一种基于数据匹配技术的审计证据获取方法*)

陈 伟[1,2] Robin Qiu[2,3] 刘思峰[2]

(南京审计学院信息管理系 南京 210029)[1] (南京航空航天大学经济与管理学院 南京 210016)[2]

(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)[3]

摘 要 联网审计是目前计算机辅助审计领域中的一个研究前沿问题。本文针对我国目前研究的联网审计的特点,以及常用审计方法的不足,提出了一种基于数据匹配技术的审计证据获取方法。该方法通过对两个数据源中的被审计数据进行数据匹配,查找相似重复实体,发现审计线索,获得审计证据。此外,采用模拟数据从定量的角度对该方法的审计风险进行了分析。最后,以一个实例验证了该方法的有效性及实用价值。研究表明:和常用的审计方法相比,该方法对业务知识依赖少,易发现被审计数据中的隐藏信息,并提高了审计效率。

关键词 联网审计,审计数据分析,数据匹配,审计证据

Audit Evidence Gathering Method Based on Data Matching

CHEN Wei[1,2] Robin Qiu[2,3] LIU Si-feng[2]

(Department of Information Management, Nanjing Audit University, Nanjing 210029, China)[1]

(College of Economics and Management, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China)[2]

(Department of Information Science, Pennsylvania State University, Malvern, PA 19355, USA)[3]

Abstract Online auditing is an active research domain in computer-assisted audit field. According to the characteristics of online auditing researched in China nowadays and the shortages of common audit methods, an audit evidence gathering method based on data matching is proposed in this paper. Through matching data from two auditee' data, this method can detect approximately duplicated data in two auditee' data, then find audit clues, and gather audit evidence. Furthermore, simulation data are used to analyze audit risk of this method from the quantitative view. Finally, the validity and applied value of this method are proved by an instance. Research result shows that comparing with the common audit methods, this method needs less business knowledge, can find information that is unknown and hidden from auditees' data easily, and the audit efficiency is also improved.

Keywords Online auditing, Audit data analysis, Data matching, Audit evidence

1 问题的提出

审计作为一种独立性的经济监督活动,对被审计单位相关经济活动具有其特有的制约和促进作用,历来受到中外国家、政府和社会的重视。信息技术的发展使得以查账为主要手段的审计职业遇到了来自计算机技术的挑战。审计对象的信息化使得计算机辅助审计成为必然。随着信息技术的发展,信息技术在审计中的应用情况也在不断变化。信息技术的发展将使得计算机辅助审计向持续、动态、实时的方向发展。持续审计[1](Continuous auditing, CA)是计算机辅助审计的一个重要发展方向。不同时期、不同技术条件下,持续审计实现的原理是有区别的。文献[2]把持续审计的实现方法分成两类:一类是子系统(sub-systems)或模块(modules)模式,它们必须被嵌入到被审计系统中去,文献[3,4]就对这类模式进行了研究;另一类是单机系统(stand-alone)模式,它能持续地监控被审计系统,从这些系统中抽取数据,和标准数据模式进行比较,并报告异常,从而达到审计的目的。这类模式的特点是它有自己的操作系统、自己的数据库,以及自己的审计软件,能有效地和被审计系统进行连接,以抽取被审计系统中的数据。文献[5-9]就对这类模式进行了研究。

目前我国正在研究的联网审计也是持续审计的一种方式。它在技术实现上属于单机系统这种模式,其原理如图1所示[9,10]。从图1可以看出,联网审计在技术实现上主要分成四个部分:

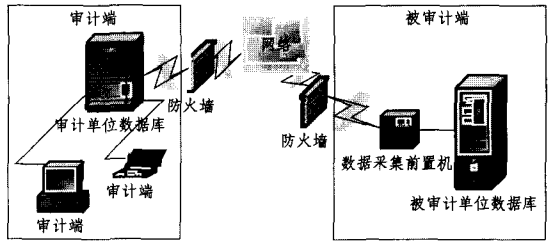

图1 联网审计实现方法的原理

(1)数据采集。要实现联网审计,必须研究如何采集被审计单位的电子数据。目前,联网审计数据采集的实现是通过在被审计单位数据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,从而完成联网审计的数据采集工作;

(2)数据传输。数据传输主要用来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;

*)国家自然科学基金项目(70701018),中国博士后科学基金项目(20060390281),江苏省高校自然科学研究计划资助项目(05KJB520054)。

陈 伟 副教授,主要研究方向为审计信息化;Robin Qiu 博士,教授;刘思峰 博士,教授,博士生导师。

(3)数据存储。对于采集到的电子数据,需要采取一定的方式来存储,武海平等人就研究了针对联网审计的海量数据存贮方式[11];

