文章编号:1001 - 9081(2010)06 - 1501 - 04

基于随机顺序的图形验证码改进算法设计

李　欢[1],高　岭[1],刘　琳[2],邢　斌[1]
(1. 西北大学 信息科学与技术学院,西安 710127;　2. 西安邮电学院 电子与信息工程系,西安 710061)
(lihuan@nwu.edu.cn)

摘　要:针对目前常用图形验证码过于简单,容易被自动化程序识别所产生的安全隐患,提出了基于随机顺序的图形验证码改进算法。该算法首先创建一张随机背景色的真彩图片,然后在特定范围内随机选择验证字符个数,在此基础上将随机字符写入随机位置并标识字符顺序。其主要特征为验证码字符数目不固定,字体不固定,字符位置不固定和验证字符输入顺序不固定。实验证明,基于随机顺序的图形验证码在健壮性和可靠性方面都有很大提高,为保证 Web 安全提供了强有力的保障。

关键词:图形验证码;网络攻击;Web 安全
中图分类号: TP391　　**文献标志码**:A

Improved algorithm for generating random CAPTCHA

LI Huan[1], GAO Ling[1], LIU Lin[2], XING Bin[1]
(1. *College of Information Science and Technology, Northwest University, Xi'an Shaanxi* 710127 *China*;
2. *Department of Electronics and Information Engineering, Xi'an University of Posts and Telecommunications, Xi'an Shaanxi* 710061, *China*)

Abstract: The current CAPTCHA is so simple that it can be easily identified by the automated procedures, which may cause many security risks. On the purpose of enhancing the security of Web applications, an improved CAPTCHA algorithm based on the random sequence was introduced. First of all, the algorithm created a true color image with random background color. Second, it determined characters and the number of character that were randomly generated in a certain range. Finally, the algorithm put characters on the image and the positions of characters were random. To determine the sequence of characters, lines were put between characters. The characteristic of the algorithm is that the number, font, position and input sequence of the characters were uncertain. The experimental results show that the CAPTCHA based on random sequence has great advantages, which can provide a strong security for Web applications.

Key words: CAPTCHA; network attack; Web security

0　引言

随着在线攻击越来越多,B/S 构架的安全性问题日趋严重。目前大多数站点都采用验证码技术来增强 B/S 构架的安全性。验证码技术,即由程序随机生成一组数字、字母或者数字与字母的组合序列,可以有效地防止来自自动识别软件或机器人软件的暴力破解,在一定程度上保证了 Web 应用的安全。这是一种人机区分的方法,因技术简单,易实现,所以被各网站特别是论坛性质和邮件性质的网站广泛使用。

目前常用的图形验证码技术已经比较成熟,但仍存在不足之处。其中:文献[1]中随机问题阅读式验证码虽在一定程度上可以阻止机器程序的攻击,但这样的验证码通用性差,验证码图库固定,灵活性和随机性差;文献[2]中选择用汉字作为字符库,有效地扩大了字符库的数量和自动识别难度,其缺点在于适应性差和程序的执行效率比较低。Google、腾讯、搜狐和百度等知名网站上所采用的验证码,其产生方法均为:将随机产生固定长度的字符串写入到固定背景色图片中,并在写入过程中将字符串扭曲变形,加入噪声干扰,以增加机器识别难度。但这些噪声同时也增加了真实用户的识别难度,而且字符数目固定、字体固定、位置固定和字符输入顺序固定等特征都给机器程序自动识别提供了便利,安全并不能得到有效保障。

鉴于以上原因,本文在对验证码 Web 攻击原理、图形验证码生成过程及图形验证码自动识别技术深入研究的基础上,提出了一种基于随机字符数目、随机顺序、随机位置的图形验证码生成算法。

1　基于验证码的一般攻击流程

黑客对于验证码的攻击通常不会干扰合法用户的正常使用,其参与实体为恶意攻击者和 Web 服务器。用 A 代表恶意攻击者,S 代表 Web 服务器,基于验证码的一般攻击流程如图 1 所示。以下是对该流程的简要介绍。

