基于改进 MD5 算法的数据篡改检测方法

林 晶,黄青松,张 晶
LIN Jing,HUANG Qing-song,ZHANG Jing

昆明理工大学 信息工程与自动化学院,昆明 650051
College of Information Engineering and Automation,Kunming University of Science and Technology,Kunming 650051,China
E-mail:LinjingL99@126.com

LIN Jing,HUANG Qing-song,ZHANG Jing.Method of data tamper detection by using improved MD5 algorithm.Computer Engineering and Applications,2008,44(33):148-150.

Abstract: This paper presents a method that can discover effectively the modification of sensitive data in database by using one-way hash algorithm that cannot be deduced inversely,because it is difficult to detect the illegal revision of sensitive data.The dual inspection approach is adopted through checking client and server for better detection effect.The MD5 algorithm,a one-way hash algorithm,is improved availably by adding a hidden 'antibody' factor to the algorithm for withstanding cribber's exhaustion search.The paper mainly describes the principle of this method based on the improved MD5 algorithm and how to realize it in the application system.Applied results demonstrate that this method has highly sensitive to intentional tampering and outstanding detection performance for data tamper detection.

Key words: hash algorithm;MD5;sensitive data;tamper detection

摘 要:针对数据库中敏感数据被非法篡改后难以检测发现的问题,提出了一种基于单向散列算法的不可逆性对敏感数据篡改的有效检测方法,并采用检测客户端与服务器的双重检测机制来加强检测效果。单向散列算法选用 MD5 算法,并通过向算法中注入隐蔽的"抗体"因子以抵抗篡改者的穷举搜索,使 MD5 算法得到有效的改进。重点阐述了上述检测方法的原理及其在应用系统中的实现。应用结果表明,该检测方法对非法篡改具有高度敏感性和优秀的数据篡改检测性能。

关键词:散列算法;MD5 算法;敏感数据;篡改检测
DOI:10.3778/j.issn.1002-8331.2008.33.046 **文章编号:**1002-8331(2008)33-0148-03 **文献标识码:**A **中图分类号:**TP31;TP39

1 引言

云南省大型科学仪器、设备协作公用网及服务平台(简称大仪网)在提高本区域内仪器共享,减少资金重复投入,促进区域科技发展中起到至关重要的作用。系统使用单位覆盖全省各地,与众多 Web 应用一样,安全性问题十分突出,重要数据容易被非法篡改且难以检测。而作为数据安全的一个研究领域,篡改检测[1,2,4,6]已是目前的一个热点课题。

当前该领域中多数研究是应用数字指纹(水印)技术对大粒度的图像及文件进行篡改检测。而对数据库中小粒度的敏感数据的篡改,特别是对来自于 DBA 的篡改,还难以检测。

要确保系统的真正安全,除了综合运用用户认证、权限控制等安全技术外, 还必须能对 DBA 绕开 DBMS 的审计功能后对敏感数据的篡改进行有效检测。为此,提出一种基于改进 MD5 算法的数据篡改检测方法:散列(数字指纹)检测法,并给出一个检测模型来解决问题。

2 关于散列检测法模型

为保证大仪网中重要数据(如用户角色、密码权限)的安全,提出一个如图 1 所示的敏感数据散列检测模型。该模型分为三层:采集层、安全层和数据层。采集层完成数据采集和维护功能。数据层保存数据,响应用户的数据请求。安全层由散列工厂、安检站和安检服务器组成。散列工厂生成初始散列值;安检站及安检服务器对敏感数据进行严格篡改检测。