(4)数据分析(处理)。这一阶段主要是采用相关审计软件对采集来的电子数据进行分析,从而发现审计线索,获得审计证据。

数据分析是关键步骤。常用的数据分析方法主要包括:账表分析、数据查询、审计抽样、统计分析等[12]。这些方法虽然能有效地对电子数据进行审计,但多是把手工的审计流程计算机化,没有充分利用目前先进的信息技术,不能从电子数据中提取一些隐藏的或未知的信息,且要求审计人员有较高的业务知识。因此,研究一些新的审计方法来发现被审计数据中有价值的信息在联网审计中具有重要的理论和应用价值。基于以上分析,本文提出一种基于数据匹配技术的审计证据获取方法。

2 方法原理分析

为了后面论述的方便,首先给出以下几个相关定义:

定义 1(相似重复记录) 相似重复记录是指那些客观上表示现实世界同一实体的但是由于在格式、拼写上有些差异而导致数据库系统不能正确识别的记录[13,14]。

定义 2(相似重复实体) 相似重复实体和相似重复记录类似,相似重复记录主要针对同一个数据表中的记录,而相似重复实体则是指那些分布在不同数据源中客观上表示现实世界同一实体,但是由于在模式级和实例级上有些差异而被认为是不同对象的数据。如果能找到数据表中合适的主键,可以使用它来解决实体异构的问题。当两个数据源中的记录没有共同的标识符时,相似重复实体检测就变得很重要。

定义 3(数据之间的相似度 S) 数据之间的相似度 S 是根据要比较的两条数据的内容而计算出的一个表示两数据相似程度的数值,$0<S<1$。S 越小,则两数据相似程度越高;若 $S=0$,则表示两条数据为完全重复数据。

定义 4(数据相似检测) 数据相似检测是指通过计算两条数据之间的相似度 S,来判定两数据是不是相似重复数据(这里相似重复数据包括相似重复记录和相似重复实体)。

定义 5(数据匹配) 数据匹配是通过对采集来自不同数据源中的数据进行匹配(包括数据相似检测),来发现不同数据源中相似重复实体的一种技术方法。

联网审计环境下,采集来的多个数据源中可能含有相似重复实体,这些相似重复实体可能就是审计分析中要查找的可疑数据。比如,数据源 A 中出现的数据不应该出现在数据源 B 中。通过数据匹配技术可以有效地发现舞弊案件。国内外对数据匹配技术的研究多用来检测数据源中的相似重复数据[13-16],达到提高数据质量的目的,直接把数据匹配技术应用于审计中的研究还不常见。

根据以上分析,作者提出一种基于数据匹配技术的审计证据获取方法[17],其原理如图 2 所示。该方法的原理描述如下:

首先,根据对两被审计数据源的分析,从两个要比较的数据表中选取公共字段。然后,对两数据表中的数据进行预处理,如标准化数据字段格式等[18]。再从算法库中调用相似检测算法,根据所选取的公共字段,执行两个数据表中数据之间的比较,并根据预定义的重复识别规则,检测出相似重复实体,即为可疑数据。最后,对检测出的每一组相似重复实体(可疑数据)由审计人员通过一定的方法进行审计专业判断,并通过对可疑数据的延伸调查,最终获取审计证据。

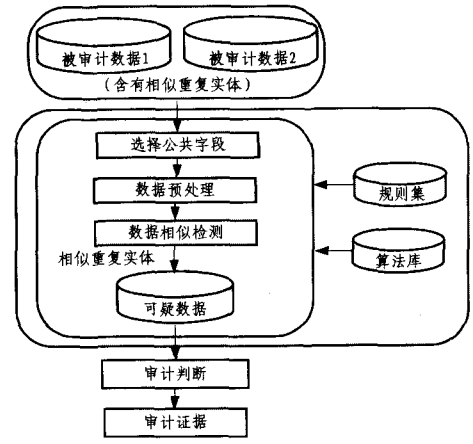

图 2 基于数据匹配技术的审计证据获取方法原理

3 关键步骤分析

基于数据匹配技术的审计证据获取方法,其关键步骤分析如下。

3.1 公共字段的选取

要比较的两个数据表在内容上是一致的,但表结构可能不相同(如字段的命名、字段的顺序、字段的个数等)。公共字段的选取是为了从两个数据表中选取要比较的字段,通过对这些公共字段的比较,来确定要比较的数据是不是相似数据。所选取的公共字段的字段名可以不一样,但字段的内容必须一致。具体来说,假设 D_1,D_2 为两个数据库,$R_1\in D_1$ 和 $R_2\in D_2$ 为两个表示同一现实对象的记录,$R_1=\{a_1,a_2,\cdots,a_m\}$,$R_2=\{b_1,b_2,\cdots,b_n\}$。$R_1$,$R_2$ 没有一个共同的标识符,R_1,R_2 公共字段的集合表示为 $\{Y_1,Y_2,\cdots,Y_k\}$,$k\in[1,N]$,N 为公共字段的数目。考虑两个实体 $a_i\in R_1$,$b_j\in R_2$;假设 $a_i(Y_K)=y_{aik}$,$a_j(Y_K)=y_{bjk}$ 为字段 Y_K 在 R_1 和 R_2 中的值,则我们需要比较的字段值就是 y_{aik},y_{bjk}。

