基于特征抽取的电子公文防伪系统的研究*

曾凡智[1,2]，赵慧民[1,3]，卢炎生[2]，方艳梅[3]
（1. 佛山科学技术学院计算机系，广东 佛山 528000；
2. 华中科技大学计算机系，湖北 武汉 430074；
3. 中山大学信息安全技术广东省重点实验室，广东 广州 510275）

摘 要：针对 web环境下存在对电子公文、电子图片的恶意伪造和删改等现象，需要对电子公文信息内容的正确性加以验证也即防伪，提出一种方便可行的基于全文特征抽取的算法。该算法抽取电子公文的各种特征值，生成特征值验证码对电子公文进行加密与防伪。实践证明，采用这种技术对一般的电子公文真伪处理可在 2 s内完成，其防伪出错率在百万分之一。

关键词：电子公文；防伪；特征抽取；散列函数
中图分类号：TN919 文献标识码：A 文章编号：0529-6579（2007）06-0044-04

在信息化数字时代，通过网络对电子公文、电子图片传输过程中经常出现恶意伪造和删改等现象。为了保证计算机用户特别是政府部门电子公文传输的可靠性和公信力，需要对电子公文信息内容的正确性加以验证也即防伪，客观上需要通过防伪技术来保证电子公文信息的正确性[1]。

目前比较常用的公文防伪方法主要是采用手写签名，防伪印章（公章）、防伪油墨，防伪印刷等物理方法，这些用于纸质公文的防伪方法，防伪能力较差。对于电子公文采用比较多的是手写电子签名、电子印章和数字签名等技术。手写电子签名和电子印章使用起来比较复杂，可靠性与防伪性有待提高。数字 CA证书虽然能较好地保障文件的真实性，但需要第三方的技术支持，对于面向公众的文件显然不合时宜[2-7]。

本文提出一种在 Web环境下方便实用的电子公文防伪与验证方法，能够利用计算机软件来完成电子公文的防伪和验证工作。实践证明，基于电子公文特征抽取从而生成验证码的电子公文防伪技术可以广泛应用于政府电子公文，各种电子档案，各种电子合同等电子文档的防伪，甚至可以通过一定的辅助手段应用于纸质公文和证书的防伪。

1 系统的工作原理

1.1 公文验证码的生成与加密

防伪系统通过抽取电子公文中统计特征信息如段落、字数等，位置特征信息如特定位置的汉字编码、汉字数，汉字笔画数等，关键词特征信息如关键词重复次数和内码等特征信息，并以单向散列函数将这些大量的数字散列为固定长度的数据信息，经过加密后就可以生成验证码。公文验证码生成和验证，首先解决的是对电子公文信息内容的分析与特征抽取；其次，如何获取这些与特征相关的验证信息加以验证；最后，如何防止这些验证特征码被伪造。

电子公文特征抽取和验证码生成可以采用图 1的工作流程来实现。首先，发文者使用防伪系统通过设定的算法抽取电子公文的特征信息，然后对特征信息进行单向散列处理与加密后生成一个不可伪造的验证码。每个验证码实际上包括三部分：一部分包含电子公文的统计特征信息，如字数、段落数等；第二部分包含电子公文的位置特征信息如汉字编码、汉字数，汉字笔画数等；第三部分包含采用散列算法抽取的关键字信息，如关键词位置、关键词重复次数、关键词字节编码等。在传递电子公文时，将生成的验证码与电子公文一起进行传输，把随公文一起传递的验证码称为外部验证码。

1.2 用户验证

验证过程与外部特征验证码生成过程是相互关联的，验证过程首先通过对接收电子公文的特征抽取及进行散列算法和加密获取内部验证码，将内部验证码与电子公文上的外部验证码进行匹配与比较，如果一致则说明接收的电子公文为真，否则为假（如图 2）。

* 收稿日期：2007－05－23
基金项目：广东省信息安全技术重点实验室资助项目；佛山市科技专项基金资助项目（200601001）
作者简介：曾凡智（1965年生），男，副教授，博士研究生；E-mail：coolhead@126.com

