第33卷 第12期
Vol.33 No.12

计 算 机 工 程
Computer Engineering

2007年6月
June 2007

·开发研究与设计技术· 文章编号：1000—3428(2007)12—0267—03 文献标识码：A 中图分类号：TP391

地方财政收费稽查选案系统的研究与设计

丁春玲[1,2]，黄永红[1]，王健华[1]，刘和洋[1]

(1. 中国科学院计算技术研究所，北京 100080；2. 中国科学院研究生院，北京 100039)

摘 要：收费稽查的计算机选案，是指根据科学的选案指标，以计算机为辅助工具，对现有的内外部收费信息进行加工处理、分析筛选，以发现存在收费异常的执收单位。该文提出了基于决策树的财政收费稽查选案模型，提出了支持收费稽查选案的元数据组织方法，阐述了稽查选案主题的提取方法和多维模型构建的方法，并将决策树方法应用于收费的稽查选案。基于收费稽查的选案模型，设计并实现了收费稽查计算机选案系统，提高了收费稽查工作的准确性和科学性。

关键词：收费；稽查选案；数据仓库；决策树；联机分析处理

Research and Design of Charge Checking Case Selection System Based on Local Finance

DING Chunling[1,2], HUANG Yonghong[1], WANG Jianhua[1], LIU Heyang[1]

(1. Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100080;
2. Graduate School, Chinese Academy of Sciences, Beijing 100039)

【Abstract】The computer selection of charge checking case uses computer as a tool, according to scientific case selecting indexes, to find out exceptional charging units by processing, analyzing and filtrating internal and external data. This paper presents a model of charge checking case selection based on decision-tree: proposes the organization method of metadata for charge checking case selection, describes how to extract charge checking case selecting subjects and constructs its multidimensional model and uses decision-tree method in charge checking case selection. Based on the case selecting model, the computer selection system of charge checking case is implemented, which improves the efficiency, veracity and scientificity of charge checking.

【Key words】Charge; Checking case selection; Data warehouse; Decision-tree; OLAP

收费稽查是指财政主管部门依法对所管辖的执收单位的收费情况和支出情况进行检查和处理，它是财政主管部门对执收单位的事后监督，通过这种监督，督促各执收单位依照法律法规正确收费，以达到制止乱收乱缴的目的。

选案是收费稽查的首要环节[2]，它的任务是根据科学的选案指标，对各类收费信息进行综合分析、归集分类，在众多的收费单位和部门中确定待查对象，为收费稽查工作提供方向和目标。

目前国内财政稽查系统的应用研究，主要集中在税务稽查领域，而未涉及到收费稽查系统的研究。税务稽查工作具有3个明显的行业特征：税务稽查主体的特定性，税务稽查对象的确定性，税务稽查依据的法律性。税务稽查数据来源比较完整，只要纳税人的涉税信息进入系统，就可对其进行稽查选案，税务稽查系统更注重业务解决方案，稽查选案主要采用人工智能的方法如专家系统[2]。与税务稽查相比，收费稽查在业务上更有难度，因为收费有如下3个特点：(1)收支两条线；(2)收费项目的类型和相关的收费标准具有模糊性；(3)收费数据的不完整行性。这就使得收费数据的收集整理成为稽查选案的关键，为此，利用如下的技术来实现收费稽查的计算机选案系统，以从根本上解决选案随意、盲目稽查的问题，为收费稽查节省人力物力：(1)基于决策树的收费稽查选案模型：基于数据仓库的数据挖掘，CLS决策树学习算法；(2)基于OLAP技术的税务稽查选案的在线分析：在决策树选案的基础上，用OLAP技术对收费数据进行多角度分析。

1 收费稽查选案业务简介

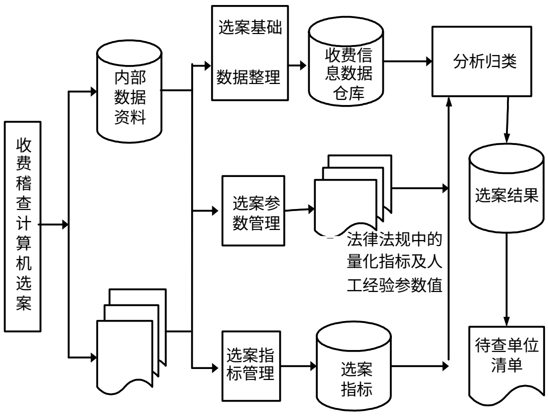

图1 收费稽查选案业务流程

收费稽查分为3类：专项稽查，专案稽查，常规稽查。专项稽查是根据上级指令及有关部门的计划安排，对特定行业或某类项目的重点稽查；专案稽查是指根据公民举报、上级或其他部门交办、转办的案件进行的稽查；常规稽查是财政主管部门利用现有的内外部资料，对执收单位进行全面综

