学校编码：10384　　分类号_____密级_____
学号：X2011230280　　UDC _____

厦门大学

工　程　硕　士　学　位　论　文

数据挖掘技术在国税普通发票代开中的应用研究

Research on the Application of Data Mining In the National Tax Invoice Issuing

李国栋

指　导　教　师：　林　坤　辉　教　授

专　业　名　称：　软　件　工　程

论文提交日期：　2013　年　4　月

论文答辩日期：　2013　年　5　月

学位授予日期：　2013　年　　月

指　导　教　师：_____

答辩委员会主席：_____

2013 年 4 月

厦门大学学位论文原创性声明

本人呈交的学位论文是本人在导师指导下,独立完成的研究成果。本人在论文写作中参考其他个人或集体已经发表的研究成果，均在文中以适当方式明确标明，并符合法律规范和《厦门大学研究生学术活动规范（试行）》。

另外，该学位论文为（　　）课题（组）的研究成果，获得（　　）课题（组）经费或实验室的资助，在（　　）实验室完成。（请在以上括号内填写课题或课题组负责人或实验室名称，未有此项声明内容的，可以不作特别声明。）

声明人（签名）：

年　　月　　日

厦门大学学位论文著作权使用声明

本人同意厦门大学根据《中华人民共和国学位条例暂行实施办法》等规定保留和使用此学位论文，并向主管部门或其指定机构送交学位论文（包括纸质版和电子版），允许学位论文进入厦门大学图书馆及其数据库被查阅、借阅。本人同意厦门大学将学位论文加入全国博士、硕士学位论文共建单位数据库进行检索，将学位论文的标题和摘要汇编出版，采用影印、缩印或者其它方式合理复制学位论文。

本学位论文属于：

（　　）1.经厦门大学保密委员会审查核定的保密学位论文，于　　年　　月　　日解密，解密后适用上述授权。

（ √ ）2.不保密，适用上述授权。

（请在以上相应括号内打“√”或填上相应内容。保密学位论文应是已经厦门大学保密委员会审定过的学位论文，未经厦门大学保密委员会审定的学位论文均为公开学位论文。此声明栏不填写的，默认为公开学位论文，均适用上述授权。）

声明人（签名）：

年　　月　　日

摘 要

随着信息技术的不断发展，税务部门信息化建设日新月异，各种税收业务系统多年来已经积累了大量的涉税数据，如何有效地利用这些宝贵的数据，为税收工作服务，成为税务部门日益关注的重点。普通发票委托代开作为税务机关的一项基本业务，承载着服务纳税人、征收零散税收、打击假发票市场的职责。为提高纳税服务质量、做到应收尽收、打击假发票市场，加强普通发票委托代开的管理工作刻不容缓。

本文以数据挖掘技术为手段，对普通发票委托代开日常管理中一些关心的问题进行探讨和研究，并结合实际工作，进行知识发现的探索。在具体应用中，选择了回归分析、K-Means 聚类分析、TwoStep 自动聚类分析、Apriori 算法分析及异常检测进行挖掘和分析。通过挖掘结果结合实际业务，从中发现一些隐藏的、未知的、有用的信息和规律，为普通发票委托代开管理工作提供决策支持。

本文的研究方法和研究结果已经开始在国税系统内部进行使用，取得了一定的效果。

关键词：数据挖掘；CRISP-DM；普通发票代开

Abstract

With the continuous development of information technology, the informatization construction of the tax authorities with each passing day, all kinds of operation system over the years has accumulated a lot of tax-related data. How to effectively make use of these valuable data to work for tax services, become the focus increasingly of the tax department. The National Tax Invoice Issuing as fundamental business tax authorities, bear the Services to the taxpayers, collection of scattered taxes and cracking down on fake invoices market. It is urgent to strengthen management of the National Tax Invoice Issuing.

In this dissertation, we make use of data mining technology to do some research about the daily management of the entrusted agency for normal invoice. We combine with practical work for the exploration of knowledge discovery. In practical application, we choose the regression analysis, the K-Means cluster analysis, automatic TwoStep clustering analysis, Apriori and mining anomaly detection and analysis. Through the mining results combined with the actual business, we discover some hidden, unknown and useful information and rules, to provide ordinary invoices entrusted agency management decision support.

The research methods and research results Involves in the dissertation already are being used within the national tax system,and has obtained the certain effect.

