文章编号:1001 - 9081(2008)08 - 2144 - 03

面向大型数据库的审计数据采集方法

陈 伟[1,2],Qiu Robin[3]
(1. 南京审计学院 信息管理系,南京 210029; 2. 江苏省审计信息工程重点建设实验室,南京 210029;
3. 宾夕法尼亚州立大学 信息科学系,美国宾夕法尼亚州 莫尔文 19355)
(chenweich@nau.edu.cn)

摘 要:计算机辅助审计是目前审计领域研究的一个热点,审计数据采集是面向数据的计算机辅助审计的关键步骤。分析了常用的审计数据采集方法,比较了各自的优缺点。在此基础上,针对我国实施计算机辅助审计的现状以及面向大型数据库的审计数据采集的特点,分析了适合大型数据库的审计数据采集方法,并以 Oracle 数据库为例,分析了该方法的应用。

关键词:计算机辅助审计;大型数据库;审计数据采集
中图分类号: TP399;F239.1 **文献标志码**:A

Study on audit data acquisition methods oriented to large databases

CHEN Wei[1,2], QIU Robin[3]
(1. *Department of Information Management, Nanjing Audit University, Nanjing Jiangshu* 210029, *China*;
2. *Jiangsu Key Constructing Laboratory of Audit Information Engineering, Nanjing Jiangshu* 210029, *China*;
3. *Department of Information Science, Pennsylvania State University, Malvern, PA* 19355, *USA*)

Abstract: Computer Assisted Audit (CAA) is an active research domain in audit field. Audit data acquisition is a key step of data-oriented computer assisted audit. Common audit data acquisition methods were analyzed, and the benefits and drawbacks of these methods were analyzed. Then, according to the conditions of computer assisted audit used in China and the characteristics of audit data acquisition methods oriented to large databases, audit data acquisition method that fitted large databases was presented. Besides, taking Oracle database as an example, how to use this method was analyzed.

Key words: Computer Assisted Audit (CAA); large database; audit data acquisition

0 引言

随着信息技术的发展,审计对象的信息化使得计算机辅助审计[1-2]成为必然。总的来说,目前我国研究的计算机辅助审计可以看成是一种面向数据的计算机辅助审计[3-4],其步骤如下:1)采集被审计对象信息系统中的数据,即审计数据采集;2)根据对这些数据的分析和理解将其转换为满足审计数据分析需要的数据格式,即审计数据预处理;3)运用相关软件对采集到的电子数据进行分析处理,从而发现审计线索,获得审计证据,即审计数据分析。由此可见,如何把被审计单位的电子数据采集过来,是开展面向数据的计算机辅助审计的关键步骤。

对于一般的中小型数据库系统,审计人员可以胜任数据采集工作。然而,当被审计单位采用的是大型数据库系统,如 Oracle 数据库系统,而审计人员一般对 Oracle 数据库不能熟练地使用,因此,如何采集大型数据库系统中的数据具有重要的实用价值。本文结合国内开展计算机辅助审计的现状及特点,分析面向大型数据库的审计数据采集方法,从而为我国实施面向数据的计算机辅助审计提供理论和实践上的指导。

1 审计数据采集的原理

简单地讲,审计数据采集就是审计人员为了完成审计任务,在进行计算机辅助审计时,按照审计需求从被审计单位的信息系统或其他来源中获得相关电子数据的过程。其原理如图 1 所示。

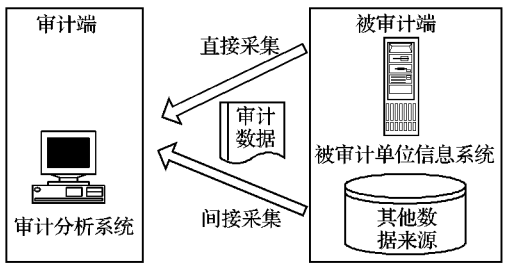

图 1 审计数据采集的原理

数据采集的对象一般是被审计单位信息系统的数据库,或数据库中的备份数据,审计人员也可以从其他来源获得被审计单位的审计数据,例如从会计核算中心、税务等部门获得审计数据。

2 常用审计数据采集方法

在数据采集过程中,审计人员常用的数据采集方法主要有以下 4 种。

1)直接复制。当被审计信息系统中的数据库系统与审计软件使用的数据库系统相同,或者虽不相同,但审计软件的数据库引擎可以直接访问被审计信息系统的数据库时,只需直接将被审计对象的数据采集到审计人员的计算机中即可,