图 1　公文外部验证码生成流程图

Fig. 1　Document certification external code process

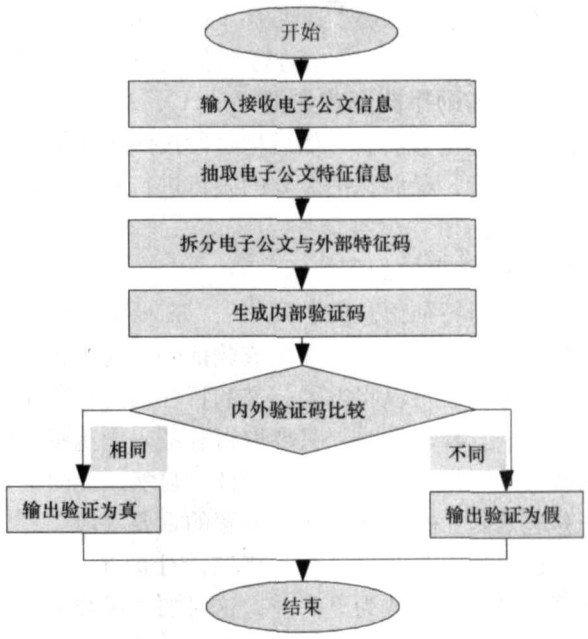

图 2　公文内部验证码生成与验证流程图

Fig. 2　Document certification internal code process

2　系统的关键技术及实现

2.1　建立关键词库

为了提高防伪的精确度，公文的关键词特征抽取是系统的核心部分。为此，需要针对各种类型的文件，建立相对应的关键词库。考虑词库的易操作性和通用性，采用文本文件来建立关建词库，每一个词库由多条关键词记录构成，每条关键词记录包括三部分：第一部分为关键词所包含的汉字个数；第二部分为关键词内容；第三部分为每个汉字所对应的 GB2312码。根据汉字编码表，每个汉字都有一个特定的四位 GB2312码。词库中每条记录中的关键词可通过统计的方法将一些经常出现的字、词进行添加，关键词的汉字编码可通过 GB2312汉字编码表得到。关键词的汉字编码、重复次数与偏移位置作为关键词特征信息参与验证码的生成过程中的运算。

2.2　特征信息的抽取技术

电子公文验证码生成和验证，最关键的技术是电子公文特征信息的抽取和加密算法。算法选取的优劣会直接关系到系统的可靠性和安全性。一般情况下，抽取的特征信息需要满足下列条件：①信息简单而易于建立数学模型。②信息具有唯一性，即所有的电子公文不可能出现同样的特征信息，这样生成的验证码才会唯一（现实中要求同码的概率非常低即可）。这是电子公文防伪的基础。③特征信息具有可重现性，即相同的电子公文内容能抽取相同的特征，从而生成与外部验证码相同的内部验证码，这是验证可靠与可信的前提。

为了描述问题的方便，特征抽取算法用特征抽取函数来表示，设特征抽取函数为 $T(X)$，原始信息矢量为 X，则特征信息矢量 S为 $S=T(X)$。

为了区分不同的特征信息，需要将公文统计特征信息，特定偏移位置的位置特征信息与关键词的特征信息分开处理。即：$S=(S_1, S_2, S_3)^T$其中 $S_1=T(X_1)$，$S_2=T(X_2)$，$S_3=T(X_3)$，其中 X_1，X_2，X_3分别为公文统计原始信息、特定偏移位置的原始信息与关键词的原始信息矢量，S_1，S_2，S_3为对应抽取的特征信息矢量。

2.3　散列函数算法

当采用某一特征抽取算法时，若按照电子公文中文字的序号进行采样与抽取，其抽取的信息量非常大，显然，这会导致验证码过长，不能直接作验证码用。因此应该采用散列函数算法压缩特征信息量[8-9]。