对于公共字段的选取,也有一些自动的方法[19],但为了准确起见,目前我们还是采取人工方式来选取公共字段。

3.2 数据预处理

在进行相似重复数据检测之前,要进行数据预处理,数据预处理主要用来完成数据标准化等。因为从不同数据源中采集来的数据在格式上可能存在差异,所以通过数据标准化可以将特定类型的数据转化成统一的格式表示,从而为审计数据分析提供方便[18]。

3.3 数据相似检测

数据相似检测是基于数据匹配技术的审计证据获取方法中的关键步骤。通过数据相似检测,可以判断两数据是不是相似重复实体。一些文献对同一数据表中的数据相似检测进行了研究[13-16]。本文吸收前人的思想,对数据相似检测算法进行改进,使其适用于两个数据表中审计数据的分析。改进后的算法中,比较的字段就是所选择的公共字段。在进行比较时,为了提高检测效率,采用经长度过滤方法优化后的相似检测算法[20]。在所要比较的公共字段权重的选取上,可由审计人员根据审计的需要来确定,从而提高了检测的精度。所采用的数据相似检测算法的伪码描述如下:

输入:两条数据 R1 和 R2,两字段相似度的阈值 δ_1,两数据相似度的阈值 δ_2(两个阈值用于判定数据是否相似)。

```
输出:True/False
(1)RSimila = 0; // RSimila 为所计算出的两数据的相似度
(2)N = GetFieldNum(R1); // 计算要比较的公共字段数,R1 和 R2
                          的字段数相同
(3)For i = 1 to N Do
   1)MinFieldSimila = MAX;
   // 初始化 MinFieldSimila 值为某一极大值,该变量为两数据字段
      间的最小相似度
   2)If R1. Field[i] == NULL OR R2. Field[i] == NULL Then
     Continue;
     // 只有当两条数据在第 i 个字段上对应的值都不为空时,才进
        行字段比较。
   3)End If;
   4)If R1. Field[i]和 R2. Field[i]是字符型字段 Then
       Dist = d(R1. Field[i],R2. Field[i]);
   //计算对应字段 R1. Field[i]和 R2. Field[i]的编辑距离
       Simila = S(Dist);// 把字符型字段 R1. Field[i]和 R2. Field
                [i]之间的编辑距离转换成相似度
   5)Else
     Simila = S(R1. Field[i],R2. Field[i]);
     // 采用相关算法计算其它类型字段 R1. Field[i]和 R2. Field[i]
        的相似度
   6)End If;
   7)If Simila < MinFieldSimila Then
     MinFieldSimila = Simila; // 根据 Simila 重新对 MinFieldSimila
     置值
   8)End If;
   9)If MinFieldSimila >δ1 Then
     Return False;
     //如果两数据任意字段间相似度大于 δ1,则它们不是相似重复
       数据
   10)Else
      RSimila = RSimila + Weight[i] * Simila;
      // 否则,数据相似度变量 RSimila 相应增加 Weight[i] * Sim-
         ila
      Weight_sum = Weight_sum + Weight[i]
      // 计算非空字段的权重和
   11)End If;
(4)End;
(5)RSimila = RSimila / Weight_sum
(6)If RSimila < δ2 Then
   Return True; // 如果数据间相似度小于 δ2,则认为它们是相似重
                   复数据
(7)Else
   Return False; // 否则,不是相似重复数据
(8)End If;
```

在数据相似检测的过程中,判定两数据是否相似,由以下两个阈值来定。

δ_1:用来判定两条数据中同一字段是否相似。如果两个字段的相似度大于 δ_1,则两字段不相似,从而也判定两数据不是相似重复数据;

δ_2:用来判定两条数据是否是相似重复数据。数据的相似度为对应字段的相似度之和。如果两条数据的相似度大于 δ_2,则两数据不是相似重复数据。

由以上算法可以看出:数据间的相似检测依赖于数据中每个字段的相似检测,因此字段的相似检测是一个相当重要的原子操作,其效率直接影响整个算法的效率。函数 $S(R1.Field[i], R2.Field[i])$ 计算两数据中相同类型字段 $R1.Field[i]$ 和 $R2.Field[i]$ 的相似度。对不同类型的字段,我们采用不同的计算方法。

