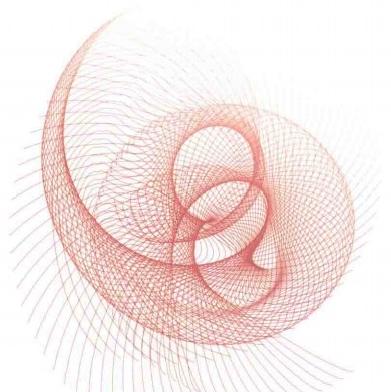
**大数据治理与应用丛书**



**Introduction to Data Quality**

**数据质量导论**

**曹建军** **刁** **兴** **春** **著**

北



**National Defense Industry Press**

大数据治理与应用丛书

**数据质量导论**

**Introduction to Data Quality**

曹建军 刁兴春 著



·北京 ·

**内** **容** **简** **介**

本书结合国内信息环境特点，系统分析了数据质量以及数据全生命周期质量 管理的内涵，构建了数据质量研究和数据清洗系统框架，并引入了数据质量管理的 并行发展模式；深入研究了实体分辨、不完整数据、不一致数据三类实例层数据质 量问题的数据清洗技术，提出了若干数据清洗技术方法；归纳了数据质量工具的发 展概况，提出了两种数据质量工具设计方法；总结提出了大数据质量面临的十大挑 战，构建了适用于国内信息环境特点的数据治理系统框架。

本书内容由浅入深，系统性强，易读性和可操作性强，既可作为数据质量领域 的入门和进阶用书，又可作为数据资源建设与利用、信息技术等相关学科的教学参 考用书。

**图书在版编目(CIP)** **数据**

数据质量导论/曹建军，刁兴春著.—北京：国防工业

出版社，2017.10

ISBN 978-7-118-11405-8

I.① 数… Ⅱ.①曹… Ⅲ.①数据处理 IV.①TP274

中国版本图书馆 CIP数据核字(2017)第243903号

※

**圆防一革二版产出版发行**

(北京市海淀区紫竹院南路23号 邮政编码100048)

三河市众誉天成印务有限公司印刷

新华书店经售

开本710×1000 1/16 印张20% 字数402千字

2017年10月第1版第1次印刷 印数1—2000册 定价79.00元

**(本书如有印装错误，我社负责调换)**

国防书店：(010)88540777

发行传真：(010)88540755

发行邮购：(010)88540776

发行业务：(010)88540717

**序** **一**

我初次涉足“数据质量”和主数据管理(Master Data Management,MDM)是 1995年，当时正致力研发算法与应用软件，以实现来自不同组织用户记录的近似 匹配。那时，数据质量行业只有一批出售数据标准和清洗工具的公司，以及一些主 张过程控制而忽视应用技术的咨询师。

当时面临的主要挑战是如何在两者之间寻求平衡——尽管能够使用工具“清 洗”数据，但如何才能明确界定“高质量数据”的真正含义呢?因此，我意识到需要 通过探究如何使用数据剖析与清洗工具，以及相关技术促进数据质量规约，进而基 于明确的数据质量期望指标来增强合规性以实现两种方法的有机结合。这也是我 2001 年专著《企业管理知识——数据质量方法》的主题。

如果向前看20年，就可以理解数据质量社区是如何出现的。有定期会议供从 业者讨论众多不同业务场景下的各类数据质量技术；建立了监督数据方针规范的 工作流程，企业使用这些方针创建、使用和保护数据资产；随着越来越多的组织任 命负责数据资产管理的首席数据官(Chief Data Officer),数据治理(Data Govern- ance) 和数据统管(Data Stewardship)得到快速普及；有许多致力推动数据质量最佳 实践的国际组织，我及其他优秀数据质量作者的书正被译成不同语言，促进信息利 用服务的全面改善。

本书是对数据质量知识体系的加强。特别值得一提的是，本书既回顾了数据 质量工具与技术相关的传统知识，又在诸如不完整数据检测，基于马尔可夫模型的 数据填补，以及改进数据清洗的机器学习算法方面提出了新思想。另外，作者还讨 论了大数据质量的时代特征，这些特征很快会成为我们面临的重大挑战。