收稿日期:2008 - 02 - 26;**修回日期**:2008 - 04 - 10。
基金项目:国家自然科学基金资助项目(70701018);中国博士后科学基金资助项目(20060390281)。
作者简介:陈伟(1976 -),男,山东菏泽人,副教授,博士,CCF 会员,主要研究方向:审计信息化; Qiu Robin(1964 -),男,美籍华人,教授,博士,主要研究方向:信息系统。

即直接复制的方式。

2)通过中间文件采集。通过中间文件采集是指被审计单位按照审计要求,将原本不符合审计软件要求的数据转换成与审计软件要求相一致的格式提供给审计人员。对于一些比较敏感的系统,审计人员可能不便于直接接触其系统和相关资料。可以在审计人员的监督下,由被审计单位技术人员将其数据转换为标准格式数据或审计人员指定格式的数据,交给审计人员。

3)通过 ODBC 接口采集。通过 ODBC 接口采集数据是指审计人员通过 ODBC 数据访问接口直接访问被审计信息系统的数据,并把数据转换成审计所需的格式。

4)通过专用模板采集。一些审计软件针对不同的被审计系统设计了相应的"专用采集模板",审计人员在进行数据采集时,通过选择相应的模板,可自动实现数据的采集,这种方式称为通过专用模板采集[6]。这种方式的优点是使用简单,自动化程度高,对审计人员的技术水平要求不高;缺点是审计软件必须为每一被审计对象(包括该软件的不同版本)设计一个专用模板,由于目前被审计单位所使用的软件各种各样,很难为每一软件以及相应的各种版本设计相应的模板,这使得模板采集法的成本相对较高。审计人员在实际的工作中,应根据被审计单位的实际情况不,有模板时用模板,没有模板时再用其他方法。

除了以上常用的 4 种数据采集方法之外,目前理论界正在研究一些通用的数据采集方法,例如,基于 XBRL 的通用审计数据采集接口。但由于目前被审计单位还不具备相应的条件,所以这些方法现在还不能应用到实践中去。

3　面向大型数据库的审计数据采集方法

常用 4 种数据采集方法的优缺点如表 1 所示。

一般来说,面向大型数据库的审计数据采集具有以下特点:

1)数据库中数据量大,不可能采集数据库中的所有数据。因此,一般不能使用直接复制的数据采集方法。

2)数据库版本多,审计软件中不可能包含针对所有版本的专用采集模板。因此,一般不能使用通过专用模板采集的数据采集方法。

3)数据库对操作人员的技术水平要求高。对于一些重要行业,如银行、税务等部门,其信息系统均采用大型数据库系统,如 Oracle 数据库。甚至一些基层部门,有时也采用大型数据库系统。然而,由于大型数据库系统的复杂性,操作人员仅能完成一般的数据备份等简单的操作,其他工作一般多依赖于软件公司提供技术支持,因此,审计人员在进行数据采集时,不能完全寄希望于被审计单位。因此,一般不能使用通过中间文件采集的数据采集方法。

根据以上分析,目前在实际的审计工作中,为了完成面向大型数据库的审计数据采集,对审计人员来说最通用的数据采集方法是通过 ODBC 接口来完成。

表 1　常用 4 种数据采集方法的优缺点分析

数据采集方法		影响使用的因素				
		动态还是静态	对被审计系统的影响	专业知识需求	对被审计单位的依赖性	灵活程度
直接复制		静态	影响小	不需要	不依赖	一般
通过中间文件采集		静态	影响小	不需要	依赖	一般
通过 ODBC 接口采集	从被审计单位信息系统中采集	动态	影响大	要	不依赖	高
	从备份数据中采集	静态	影响小			
通过专用模板采集	从被审计单位信息系统中采集	动态	影响大	需要	不依赖	低
	从备份数据中采集	静态	影响小			

4　实例

本节以如何采用 Access 采集 Oracle 数据库中的数据为例,分析"通过 ODBC 接口"采集大型数据库中的数据的方法。

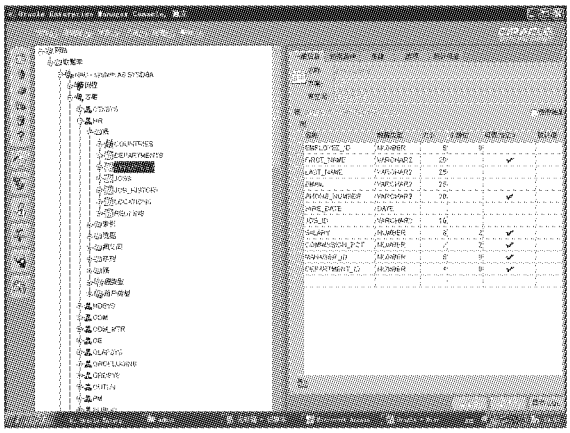

图 2　某 Oracle 数据库界面

假设图 2 为 Oracle 数据库的界面,现把其中的部分数据采集到 Access 中去,该方法的操作过程为:

1)打开 Access,点击菜单"文件 → 获取外部数据 → 导入",出现如图 3 所示的界面;

图 3　导入数据源类型选择界面

2)在图 3 所示的"导入"界面中,文件类型选择"ODBC Databases",弹出"选择数据源"的界面,在该界面中选择"机器数据源"选项卡,如图 4 所示;

3)在图 4 所示的"选择数据源"界面中,如果没有所需要的数据源,则需要创建新的数据源,如图 5 所示;

4)在图 5 所示的"创建新数据源"界面中,驱动程序选择

“Oracle in OraHome92”选项,然后,进入新数据源配置界面,如图6所示。

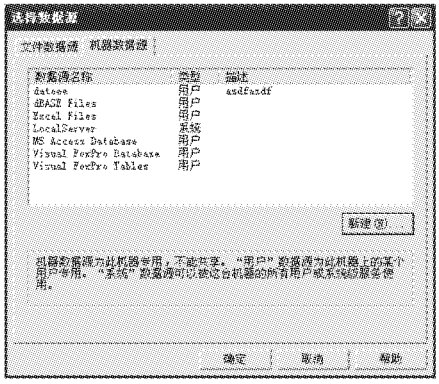

图4 “机器数据源”选项卡

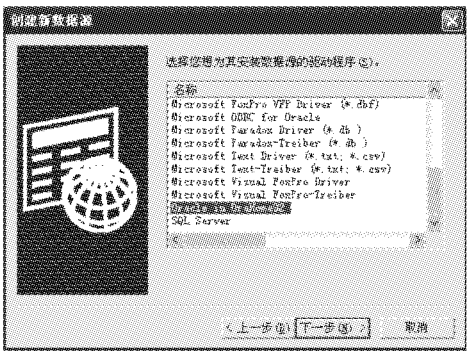

图5 驱动程序选择界面

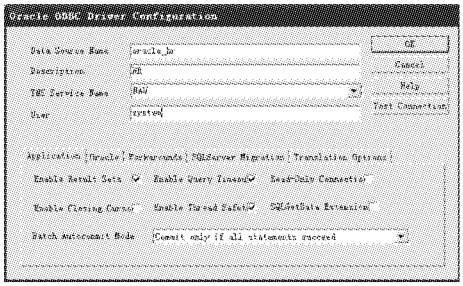

图6 Oracle新数据源创建的配置界面

5)在图6中,输入要配置的数据源参数。其中,Data Source Name为所定义的数据源名称,TNS Service Name为创建的数据源所要连接的数据库实例名称,User为登录用户名称,点击“Test Connection”按钮,测试配置数据源的连接情况,如图7所示;

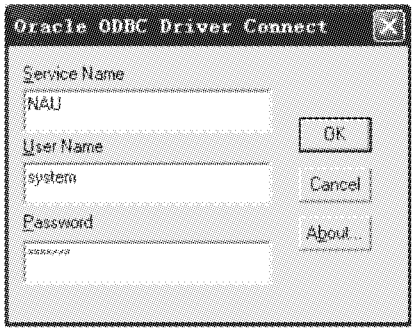

图7 配置数据源连接情况的测试

6)在图7中,输入各参数测试配置数据源的连接情况。其中,Service Name为创建的数据源所要连接的数据库实例名称,User Name为登录用户名称,Password为登录用户密码;若测试成功,则弹出如图8所示的界面,从图8中可以看出,所配置的新数据源已显示在界面中;

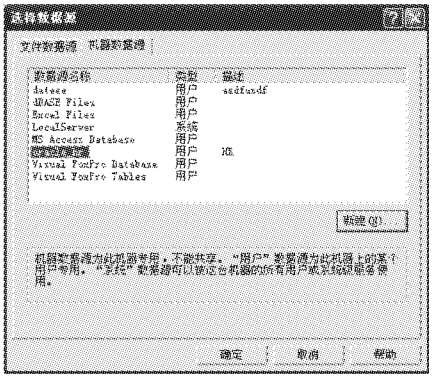

图8 含有新数据源的“机器数据源”选项卡

7)在图8中,选择所配置的数据源,并登录所配置的数据源,进入“导入对象”界面,如图9所示,要连接的数据库实例中所有的表都显示在该界面中;