(1)布尔型字段相似度计算方法。对于布尔型字段,如果两字段相等,则相似度取 0;如果不同,则相似度取 1。

(2)数值型字段相似度计算方法。对于数值型字段,可采用计算数字的相对差异算法:

$$S(s_s,s_2)=\frac{|s_1-s_2|}{\max(s_1,s_2)}$$

其中,s_1,s_2 为数值型字段。

(3)字符型字段相似度计算方法。对于字符型字段,一个字段可以看成是一个字符串。字符串的相似检测也称字符串匹配,它是计算机科学中的一个最重要的研究问题,最主要的方法是基于编辑距离算法[21,22]。通过采用编辑距离算法,可以计算出两个字段间的编辑距离。由于编辑距离值为整数,为了把字段间的编辑距离转换成字段间的相似度,我们提出以下转换方法,如表 1 所示。

表 1 编辑距离和相似度的对应关系定义

编辑距离	相似度
1	0.9
2	0.8
3	0.7
4	0.6
…	…

表 1 中的对应关系也可以根据对数据源的分析进行调整,从而更准确地检测相似重复数据。

编辑距离算法使用动态程序来实现,它运行 $O(m\times n)$ 次,其中 m 和 n 是两条数据的长度。作者在文献[23]中对编辑距离算法的原理及实现做了分析和研究。由于编辑距离算法的计算比较复杂,其原理及实现过程描述如下:

定义 6(编辑距离) 两个字符串 x 和 y 之间的编辑距离 $d(x,y)$ 定义为:把一个字符串转换成另一个字符串时在单个字符上所需要的最小编辑操作(比如插入、删除、代替)的代价数。

假设 A 是一个有限的符号字母表,A^* 是 A 上所有字符串的集合;ε 表示空符号,$|x|$ 表示字符串 x 的长度,$|\varepsilon|=0$。一个编辑操作就是以下的任何一个:

$a\to b, a\to\varepsilon, \varepsilon\to a$

这里 $a,b\in A$,称 $a\to b$ 为一个代替操作,$a\to\varepsilon$ 为一个删除操作,$\varepsilon\to a$ 为一个插入操作。如果 $a=b$,则 $a\to b$ 称为一个同一的代替操作,否则称为不同一的代替操作。一个代价函数就是一个对每个编辑操作指派一个非负实数值的函数。令 $c(a\to b)$ 表示代替操作 $a\to b$ 的代价,$c(a\to\varepsilon)$ 表示删除操作 $a\to\varepsilon$ 的代价,$c(\varepsilon\to a)$ 表示插入操作 $\varepsilon\to a$ 的代价,假设 $S=e_1,\cdots,e_k$ 为一个编辑操作的序列,它的代价被定义为

$$c(S)=\sum_{i=1}^{k}c(e_i)$$

根据以上定义,两个字符串 x 和 y 的编辑距离 $d(x,y)$ 可被定义为转换 x 到 y 所需的最小操作序列的代价数,即

$$d(x,y)=\min\{c(S)\}$$

其中,S 是一个转换 x 到 y 的编辑操作序列。

计算编辑距离 $d(x,y)$ 的标准算法是基于一个动态程序,它使用以下递归公式来计算维数为 $(n+1)\times(m+1)$ 的二维编辑矩阵 $D(i,j)$ 中的元素:

$$D(0,0)=0$$

$$D(0,j)=D(0,j-1)+c(\varepsilon\to y_j)\quad j=1,\cdots,m$$

$$D(i,0)=D(i-1,0)+c(x_i\to\varepsilon)\quad i=1,\cdots,n$$

$$D(i,j)=\min\begin{Bmatrix}D(i-1,j-1)+c(x_i\to y_j),\\ D(i-1,j)+c(x_i\to\varepsilon),\\ D(i,j-1)+c(\varepsilon\to y_j)\end{Bmatrix}$$

$$i=1,\cdots,n\ j=1,\cdots,m$$

可以看出:$d(x,y)=D(m,n)$。

在本文中,计算编辑距离 $d(x,y)$ 的算法描述如下:

```
输入:要比较的两个字符串 X,Y
输出:两个字符串的编辑距离
(1)求 X 的长度 N,Y 的长度 M,如果 N 为 0,返回 M 并退出;如
果 M 为 0,返回 N 并退出;
(2)构建一个 M 行 N 列的矩阵 D[M][N],初始化第一行为:0 到
N;第一列为:0 到 M;
(3)i 从 1 到 N 检测 X 中的每一个字符;
(4)j 从 1 到 M 检测 Y 中的每一个字符;
(5)如果 X[i]= Y[j],则操作代价 COST 为 0,否则,操作代价
COST 为 1;
(6)使 D[i][j]为 Minimum(d[i-1][j]+1,d[i][j-1]+1,d[i-
1][j-1]+cost),其中,Minimum()为求最小值函数;
```

(7)返回 D[M][N];
(8)结束。

3.4 规则集与算法库

规则集与算法库是基于数据匹配技术的审计证据获取方法的一个重要部分。其中,规则库用来保存关于数据匹配的规则,主要包括:

