# 从数据采集到数据挖掘

谢邦昌　李扬　匡宏波　北京商智通团队　编著

FROM CATI TO DATAMINING

中国统计出版社
China Statistics Press

# 从数据采集到数据挖掘

谢邦昌　李扬　匡宏波　北京商智通团队　编著

FROM CATI TO DATAMINING

中国统计出版社
China Statistics Press

(京)新登字 041 号

图书在版编目(CIP)数据

从数据采集到数据挖掘/谢邦昌,李扬,匡宏波,北京商智通团队编著.
—北京:中国统计出版社,2009.3
ISBN 978-7-5037-5643-6

Ⅰ.从…
Ⅱ.①谢… ②李… ③匡… ④北…
Ⅲ.数据采集
Ⅳ. TP274

中国版本图书馆 CIP 数据核字(2009)第 024058 号

从数据采集到数据挖掘

作　　者/谢邦昌　李扬　匡宏波　北京商智通团队
责任编辑/吕　军
装帧设计/艺编广告
出版发行/中国统计出版社
通信地址/北京市西城区月坛南街 57 号　邮政编码/100826
办公地址/北京市丰台区西三环南路甲 6 号
网　　址/www.stats.gov.cn/tjshujia
电　　话/邮购(010)63376907　书店(010)68783172
印　　刷/河北天普润印刷厂
经　　销/新华书店
开　　本/880×1230mm　1/32
字　　数/180 千字
印　　张/6.75
版　　别/2009 年 3 月第 1 版
版　　次/2009 年 3 月第 1 次印刷
书　　号/ISBN 978-7-5037-5643-6/TP·47
定　　价/24.00 元

# 序

现在的时代已经进入“信息时代”，对数据的采集与挖掘成为一项十分重要、同时又具有挑战性的任务。计算机辅助电话调查方式的出现，顺应了历史发展的需求，充分利用现代科学提供的工具和技术手段，提高了社会、经济领域研究中数据采集的效率与质量，已经成为目前国际上发达国家采集数据的标准模式。同时，在计算机辅助电话调查不断发展的过程中，也给我们带来了一系列值得研究与思考的问题。随着这种调查手段的不断普及与发展，数据的采集越来越便利，数据量越来越大。如何运用这些数据，挖掘其背后隐含的信息，是对理论研究者和实际工作者的又一挑战。

谢邦昌教授是市场调查的专家，也是数据挖掘的专家，多年来一直从事相关理论与应用实践研究。谢教授与大陆统计学界有密切交往，始终不渝地、积极热情地推动海峡两岸的学术交流。在谢教授的关注与指导下，中国人民大学统计学院在全国高校中率先成立了计算机辅助电话调查实验室与数据挖掘中心，对教学和科研的发展起到了积极的辅助和推动作用，取得了十分显著的成效。今天，由谢邦昌教授和中国人民大学统计学院的研究团队共同完成这本著作，就成为一件水到渠成的事情。

《从数据采集到数据挖掘》一书共分八个章节，从计算机辅助电话调查的起源出发，按照“初识 CATI→基础理论→操作技巧”的认知思路系统介绍了 CATI 的概念、发展与操作流程。

通过贯穿全书的实际案例演示，深入浅出的向读者展现了计算机辅助电话调查的实施流程与技巧，具有非常重要的实践意义。使得本书既可以作为初学者了解相关知识的入门教材，同样也可以作为实际工作者的操作技术手册。与其他相关书籍不同的是，作者特别在第八章“计算机辅助电话调查的数据挖掘实践”中立足实际案例，对数据挖掘技术与计算机辅助电话调查的结合研究展开讨论，对市场研究未来的发展方向是一个重要且可行的启示。

金勇进

2009 年 2 月 15 日

# 前　言

自20世纪70年代诞生以来，计算机辅助电话调查以其可控性高、时效性强等特点越来越为研究者所接受。在信息挂帅的今天，CATI系统更被视为收集资料、分析数据的利器，在商业、学术以及政府调研行为中得到了广泛应用。本书以计算机辅助电话调查的过去、现在和未来为主线，结合CATI调研项目实践操作，以ATHENACATI为工具向读者介绍了从数据采集到数据挖掘的过程。

本书适合多层次多专业如统计、市场营销管理类专业的研究生、本科生、专科生学习，可作为相关专业学生课外数据采集实践的操作手册，还适合于从事相关数据采集、数据分析的人员阅读。