很荣幸受邀为本书作序，希望作者的工作对世界范围的数据质量从业者产生 深远影响。

David Loshin

Website:www.knowledge -integrity.com Email:loshin@knowledge -integrity.com

LinkedIn:David Loshin

Twitter:@davidloshin

2017年2月13日

**David** **Loshin:**被誉为信息管理行业的思想引领者，著有“The Practitioner's

*Guide to Data Quality Improvement”(2011) 、“Master Data Management”(2008)、*

“Business Intellige~~nce~~—The Sawy Manager's Guide”(2003)等，开设了多门数据管理 最佳实践课程，并多次做相关专题报告。

**序** **二**

智能设备、云计算和物联网，这些技术是如此让人钟爱!它们就像闪闪发光的 金币一样吸引眼球。然而，事实上这些技术的价值源自对所创建、存储及分发数据 与信息的利用。立足数据与信息的新应用，各类组织正在谋求更大的成功。为此， 它们必须知道所拥有的数据是正确的并且是可信的。那么,是谁在关注这些炫目 科技中的数据与信息呢?

众所周知，所有组织都有诸如财务(Finance) 、人力资源(Human Resources)管 理系统，那么是否有数据与信息管理系统呢?在过去，这样的系统称为信息技术 (Information Technology)或直接称作 IT 。IT 作为一项重要技术，只是故事的一部 分。我们需要一个和数据与信息自身相关的管理系统，不仅包括技术，而且涉及人 员、组织、流程、标准方方面面。以数据质量为中心是数据管理系统的基本特征。 不管是以营利为目的的企业，还是政府、教育、医疗、慈善，及其他非营利机构，任何 组织的成功都离不开信息。用户、客户、雇员和商业伙伴使用数据进行决策、处理 事务、相互协作，并提供产品与服务。他们规划、创建、更新、转换、存档和删除数据 时，也会影响数据质量。自动化系统基于高质量数据顺利完成交易，无需人为干 预。自然，以数据质量为中心是为了确保上述所有组织和人员满怀信心地使用数 据。再次重申，数据质量对信息利用和期望结果至关重要。

高质量的数据与信息是适用的(Fit for Purpose)。这意味着能够找到(能够获 取并访问)所需要的数据或信息，当需要时是可用的(及时而不滞后),包含所需要 的全部信息(无缺失),是安全的(未经授权不得访问与操作),是可理解的(能够对 其解释),并且是正确的(是对真实世界的准确反映)。这样才能确保所获取的信 息是可信的，并且满怀信心地利用它们。为了拥有高质量数据与信息，需要综合考 虑多种因素。只有在数据全生命周期过程中优化协调大量流程、技术、人员、组织 以及各种各样的数据，才能提供组织所需的高质量可信信息。尽管这一过程错综 复杂，但从数据质量的视角，能够综合考虑多种因素，进而找出全新解决方案。

能够从事数据质量职业是令人兴奋的。作为第一本译成中文的数据质量专 著——《数据质量工程实践——获取高质量数据和可信信息的十大步骤》的作者， 很荣幸受邀为中国的第一本数据质量专著作序。我最近一次去中国是2016年11 月，当时作为特邀国际演讲专家参加了“首届中国数据标准化及治理大会”。看到 中国对相关领域表现出的热情以及所做的工作非常振奋。期间，我还与大数据专

业的 MBA、管理数据的技术用户、致力于增强数据质量的公司，以及乐于分享新闻 的媒体进行了交互。

作为一个职业和专业领域，数据质量还不如计算机科学、工程学、法学、会计学 这些职业领域成熟。任何新兴职业在成为主流并为众人所知之前，都要经历一个 漫长的发展过程。对于数据质量研究者来说，这是一个激动人心的时代，因为我们 已经从前人的工作中获得了许多经验。尽管在很多地方取得了重大突破，但仍有 广阔的创新和发展空间。于是该书应运而生。如前所述，这是中国第一本数据质 量专著;它建立在对数据质量的认知之上，并会为数据质量领域的发展作出贡献； 它还讨论了中国信息环境的独特性。本书为技术研发等数据质量管理工作提供专 业知识。