基金项目：科技部中小企业技术创新基金资助项目“基于Web的票据定制及交费监管系统”(03C26212110499)

作者简介：丁春玲(1980－)，女，硕士生，主研方向：电子政务；黄永红，副研究员；王健华，高工；刘和洋，工程师

收稿日期：2006-07-23 **E-mail：**dingchunl@sohu.com

合的检查处理，既有广度又有深度。

计算机选案主要用于常规稽查，它是通过全面的内外部资料，利用科学的选案指标，对收费单位信息进行分析归类，发现存在收费异常的执收单位。收费稽查计算机选案的业务流程如图 1 所示。

2 收费稽查选案元数据的组织方法和多维模型

2.1 收费稽查选案信息组织

由稽查选案业务流程可见，收集收费数据信息，建立支持收费稽查选案的元数据是进行收费稽查选案的前提，元数据组织的好坏直接影响到选案的执行及系统性能。

收费稽查选案元数据来源于内部数据和外部数据资料，内部数据是指财务主管部门已经掌握的执收单位的收费信息及相关法律法规文件，主要是收费管理信息系统长期积累的历史数据及收费项目立项文件，如：执收单位的统计上报数据，执收单位的上缴核销票据，执收单位的票据打印所产生的收费数据。外部资料是指从其他财务部门或外部有关职能部门获取的执收单位相关信息，如代收银行提供的财政账户进出账信息、稽查结果反馈信息、物价局提供的平均消费水平信息等。

收费稽查选案系统核心数据主要包括 4 个方面：(1)收费项目源于财政主管部门的立项文件，每个收费项目都必须经过财政部门审批才能立项收费，每个项目的收费标准由财政部门协同物价局统一制定；(2)财政票据由财政主管部门统一印制，作为收费的唯一凭证；(3)执收单位是已执有收费许可证获准收费的行政事业性单位；(4)收费数据是执收单位提交的日常收费信息。这些数据之间是相互关联的：每个单位只能就特定项目收费，每种票据用于收取特定项目的费用，每个单位只能使用特定种类的票据；收费数据、财政票据必须源自收费项目；收费项目针对不同的缴费对象有多条收费数据；财政票据包含了某个缴费对象的多个项目的收费信息；这种约束关系可用图 2 表示。

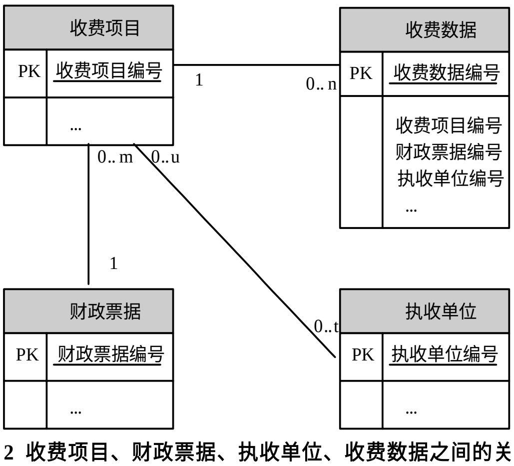

图 2 收费项目、财政票据、执收单位、收费数据之间的关系

另外，为了辅助选案，除了收费信息核心数据外，还需要补充一些附加信息表，如选案指标表、选案变量表、指标变量对照表、选案数据汇总表、选案结果表、选择指标评估结果表等。这些表用于存储从外部数据资料中提取的有用信息、由核心数据衍生出来的汇总信息以及选案产生的中间数据。

2.2 收费稽查选案的数据仓库模型

鉴于收费的 3 个特点，采用数据仓库技术来实现收费稽查计算机选案。数据仓库的构建是数据挖掘和OLAP分析的基础。数据仓库是按主题组织的多维结构[5]，主题是人们所关心问题的主要方面，一个主题对应一个多维模型，必须先确定系统的主题域，再按主题构建数据仓库多维模型。

当通过数据挖掘发现了异常执收单位之后，稽查人员往往需要对收费信息进行多角度观察分析，以深入了解问题所在方面。对于收费稽查人员来说，他们需要查询历史的违法定性数据，了解执收单位的票据使用情况，对执收单位的支出项目及金额进行审查，具体可以分析以下问题：