散列函数可以将大量的数据信息散列为一小段的数据信息。实际上，散列函数通常选择单向散列函数。由于单向散列函数具有不可逆性，在已知 $u=H(s)$时很难根据 u求得 S，这样可以有效防止特征抽取算法过程被破解。当散列函数为 $H(x)$时，取 $u=H(S_1)=H(T(X_1))$，$v=H(S_2)=H(T(X_2))$，$w=H(S_3)=H(T(X_3))$，其中 u，v，w分别为公文统计特征信息、位置特征信息与关键词的特征信息的散列值。

本文选取的散列函数 H（x），散列过程为：对于抽取的信息量较大的特征信息 X_1，X_2，X_3，设立一个辅助矩阵，首先把 X_1，X_2，X_3 分别按照矩阵的列进行分段后，然后把每段按照次序填充到矩阵每一行，最后，把矩阵的各行进行异或叠加形成 u，v与 w。

2.4 特征信息的加密技术

在获得特征信息的散列值后，需要对这些信息进行加工和加密处理，并生成验证码。通过各种加密编码算法就可以将电子公文的特征信息蕴藏于验证码中，这样验证时，只要取得验证码中蕴藏的信息就可以运用于电子公文的验证。显然，此时的加密算法是防止验证码被恶意破解的重要手段[10−11]。

提出加密分两步进行：首先，对 u，v，w分别加密，设加密函数为 P（x），这时加密函数 P（x）不必求出其逆函数 P^{-1}，所以可以选择较复杂的加密算法。即有：p=P（u），q=P（v），h=P（w）然后，再对 p，q，h进行第二次加密，得到外部验证码。设此加密函数为 Φ，则验证码 ω为：ω=Φ（p，q，h），设 Φ存在逆函数，则有：$(p,q,h)^T=\Phi^{-1}(\omega)$。由于采用了多维特征抽取技术、散列技术和多重的加密技术，增加了恶意破解的难度。验证时，验证时对接收到的电子公文按上述计算过程分别得出 $(p',q',h')^T$ 和 $(p,q,h)^T=\Phi^{-1}(\omega)$，比较两者是否相同就可以得到验证结果的真伪。本文采用 64位 DES加密算法来对散列信息 u，v和 w进行加密处理。

2.5 核心算法

针对以上关键技术的分析，系统实现的核心算法如下：

算法 1 获取电子公文的外部验证码。

PROCEDURE GetCert_ID（fpdoc，pwords，encr_f）

//fpdoc：公文文件指针；pwords：关键词库指针；encr_f：加密函数指针）

STEP1：获取公文特征值：U=get_f1（fpdoc），V=get_f2（fpdoc），W=get_f3（fpdoc，pwords）；//获取统计特征信息，位置特征信息，关键字特征信息

STEP2：对特征信息进行散列处理：P=hash_f1（U），Q=hash_f2（V），H=hash_f3（W）；

STEP3：填充特征信息散列矩阵：M=M（P，Q，H）；

STEP4：加密获取验证码：Cert_ID=（*encr_f）（M）；

End

算法 2 用外部验证码进行电子公文真伪的验证。

PROCEDURE Check_ID（fpdoc）

//fpdoc：接收公文文件指针

STEP1：拆分获取打包的外部特征验证码：Cert_ID=Detach_ID（fpdoc）；

//Detach_ID为拆分外部特征验证码函数

STEP2：计算内部验证码 Cert_ID1=GetCert_ID（fpdoc，pwords，encr_f）；

STEP3：比较 Cert_ID，Cert_ID1，输出电子公文防伪验证结论；

End

3 应用实例

基于以上讨论的算法过程，研究开发了一套基于网络环境下的电子公文验证码生成、验证软件系统，系统的文件核心处理部分采用 C语言编写[12]。系统的功能结构如图 3所示。