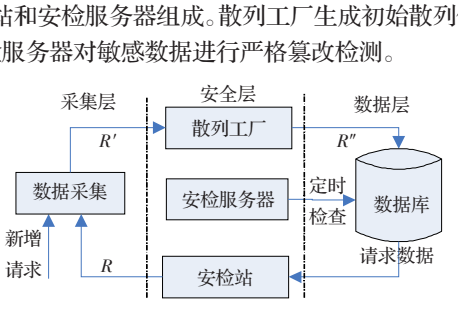

图 1 敏感数据散列检测模型

2.1 散列检测法的思想

匹配由散列算法对相同数据先后两次产生的散列值(数字

基金项目:云南省大型科学仪器、设备协作公用网及服务平台(No.2006PT06)。
作者简介:林晶(1970-),男,硕士研究生,研究方向:智能信息系统与信息安全;黄青松(1962-),男,教授,硕士生导师,主要研究领域:智能信息系统、人工智能;张晶(1974-),男,在职博士,副教授,研究方向:Web 技术、软件工程、实时控制软件。
收稿日期:2007-12-13 **修回日期:**2008-02-25

指纹)来检测数据是否被篡改。在散列工厂生成合法的初始散列值后(颁证)，安检站与安检服务器比较对当前数据计算的散列值和初始散列值(验证)，发现差异并报警。此外初始散列值也可被合法更新(换证)。

2.2　敏感数据记录格式

大仪网中的敏感(重要)数据的记录 R 格式设计如图 2 所示。其中敏感数据 D 由各敏感信息字段组成；散列值字段 H 用来存放敏感数据记录的数字指纹。

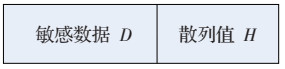

图 2　敏感数据记录格式

3　散列检测模型安全层的实现

由于 MD5 算法的单向性[3,5]，极难找到两个随机信息使其产生相同摘要(数字指纹)，因此它非常适合快速安全的产生摘要。

3.1　MD5 算法的改进

大仪网中，由于对这些重要用户数据并未采用加密措施，以明文形式保存在数据库中，对能够直接接触数据库中原数据的人来说，很容易根据散列字段值的长度猜知所使用的散列算法，从而对这些数据的安全造成威胁。因此作者从数据记录的组合着手，植入隐蔽“抗体”因子对 MD5 算法进行改进以增强算法安全强度。一般散列值的生成按公式 $H=f(m)$ 进行。其中 H 是所产生的散列值，f 是所选用的散列算法，m 是输入的消息流。如果 m 是一固定格式的数据记录，篡改者还是可以在短时间内以穷举的方法寻找到这一记录格式，从而达到任意篡改的目的。现在对 m 施以某种变换 $m'=g(m)$ 进行改进使 $H=f(g(m))$。改进的变换规则如下：

(1)在原记录信息中植入隐蔽“抗体”因子以抵抗篡改者的穷举搜索；

(2)按记录字段和“抗体”因子的某种次序来排列组合信息 m；

(3)将变换函数 $g(m)$ 的输出 m' 作为 MD5 函数的输入来产生散列值 H。

若取“抗体”因子为 $2k$ 字节长度的 ASCII(8 bit)集字符串($k\geqslant 2$ 为整数)，敏感数据的字段数为 n。则根据排列组合原理，现在要匹配敏感数据记录 $g(m)$ 成功所需的次数为 $C=(n+1)!*(256^{2k})$。按最简单的情况，取 $n=4$，$k=2$。计算得 $C=103\ 079\ 215\ 104$ 次。以一年 365*24*3 600 s 计。假设按这种组合规则完成一次匹配耗时为 1 ms，则匹配成功耗时将超过 16 年。如果再增加“抗体”因子长度和敏感数据的字段数，以及扩大“抗体”因子的取值范围，使其可取 Unicode(16 bit)字符集字符，那么破解 $g(m)$ 算法寻找“抗体”因子的耗时将是天文数字，代价极大，说明改进算法是很安全的。改进的 MD5 算法接口：

```
public interface BetterMD5 {
    public static byte[] recToMsg(String record);//实现改进的变换 m'=g(m)
    public static byte[] MD5Array(byte[] input);//产生二进制摘要
public static string betterMD5(byte[] message);//调用 MD5Array 产生字符摘要
};
```

