文章编号：1007－5399（2013）02－0008－03

应用产生式规则的邮政地址标准化方法研究

谷 斌，靳艳峰，张 昶

（中国邮政集团公司商函研究中心，河北 石家庄 050021）

摘 要：文本结构化是中文处理的重要内容，包括分词、关键字提取等都与英文处理有较大差异。在行业应用中，中文短句特别是地址处理和标准化问题更是基于地址整合客户数据的核心问题。专家系统作为人工智能最重要的应用领域，包含了领域专家的大量知识。文章在分析地址数据结构化方法的基础上，提出一种基于专家系统的地址标准化方法。该方法将混乱的中文地址处理为完整、标准、可用于精确匹配的地址。

关键词：文本挖掘；中文分词；关键字；专家系统；地址；标准化

中图分类号：F61 **文献标识码**：A

专家系统是人工智能中最重要、最活跃的应用领域，它实现了人工智能从理论研究走向实际应用、从一般推理策略探讨转向运用专门知识的重大突破。1965 年，世界上第一个专家系统 dendral 问世，可以推断化学分子结构，目前，一些人把 dendral、macsyma 称为第一代专家系统。20 世纪 70 年代中期，mycin、casnet、prospector、hearsay 等一批卓有成效的专家系统相继研制成功，人们称这一批专家系统为第二代专家系统。20 世纪 70 年代后期，专家系统已基本成熟。

第三代专家系统属多学科综合型系统，采用多种人工智能语言，综合采用各种知识表示方法和多种推理机制及控制策略，并开始运用各种知识工程语言、骨架系统及专家系统开发工具和环境来研制大型综合专家系统。在总结前三代专家系统设计方法和实现技术的基础上，目前已开始采用大型多专家协作系统、多种知识表示、综合知识库、自组织解题机制、多学科协同解题与并行推理、专家系统工具与环境、人工神经网络知识获取及学习机制等最新人工智能技术来实现具有多知识库、多主体的第四代专家系统。

1 中文地址概述

地址是人员信息中重要的组成部分，是联系客户、分析客户家庭情况和其他信息的入手点。企业中的客户数据经常分散在不同来源的系统中，造成所谓信息孤岛。不仅如此，由于同一客户的信息无法整合在一起，也会造成单一数据来源的客户信息常常缺乏足够用于分析的属性和特征。在日益看重客户资源和客户关系管理的现今时代，拥有大量客户信息，但无法实际分析客户属性的情况难以被接受。通常情况下，用于连接不同表或数据库，将同一个客户或同一个家庭信息整合起来最直接、最方便的属性是地址。

在邮政行业内，一般将未划分过的非结构化地址信息称为行地址。行地址是最常见的原始客户地址数据形式，可以将地址看成非结构化、半结构化的文本数据，即将行地址划分为不同的地址段，分别用于不同处理。通常完整的、标准的地址较长，用户留存的时候往往有缩写、错写情况，地址在录入业务系统时有时也会出现错误。因此，虽然每个地址不完全相同，但可以判断出有些记录对应的是一个客户，有些记录对应的是一个家庭。类似情况在绝大部分企业的客户表中是大量存在的。在此种情况下，使用连接操作、like 模糊查询等 SQL 命令无法完成根据地址属性整合个人或家庭信息的工作。

为了解决这一问题，在处理之前，首先必须对客户地址进行数据清洗，补充缺失的部分，将客户地址整理为标准地址格式，这样才可能实现利用地址连接不同数据表和不同家庭成员的任务。

地址数据清洗在邮政企业内是通过收集行政区划范围内完整的、标准的地址写法，然后对客户地址加以逐段匹配实现的。但是，即使拥有最全面的街道、邮编、小区名，数据库也无法全面处理所有可能出现的缩写、错写情况。在处理整个行地址时，经常出现由于部分地址段无法匹配而造成整条地址无法处理的情况，因此仅依靠标准数据库实现客户地址清洗、标准化的方法难以达到目的。

本文旨在提出一种使用产生式规则定义地址分段的方法，在清理地址前将非结构化的原始客户地址结构化，并在此基础上完成地址清理和标准化工作。

2 地址的结构化方法研究

地址数据由自然地理的不同部分相互结合构成。通常标准地址包括以下部分。

P/C1/C2/R1/R2/N/X1/X2/B/F/D/Rn（2－1）

P 省名	C1 城市名
C2 区县名	R1 主要道路名
R2 附属道路名	N 门牌
X1 小区名称 1	X2 小区名称 2
B 楼号	F 楼层号
D 单元号	Rn 房门号

其中部分地址段在日常书写时经常省略，如：省名、区县名、附属道路名、楼层号。其他部分根据用户所在地区的习俗不同也会出现省略和其他写法。正是由于地址段的灵活取舍造成了相同地址千差万别的写法。