(1)重复识别规则。重复识别规则用来指定两条数据为相似重复数据的条件,比如字段相似度阈值 δ_1、数据相似度阈值 δ_2 等。

(2)相似度关系规则。相似度关系规则用来保存编辑距离和相似度的对应关系,供执行相似重复数据检测时调用。

(3)警告规则。警告规则用来指定对一些特殊事件的处理规则及相应提示信息。

在对被审计电子数据进行分析时,可根据具体的业务,在规则集中定义相应的规则,或者修改已有的规则,从而使该系统适用于不同的业务数据,具有较强的通用性和适应性。算法库用来存放相似数据检测算法。多种算法存放在算法库中,供审计数据分析时根据不同的情况来选用相应的合适算法。

4 软件系统的实现

根据以上分析,作者采用 Delphi 7.0 实现了以上研究的基于数据匹配技术的审计证据获取方法,其主界面如图 3 所示。

图 3 软件系统主界面

5 方法的审计风险分析

5.1 审计风险评价指标的定义

审计风险是审计人员在审计过程中采用了没有意识到的不恰当的,审计程序和方法,或者错误地估计和判断了审计事项,做出了与事实不相符合的审计结论,进而受到有关利害关系人或潜在的利害关系人的指控,乃至承担法律责任的可能性。国际审计和保证准则委员会(IAASB)把审计风险的模型定义为

审计风险 = 重大错报风险 × 检查风险

在审计风险模型中,审计人员所能控制的只有检查风险,重大错报风险与被审计单位有关,审计人员对其无能为力,只能对其水平进行评估,以便确定可接受的检查风险水平。根据以上审计风险模型,我们可以看出,可以通过采用合适的审计方法来降低检查风险。

随着计算机审计技术的应用,审计技术和方法改变的同时,也带来了新的审计风险。所以,信息化环境下审计风险的控制仍然是一个重要的问题。但总的来说,目前,国内对信息化环境下计算机审计风险的研究多是从理论层面分析计算机审计风险的成因与规避,在审计风险控制这方面的研究也多是从定性的角度进行分析,没有从定量的角度对其进行深入的研究。本文从用于审计取证的审计数据分析算法(数据匹配算法)入手,从定量的角度分析基于数据匹配技术的审计证据获取方法的审计风险。其原理说明如下:

针对用于审计数据分析的数据匹配技术,定义相应的查全率 R(Recall)和查准率 P(precision),分别为:

(1)查全率 R(Recall):相似重复数据被正确识别的百分率

$$R=\frac{\text{正确识别出的可疑数据}}{\text{实际的可疑数据}}$$

(2)查准率 P(Precision):识别相似重复数据的正确率

$$P=\frac{\text{正确识别出的可疑数据}}{\text{识别出的可疑数据}}$$

通过以上两个指标,来评价基于数据匹配技术的审计证据获取方法的审计检查风险。

5.2 审计风险的实验分析

表 2 实验数据

	表 1 中的记录个数	表 2 中的记录个数
第 1 组数据	1 万	1 万
第 2 组数据	2 万	2 万
第 3 组数据	3 万	3 万
第 4 组数据	4 万	4 万
第 5 组数据	5 万	5 万

为了对基于数据匹配技术的审计证据获取方法的检查风险进行定量分析,需要一些实验数据。但是,由于实际数据的规模有限,而且不能容易地按测试目的得到相应的实际数据,因而无法按要求做一些检测实验。为了解决这个问题,我们采用了自己设计的一个基于 Visual Basic 6.0 的模拟数据生成系统[24],该系统能用来生成关于供应商信息的数据。通过对数据生成系统设置相应的参数,可生成所需的面向各种不同测试目的和不同规模的测试数据。采用这些模拟数据可以对基于数据匹配技术的审计证据获取方法的检查风险进行定量分析,并根据分析结果,不断完善该方法,从而为降低审计风险提供了保障。由于篇幅所限,只对其中的一个实验过程分析如下。

我们采用所研制的模拟数据生成系统生成五组数据,每组数据中含有设定个数的相似重复数据,如表 2 所示。以"供应商名称"和"供应商地址"两字段为比较字段,对"供应商信息"表中的相似重复数据进行检测,并和设定个数进行比较。系统运行在 PC 工作站上,工作站的硬件配置为:CPU P4 1.8G,256M DDR;操作系统为 Windows2000 Server。测试指标和实验结果如图 4 所示,其中 G 为数据的组。从图 4 可以看出,该系统能较好地完成两数据表中相似重复数据的检测工作,具有较低的检查风险。

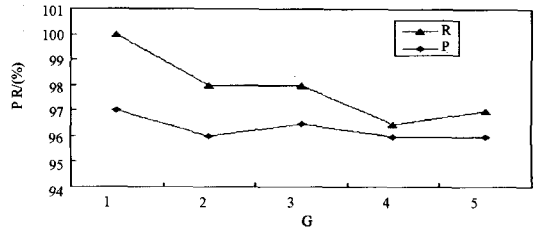

图 4 方法的检查风险实验结果

另外,通过灵活地设置字段和数据的阈值,以及字段的权重,可以改变系统的查全率和查准率,从而可以控制系统的检查风险。需要指出的是,当数据量较大时,系统在 PC 工作站