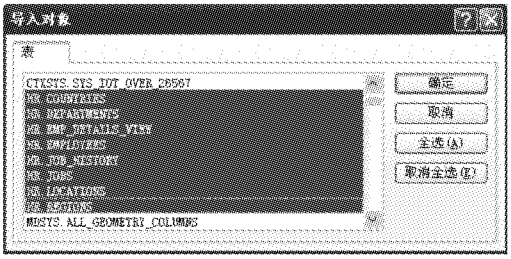

图9 选择需要采集的Oracle数据表

8)在图9中,选择需要采集的数据表,即可完成所需要的数据采集操作。

5 结语

如何采集大型数据库系统中的数据是审计人员在开展面向数据的计算机辅助审计过程中常常遇到的问题。本文针对我国实施计算机辅助审计的特点,分析了适合大型数据库系统的审计数据采集方法,从而为开展计算机辅助审计提供了指导。

联网审计是计算机辅助审计的一个重要发展方向[7-8]。联网审计环境下,对于大型数据库数据的采集一般是通过在被审计单位数据库服务器端放置一台称之为“数据采集前置机”的服务器,通过在“数据采集前置机”上安装数据采集软件(在技术实现上,也可采用基于交互式数据迁移技术的数据采集方法[9]),把审计需要的财政财务数据和相关经济业务数据采集到部署在本地的审计数据采集服务器(前置机)中。然后根据需要,把采集来的数据通过网络传输到审计单位中去,以供审计分析使用。因此,在联网审计系统的开发上,为了降低联网审计的实施成本,如何开发具有一定实用性、通用性和可移植性的联网审计数据采集接口(例如,基于XBRL的通用审计数据采集接口,可重构的数据采集系统等)是今后研究的一个重要方面。

参考文献:

[1] ROBERT L B, HAROLD E D. Computer-assisted audit tools and techniques: analysis and perspectives [J]. Managerial Auditing Journal, 2003, 18 (9): 725 - 731

[2] 陈伟,张金城,QIU R. 计算机辅助审计技术(CAATs)研究综述[J]. 计算机科学, 2007,34(10): 290 - 294.

(下转第2149页)

Algorithm)和 BIA(Block Match Immune Algorithm)两种算法的搜索质量和 SAD 运算的花消进行比较。这里使用了 3 个测试文件,共 6 个部分用于测试,对不同性质的图像进行仿真。这 6 个部分分为 QCIF 格式和 CIF 格式两种,每种格式各 3 各部分。QCIF 格式包括是:第 1 部分,取自标准测试序列 Foreman 的 1 ~150 帧,镜头不动,前景做中等程度的运动;第 2 部分,取自标准测试序列 Foreman 的 151 ~250 帧,背景快速的变化;第 3 部分,取自 Miss America,背景不变,前景做轻微的运动。而 CIF 格式的包括:第 1 部分,取自标准测试序列 Foreman 的 1 ~150 帧,镜头不动,前景做中等程度的运动;第 2 部分,取自标准测试序列 Foreman 的 151 ~250 帧,背景快速的变化;第 3 部分,变速运动的 Stefan。

传统基于块匹配的免疫算法和改进算法都采用 I-P 帧格式,逐一对相邻两帧进行运动估计。仿真的视频序列宏块大小为 16 × 16,搜索距离 d_x 和 d_y 在 QCIF 格式下都设为 7,在 CIF 格式下都为 15。参数设置为:$T = 3$, $Thh\_SAD = 0$, $N = 6$, $r_{CS} = 0.33$, $M = 6$, $d_{\max} = 15$(QCIF)或 $d_{\max} = 15$(CIF), $d_{\min} = 2$(QCIF)或 $d_{\min} = 2$(CIF), $\sigma = 2$, $N_{\max} = 6$。

传统算法的运算量可以用每帧平均 SAD 计算次数来表征,而改进算法则除了每帧平均的 SAD 计算次数以外,还由于预处理算法的增加,相对传统算法会多出一定的加法 ADD 次数。而从 SAD 计算定义和改进算法具体实现过程可以知道,一次 SAD 的运算量是一次加法 ADD 的两倍。定义参数 IP 来表征改进算法相对于传统算法的运算量减少百分比,即:

$$IP = \left(SAD_{\text{old}} - SAD_{\text{new}} - \frac{ADD_{\text{new}}}{2}\right) \Big/ SAD_{\text{old}} \tag{8}$$

其中,SAD_{old} 表示传统算法 SAD 次数,SAD_{new} 和 ADD_{new} 分别表示改进算法的 SAD 和 ADD 次数。从表 1 中可以看出,改进算法总体运算量相对传统算法有所减少,但不同性质的图像数据其效果并不相同。