数据质量领域历经近24年仍在不断发展，而我自诩该领域第二代先锋。很庆 幸自己曾经与几位首先提出数据质量重要性的先驱们并肩工作。我不断向不同领 导者、管理者、从业者，以及咨询师请教学习。你是组织中少数数据专家之一吗? 你理解数据的重要性吗?你想让你的组织认识到数据的重要性并想得到数据质量 管理的资源吗?你是公司的致力于保证高质量数据的第一位首席数据官吗?你想 引领数据质量潮流吗?如果是，那么你也是一名数据质量先锋。

成为数据质量先锋并非易事，要不畏挑战，勇往直前。要让更多人认识到数据 质量的重要性，并提升相关技能。纵然要不断开拓进取，但能成为世界范围的弄潮 儿又倍感欣慰。数据质量的理念超越了语言、文化和国家的界限。很高兴所到之 处皆能与志同道合的人士并肩战斗，很高兴数据质量原理正在产生积极影响。 数据质量在我心中是一股团结的力量。我们因数据质量走到一起，团结协作， 和谐发展。我在世界范围内结交了众多同道中人，并得到了国际数据质量社 区的高度赞赏。这个世界真小，不论走到哪里，都能学习他人的工作动机及协 作精神，对他们表现出的友善我心存感激。这也正是数据质量社区的内在 精神!

那些真正理解数据质量工作的人还认识到了“人的因素”对成功的重要 性，包括工作过程中的沟通(Communication) 、协调(Coordination) 、互助(Coop- eration)、合作(Collaboration), 还包括另外一个 C——勇 气(Courage) 。 尽管在 数据质量领域很少提及“勇气”,但是创新、改革、进取的确需要勇气，放弃安 逸、提升自我、不断突破也需要勇气。要鼓励公司相关成员像管理人员与财务 一样对数据进行管理。

尽管投身数据质量领域多年，但我依然热情不减，坚信这项重要工作切 实影响组织发展和人们的日常生活。人生苦短，希望热情和理想激励我们 路前行。希望读者利用从本书中汲取的营养实现组织受益并为社会作出 贡献。

感谢信息质量研究组(Information Quality Research Group,IQRG)以及本书 的所有贡献者，感谢你们的分享。鉴于本人不谙中文，感谢曹建军博士将本序 言译成中文，感谢御数坊(Data Governance Workshop)的刘晨先生对本书的 评论。

Danette McGilvray Website:www.gfalls.com

Email:danette@gfalls.com

LinkedIn:Danette McGilvray

Twitter:@Danette\_McG 2017年1月31日

**Danette McGilvray:**国际数据标准化、治理与质量领域知名专家，自诩为数据 质量领域的第二代先锋，著有“Executing Data Quality Projects—Ten Steps to Quality Data and Trusted Information”(2008),御数坊“数据质量十步法”在线课程主讲专 家，2012年企业数据世界(Enterprise Data World,EDW)大会最受欢迎讲师。

**前** **言**

大数据战略进展如火如茶，数据质量问题日益突显。好产品的典型特征是具 有较好的自身守恒能力，能够稳定保持用户期望的产品使用价值，较之其他有形产 品或软件产品，数据产品的这种能力恰恰较差。同时，数据的价值主要体现在“流 通”,而非“存储”,所以，数据质量问题较传统产品质量面临更多挑战。

信息质量研究组(Information Quality Research Group,IQRG)成立于2008年， 以结合我国信息环境特点系统开展数据质量研究与实践为己任，随着相关工作的 深入推进，对国内数据现状及特点的认识也逐渐清晰。