上运行时间较长，建议在小型机或服务器上运行。

6 应用实例

本节以如何查找重复发放失业金的人员为例，来分析基于数据匹配技术的审计证据获取方法的应用。假设“失业人员登记表 1”和“失业人员登记表 2”为联网审计环境下从某两个劳动局采集来的失业保险数据，数据格式为 MS Access 数据表。采用本文提出的方法对两数据源进行分析，其关键步骤如下：

(1)根据对被审计数据的分析，设置相似度对应关系以及字段阈值，如图 5 所示。

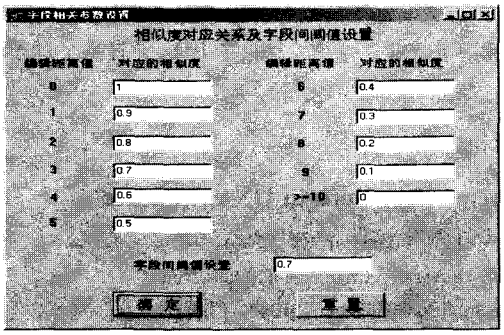

图 5 相似度对应关系以及字段阈值设置界面

(2)把两 Access 数据库中的数据分别采集到软件系统中来，如图 6 所示。

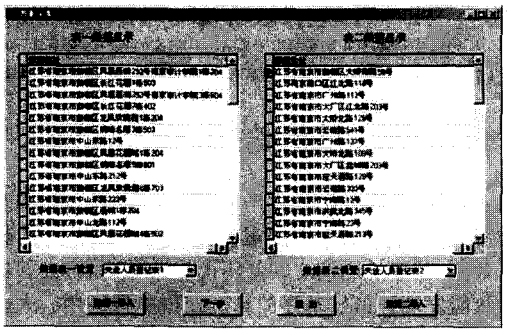

图 6 系统数据采集界面

(3)设置相应的匹配参数，如图 7 所示。

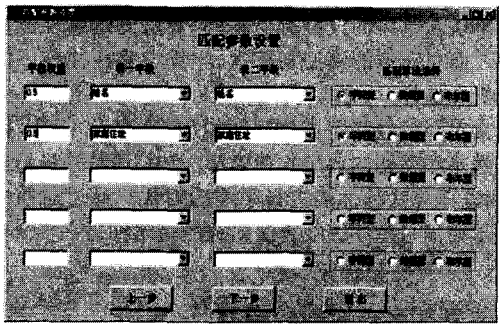

图 7 匹配参数设置界面

(4)设置数据阈值，运行检测，结果如图 8 所示。

(5)通过系统的检测，发现两对相似重复实体。对检测结果进行延伸调查，确认检测结果的正确性，并最终获取审计证据。

由以上应用可以看出：本文所研究的方法能有效地发现重复发放失业金的人员。

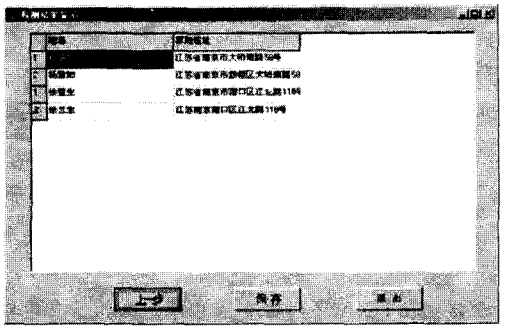

图 8 系统检测结果界面

结束语 本文根据我国联网审计的特点及实践，提出了一种基于数据匹配技术的审计证据获取方法。和常用的审计方法相比，该方法具有以下优点：

(1)能发现被审计数据源中的“隐藏”信息。常用的审计数据分析方法多是仅仅把手工的审计流程计算机化，没有充分利用目前先进的信息技术，不能从电子数据中提取一些隐藏的或未知的信息。而本文研究的方法能有效地发现被审计数据源中的“隐藏”信息。

(2)能充分利用采集来的电子数据。除了采用常用的数据分析方法对采集来的同一单位内的审计数据进行分析外，采用本文研究的方法对不同单位间的相关审计数据进行分析，可发现隐藏于不同被审计单位数据中的审计线索，从而可对采集来的数据进行充分利用，较好地适应我国面向数据的联网审计的特点。

(3)为审计数据分析提供了新方法。常用的审计数据分析方法“重号分析法”仅能用来查找被审计数据中的相同数据(例如，检查一个数据表中是否存在相同的发票被重复多次记账)，而本文提出的方法可针对选择的多个字段完成相似检测，更能检测出被审计数据中的“舞弊”信息，具有更广泛的应用范围。