Key Words: Data Mining; CRISP-DM; National Tax Invoice Issuing

目录

Contents

第一章 绪论

1.1 课题研究背景及意义

数据挖掘（Data Mining, DM）是一种决策支持过程，它主要基于人工智能、机器学习和数理统计等技术，高度自动化地分析原有数据，做出归纳性推理，从中挖掘出潜在的模式，从而将数据资源转换成有用的信息。数据挖掘与其他的数据仓库应用不同，它不是查询已知的、明显的信息，其本质是探测性的，试图获得隐藏的而不是明显的信息。

发票是指单位和个人在购销商品、提供或者接受服务已经从事其他经营活动中，开具、收取的收付款凭证[1]。它不仅是财务核算的原始凭证，同时也是税务机关进行税源监控和税务检查的重要依据，管理好用好普通发票，对于提高税收征管的质量和效率，堵塞税收流失漏洞意义重大。

普通发票代开是由税务机关或者税务机关委托的其他单位根据收款方或者提供劳务服务方的申请，依照法规、规章以及其他规范性文件的规定，代为向付款放或者接受劳务服务方开具普通发票的行为。普通发票代开业务的开展方便了广大零散纳税人，提升了纳税服务质量，完善了“以票控税”的税收管理模式，有利于税务机关加强零散税源的管理，堵塞税收征管漏洞，维护良好的社会经济秩序。

相对于社会、经济的快速发展，我国普通发票代开相关制度和保障措施的建设明显落后，给一些不法分子有可乘之机，虽然国税机关不断努力，但是普通发票代开的管理仍然存在一些问题和风险。这些问题和风险产生的原因比较复杂，不能仅仅局限于普通发票代开业务本身，而应从更加全面的角度理性、客观地进行分析。

随着信息技术的不断发展，我国税务系统信息化建设不断创新突破，运用现代信息技术手段对税务部门各个业务环节中积累下来的数据进行深度数据挖掘，为税收管理者和决策者提供更为专业和科学有效的决策依据，已经成为税务信息化建设的当务之急。在实施金税工程、网上申报、网上行政审批、网络开票等一系列信息化项目后，国税系统的日常管理已经从遍地撒网式的粗放型管理转变为以信息化技术为依托的精细化管理。作为日常工作中工作量比较大的普通发票管理，更是受到各级税务机关的重视。然而在实际管理过程中，目前国税系统对普

通发票业务的管理基本上处于较低层次的数据采集、比对、分析阶段，管理具有滞后性及主观性。通过科学分析的手段，利用数据挖掘等现代信息技术手段对普通发票业务信息进行知识挖掘，获取该业务中隐藏的、未知的、有用的信息，对于普通发票的管理工作有着积极重要的意义。

1.2 国内外研究现状

在国外，数据挖掘技术在税务领域成功应用的案例为数不少。比如早在 1998 年美国加州税务局就启动了基于 IBM DB2 数据库软件的综合逃税人监察项目数据仓库解决方案（INC）项目，使加州税务能够在超过 2.2 亿项的独立信息中利用商业智能技术进行业务分析。又比如 NCR Teradata 成功地实施了包括美国国家税务局（IRS）、澳洲国家税务局（ATO）等在内的数据仓库和数据挖掘项目。利用数据挖掘技术，1996 年美国国家税务局追回补交税款两亿笔，增收 200 亿美元的税金和罚款，并进行了 120 万笔帐目审计。

目前就全国税务系统的数据挖掘应用来说，还存在较低的应用层次，存在很多不足之处，例如不同应用系统生产数据缺少统一完善的数据标准和规范；信息来源多头，数据定义口径不一致；缺少集中的数据应用服务，各核心业务系统仍需承担大量的查询、统计等应用，无法“瘦身”，影响了系统运行效率；应用系统之间数据交叉、重复，缺乏数据共享机制；个别省市建立了以核心业务数据为主的数据仓库，但数据仓库的应用水平低，用户范围窄等。但可喜的是国家税务总局已经将数据分析平台纳入金税三期建设中来。预计不久的将来，全国税务系统将建成省级综合数据分析平台、总局综合数据分析平台、南海综合数据分析平台的三级综合数据分析平台。