本书在编写及出版过程中，得到了中国人民大学统计学院金勇进院长及相关教授指导及鼓励与调查技术研究所、中国人民大学数据挖掘中心、台湾辅仁大学统计资讯学系、中国统计出版社以及北京商智通信息技术有限公司的大力支持。并特别感谢中国人民大学统计学院张晓牟同学、张兰兰同学、刘冬同学，首都经济贸易大学统计学院王晓雪同学、徐竞雄同学付出的辛苦劳动。

计算机辅助电话调查具有相当广泛的应用前景与价值，本书难以全部阐明。谨以此书作为向多年实践工作的献礼，提供给诸位参考。书中难免有疏漏之处，恳请读者多提宝贵意见，以便今后进一步修改与完善。

编　者

2008年12月

# 目　录

# 第1章　计算机辅助电话调查概述

## 1.1　调查概述

19世纪以来，随着社会工业化和现代化的迅速发展，人们越来越迫切地感到，只有对社会和人类行为有更深入的认识，才能有效地制定社会政策、进行社会管理、预测社会发展，才能解决由急剧的社会变迁所产生的一系列社会问题。在这一要求的推动下，社会调查作为人们认识社会、改造社会的一种手段，逐渐在社会科学和各个工作部门得到广泛应用。目前，它已成为研究社会现象的主要方法之一。

### 1.1.1　社会调查的含义及特征

1. 社会调查的含义

所谓社会调查，就是人们有目的、有意识地通过对社会现象的考察、了解，收集资料并分析、研究，以认识社会生活的本质及其发展规律的一种科学活动。社会调查以各种社会现象为研究对象，收集相关资料，运用包括统计学、经济学、社会学、逻辑学、社会心理学等多方面的知识，透过现象揭示事物的本质和发展变化规律，并进而寻求改造社会的途径和方法。

2. 社会调查的特点

社会调查的特点是：(1)直接在社会生活中系统地收集资料；(2)利用第一手资料进行分析和研究；(3)以分析和研究社会现象为目的；(4)在实践中形成理论并检验理论。这些特点将社会调查研究与其他认识活动和其他研究方式区分开来。

3. 社会调查的本质特征

一是社会目的性。社会调查的目的归纳起来不外乎以下几个方

面：服务于政府，提供决策参考；服务于企事业单位，提供管理参考；服务于社会，提供社会生活及个人行为准则之参考；服务于科学理论，作为科学理论的佐证。由于社会调查的目的在调查之前是明确的，因此，反映的事物必然带有方向性或倾向性，选择的对象也会具有典型性。

二是系统性。人类社会是一个复杂而庞大的系统，是一个不断变化和发展的有机体，因而我们调查任何一个社会问题都不可能是单方面的。在社会调查中，不仅要考虑调查对象和调查内容的系统性，还要考虑调查方法的系统性。如，在运用典型调查、个案分析时，也要同时考虑是否与抽样调查、整体分析结合起来。否则可能会出现研究结论的误差。

三是方法上的先进性。现代社会现象的多样性和复杂性决定了社会调查方法必须具有先进性。调查手段必须是多种方式的结合；对调查资料的分析和研究采用定性分析和定量分析相结合。而其核心是电子计算机及其 SPSS(SPSS 是软件英文名称的首字母缩写，原意为 Statistical Package for the Social Sciences，即“社会科学统计软件包”。但是随着 SPSS 产品服务领域的扩大和服务深度的增加，SPSS 公司已于 2000 年正式将英文全称更改为 Statistical Product and Service Solutions，意为“统计产品与服务解决方案”，标志着 SPSS 的战略方向正在做出重大调整。)等统计分析软件包的广泛运用。现代社会调查中，定量分析越来越复杂化，规模也越来越大，复杂数据的处理必须依靠计算机才能最后完成，如回归分析、聚类分析、因子分析等。

四是整体性。社会调查不仅仅是一项收集资料的工作，它包括资料收集和资料的加工分析研究两个环节，因而调查与研究二者不能割裂开来，而且，在现代社会调查中，研究性更是其突出的特点。不仅资料的分析具有研究性，在资料收集的过程中也具有研究性，因为，在调查对象的选择、直接收集事实资料的过程本身，也必须包括人们的思维加工。

### 1.1.2 社会调查研究的分类

社会调查研究可以从各种角度、按不同的标准划分为不同的类

型。各种类型具有各自的特点，它们在调查方式、方法、步骤、程序、适用范围等方面都有所不同。一项调查研究应当首先根据调查任务和调查课题来选择和确定适当的调查研究类型，这样才能有效地制定调查方案，确定调查对象、调查方法和调查程序。