(1)哪些地区的违法定性情况比较严重，主要集中在哪些部门或单位；

(2)违法金额的分布情况，主要集中在哪些项目；

(3)哪些单位违法比较严重，主要在哪些指标上；

(4)违法现象跟哪些票据有关联，关联程度；

(5)哪些单位可以收取哪些费用，使用哪些票据，票据使用是否正常，收支是否平衡。

根据上述需求确定以下主题：违法定性主题，选案指标分析主题，票据使用查询主题，财务收支平衡主题。其中违法定性主题的元数据描述如表 1。

表 1 收费稽查违法定性主题元数据描述

主题名			属性
违法定性	事实表	违法定性主题	案件编号、执收单位、收费所属日期、收费项目、所用票据、违法类型、指标分数、应收金额、实收金额
	维表	稽查时间维度	收费所属时期、稽查年度、稽查季度、稽查月份
		稽查地域维度	执收单位代码、稽查地域、所属省份、所属市、所属地区(县)
		收费项目维度	项目代码、项目名称、项目级次、上级项目、执收级次、管理形式、启用日期、当前状态
		执收单位维度	单位代码、单位名称、单位主管、上级单位、主管部门 …
		收费票据维度	票据代码、票据名称、票据类型、票据用途、票据性子、是否核销、是否跨年度使用

建立了收费稽查主题的多维模型之后[5]，就可以用OLAP技术对数据进行分割查询，比如按年度查询违法定型情况，结果如表 2。

表 2 收费稽查数据分片

2000 年	2000 年收费稽查数据分片
2001 年	2001 年收费稽查数据分片
2002 年	2002 年收费稽查数据分片
…	…

收费稽查中的主题是由多个表来实现的[3]，这些表通过外键联系在一起，形成一个完整的星型模型，违法定性主题的星型模型如图 3 所示。

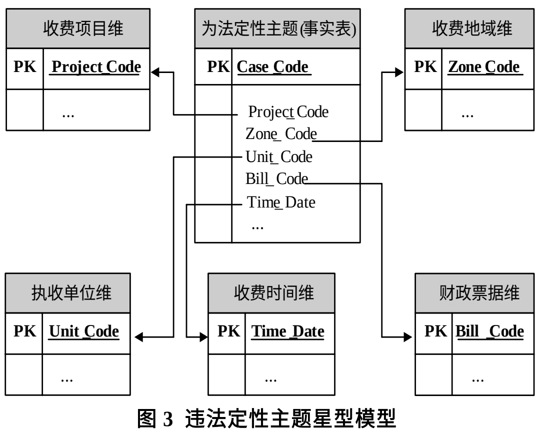

图 3 违法定性主题星型模型

3 基于决策树的收费稽查选案业务模型

收费稽查选案模型是系统的核心，稽查选案过程就是将目标对象的选案属性变量与相应的选案指标进行比较、判别分类的过程，它属于数据挖掘任务中的分类知识发现，把决策树方法用于收费稽查选案，形成基于选案指标的树形选案模型。

决策树本质上是通过一系列规则对数据进行分类的过程，它采用自顶向下的递归方式[5]，在决策树的内部结点进行属性值的比较，判断从该结点向下的分支，在决策树的叶结点得到结论，从根到叶结点的一条路径就对应着一个选案方法。

树的高度越小则树的决策能力越强，构造好的决策树的关键在于选择好的测试属性，收费稽查选案的测试属性就是与选案指标相对应的选案属性。选案指标是基本的选案元素，它是从某一方面评价收费异常与否的标准，所有的指标并非重要程度都一样，它们通过权值来体现，权值越大指标越重要，指标的决策能力也就越强，所以把指标权值作为测试属性选择标准。

目前比较成熟的决策树学习算法有 CLS、ID3 等，本文采用的是 CLS 算法，并对其进行改进，基本原理如下：

(1)生成一棵空决策树和一张训练样本属性表 S(选案指标表)，选择训练样本集 T；

(2)若属性表 S 非空并且样本集非空，则根据指标权值从 S 中选择指标 A 作为测试属性，并生成节点 A，否则结束学习算法；

(3)在节点 A 将样本集分为异常收费 T1 和非异常收费 T2 两个子集；

(4)从表 S 中删除属性 A；

(5)对于 T1，生成叶结点 T1，对于 T2，转到第(2)步，递归调用算法。

另外，为了对收费异常程度进行量化，为每个指标节点添加一个分数，叶结点的分数就是从树根到叶结点路径上的节点分数与相应权值乘积之和，某次选案结束后，根据叶结点的分数对选案对象进行排序，分数高者异常程度越严重，可以作为稽查重点。收费稽查选案决策树模型如图 4 所示。