1.3 普通发票代开现状

代开发票，是指由税务机关根据收款方（或提供劳务服务方）的申请，依照法规、规章以及其他规范性文件的规定，代为向付款方（或接受劳务服务方）开具发票的行为[4]。普通发票代开分为办税服务厅代开和委托代开两种方式，本文如无特殊说明，普通发票代开是指税务机关委托其他单位（或部门）代开增值税普通发票的委托代开行为。

以西南某市国家税务局普通发票代开系统为例，截止 2012 年 12 月 31 日全市 5 个区局共签订委托代征协议单位 35 个，设置代开点 72 个。委托代征单位有

3 种类型：1 是办事处；2 是专业市场；3 是金融机构。目前代开系统将机构类型设置为：办事处和专业市场两种类型（金融机构设置为办事处）。其中办事处代开范围仅限临时经营个人行为；专业市场除临时经营个人行为外还包括主管税务机关辖区内在管户中的个体户和个人。

系统自 2007 年上线运行以来，累计代开普通发票 108 万余份，征收税款 7.78 亿元。系统经历了从单机版到网络版的演变，从加强征管、堵塞税收漏洞、规避法律风险等方面有了一定的提高。2012 年，全市通过代开系统代开普通发票 306292（不含作废）份，征收税款 2.8 亿元，比上年度增长 17%。

2012 年，全市普通发票代开总计征收税款 3.315 亿元，其中通过委托代征的方式征收税款 2.83 亿元，大厅代开征收 0.48 亿元。2010-2012 年期间，代开征收税款增长 112%，其中委托代征税款增长 163%，如图 1-1 所示。