信息质量研究组成立以来，我们陆续出版了译著《数据质量工程实践》、《信息 质量》和《数据质量改进实践指南》,后两者受到了装备科技译著出版基金的资助。 “御数坊”在介绍第20届企业数据世界(Enterprise Data World)大会(加利福尼亚 州圣迭戈，2016年4月17—22日)时，向关注数据质量的同学推荐了《数据质量工 程实践》。三本译著在国内普及数据质量理论与实践体系、提升数据质量认识层 次上发挥了积极作用。为了有计划地推出研究成果，立足我国信息环境特点逐步 构建数据治理与应用理论技术体系，2016年上半年，受国防工业出版社之邀，信息 质量研究组启动了“大数据治理与应用丛书”出版工作，译著《数据质量改进实践 指南》是丛书开卷，本书是此丛书的第二个成员。

本书共分12章。第1章至第3章是本书的总述部分。第1章为绪论，引出数 据质量问题，介绍了数据质量以及数据全生命周期质量管理的含义，分析了数据质 量问题的来源并归纳其研究发展简况；第2章分析构建了数据质量研究和数据清 洗系统框架，引入了数据质量管理的并行发展模式，构建了数据质量控制层次框 架，分析了其实现所涉及的关键问题，在进一步辨析数据清洗概念的基础上，构建 了数据清洗的一般性系统框架；第3章综述了典型数据清洗技术的发展动态，系统 归纳了实体分辨、不完整数据、不一致数据三类实例层数据质量问题的数据清洗技

术发展动态。第4章至第10章是以上三类数据清洗技术的研究成果。第4章研 究了实体分辨中的数据分块问题，第5章研究了实体分辨中的相似度算法，第6章 研究了基于关系的实体分辨；第7章研究了不完整数据的分类与检测，第8章研究 了不完整数据的估计与填充；第9章研究了条件函数依赖挖掘及其优化方法，第 10章研究了基于规则的不一致数据检测与修复方法。第11章研究了数据质量工 具的发展概况及设计方法，分别研究了基于表达式树的数据质量工具设计和基于

流程的数据质量工具设计方法。第12章研究了大数据与大数据质量问题，归纳了 大数据时代的特征，总结提出了大数据质量面临的十大挑战，构建了适用于我国信 息环境特点的数据治理系统框架。

本书由曹建军、刁兴春全面筹划，并负责了第1章至第3章、第12章的研究撰 写工作，指导参与了其他各章的研究撰写；谭明超、周星负责了第4章至第6章的 研究撰写；郑奇斌、谭明超负责了第7章的研究撰写；郑奇斌、谭明超、陈爽负责了 第8章的研究撰写；周金陵负责了第9章的研究撰写；高科负责了第10章的研究 撰写；江春、翁年凤、高科负责了第11章的研究撰写。许永平参与了第9章、第10 章的编辑整理，刘艺、冯钦参与了部分章节的编辑整理。江春、彭琮负责了全书的 文字编辑润色；尚玉玲、刘艺、李红梅、张磊、冯钦负责了全书的规范性审核与修改 工作。

感谢两位国际著名数据质量领域专家 David Loshin、Danette McGilvray 为本书 拨冗作序，感谢二位对信息质量研究组相关工作的支持与肯定。

本书是作者在数据质量领域研究成果的梳理小结，试图传递三个信息： 一是国 内数据质量领域的发展模式要紧贴国内信息环境特点与数据应用实际；二是数据 质量控制技术研究要紧贴国际前沿；三是数据质量管理实践既要重视具体的数据 质量工具又要重视体系化的数据治理平台。通过阅读本书，甚望读者能够在概念 层面对数据质量有全面客观的认识，在技术层面能够管中窥豹，在实践层面获得可 用参考。

本书可作为数据资源建设与利用、信息技术等领域科研和工程技术人员进行 数据质量研究与实践的入门指导及工程参考用书。

在本书内容的研究整理过程中，广泛参考了国内外相关成果，并与多家兄弟科 研团队及专家同仁进行有益的经常化交流研讨，在此一并致以诚挚的谢意。

受水平所限，书中若有错误和不妥之处，恳请广大读者批评指正，并欢迎与作 者直接交流。

作者

2016年10月

**目** **录**

**第** **1** **章** **绪** **论** [1](#bookmark1)

1.1 引言 [1](#bookmark2)