划分类型的标准是多种多样的，包括：

——根据调研任务的性质，或划分为理论性调查研究和应用性调查研究（简称理论研究与应用研究）；

——根据调查研究对象的范围，可分为普查（或整体调查、全面调查）、抽样调查、典型调查和个案调查；

——根据调查研究的作用和目的可分为探索性调查研究、描述性调查研究和解释性调查研究；

——根据调查的时间性，可分为横剖式调查研究和纵贯式调查研究（简称横剖研究与纵贯研究）；

——根据调查的基本方式方法，可分为统计调查（或问卷调查）与实地研究（或蹲点调查）；

——根据调查研究的层次，可分为宏观调查研究与微观调查研究（简称宏观研究与微观研究）；

——根据调查的区域性，可分为农村调查与城市调查、地区性调查与全国性调查等；

——根据调研题目的范围，可分为综合性调查与专题性调查，前者的内容比较广泛，涉及的领域较多；后者内容比较单一，针对性较强；

——根据调查研究的领域，可分为各种专题调查，如家庭调查、舆论调查（民意测验）、人口调查、企业调查、市场调查、犯罪调查、劳动问题调查、教育问题调查、民族问题调查、社会福利调查、社区调查等等；

——还可以根据数据分析方法，将调查研究区分为定性研究与定量研究。前者是采用观察、访问等方法收集文字资料，然后对材料进行定性分析；后者是对由问卷、调查表、统计报表收集来的数据资料进行定量分析。当然，这种划分不是绝对的，因为采用统计分析方法并非是不对资料作定性分析和理论分析。实际上，许多利用统计资料的研究是既有定量分析，也有定性分析。

由以上分类标准可以看出，社会调查研究的分类是多角度或多维的，每一项具体的调查研究都可以按各种分类标准归为多种类型。例如，调查城市居民对物价上涨的态度，既是一种应用性研究和描述性研究，也是一种城市调查和民意调查。又如，抽取不同企业调查各种所有制形式与工人的生产积极性之间的关系，就属于理论性研究、解释性研究、抽样调查、企业调查等。

### 1.1.3 社会调查方法

1. 社会调查方法

社会调查方法是一种采用实证方式获取社会信息的手段，它通过直接实地调查，收集实在的数据，进行统计分析，进而推出结果。具体地说，就是通过向被访者询问问题来搜集资料，然后对资料进行统计分析的社会研究方法。在现代社会，社会调查是一种具有一定社会服务目的，运用现代科学方法，系统地了解社会现象、收集经验材料并对其进行分析研究，得出规律性结论的过程。

2. 社会调查的特点

对于调查研究方法，可以从以下三点来理解：第一，询问作为调查研究的基本要素，是一个科学测量的过程。在调查研究中，对于询问的第一个步骤，都要进行理论上的检验，而要想实现对询问过程的理论检验，必须先确立标准化询问规范。无论是当面访问或是电话访问，要检验的其实就是实际询问过程对标准化询问规范的偏离程度。第二，选取有代表性的被访者。在调查研究中，如果不是无一遗漏地询问每一位被访者，而是从研究总体中抽出一部分来询问，就存在一个合理挑选样本的过程。如果抽出的样本能代表总体，那么这个抽样过程就是合理的；如果不能代表总体，则是不合理的。由于大部分研究都需要进行抽样，因此，抽样调查(sampling survey)几乎成了调查研究的同义语。第三，数据的统计是完成调查研究的必要环节。

由上可见，调查研究是一项综合了多项技术的研究方法。调查本质上是一个测量过程，抽样和统计分析技术的完善，进一步扩展了调查的应用范围。由于概率抽样也是统计学的一个分支专业，因此，当代调查方法比较贴切的名称应该是统计调查。

3. 收集资料方法

在传统的询问调查中收集资料的方法主要有三种，即邮寄访问、面访和电话访问，其中电话访问是近30年才迅速发展起来的。目前国内外资料一致证实：传统的面访日益困难，因为总体资料不易取得，使得访问率下降，调查质量日益低落，成本高且费时。邮寄访问法回收率偏低，较难掌握受访者，数据质量不稳定，通常只适合作为参考之用。电话访问成本低、效率高、时间容易控制，目前被广泛采用，已成为调查研究的主要方法。

随着电话访问的迅速发展，为应对大规模统计调查和实时性民意反应，解决调查中种种统计问题，计算机技术与电话访问之结合的产物——计算机辅助电话调查系统(CATI)应运而生。