1)$A \rightarrow S$,HTTP 请求 r_1。

2)S 验证 r_1 所请求 URL 的保护级别。若需要认证,则 $S \rightarrow A$,返回用户认证表单及验证码,并将验证码信息写入会话。

3)A 仿造表单内容并提取验证码,识别程序识别验证码并将结果和表单一起自动提交给 S。

收稿日期:2009 - 12 - 24;**修回日期**:2010 - 03 - 06。
基金项目:国家科技支撑计划资助项目(2007BAH08B01);陕西省自然科学基金资助项目(2005f36)。
作者简介:李欢(1984 -),女,陕西西安人,硕士研究生,主要研究方向:网络信息安全;　高岭(1964 -),男,陕西绥德人,教授,博士生导师,博士,主要研究方向:计算机网络;　刘琳(1981 -),女,陕西西安人,助教,硕士,主要研究方向:远程教育、射频识别;　邢斌(1984 -),男,陕西咸阳人,硕士研究生,主要研究方向:网络信息安全。

4)S 验证表单内容合理性或比较其合法性,并将验证码识别结果与会话信息进行比较,有以下两种可能:

① 表单信息合法且验证码与会话信息一致,响应 r_1 同时返回所请求的 URL;

② 若验证信息不一致,提示错误,会话信息失效,生成新的验证码并请求重新验证。

在整个攻击过程中,A 不断发送请求,自动提交表单。如果验证码识别正确,验证成功,那么就可以短时间内注册大量用户并同时登录,发送大量垃圾邮件或回复成千上万的垃圾帖子来降低服务器速度或使服务器崩溃,或通过反复登录破解用户或管理员密码,泄露用户隐私信息或查看修改服务器相关内容,造成了巨大的安全隐患[3]。但是如果验证码难以识别,验证失败,那么上述攻击就很难顺利进行,Web 安全得到了有效的保障。由此可见,验证码设计和应用的重要性。从验证码识别过程和常用验证码的特征分析两方面,论证了常用验证码的易识别性和脆弱性及由此带来的安全隐患。

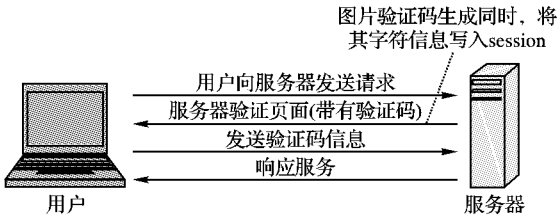


图 1 基于验证码的恶意攻击流程

2 一般图形验证码的识别及特征分析

目前大多数 Web 应用所用的图形验证码都是在一张背景颜色固定的图片上,写入包含若干个字符的字符串,然后加入随机点和随机线段等干扰噪声而形成的。随着图像处理技术的进一步发展,图像灰度化、二值化以及去噪声等图像预处理技术的进一步成熟,这种传统的图形验证码的缺点和脆弱性也越来越明显,劫持会话的程序和软件使 Web 安全问题进一步严重。为了证实这个问题,本文在研究验证码识别技术的基础上,设计实现了一个基于 PHP 图形验证码的识别程序。具体介绍如下。

2.1 识别流程

图形验证码的识别一般涉及图像预处理、图像分割、特征提取、识别等技术。其中,图像预处理工作一般包含图像灰度化、图像二值化和图像去噪 3 大过程。基于 PHP 的验证码识别程序由灰度化、二值化、去噪、分割、特征提取、识别 6 大部分构成,其识别流程如图 2 所示。

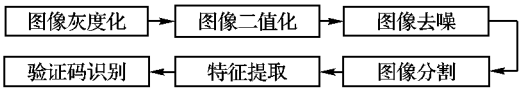


图 2 识别流程

一般情况下,从网上获取的验证码图片都是 RGB 彩色图片,RGB 彩色模型是由 3 个图像分量组成的,每个分量图像都用一定的值(0~255)代表一定的原色图片,当被送入 RGB 监视器的时候,这 3 幅图像会合成一张彩色图片。由于彩色图片中的数据信息非常大,如果直接对彩色图片进行处理,需要很大的运算量,因此识别要做的第一步工作就是要对这些 RGB 图片进行灰度化处理。利用 RGB 图像 3 种颜色的特点,该验证程序中使用加权平均值法,采用下列公式,对图片进行灰度化处理。具体公式[4]如下:

$$V = 0.299R + 0.587G + 0.114B \quad (1)$$

其中:R、G、B 为彩色图像的 3 色分量。灰度效果如图 3 所示。

通过阈值将灰度化图像处理成二值图像的过程,称为数字图像的二值化[5]。灰度图像的二值化处理不但可以有效地减少数据存储容量,而且很大程度上减轻了后续处理的复杂性。本程序中采用整体阈值二值化,它是由像素点(i,j)的灰度值$f(i,j)$确定阈值的方法[6],其中阈值 T 表示为 $T = T(F(i,j))$。该验证程序中使用灰度级直方图确定整体阈值[7],经过二值化处理后的图像如图 4 所示。