通过标准地址段自由组合，可以大致覆盖用户地址的不同写法，但是这个方法在大规模客户地址分析时效率较低，特别是客户地址存在部分错写时，很难实现精确匹配。

国内地址虽然没有类似美国约定俗成的分界符可用于地址段的划分，但是部分地址段存在一些具备典型语义特征的字符可以用来完成划分。例如，大部分城市的街道命名规则均以“路、道、街”为主，辅以“里、弄、条”等关键字。门牌通常是数字或带有“号”字，有时存在“付 X 号”的特殊写法。根据这些具备特殊语义的字符可以大致划分主要地址段，并对划分出的段进行段含义的定义，即标定此地址段是哪一段信息。

3　地址的分段与表示方法

领域知识的表示方法是专家系统实现的关键环节，专家系统的知识表示模型既要有利于计算机的存储和使用，又要能有效描述人类专家在专业领域的理论和经验知识。产生式专家系统采用产生式规则知识表示方法。产生式规则是一个“if 前件 then 后件”形式的语句，前件为规则的条件，后件为规则的结果。产生式规则具有结构简单、表达便利、应用广泛的特点。

在本系统中典型的分段及段定义规则描述如下。

If 包含段关键字 ∨ 标准地址中不存在对应段，then 当前信息为指定段信息。

典型的路名段关键字包括：路、道、里、街、巷等。其他段也具有相应关键字。

实际上，除不同地址段拥有不同段关键字外，地址信息还是一种特殊的有序数据。地址序列在东西方的文化中非常明显。典型的西方地址写法是从小范围到大范围，即从最小区域段到最大区划段，而国内标准写法则与之相反。在设计规则时，应考虑到国内通常是以区划名、街道名、门牌、小区名、楼号、单元号、房门号等顺序填写地址。在处理地址分段赋予段以相应含义的时候，还应依据地址段的序列信息来判断，即按中文地址书写顺序来判断段的含义。特别是在部分段无法直接由关键字分辨的情况下，序列顺序则是重要的判定依据。

例如，建华南大街 10 号省种子公司宿舍 2 号 1－604（3－1）即为典型的情况。

门牌号、楼号、房门号通常由数字或带“号”关键字构成。仅通过“号”关键字无法判断 10 号和 2 号分别表述什么信息。但是由于 10 号在前，则此段为门牌号，而 2 号为楼号。类似的，国内在书写楼号、单元号、楼层号、房门号时，经常使用“－”分隔不同段，缩写成：建华南大街 10 号省种子公司宿舍 2－1－604（4－2）。

这种情况下可以分为两种模式处理，即 A－B－C 和 A－B－C－D 模式。第一种模式代表了楼号－单元号－房门号，而第二种模式代表了楼号－单元号－楼层号－房门号。

通过以上设计方法可以得到对常见地址的标准处理规则，但是由于我国幅员辽阔，地址千差万别，存在大量与分段处理关键字重合的特殊地址信息。

例如：朝阳区三间房西里 6 楼 2 门 302（4－3）依据标准段处理关键字分段、段定义后得到的结果见表 1。

表 1　地址分段结果

段序号	地址段
1	朝阳区
2	三间房
3	西里
4	6 楼
5	2
6	302

但是已知三间房西里是一个完整的街道、小区名，表 1 将其分为“三间房”和“西里”是因为“房”字为一个房门号的分段处理关键字。

解决这个问题可以单独保存特定地址信息为关键字，在分段时单独处理，见表 2。

表 2　分段地址表

段定义	地址段
行政区划	朝阳区
街道、小区名	三间房 西里
楼号	6 楼
单元号	2
房门号	302

4　地址的系统验证

为验证系统处理地址的质量，分别选取河北省石家庄市 4.6 万条和北京市 2 万条用户地址进行地址分段和段定义。在石家庄市地址中，选择长安区作为试点，专门设计实现了特殊关键字库。

石家庄市的数据结果表明，在使用特殊关键字库的情况下，分段、段定义准确率要高于仅通过标准关键字处理的方法。考虑到存在部分特殊关键字覆盖石家庄市所有区划的情况，所以石家庄市的数据正确率受到影响，仅有部分提高。

在得到结构化标准客户地址后，按地址整合个人地址和家庭地址。查找出的重合个人和家庭数量要远远高于直接按原始客户地址查找的方法。在石家庄市长安区 8 670 条客户记录中，共发现地址相同的记录 1 262 条，而直接在客户地