## 1.2 计算机辅助电话调查及应用范围

### 1.2.1 机辅助电话调查(CATI)概述

1. CATI 系统

计算机辅助电话调查系统(Computer Assisted Telephone Interviewing System)，简称 CATI 系统，是利用计算机辅助电话调查而开发的调查访问操作系统，是近年来高速发展的通讯技术及计算机信息处理技术与传统的电话调查相结合的产物。CATI 系统自20世纪70年代在美国诞生以来，以其无可比拟的优势在美国、西欧等地得到迅速推广应用，在许多调查中已取代传统的面对面访问方式，成为一种重要的调查方法。CATI 系统在我国的应用虽然较晚，但已有不少信息咨询公司和官方统计机构运用 CATI 系统开展调查，正得到越来越广泛的应用。

在市场经济条件下，经济主体要求以最低的费用、最快的速度获取高质量、全方位的信息。实践经验证明，要想以最低的成本、最快的速度、最高的效率掌握一定范围内大多数人最可靠的意见，最好的方法就是采用计算机辅助电话调查。CATI 系统是一个由电话、计算机、访员组成一体的访问系统。其整套系统的硬件包括：一台起总控作用的计算机总机，若干台与主机相连的 CRT(Cathode Ray Tube)终端，耳机式电话和鼠标，若干台起监视作用的计算机和配套的音像设备等。整套系统的软件包括：预览式呼出和智能预拨号系

统，项目管理系统，问卷设计系统，项目监控，工作站监控（监听、录音、发消息）和简单统计系统等。

2. CATI 系统的工作方式

CATI 系统通常的工作形式是：开始工作时，计算机首先会自动拨号并保存拨号记录。其次，如果拨号成功，终端屏幕会提示访问员筛选被访者。对于被访者不在家或当时没有空的情况，CATI 系统会自动地储存该被访者的电话号码和下次访问的时间，届时该号码会自动地出现在拨号系统中。第三，访问开始时，终端屏幕会显示出问题，经访员向电话对面的被访者读出问题，并将受访者的回答结果输入计算机，此后，计算机会根据答案自动地跳到下一道相关的题目。第四，在访问进行中，督导员可以通过监视终端随时了解访问进展的情况，包括调查结果和每个访问员的工作情况，有针对性地提出督导意见。

### 1.2.2 计算机辅助电话调查的应用范围

CATI 系统的应用范围很广，几乎涉及到任何内容任何范围的调查，但运用最

多的还是市场调查和民意调查，主要有如下几方面：

1. 市场调查

主要针对产品品牌知名度、产品渗透率、品牌市场占有率、产品广告到达率、广告投放后的效果跟踪研究，居民消费观念、消费习惯研究，消费者生活形态研究，顾客满意度研究，服务质量跟踪调查，媒体覆盖率等研究。主要委托者是工商企业、媒体和研究机构。

2. 社情民意调查

主要包括居民对市政建设、环境治理、治安情况以及就业、教育、住房现状的评价。此类调查主要由各级政府和相关部门委托。

3. 行业行风调查

包括政策透明度、办事程序和办事效率以及办事人员工作态度等。主要服务对象为党政部门、行业主管部门和大型企业集团，如工商、税务、公安、银行、电信等。

4. 热点问题或突发事件的调查

例如，对世界性比赛的评价，对某国领导人访华的反映，禽流感

对公众和政府的影响等。

5. 对一些特殊群体的调查

例如，新闻记者对塑造企业形象的看法，政府官员对扶植国内名牌的态度，投资者对近期投资意向的打算等。

6. 对公众健康的调查

## 1.3 计算机辅助电话调查的优点

### 1.3.1 调查范围广

CATI系统具有覆盖面广、不受地理限制的优点。访问员可以跨越的地域更为广泛，不再受到地理的限制，可以对任何有电话的地区、单位和个人进行调查。比如电话调查可以跨越一些人为障碍（如公寓楼的保安系统）可能访问到一些不愿意陌生人进入他们居所的受访者等。

### 1.3.2 调查可控性高

一方面，被调查者的参加过程是没有适当的控制和严格的监督的。虽然调查者无法像当面访问一样根据自己的观察来判断被调查者所提供的资料的真实性与准确性，但同时被调查者反应的实验者效应和社会称许效应出现的可能性也就相减小了。依照戈夫曼的“拟剧理论”，一个人的行为深受在场的“观众”影响，而电话调查员的“不在场”可以将这一影响降低许多，使调查更加客观。另一方面，电话调查对调查员可以做到严格的监督和控制，不会出现调查员作弊等严重影响数据质量的问题。

### 1.3.3 调查样本特点

首先，电话调查能够对由不同的访问员完成的样本进行即时汇总分析，准确及时地掌握样本的构成情况，因此可以及时调整对样本的取舍，这在某些配额抽样调查中是十分必要的。其次，电话调查的样本中能够包含一些通过面谈访问很难接触到的个体，有些地位较高的被调查者由于工作繁忙等原因，个人面谈方式不易接纳，相对比较短暂的电话访问则可能被接受，因此在一定程度上提高了样本的随机性。第三，现在部分采取当面访谈的等概率承机抽样的调查报告在样本上都存在着女性样本多于男性样本、中青年样本偏少、下