(a) 灰度化前 (b) 灰度化后

图 3 灰度效果图

图 4 二值化后的图像

随机干扰噪声可以有效阻止光学字符识别(Optical Character Recognition, OCR)技术对验证码的识别,增加 B/S 架构安全性,但是这些干扰噪声恶化了图像的质量,使字符变得模糊不清,部分字符的特征被掩盖,字符分割和字符特征的提取难度增大,难以识别。因此,去除干扰噪声即将二值化处理后的图片恢复成原始图片,是识别验证过程中一个必不可少的重要环节。该验证程序中采用连通去除噪声[8]的方法去除噪声。将单个字符从背景中提取出来,以供下一步字符识别,称为字符的分割,该验证程序中采用垂直投影字符分割算法[9]。至此,对图形的预处理工作全部结束,然后提取字符特征并与特征库比较,最终实现对该字符的识别。

在识别过程中,灰度化、二值化和去噪均采用常见识别程序中所采用的方法,作者获取了几个知名网站上所采用的验证码,并使用此识别程序进行识别,均得到了较高的识别率(具体数据详见后文)。由此可见,常见图形验证码在设计实现还存在着较大的漏洞,给机器程序留下了极大的攻击机会,也给应用系统带来了极大的威胁。图形验证码识别的界面如图 5 所示。

图 5 程序运行界面图

2.2 常见图形验证码示例及特征分析

验证码的发展经历了很大的变化。一方面,形式从最初的文字验证码到现在普遍使用的图形验证码;另一方面,设计从简单到复杂等。常见的几种图形验证码如图 6 所示。

首先,这几种验证码中的验证字符数目、字符初始位置和字符间的间隔距离都是固定的,字符排列成行,验证码字体单一。虽然对于不同类型的图形验证码,其字符串长度、字符初始位置、字体和字符间间隔距离有所不同,但是对于同类图形验证码,上述特征都是相同的,是在验证码生成程序中固定好的。恶意程序很容易通过多次反复在线攻击,提取类型特征,建立识别特征库,从而不断提高其识别率。其次,验证码的背景固定,颜色变化少,干扰噪声比较简单,干扰效果弱。这简

化了图形预处理工作,降低了验证码的识别难度。最后,验证字符串的输入顺序固定,这是这些图形验证码的共同缺陷。

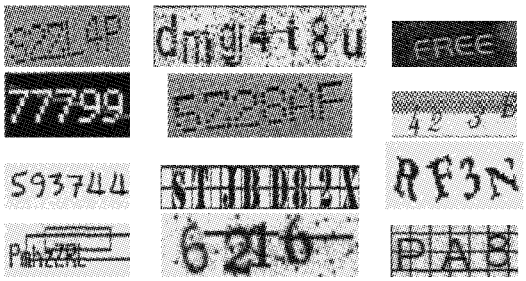


图 6　常见的图形验证码

识别程序的实验结果和对常见图形验证码的特征分析充分说明了图形验证码的严重弊端。基于此,本文提出了一种基于字符数目和随机位置、随机顺序相结合的新的图形验证码生成算法。新的图形验证码主要是利用了字符个数和位置不确定性,增加了图形分割和特征提取的难度,提高了攻击程序识别难度,打破了一般图形验证码字符从左到右的输入顺序,以随机字符出现在随机位置上的顺序作为输入顺序,从而可以有效防止恶意程序的自动识别,提高 Web 应用的安全性。

3　图形验证码生成的改进算法及分析

3.1　改进算法

1) 创建一张宽为 W、高为 H 的真彩图片,在 *Blackgroundcolor* 二元组中随机选择图片背景色,*Blackgroundcolor* = (*num*, *color*),其中 *num* 代表已加载的颜色编号集合,*color* 代表具体的颜色集合。

2) 在 2 ~ 6 随机选择整数 i(i 代表验证码中字符个数)。

3) 将宽为 W、高为 H 的图片分为 $M \times N$ 方格区域,其中:$M = H/P_M$,$N = W/P_N$(P_M 和 P_N 分别代表一行或者一列所占的像素数)。在 0 ~ $M-1$ 和 0 ~ $N-1$ 分别选择 p_1、q_1,$P_1(p_1, q_1)$ 表示字符的写入位置,然后将随机产生的字符写入到第 p_1 行,第 q_1 列。$P_1(p_1, q_1)$ 为初始字符区域,将此区域用特殊颜色标识。