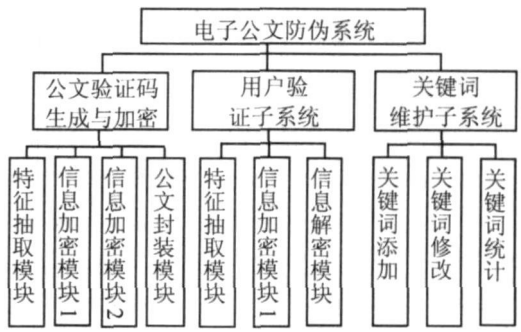

图 3 电子公文防伪系统功能结构图

Fig.3 Electronic document falsification system structure

针对一个文本文件处理的效果及获取的验证码如图 4。

图 4 文本文件及获取的验证码

Fig.4 Document files and getting test codes

系统在处理文本文件时，抽取了以下特征信息：如版本号、段落数、文字总数、特殊位置的汉字编码、汉字笔画数，关键词重复次数及内码等。

然后，通过单向散列函数将这大量的数字散列为固定长度的数据信息，加密这些信息就可以生成验证码。实验证明，该套系统对本文电子公文的加密防伪，其方便实用性得到极大提高，一般公文信息处理在 2 s内完成，防伪出错率在百万分之一以内，达到较好效果。

4 结　语

针对 Web环境下电子公文防伪问题，通过关键词库的建立，提出了一种基于全文特征的抽取技术和构造单向散列函数技术从而生成公文特征验证码的算法，并对算法进行深入的研究与实现。在网络环境下，此方法的研究与实现技术对电子公文进行防伪识别提出了一种新的思路，并且识别率较高，算法简单、实用。将来通过更深入的研究，这种防伪验证原理也可以用于重要文件和各种合同文本的等纸质公文的防伪，具有广阔的应用前景。

参考文献：

[1] 刘浩翰. 计算机防伪系统的设计与实现[J]. 电脑学习，2000，16(2)：7－8.

[2] 戴跃伟，施燕，王执铨. 一种用于电子公文防伪的鲁棒图像水印算法[J]. 机器人技术与应用，2002，6：42－46.

[3] 李子臣，杨义先，吴伟陵. 一类基于数字签名的密钥认证方案[J]. 电子学报，2000，28(4)：115－116.

[4] BRUCE Schneier. Secrets and lies：digital security in a networked world[M]. 吴世忠，马芳，译. 北京：机械工业出版社，2001.

[5] 沈世镒. 近代密码学[M]. 南宁：广西师范大学出版社，1998.

[6] 冯晖. 计算机密码学[M]. 北京：中国铁道出版社，1999.

[7] RIVEST R，SHAMIR A，Adleman L. A method for obtaining digital signatures and public-key cryptosystems[J]. Communications of the ACM，1976，21(2)：120－126.

[8] 辛运帏，廖大春，卢桂章. 单向散列函数的原理、实现和在密码学中的应用[J]. 计算机应用研究，2001(2)：25－27.

[9] 严蔚敏，吴伟民. 数据结构[M]. 北京：清华大学出版社，2006.

[10] WANG Huiqin，LI Renhou. An encrypted fuzzy image-adaptive watermarking algorithm based on HVS[J]. Journal of Electronics，2002，11(4)：510－514.

[11] YEN S，LAIH C. New digital signature scheme based on discrete logarithm[J]. Electronics Letters，1993，29(12)：1120－1121.

[12] 曾凡智. Web环境下电子公文防伪系统的研究与开发[J]. 佛山科学技术学院学报，2007，25(1)：46－50.

A Research of Electronic Document Certification Testing System Base on Characteristics Getting

ZENG Fan-zhi[1,2]，ZHAO Hui-min[1,3]，LIU Yan-sheng[3]，FANG Yan-mei[1]

(1. Department of Computer Science，Foshan University，Foshan 528000，China；
2. Department of Computer Science，Huazhong University of Science and Technology，Wuhan 430074，China；
3. Guangdong Key Laboratory of Information Security Technology，Sun Yat-sen University，Guangzhou 510275，China)

Abstract：In the web environment，there is a novel algorithm of electronic documents falsification，which uses the document whole characteristics getting to certify the reliability of the information. To avoid malicious attacks，the algorithm extracts various statistic characteristics of the document，and utilizes the characteristics getting for security and falsification. The results show that the method can differentiate all kinds of electronic documents truthfulness.

Key words：electronic document；certificate testing；characteristics getting；hash function