Unicode 编码的数字水印分密级公文系统, 即实现了对公文文本编码级的加密同时通过编码数字水印实现分密级管理, 且效果良好, 适用广泛.

设计的系统通过文本 Unicode 编码嵌入使用密钥加密后的水印信息并附加密级编码、奇偶校验码和纠错码等进行编码实现的文本数字水印, 通过数字水印和密钥配置实现涉密文件的分密级管理, 并在文件流转全过程中安全管理.

参考文献

1 蒲天银.安全隔离网闸技术发展探讨.计算机时代,2006,(6):18.

2 蔡谊,沈昌祥.PKI 技术在电子政务中的应用.计算机应用研究,2002,19(10):11–13.

3 雷震声.计算机网络管理及系统开发.北京:电子工业出版社,2002:156–160.

4 Awrangjeb M. An overview of reversible data hiding. ICCIT 2003. Bangladesh. 2003. 75–79.

5 白剑,杨榆,徐迎晖等.基于文本的信息隐藏算法.计算机系统应用,2005,14(4):32–35.

6 陆绿,方勇.基于字符 Unicode 奇偶性的数字水印设计与实现.计算机技术与发展,2010,20(8):176–179.

7 黄国超,王衍波,张凯泽.基于 Unicode 编码的信息隐藏算法研究与设计.计算机技术与发展,2011,21(10):233–236.

8 梁红玉,陈冬梅.扩展汉明码的交织重排算法研究及其实现.计算机应用,2012,32(S1):85–87.

9 冀斌.Windows 平台下应用软件多语言支持.计算机工程,2004,30(12):163–165.

10 房婧婧.基于不可见水印的纸质文档泄密源头管理系统的设计与实现[学位论文].北京:北京邮电大学,2010.

(上接第 165 页)

3 楚晓丽.K-Means 聚类算法和人工鱼群算法应用于图像分割技术.计算机系统应用,2013,22(4):92–94.

4 于海涛,贾美娟,王慧强等.基于人工鱼群的优化 K-means 聚类算法.计算机科学,2012,39(12):60–64.

5 王联国,韩晓慧,宋磊.基于改进混合蛙跳-K 均值聚类算法的无功电压控制分区.传感器与微系统,2013,32(6):18–21.

6 刘衍民,隋常玲,赵庆祯.基于 K-均值聚类的动态多种群粒子群算法及其应用.控制与决策,2011,26(7):1019–1025.

7 鲍新中.基于粒子群的 K 均值算法和粗糙集理论的财务预警.系统管理学报,2012,21(4):461–469.

8 Kalyani S, Swarup KS. Particle swarm optimization based K-means clustering approach for security assessment in power systems. Expert Systems with Applications, 2011, 38(9): 10839–10846.

9 Tsai CY, Kao IW. Particle swarm optimization with selective particle regeneration for data clustering. 2010 Elsevier Ltd. 2011, 38(6): 1–6.