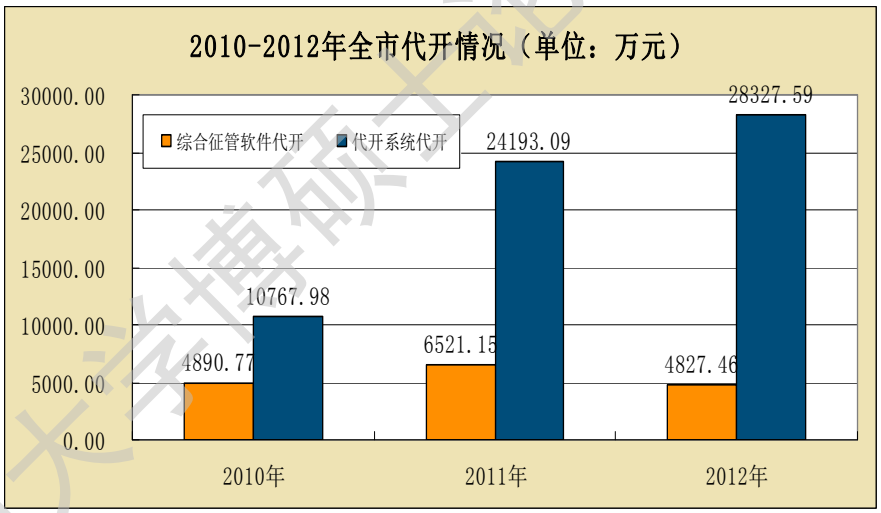

图 1-1 2010-2012 年某市普通发票委托代开情况

综合全市，普通发票代开委托代征税款由 2010 年的 69%提高到 2012 年的 85%，同比提高 16 个百分点，如图 1-2 所示。

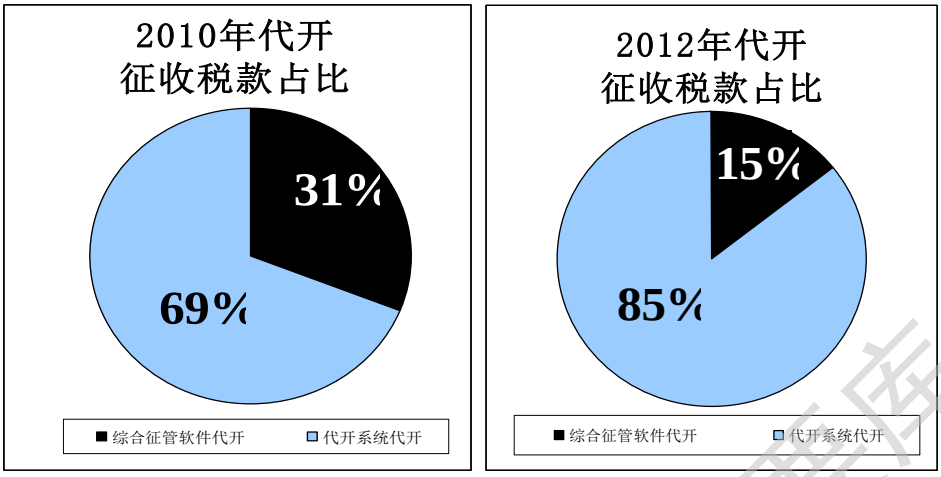

图 1-2 某市 2010 年与 2012 年代开征收税款占比比较

1.4 论文的研究内容与结构

本文以某市国家税务局普通发票代开系统数据为研究对象，参照 CRISP-DM 模型的“商业理解、数据理解、数据准备、建立模型、模型评估、结果部署”六个步骤，选择了回归分析、K-Means 聚类分析、TwoStep 自动聚类分析、Apriori 分析及异常检测进行挖掘和分析。

本文共分为五章。

第一章绪论，主要介绍课题的背景和研究的方法及内容。

第二章主要研究和介绍了数据挖掘技术的相关知识和应用现状，并阐述了数据挖掘过程的一些模型和研究方法。

第三章主要对普通发票代开业务进行介绍和分析，指出普通发票代开具有一定的风险性。

第四章以第二、三章知识点为基础，结合实际，选择了二项 Logist 回归分析、K-Means 聚类分析、TwoStep 自动聚类分析、Apriori 分析及异常检测进行挖掘和分析。

第五章总结前四章，并对数据挖掘技术在税务系统中的应用进行了展望。

第二章 系统相关技术简介

普通发票代开业务的数据分析主要采取数据挖掘的技术进行研究。本文中具体的技术及算法包括 CRISP-DM 模型、回归分析、K-Means 聚类分析、TwoStep 自动聚类分析、Apriori 分析及异常检测等，采用的数据挖掘工具是 SPSS Clementine。

2.1 数据挖掘的概念

2.1.1 数据挖掘的定义

简单地说，数据挖掘是指从大量数据中提取或“挖掘”知识。

数据挖掘，又被称为数据库知识发现（Knowledge Discovery from Database KDD）,它是一个从大量数据中提取、挖掘出未知的、有价值的模式或规律等知识的复杂过程[2]。它通常与计算机科学有关，并通过统计、在线分析处理、情报检索、机器学习、专家系统（依靠过去的经验法则）和模式识别等诸多方法来实现上述目标。

许多人把数据挖掘视为 KDD，而另一些人只是把数据挖掘视为知识发现过程的一个基本步骤。本文所指的数据挖掘是：数据挖掘是从存放在数据库、数据仓库或其他信息库中的大量数据中发现知识的过程[2]。

2.1.2 数据挖掘的产生与发展

人们在长期的社会生活和时间中总结了许多经验和教训，这些内容对于人们深层次地了解事物的发展规律有着重要的指导意义。因此人们将生活和实践的经验以及这些经历中产生的的经验和教训以各种形式记录下来，如以文本的形式、音频的形式、视频的形式等。从广义上来说，这些被记录下来的内容就是数据。

随着人类社会的飞速发展，人类社会活动越来越频繁和多样化，人们工作生活所积累的数据和信息也日益增长。人类近 30 年来所掌握的信息量占有史以来积累总量的 90%。加之计算机网络的迅猛发展，数据积累呈爆炸式增长。以商业领域为例，美国著名的连锁超市 Wal-Mart 的数据库中已积累了 TB 级以上的顾客购买行为数据和其他销售数据。在电子商务领域，各类网上书店、网上银行、网上营业厅和网上商城等积累的 Web 点击流数据，存储容量也多高达 GB 级。

面对海量数据，如何充分和有效的利用成为当务之急。大规模海量数据的整合处理和深层次量化分析的实际需求，直接孕育了 20 世界 90 年代初期的两项重

大技术，这就是数据仓库技术和数据挖掘技术。

2.1.3 数据挖掘的过程

数据挖掘是通过自动或半自动化的工具对大量数据进行探索和分析的过程，其目的是发现其中有意义的模式和规律。

目前数据挖掘领域最权威的过程模型是 CRISP-DM 模型，它是目前事实上最权威的行业标准。CRISP-DM（Cross-Indusry Standard Process for Data Mining），即“跨行业数据挖掘过程标准”。