(4)减少了审计风险。根据建立的审计检测风险评价指标，可以定量地评价基于数据匹配技术的审计证据获取方法，并通过对数据匹配算法进行优化或设置合理的字段和数据的阈值，使其适用于审计数据分析，从而为减少审计实施过程中的审计风险提供了保障。

当然，该方法不能够解决联网审计环境下所有的审计问题。但通过和其它审计方法一起使用，能在很大程度上提高审计效率，降低审计风险。下一步就是不断完善这种方法(如提高该方法的检测速度)，并将其应用到审计实践中去，从而成为一种实用的审计方法。

参 考 文 献

[1] Alexander K, Ephraim F S, Miklos A V. Continuous online auditing: a program of research[J]. Journal of Information Systems, 1999, 13(2): 87-103

[2] Du H, Roohani S. A framework for independent continuous auditing of financial statements[A]// American Accounting Association 2006 Annual Meeting [C]. Washington: http://aaahq.org/AM2006/ abstract. cfm? submissionID=1783, 2006

[3] Groomer S M, Murthy U S. Continuous auditing of database applications: an embedded audit module approach [J]. Journal of Information Systems, 1989, 3(2): 53-69

(下转第 194 页)

biopolymers using expectation maximization [J]. Machine Learning, 1995, 21(1/2): 51-80
[4] Hertz G, Stormo G. Identifying DNA and protein pat-terns with statistically significant alignments of mul-tiple sequences [J]. Bioinformatics, 1999, 15: 563-577
[5] Styczynski M P, Jensen K L, Rigoutsos I, et al. An ex-tension and novel solution to the (l,d)-motif challenge problem [J]. Genome Informatics, 2004, 15: 63-71
[6] Buhler J, Tompa M. Finding Motifs using random pro-jections [J]. J Comput Biol, 2002, 9(2): 225-242
[7] Leung H C M, Chin F Y L. Generalized Planted (l, d)-Motif Problem with Negative Set [A]// WABI [C]. 2005: 264-275
[8] Leung H C M, Chin F Y L. An Efficient Algorithm for the Extended (l,d)-Motif Problem with Unknown Number of Binding Sites [A]// Proceedings of the IEEE 5th Symposium on Bioinformatics and Bioengi-neering [C]. Min-neapolis, Minnesota, USA: IEEE Press, 2005: 11-18
[9] Leung H C M, Chin F Y L. Finding Exact Optimal Motif in Matrix Representation by Partitioning [J]. Bio-informatics, 2005, 21(2): 86-92
[10] Chin F Y L, Leung H C M. Voting algorithm for discovering long Motifs [A]// Proc. of Asia-Pacific Bio-informatics Conference[C]. 2005: 261-272
[11] Davila J, Balla S, Rajasekaran S. Space and Time Efficient Algorithms for Planted Motif Search [A] // Second International Workshop on Bioinformatics Research and Applications (IWBRA) [C]. Accepted
[12] Sinha S, Tompa M. Discovery of novel transcription factor binding sites by statistical overrepresentation [J]. Nucleic Acids Res, 2002, 30(24): 5549-5560
[13] Eskin E, Pevzner P A. Finding composite regulatory patterns in DNA sequences [J]. Bioinformatics, 2002, 18(1 Suppl): 354-363
[14] Rajasekaran S, Balla S, Huang C H. Exact algorithms for planted motif challenge problem [A]// APBC[C]. 2005: 249-259
[15] Pavesi G, Mauri G, Pesole G. An algorithm for finding signals of unknown length in DNA sequences [J]. Bioinformatics, 2001, 17 (1. Suppl): 207-214
[16] Price A, Ramabhadran S, Pevzner P A. Finding subtle Motifs by branching from sample strings [J]. Bio-informatics, 2003, 19(2. Suppl): 149-155
[17] Keich U, Pevzner P A. Finding motifs in the twilight zone [J]. Bio-informatics, 2002, 18(10): 1374-1381
[18] Fratkin E, Naughton B T, Brutlag D L, et al. MotifCut: regulatory motifs finding with maximum density sub-graphs [J]. Bio-informatics, 2006, 22: 150-157
[19] Wang J X, Yang D. UPNT: Uniform Projection and Neighbourhood Thresholding Method for Motif Discovery [J]. International Journal of Bioinformatics Research and Applications (IJBRA). Accepted
[20] Liu X S, Brutlag D L, Liu J S. An algorithm for finding protein-DNA binding sites with applications to chroma-tin-immunoprecipitation microarray experiments [J]. Nat. Biotechnol, 2002, 20 (8): 835-839
[21] Liang S, Samanta M P, Biegel B A. cWINNOWER algorithm for finding fuzzy dna motifs [J]. J. Bioinfo. Comput. Biol, 2004, 2 (1): 47-60
[22] Yang X, Rajapakse J C. Graphical Approach to Weak Motif Recognition [J]. Genome Informatics, 2004, 15(2): 52-62
[23] Raphael B, Liu L T, Varghese V. A uniform projection method for Motif discovery in DNA sequences [J]. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2004, 1 (2): 91-94
[24] King J, Cheung W, Hoos H H. Neighbourhood Thresh-olding for Projection-based Motif Discovery [J]. Bioinformatics, Accepted
[25] Davila J, Rajasekaran S. Extending Pattern Branching to Handle Challenging Instances [A] // Proceedings of the Sixth IEEE Symposium on Bio-Informatics and BioEn-gineering (BIBE'06) [C]. 2006: 65-69
[26] Jensen K L, Styczynski M P, Rigoutsos I, et al. A generic motif discovery algorithm for sequential data [J]. Bioinformatics, 2006, 22(1): 21-28
[27] Yangb X, Rajapakse J C. Robust Algorithm for Finding Weak Motifs [A] // Proceedings of the 2005 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology[C]. 2005: 1-6
[28] Tompa M, Li N, Bailey T L, et al. Assessing computational tools for the discovery of transcription factor binding sites [J]. Nat Biotechnol, 2005, 23(1): 137-144
[29] 王建新，黄元南，陈建二. 一种基于彩色编码的基序发现算法 [J]. 软件学报，2007，18(6)：1298-1307