4) 在 0 ~ $M-1$ 和 0 ~ $N-1$ 分别选择 p_2、q_2,并判断 $P_1(p_1, q_1)$ 和 $P_2(p_2, q_2)$ 位置是否重复,若不重复则作为第 2 个字符的写入位置,否则重新选择。同样方法依次产生 (p_2, q_2),(p_3, q_3),…,(p_i, q_i)。

5) 字符连接位置的确定。$C_1(x_1, y_1)$,$C_2(x_2, y_2)$,$C_3(x_3, y_3)$,…,$C_i(x_i, y_i)$ 分别为字符连接位置坐标,其中 $x_i = (q_i - 1/2) \times P_N$,$y_i = p_i \times P_M$。

6) 字符顺序指示位置的确定。$D_{12L}(x_f, y_f)$,$D_{12R}(x_r, y_r)$,$D_{23L}(x_f, y_f)$,$D_{23R}(x_r, y_r)$,…,$C_{(i-1)iL}(x_f, y_f)$,…,$C_{(i-1)iR}(x_r, y_r)$ 分别为字符顺序指示位置的坐标,其数学运算公式为:

$C_{(i-1)iL}(x_i + r\cos(a), y_f + r\cos(a))$

$C_{(i-1)iR}(x_i + r\cos(2a), y_f + r\cos(2a))$

最后依次连接 C_1C_2、C_2D_{12L}、C_2D_{12R}、C_2C_3、C_3D_{23L}、C_3D_{23R}、…、$C_{i-1}C_i$、$C_iD_{(i-1)iL}$、$C_iD_{(i-1)iR}$。

主要的实现代码如下:

```
$pointnum = rand(2,6);
$img =@ imagecreatetruecolor(width, height) or die("创建图片失败");
$background_color = array();
$font_color = array();
$font_style = array();
$first_font_color = imagecolorallocate($img, 255, 0,0);
imagefill($img, 0, 0, $background_color);
$pointcoordinate = array();
//生成主要过程
while ($pointnum >0) {
  $m = rand(2, height/Pm);
  $n = rand(2, width/Pn);
  $a = ($m -1) * Pm;
  $b =($n -1) * Pn
 //重复位置判断
 repateposition();
 //随机字符生成
 $imgcode = get_imgcode();
 //随机字符写入随机位置及初始字符标注
 if ($sign == false) {
   imagestring($img,5, $a, $b, $imgcode, $font_color);
    $sign = true;
 }
 else {
  //初始位置标识
  imagestring($img,5, $a, $b, $imgcode, $text_color);}
   $numcode --;
 }
 //顺序标注
 charorder($img, $pointcoordinate[$j][0], $pointcoordinate[$j]
   [1], $pointcoordinate[$j +1][0], $pointcoordinate[$j +1][1],
    $allowlength, $aloowwidth, $text_color)
  //干扰噪声加入
  for ($i =0; $i <50; $i ++) {
    disturbpixel($img, rand() %140, rand() %78, $color);
  }
```

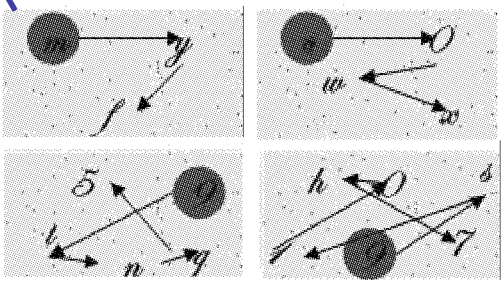


图 7　基于随机顺序的图形验证码示例

3.2　识别测试结果

为了证明基于随机顺序的图形验证码的健壮性、可靠性和安全性,本文在新的图形验证码实现的基础上,使用基于 PHP 的验证码识别程序,对基于随机顺序的图形验证码和几类取自国内一些知名站点上的图形验证码进行识别,识别结果如表 1 所示。

3.3　算法分析

实验数据表明,基于随机顺序的图形验证码在可实现性、实用性和安全性方便都具有明显的优势。在基于随机顺序的图形验证码的生成算法中,验证字符的数目,验证字符串的生成过程、字符的写入位置和字符输入顺序都与一般的图形验证码生成算法不同。也正是这些不同使得基于随机顺序的图形验证码具有更大的识别难度,有效阻止了恶意攻击,保障了 Web 应用的安全。