3.2　散列工厂的实现

当敏感数据记录 $R'(=H+D)$ 输入后，先分离出数据 D 产生数字指纹 H'，再与数据 D 进行组装，最后输出带数字指纹的记录 $R''(=H'+D)$。散列工厂的核心算法采用改进的 MD5 算法。散列工厂的实现如下：

```
public class MD5Shop {
  public String* md5Shop(String* record) {
      String data=StringUtil.getData(record);//分离数据
      BetterMD5 bmd5=new BetterMD5();
      String hash1=bmd5.betterMD5(bmd5.recToMsg(data));
      Return StringUtil.setHash(record,hash1);//设置散列值}
};
```

3.3　安检站工作机制

安检站对从数据库请求的敏感数据进行严格的安全检查。当带有有效数字指纹的敏感数据记录 $R'=D+H$ 输入后，先分离出敏感数据 D 输送到散列工厂，制作出数字指纹副本 H'，再同原记录的数字指纹进行匹配。匹配成功就响应请求，检测到非法篡改时，除了向系统报警外，还将检测到的被篡改的数据记录写入日志文件，为事后人工或自动检查提供依据，并拒绝请求。因此无论是篡改敏感数据，还是制造伪证(修改数字指纹)，都不可能逃脱检测，确保系统在安全状态下工作。安检站设计成可复用类 TamperCheck，通过类中的 tamperCheck(String*)方法实现图 3 所示的工作流程。

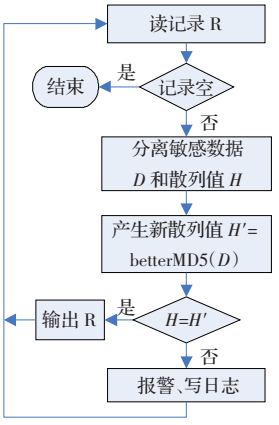

图 3　安检站工作流程图

3.4　安检服务器

安检站工作在客户端，是一种被动式的工作方式，系统只有在对敏感数据有请求时才能发现对它的篡改。利用可复用的安检站构件，扩充设计成安检服务器，定时唤醒安检程序对敏感数据例行检查，及时主动地发现非法篡改行为。从而形成安检服务器和安检站客户端的双重检测机制，增强了应用系统检测非法篡改行为的能力。安检服务器的工作过程描述如下：

```
main() {
/* 定时唤醒安检站对敏感数据例行检测；interval：唤醒时间间隔 */
    Boolean start=false;//唤醒安检站工作的标志
   setInterval(integer interval);//设定唤醒时间间隔
   timer(1);//时间到 start=true；1：启动定时器；0：关闭定时器
   if(start){  //调用安检站检查；msg_ptr 为数据表指针
     new TamperCheck().tamperCheck(String* msg_ptr);
     start=false;}
}
```

4　测试实验和结果分析

4.1　黑盒测试实验

模拟应用环境进行以下实验：在 Oracle 环境下以 DBA 的

身份对大仪网的用户信息表 Reguser(含主键、删除标记、散列值、用户 ID、姓名、口令、角色等 14 个字段,2 000 条记录)进行篡改检测;在 2 000 条记录中随机抽取 500 条进行非法修改检测,其中修改统计情况如表 2 所示。任意添加 200 条记录做非法增加检测。随机删除 100 条记录进行了非法删除检测。分别运行安检站和安检服务器程序进行篡改检测,测试统计结果如表 1 和图 4 所示。

表 1 测试统计

方式	实验					
	增加记录/条		修改记录/条		删除记录/条	
	增加	检测	修改	检测	删除	检测
服务器	200	200	500	500	100	100
客户端		200		500		0
成功率	100%		100%		S:100%;C:0%	