(上接第 187 页)

[4] Debreceny R S, Gray G L, Ng J, et al. Embedded audit modules in enterprise resource planning systems: implementation and functionality [J]. Journal of Information Systems, 2005, 19(2): 7-27
[5] Vasarhelyi M A, Halper F B. The continuous audit of online systems[J]. Auditing: A Journal of Practice and Theory, 1991, 10 (1): 110-125
[6] Woodroof J, Searcy D. Continuous audit - Model development and implementation within a debt covenant compliance domain [J]. International Journal of Accounting Information Systems, 2001, 2(3): 169-191
[7] Rezaee Z, Sharbatoghlie A, Elam R, et al. Continuous auditing: building automated auditing capability[J]. Auditing: A Journal of Practice and Theory, 2002, 21(1): 147-163
[8] Murthy U S, Groomer S M. A continuous auditing web services model for XML-based accounting systems [J]. International Journal of Accounting Information Systems, 2004, 5(2): 139-163
[9] Chen Wei, Wang Hao, Zhu Wenming. Study on data-oriented IT audit used in China[A]// Zhu Qingsheng, ed. Proceedings of the 11th Joint International Computer Conference[C]. Singapore: World Scientific Publishing, 2005: 666-669
[10] 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
[11] 武海平，余宏亮，郑纬民，等. 联网审计系统中海量数据的存储与管理策略[J]. 计算机学报，2006，29(4)：618-624
[12] 陈伟，刘思峰，Qiu Robin. 审计数据处理方法研究综述[J]. 统计与决策，2007(3)：129-131
[13] Verykios V S, Elmagarmid A K, Houstis E N. Automating the approximate record matching process[J]. Journal of Information Sciences, 2000, 126(1/4): 83-98
[14] 邱越峰，田增平，季文赟，等. 一种高效的检测相似重复记录的方法[J]. 计算机学报，2001，24(1)：69-77
[15] Monge A E. Matching algorithms within a duplicate detection system[J]. IEEE Data Engineer Bulletin, 2000, 23(4): 14-20
[16] 俞荣华，田增平，周傲英，等. 一种检测多语言文本相似重复记录的综合方法[J]. 计算机科学，2002，29(1)：118-121
[17] 陈伟，刘思峰，邱广华. 计算机审计中数据处理新方法探讨[J]. 审计与经济研究，2006，21(1)：37-39，48
[18] 张进，易仁萍，陈伟. 计算机审计中电子数据的清理研究[J]. 审计研究，2004(6)：21-25
[19] Li W S, Clifton C. Semantic integration in heterogeneous databases using neural networks [A]// Bocca J B, Jarke M, Zaniolo C, eds. Proceedings of the 20th International Conference on Very Large Data Bases[C]. Santiago: Morgan Kaufmann, 1994: 1-12
[20] 陈伟，丁秋林，谢强. 交互式数据迁移系统及其相似检测效率优化[J]. 华南理工大学学报(自然科学版)，2004，32(2)：58-61
[21] Navarro G. A guided tour to approximate string matching [J]. ACM Computing Surveys, 2001, 33(1): 31-88
[22] Vidal E, Marzal A, Aibar P. Fast computation of normalized edit distances [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1995, 17(9): 899-902
[23] 陈伟，丁秋林. 数据清理中编辑距离的应用及 Java 编程实现 [J]. 电脑与信息技术，2003，11(6)：33-35，60
[24] 陈伟，张金城，Qiu Robin. 审计数据处理实验中模拟数据生成系统的研究[J]. 计算机工程，2007，33(19)：54-56