岗、无业者样本偏多的问题，除去抽样误差，造成这一现象的客观原因就是调查时间与被调查者工作时间的冲突。相比较而言，电话调查样本的自然属性（年龄、性别等）和社会属性（职业、文化程度）分布则比较合理。

### 1.3.4 调查结果更真实

对一些涉及个人隐私或比较敏感的问题，如教育水平、个人存款等，在面谈的情况下，被访者有时会感到窘迫或心存顾虑，而在电话访问中，由于存在着较大的距离感，往往可能获得较真实的回答，也能在一定程度上克服问卷调查的“心理二重区域”。

### 1.3.5 调查的时间短、费用低

调查速度快是电话访问最大的优点，由于省去了往返调查现场，访问员可以迅速联络上距离极为遥远的被访者，因此，电话调查在时效性上具有明显的优势。例如，南京大学社会学系在2003年5月1—4日的短短4天时间内，通过电话调查，成功地对北京、上海、广州、重庆与南京五大城市的1030个家庭进行了有关“非典”的舆情调查，了解了中国普通民众对其认知状况、行为反应以及对政府控制疫情所表现出来的基本态度、采取措施及其效果的评价等。显然，如果采用当面访问的调查方法，需要花的时间要多得多，而面对这样的突发事件，电话调查所表现出来的时效性优势就显得非常重要了。另一方面，电话调查在费用上的优势也非常明显。目前，电话调查的成功率还没有一个比较明确的数据，但在10%－20%之间是被大多数研究人员所认可的，而一般一个成功的电话调查在15－20分钟，因此完成一个电话调查所需费用在2～4元之间（按市话计），而采用当面访问方法调查同样一份问卷的费用就要高出许多了。

### 1.3.6 容易编码、录入

传统调查的数据编码与和录入往往是一个很费时费力的过程，而电话调查配合计算机使用可以在访问结束后很快地得到分析结果，这在某些对时效要求较高的调查中尤其体现出其优越性。另一方面，可以对数据进行即时检查，最简单的是以对取值范围进行检查，例如，如果某个问题可能的答案编码为1到5，而调查员误输入6，那么计算机将不会接受，并提醒改正错误。在传统的调查方法中，

这种错误只有在数据编码核查阶段才能发现，已经无法返回调查现场进行更正，只能作为丢失数据处理。而电话调查由于能够即时发现错误，有机会进行修改，因此完全可以避免这类错误。

### 1.3.7 比较容易控制访问员的误差

通常，访问员是在计算机房或专门的电话访问实验室里集中进行电话访问的。在调查现场的研究人员或督导员可以对访问员的访问工作进行即时监督，以确保访问员能将研究者设计的问卷题意清楚且正确地表述给被访者，同时将被访者的回答忠实完整地记录在电脑中。一旦发现访问员曲解题意，或误解被访者答案内容，则立即纠正访问员的偏差行为，进而确保调查资料的质量。

### 1.3.8 采集的数据可迅速用于分析

计算机辅助调查还有一大优点就是，数据一旦被采集就很快可以用于分析和处理其他资料。因为访问员将所有回答直接键入计算机中(没有问卷的书面存档)且一定要输入到机读文件中。这样，所有的调查一完成，程序员就可以很快地创建 SAS、SPSS 或者其他统计分析的全套文件。计算机辅助调查还使调查组对问卷范本特点的测定更为容易。

## 1.4 计算机辅助电话调查的缺点

### 1.4.1 抽样总体目标与目标总体不一致

在电话访问中，电话号码簿涵盖不足的问题已成共识，因此当前大部分电话访问都采用随机拨号抽样。这样虽然解决了电话号码簿不足的问题，但仍然没解决抽样总体与研究总体不一致的问题。抽样总体实际是全体电话用户，而研究总体则可能包括那些没有安装电话的调查对象。目前我国大中城市的电话家庭拥有率一般都可以达到 80%以上，有些城市还会更高，所以在这些城市中使用 CATI 系统调查问题不大。但在一些边远地区和农村，电话的家庭普及率仍然较低，因此样本的代表性就会有偏差。另外，在电话普通率很高的大城市也存在样本代表性问题。因为许多人都拥有不止一个电话号码，这样不同的访问员在进行随机拨号抽样时可能会对同一个人重复访问。此外，在一般家庭中，年龄较大的成员通常不愿意接听电