首先,基于随机顺序的图形验证码中,验证字符显示位置是随机的。验证字符的位置是在单个随机字符生成后,由随机函数在特定范围内随机产生。验证字符串既无固定的开始位置,也无固定的字符间隔,提升了机器识别难度。

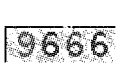
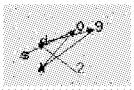
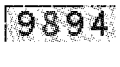
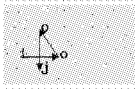
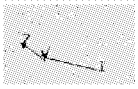
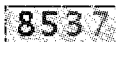
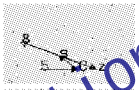
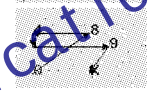
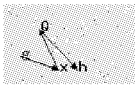
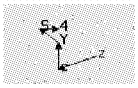
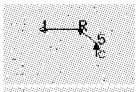
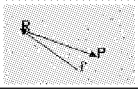
其次,基于随机顺序的图形验证码中,验证字符串输入顺序也是随机的。字符输入顺序只有在其生成之后才能确定,

输入顺序需要合法用户通过观察和简单推理才能得到，而 OCR 或者其他的识别程序很难完成推理，极大地提高了 Web 应用系统的安全性。

第三，基于随机顺序的图形验证码中，验证字符数目也是随机的。常规图形验证码中验证字符为 4 个或 6 个或者更多的固定长度。一般的机器程序在程序设计时，对于待识别的验证码，其字符数目已经确定，字符数目不确定性有效减小了机器程序识别验证码的可能性。

最后，通过程序实现和相应的实验数据证明，新的图形验证码不但可以有效阻止黑客的非法攻击，而且在图形验证码生成过程中，间隔时间短且对内存消耗小，表现出了极强的可用性和广泛的高效性。

表 1 常用图形验证码与基于随机顺序图形验证码识别结果对比

编号	类型一 http://www.■■.■■■■■■■■■.com/		类型二 http://www.■■■■■■■■.com/member/register.aspx		类型三 http://www.■■■■■.com/users/reg.asp		类型四 基于随机顺序的图形验证码	
1		9666		9828		17		8d4o92
2		9894		4751		307		1ojo
3		3360		6984		861		7yi
…	…	…	…	…	…	…	…	…
20		8537		3241		402		858Gz
21		0904		2329		376		4108k9
…	…	…	…	…	…	…	…	…
50		1955		3537		900		goxh
51		1939		1747		053		S4ez
…	…	…	…	…	…	…	…	…
98		5118		2735		937		1B8
99		2914		1675		867		xGcj
100		4718		4591		682		Bfp
统计结果	正确：71 错误：29 识别率：71%		正确：78 错误：22 识别率：78%		正确：57 错误：43 识别率：57%		正确：2 错误：98 识别率：2%	

4 结语

基于随机顺序的图形验证码生成算法的字符数目、显示位置、输入顺序均为随机产生，特别是字符输入顺序由随机位置决定。与常规的验证字符从左到右的输入顺序相比，具有安全、可靠的优点，提高了 Web 应用的安全性。但这种新的图形验证码可读性相对较低，一定程度上增加了合法用户的阅读难度，同时缺少了对图片预处理的难度设计，这将是我们下一步工作需要思考和进一步改进的问题。

参考文献：

[1] 胡金蓉，王玲. 基于字符特征的随机问题阅读式验证码技术[J]. 计算机工程与设计，2008，29(7)：1619－1621.

[2] 陈占芳，冯欣，张伟. 随机中文汉字验证码的生成及应用[J]. 电脑知识与技术，2007(16)：1096－1097.

[3] 吉治钢. 基于验证码破解的 HTTP 攻击原理与防范[J]. 计算机工程，2006，32(20)：170－172.

[4] 李颖. Web 验证码的识别与反识别[D]. 南京：南京理工大学，2008.

[5] TRIER O D，TAXT T. Evaluation of binarization methods for document image [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence，1995，17(3)：312－315.

[6] SAHOO P K，SOHANI S，WONG A K C，*et al*. A survey of threasholding techniques [J]. Computer Vision，Graphics and Image Processing，198，41(2)：233－260.

[7] 陈福忠. 面向 Web 代理的验证码图片识别[D]. 南京：南京理工大学，2007.

[8] 陈柏生. 一种二值图像连通区域标记的新方法[J]. 计算机工程与应用，2006，42(25)：46－47.

[9] 刘明军，谢宏霖，孙雪松，等. 车牌字符分割算法的比较研究[J]. 济南大学学报：自然科学版，2006，20(3)：245－248.