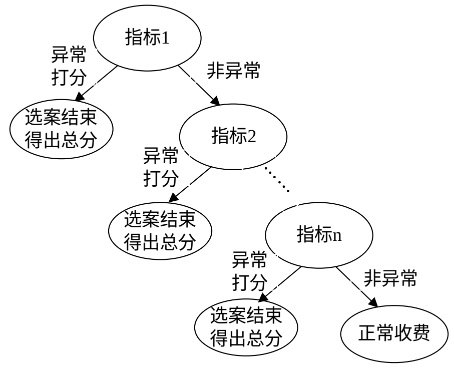

图 4 收费稽查选案决策树模型示意图

选案指标的重要程度会随着稽查结果的反馈信息和选案业务需求的变化而变化，相应的权值也会改变，所以决策树的结构也要随着调整。

4 收费稽查选案系统总体设计

收费稽查选案系统以 Oracle9i 为平台，以 JBuilder9 为开发工具，充分利用了 Oracle9i 数据仓库解决方案对 OLAP 的支持，及其与 Java 语言的无缝结合。系统的总体结构如图 5。

系统各模块功能如下：

(1)数据采集预处理模块。根据选案需求，从所有执收单位的信息中抽取与本次选案有关的资料进入数据仓库；

(2)选案方法维护模块。选案方法维护人员通过它对选案方法知识库进行维护；

(3)选案执行。从数据仓库中提取数据，利用选案方法数据库中的选案模型进行选案；

(4)选案指标评估。将计算机选案结果与实地稽查结果进行比较，验证选案指标的准确性，以便适时调整选案方法；

(5)OLAP 查选分析。根据选案结果，用 OLAP 技术对目标对象进行分析查选，以深入了解违法所在方面。

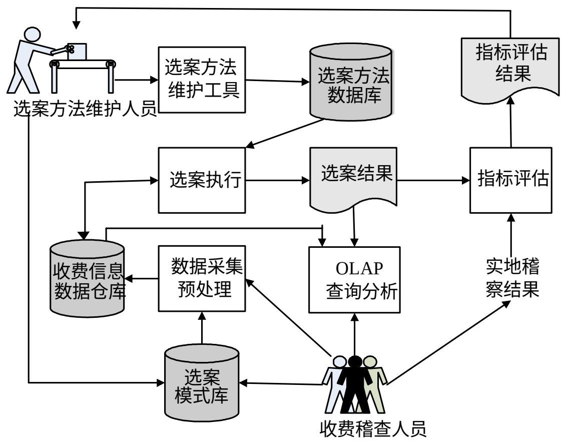

图 5 收费稽查选案系统总体结构

系统有 3 个库，选案模式库存储粗选模式，用于从多个执收单位中选出待选对象；收费信息数据仓库是系统的核心，用于存储系统核心收费数据，以供数据挖掘和 OLAP 分析；选案方法数据库存储选案方法相关信息，如指标信息等。

目前，笔者已成功构建收费稽查数据仓库，实现了基于决策树的选案模型和 OLAP 查询分析模块，基本实现了收费稽查选案系统。指标评估和选案方法维护模块是收费稽查选案系统的重要组成模块，也是提高系统性能的关键模块，是下一步工作的重点。

5 总结

本文在分析收费稽查选案业务的基础上，提出了支持收费稽查选案的元数据组织方法，和收费稽查主题的提取及其多维模型的构建方法，把数据仓库 OLAP 技术用于稽查选案的数据分析；把决策树方法用于收费稽查选案，并对其进行改造，建立了基于指标打分的决策树收费稽查选案模型；最后给出了收费稽查选案系统的总体设计，基本上实现了收费稽查计算机选案系统，并指出了下一步的工作方向。测试结果表明系统具有准确的选案性能，并将在沈阳市地方财政局进行试点。

参考文献

1 赵　璇，史忠植. 税务稽查选案专家系统的研究[J]. 计算机系统应用, 1999(10): 12-15.

2 安志勇. 税务稽查选案技术方法研究[D]. 太原: 太原理工大学, 2004.

3 陈　颖. 数据仓库技术在税务稽查系统中的研究[D]. 天津: 天津大学管理学院, 2004.

4 林杰斌，刘明德，陈　湘. 数据挖掘与 OLAP 理论与实务[M]. 北京: 清华大学出版社, 2003.

5 张云涛，龚　玲. 数据挖掘原理与技术[M]. 北京: 电子工业出版社, 2004.