表 1　每宏块的平均 SAD 和 ADD 计算次数

标准测试序列		传统算法	改进算法		IP/%
		SAD	SAD	ADD	
QCIF	Foreman1	26.5	11.3	27.2	6.04
	Foreman2	26.2	11.5	19.8	18.32
	Miss	26.3	10.6	32.1	6.27
CIF	Foreman1	28.1	12.9	25.1	9.43
	Foreman2	28.7	14.6	21.9	11.50
	Stefa_	27.9	12.8	26.9	5.91

表 2 为改进算法相对于传统算法在不同仿真测试条件下的 PSNR 和估计准确率。由于通过对候选块预处理,去掉了不可能匹配的块,则改进算法准确进行块匹配的可能性提高,这直接体现在 PSNR 的提高和估计准确度上。表 2 中,假设 FS 算法的匹配准确度为 1。

表 2　估计的准确率

测试序列		传统免疫算法		改进算法	
		准确度	PSNR /dB	准确度	PSNR /dB
QCIF	Foreman1	0.9709	32.78	0.9698	32.81
	Foreman2	0.9694	28.46	0.9698	28.49
	Miss	0.9500	41.34	0.9510	41.34
CIF	Foreman1	0.9369	33.43	0.9374	33.50
	Foreman2	0.9120	30.09	0.9210	30.10
	Stefa_	0.9466	35.46	0.9469	35.60

6　结语

在运动估计中的基块匹配的免疫算法,其匹配质量要好于传统经典算法,但是其计算量过大,这将制约其应用。改进算法通过对不可能匹配的候选块的预消除,可以使得搜索质量不降低的条件下,减少了运算量,增加了算法的实用性。

参考文献:

[1] ZHU JUN, ZHU BINGLIAN. A novel fast block-matching motion estimation algorithm based on artificial immune system[C]// IEEE International Conference on Integration Technology. Shenzhen, China: IEEE. 2007:579 -583.

[2] XUAN J, CHAU L P. An efficient three - step search algorithm for block motion estimation[J]. IEEE Transactions on Multimedia, 2004, 6(3):435 -438.

[3] ZHU SHAN, MA KAI KUANG. A new diamond search algorithm for fast block-matching motion estimation[J]. IEEE Transactions on Image Processing, 2000, 9(2):287 -29.

[4] KIM J N, BYUN S C, KIM Y H. Fast full search motion estimation algorithm using early detection of impossible candidate vectors[J]. IEEE Transactions on Signal Processing, 2002, 50(9):2355 -2365.

[5] LI W, SALARI E. Successive elimination algorithm for motion estimation[J]. IEEE Transactions on Image Processing, 1995, 4(1):105 -107.

[6] AHN T G, MOON Y H, KIM J H. Fast full-search motion estimation based on multilevel successive elimination algorithm[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2004, 14(11):1265 -1269.

[7] CHEN W G, LING Y. Noise variance adaptive successive elimination algorithm for block motion estimation: application for video surveillance[J]. IET Signal Processing, 2007, 1(3):150 -155.

[8] 刘芳, 潘晓英. 基于免疫克隆选择的块匹配运动估计[J]. 软件学报, 2007, 18(4):850 -860.

(上接第 2146 页)

[3] 国家 863 计划审计署课题组. 计算机审计数据采集与处理技术研究报告[M]. 北京: 清华大学出版社,2006.

[4] CHEN WEI, WANG HAO, ZHU WEN-MING. Study on data-oriented IT audit used in China[C]// Proceedings of the 11th Joint International Computer Conference (JICC2005). Singapore: World Scientific Publishing, 2005:666 -669.

[5] 中华人民共和国审计法[EB/OL]. [2007 -08 -23]. http://www.gov.cn.

[6] 中华人民共和国审计署《AO》研发项目组. 现场审计实施系统实用手册[M]. 北京:中国时代经济出版社,2005.

[7] CHEN WEI, ZHANG JIN - CHENG, JIANG YU - QUAN. One continuous auditing practice in China:data-oriented online auditing (DOOA)[C]// The 7th IFIP International Conference on E-business, E-services, and E-society (I3E2007). Boston: Springer, 2007: 521 -528.

[8] ALLES M G, KOGAN A, VASARHELYI M A. Feasibility and economics of continuous assurance [J]. Auditing: A Journal of Theory and Practice, 2002, 21 (1): 125 -138.

[9] 陈伟, 王昊, 陈丹萍. 一种基于交互式数据迁移技术的数据采集方法[J]. 计算机工程, 2006,32(9): 62 -63,66.