1.2 数据工程建设概述 [2](#bookmark3)

1.2.1 数据处理与应用的发展简况 [2](#bookmark4)

1.2.2 信息系统建设中的数据工程 [3](#bookmark5)

1.2.3 我国数据工程建设面临的问题 [9](#bookmark6)

1.3 数据质量概述 [10](#bookmark7)

1.3.1 数据质量的含义 [10](#bookmark8)

1.3.2 数据全生命周期质量管理 [12](#bookmark9)

1.3.3 数据质量问题的来源 [13](#bookmark10)

1.3.4 数据质量研究发展简况 [14](#bookmark11)

1.4 本书内容结构安排 [16](#bookmark12)

参考文献 [18](#bookmark13)

**第** **2** **章** **数据质量研究和数据清洗系统框架** [**20**](#bookmark14)

2.1 引 言 [20](#bookmark15)

2.2 数据质量研究框架 [20](#bookmark16)

2.2.1 典型的数据质量框架 [20](#bookmark17)

2.2.2 数据质量的研究主题 [25](#bookmark18)

2.2.3 数据质量的研究方法 [30](#bookmark19)

2.3 对数据质量管理的思考 [33](#bookmark20)

2.3.1 数据质量管理的发展模式 [33](#bookmark21)

2.3.2 数据质量管理问题剖析 [35](#bookmark22)

2.4 典型数据质量控制的框架 [38](#bookmark23)

2.4.1 层次结构数据质量控制框架 [38](#bookmark24)

2.4.2 层次结构数据质量控制所涉及的关键问题 [40](#bookmark25)

2.4.3 数据清洗技术简介 [42](#bookmark26)

2.4.4 数据清洗的概念辨析 [42](#bookmark27)

2.4.5 数据清洗的一般性系统框架 [45](#bookmark28)

2.5 本章小结 [47](#bookmark29)

参考文献 [48](#bookmark30)

**第** **3** **章** **典型数据清洗技术的发展动态** [51](#bookmark31)

3.1 引 言 [51](#bookmark32)

3.2 实体分辨技术的发展动态 [51](#bookmark33)

3.2.1 数据分块算法 [52](#bookmark34)

3.2.2 记录比较算法 [54](#bookmark35)

3.2.3 匹配决策模型 [55](#bookmark36)

3.2.4 基于关系的实体分辨 [57](#bookmark37)

3.2.5 实体分辨中的训练和测试数据集 [59](#bookmark38)

3.2.6 实体分辨面临的挑战 [61](#bookmark39)

3.3 不完整数据清洗技术的发展动态 [63](#bookmark40)

3.3.1 数据完整性及其评价方法 [63](#bookmark41)

3.3.2 不完整数据的分类 [64](#bookmark42)

3.3.3 不完整数据清洗技术 [65](#bookmark43)

3.4 不一致数据清洗技术的发展动态 [66](#bookmark44)

3.4.1 针对一致性的数据依赖理论 [67](#bookmark45)

3.4.2 典型数据依赖(规则)挖掘方法 [72](#bookmark46)

3.4.3 基于数据依赖的数据一致性保证 [75](#bookmark47)

3.5 本章小结 [79](#bookmark48)

参考文献 [80](#bookmark49)

**第** **4** **章** **实体分辨中的数据分块方法** [**86**](#bookmark50)

4.1 引 言 [86](#bookmark51)

4.2 基于冗余的数据分块 [86](#bookmark52)

4.3 基于倒排索引消除冗余记录对 [87](#bookmark53)

4.3.1 数据块排序索引 [88](#bookmark54)

4.3.2 记录倒排索引 [89](#bookmark55)

4.3.3 冗余记录对识别 [89](#bookmark56)