表 2 修改统计

实验	字段			
	主键	删除标记	散列值	其它(11 个)
修改	0	50	200	1 000
检测	0	50	200	1 000

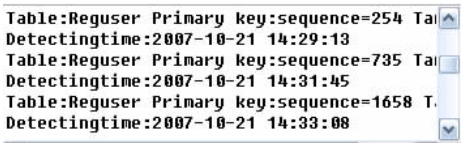

图 4 篡改检测日志

4.2 测试结果分析

(1)对任意增加的 200 条记录,都能被应用程序客户端和检测服务程序检测到被篡改,成功率都为 100%。

(2)随机删除了 100 条记录,应用程序客户端对删除记录的检测命中率为 0,因为客户端是对用户的查询请求结果进行检测,被删除记录不会出现在查询结果中,故无法检测得到。检测服务程序的检测命中率 100%, 原因是在表设计时主键采用连续序列号,应用程序对无效记录只做删除标记,不做物理删除。若允许物理删除记录,那么非法删除就难以检测。

(3)随机抽取 500 条记录进行非法修改,应用程序客户端和检测服务程序的检测成功率都是 100%。对 500 条记录的删除标记、散列字段、11 个其它字段按各种组合分别进行了 50 次、200 次、1000 次非法修改,根据主键特性对主键未做修改。

测试结果表明,数字指纹对数据篡改高度敏感,该散列检测方法有优秀的篡改检测能力。

5 结束语

提出的数据篡改检测方法, 通过在 MD5 算法中植入隐蔽的“抗体”因子对算法加以改进,在数据采集端生成数字指纹(散列),采用安检客户端和服务器的双重检测机制来加强检测能力。测试结果表明它能够及时有效地检测到对数据库中敏感数据的任何非法篡改,具有优秀的篡改检测性能,并结合大仪网系统给出了具体实现, 并在该系统的应用中取得良好的效果。它可以作为数据安全应用技术的一种补充。但是数据的篡改定位,目前还没有合适的方法;实时性检测方面未做探讨。这些都有待做进一步的研究。

参考文献:

[1] 唐承亮,肖海青,向华政.基于文字 RGB 颜色变化的脆弱型文本数字水印技术[J].计算机工程与应用,2005,41(36):6-8.

[2] WANG Yan-hui,WANG Xiang-hai.Overview on digital watermarking for image authentication[J].Computer Engineering and Applications,2007,43(2):33-37.

[3] Rivest R.Network Working Group RFC 1321,MD5 message-digest algorithm[S].1992-04:1-21.

[4] Mead S.Unique file identification in the national software reference library[J].Digital Investigation,2006,3(3):138-150.

[5] Cid C.Recent developments in cryptographic hash function: Security implications and future directions[R].Information Security Technical Report,2006,11(2):100-107.

[6] Snodgrass R T,Yao S S,Collberg C.Tamper detection in audit logs[C]// Proceedings 2004 VLDB Conference,2004:504-515.

(上接 75 页)

最大地满足随机性、试验测试用例空间的大小以及先验数据的规模都将在一定程度上影响算法的有效性,这也将是下一步研究的方向。

参考文献:

[1] The Standish Group.The Standish Group Report:CHAOS[R].1995.

[2] 程火旺,王戟,董威.高可信软件工程技术[J].电子学报,2003,12A:2-5.

[3] Fenton N E,Lawrence S.软件度量:严格而实用的方法:影印版[M].2 版.北京:清华大学出版社,2003:364-401.

[4] 杨仕平,熊光则.高可信软件的防危性评估研究[J].计算机工程与设计,2004,25(2):2-3.

[5] 覃志东,雷航.安全关键软件可靠性验证测试方法研究[J].航空学报,2005,26(3):2-4.

[6] Futrell R T,Shafer D F.高质量软件项目管理[M].袁科萍,译.北京:清华大学出版社,2006:200-360.

[7] Littlewood B,Strigini L.Assessment of ultra-high dependability for software-based systems[J].Communications of the ACM,1993,36(11):70-80.

[8] 郑人杰,殷人昆,陶永雷.实用软件工程[M].2 版.北京:清华大学出版社,1997:70-200.