CRISP-DM 模型定义了 6 个阶段，分别是商业理解（Business Understanding）、数据理解（Data Understanding）、数据准备（Data Preparation）、建立模型（Modeling）、模型评估（Evaluation）、结果部署（Deployement）[3]。

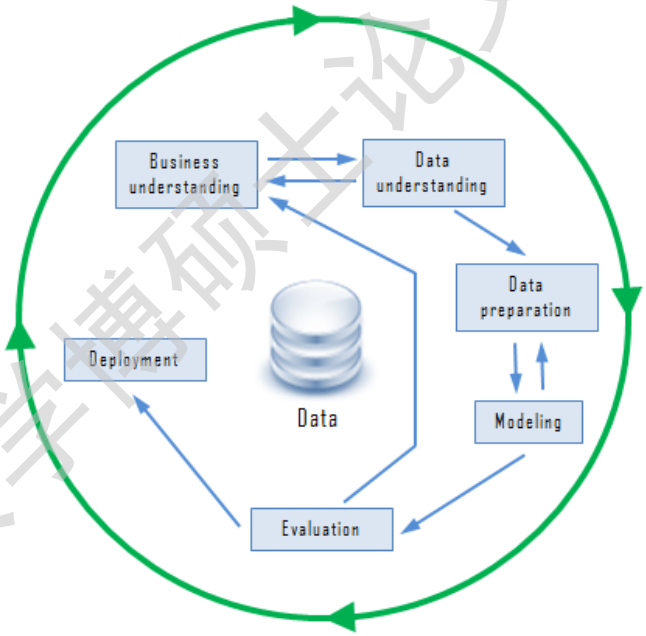

图 2-1 CRISP-DM 模型

CRISP-DM 模型为一个 KDD 提供了一个完整的过程描述。一个数据挖掘项目的生命周期包含 6 个阶段。这 6 个阶段的顺序是不固定的，经常需要前后调整这些阶段，这依赖每个阶段或是阶段中特定任务的产出物是否是下一个阶段的必须的输入。图 2-1 中箭头指出了最重要和依赖度高的阶段关系。

图 2-1 的外圈象征数据挖掘自身的循环本质——在一个解决方案发布之后的一个数据挖掘过程才可以继续。在这个过程中得到的知识可以触发新的过程，经

常是更聚焦的商业问题。后续的过程可以从前一个过程得到益处。

1. 商业理解

最初阶段的阶段集中在理解项目目标和从业务的角度理解需求，同时将这个知识转化为数据挖掘问题的定义和完成目标的初步计划。

2. 数据理解

数据理解阶段从初始的数据收集开始，通过一些活动的处理，目的是熟悉数据，识别数据的质量问题，首次发现数据内部属性，或是探测引起兴趣的子集去形成隐含信息的假设。

3. 数据准备

数据准备阶段包括从未处理的数据中构造最重数据集的所有活动。这些数据将是模型工具的输入值。这个阶段的任务有的需要执行多次，没有任何规定的顺序。任务包括表、记录和属性的选择，以及为模型工具转换和清洗数据。

4. 建立模型

在这个阶段，可以和应用不同的模型技术，模型参数被调整到最佳的数值。一般，有些技术可以解决一类相同的数据挖掘问题，有些技术在数据形成上有特殊要求，因此需要经常跳回数据准备阶段。

5. 模型评估

到项目的这个阶段，已经从数据分析角度建立了一个高质量的模型。在开始最后部署模型之前，需要彻底地评估模型，检查构造模型的步骤，确保模型可以完成业务目标。这个阶段的关键目的是确定是否有重要业务问题没有被充分考虑。在这个阶段结束后，将决定一个数据挖掘结果是否可以付诸使用。

6. 结果部署

通常，模型的创建不是项目的结束。模型的作用是从数据中找到知识，获得知识需要以便于用户使用的方式重新组织和展现。根据需求，这个阶段可以产生简单的报告，或是实现一个比较复杂的、可重复的数据挖掘过程。在许多案例中，这个阶段是由客户而不是数据分析人员承担部署的工作。

2.1.4 数据挖掘的方法

数据挖掘所涉及的学科领域和方法很多，目前比较成熟且应用广泛的方法主要有分类方法、聚类方法、关联规则分析、异常检测、预测方法等。