精细化管理提升存货管理水平

存货管理水平的好坏，直接影响着存货品质是否优良、规模是否合理、流通是否有序。如何才能提升存货管理水平，提高企业的经营效益，笔者认为应从以下两方面坚持精细化管理。

1　细化职责，确保管理到位

职责是否明确关系到人员是否越位、空位、缺位。应坚持仓库保管人员与存货管理人员职务相分离，存货采购与存货验收、保管人员相分离，采购工作与采购计划审批相分离。具体来说，存货管理人员应负责存货出、入库台账登记，定期组织盘点存货，编制各类统计报表，保证存货台账与实物一致。对于纳入业务系统管理的专业存货，如通信票、集邮票品、普通邮资封片、分销配送商品等，业务、财务双方相关人员应核对进销存、收入、成本和用户欠费，确保信息系统数据相符。若发现数据不相符，财务部门应积极会同业务部门共同查找原因：如因账务处理错误，本着“谁出错谁调整”的原则，由具体办理人员写清事情原委，经双方负责人签字确认后，及时进行账务调整；如因库管人员失职发生存货短少，应由库管人员按价全额赔偿。仓库保管员应负责各类实物的接运、验收、入库存储、发放等具体执行工作，建立仓库实物台账登记；按照仓库现场管理的要求做好相关现场管理工作。

2　细化流程，确保程序科学

一是采购环节，存货采购应按照上级部门下发的相关文件规定，由本单位统一组织采购，属上级部门采购范围内的，应认真做好物品的汇总、上报和要数工作。二是入库环节，应严格执行到货验收交接制度，货物送到后，存货管理人员与仓库保管员必须按供货合同、清单等查验品种规格、检验货物质量，核准数量，实际到货核实后仓库保管员在相关单据上签字确认。未经验收的物品应拒绝入库列账，如发现数量短少、质量残次、品种不符等问题，及时与供货单位交涉，并妥善处理。三是调拨环节，坚决摒弃部门之间自行调拨库存商品的行为；确需调拨的，应按相关文件报请上级部门批准后方可调拨。四是出库环节，领用人员到相关管理人员处开具《请领单》，由相关部门负责人签字审批，仓库保管员根据审批后的《请领单》，按照“先进先出”的原则发放实物。领取时，领用人员应当面核准数量、规格、型号等，仓库保管员和领用人员在《请领单》上签字确认。五是退回环节，库存商品发生退回时，存货管理员、仓库保管员、退货人员三方应当面核准数量、质量情况、规格型号等，由存货管理员开具《入库单》，注明退回原因，三方共同在《入库单》上签字确认。六是盘点环节，库存管理人员应会同相关人员定期对仓库实物进行全面清查，就盘存情况编制《库存商品盘点表》，经盘点人和监盘人签字确认。对盘点出的盈、亏情况，报主管部门批准后按规定及时处理，涉及仓库保管员责任短缺的，要追究其赔偿责任。七是记账环节，存货管理人员应根据《入库单》、《请领单》等逐日逐笔登记《库存商品台账》，建立仓库实物台账，以利于实物的核对和发放。而且，存货管理人员、仓库保管员应定期核对账目，以保证账账相符、账实相符。

（安徽省亳州市邮政局　常冠林）

址中仅能查找到 318 条同地址记录。

5　结论

本文按照中文地址的语义关键字和语序结构设计了地址分段、段定义规则，实现了地址处理专家系统。经过石家庄和北京 6 万条客户地址的测试，本系统能以较高的准确率将非结构化的行地址处理为结构化的标准地址信息。通过本系统，可以更多、更准确地发现分散在企业不同客户数据库中的相同用户和用户家庭信息，为企业整合客户信息和开展深入的客户发掘提供了数据基础。

参 考 文 献

1　An efficient approach computing edit distances and edit paths，Zou xukai，mini-micro systems，jul，1996，vol. 17，No. 7

2　A Chinese organization full name and abbreviation matching algorithm based on edit-distance，Huang lin-sheng，Deng zhi-hong etc.，journal of shandong university（natural science），May 2012，vol. 47，No. 5

3　ROB CALLAN. 人工智能. 北京：电子工业出版社，2004

4　Thomas Dean，James Allen，Yiannis Aloimonos. 人工智能——理论与实战. 北京：电子工业出版社，2004

5　何新贵. 知识处理与专家系统. 北京：国防工业出版社，1990

6　张金寿，周建峰. 专家系统建造原理及方法. 北京：中国铁道出版社，1992

7　王碧泉，范洪顺，陈佩燕，王春珍. 专家系统及其在地震预报中的应用. 北京：中国科学技术出版社，1993

8　戴汝为. 人工智能. 北京：化学工业出版社，2002

收稿日期：2012－10－19

作者简介：谷斌（1978～），男，河北石家庄人，硕士，讲师，主要从事数据挖掘、人工智能、电子商务研究；靳艳峰（1981～），男，河北灵寿人，硕士，讲师，主要从事数据挖掘、数据分析研究；张昶（1986～），男，河北石家庄人，硕士，主要从事数据挖掘、数据分析研究。

注：本文研究成果取自河北省教育厅项目《名址信息智能分析方法研究》，项目编号 Z2010313