4.3.4 实验分析 [90](#bookmark57)

4.4 基于空间映射的数据块约减 [94](#bookmark58)

4.4.1 数据块映射 [95](#bookmark59)

4.4.2 数据块约减 [96](#bookmark60)

4.4.3 实验分析 [98](#bookmark61)

4.5 基于Canopy 聚类的数据分块 [101](#bookmark62)

4.5.1 整体流程 [102](#bookmark63)

4.5.2 ID指定 [102](#bookmark64)

4.5.3 B K 生 成 [103](#bookmark65)

4.5.4 Canopy 聚类 [103](#bookmark66)

4.5.5 候选对象获取 [105](#bookmark67)

4.5.6 复杂性分析 [105](#bookmark68)

4.5.7 实验分析 [105](#bookmark69)

4.6 本章小结 [108](#bookmark70)

参考文献 [109](#bookmark71)

**第5章** **实体分辨中的相似度计算方法** [111](#bookmark72)

5.1 引言 [111](#bookmark73)

5.2 基于多编辑距离融合的相似度计算 [111](#bookmark74)

5.2.1 相似特征定义及其标准化 [112](#bookmark75)

5.2.2 编辑距离 [113](#bookmark76)

5.2.3 中西文混合字符串的编辑距离 [114](#bookmark77)

5.2.4 多编辑距离字符串相似度融合 [116](#bookmark78)

5.2.5 实验分析 [117](#bookmark79)

5.3 属性相似度与函数依赖的关系 [119](#bookmark80)

5.4 基于函数依赖的属性相似度调整 [122](#bookmark81)

5.4.1 属性相似度划分 [122](#bookmark82)

5.4.2 属性相似度调整 [123](#bookmark83)

5.4.3 算法描述 [126](#bookmark84)

5.4.4 实验分析 [127](#bookmark85)

5.5 本章小结 [133](#bookmark86)

参考文献 [134](#bookmark87)

**第** **6** **章** **基于关系的实体分辨** [136](#bookmark88)

6.1 引言 [136](#bookmark89)

6.2 基于云模型的实体分辨记录对划分 [137](#bookmark90)

6.2.1 云模型简介 [137](#bookmark91)

6.2.2 记录相似度的分布 [138](#bookmark92)

6.2.3 记录相似度的云模型表示 [139](#bookmark93)

6.2.4 划分方法 [140](#bookmark94)

6.2.5 结果分析 [142](#bookmark95)

6.3 基于邻域粗糙集的实体分辨记录对划分 [143](#bookmark96)

6.3.1 邻域粗糙集 [144](#bookmark97)

6.3.2 基于邻域粗糙集的记录对划分 [145](#bookmark98)

6.3.3 实验分析 [146](#bookmark99)

6.4 基于关系类型的自适应实体分辨 [150](#bookmark100)

6.4.1 路径权重 [150](#bookmark101)

6.4.2 路径概率 [151](#bookmark102)

6.4.3 连接强度 [152](#bookmark103)

6.4.4 自适应关系类型权重学习 [153](#bookmark104)

6.4.5 实验分析 [154](#bookmark105)

6.5 本章小结 [159](#bookmark106)

参考文献 [159](#bookmark107)

**第** **7** **章** **不完整数据的分类与检测** [161](#bookmark108)

7.1 引言 [161](#bookmark109)

7.2 基于位运算的不完整数据分类与检测 [162](#bookmark110)

7.2.1 不完整数据及其分类 [162](#bookmark111)

7.2.2 记录的二进制表示 [164](#bookmark112)

7.2.3 不完整记录的位运算分类检测方法 [164](#bookmark113)

7.2.4 应用实例 [166](#bookmark114)

7.3 基于统计关系的不完整数据分类 [167](#bookmark115)

7.3.1 数据缺失模式分类 [167](#bookmark116)

7.3.2 数据缺失机制分类 [169](#bookmark117)

7.4 本章小结 [171](#bookmark118)

参考文献 [171](#bookmark119)

**第** **8** **章** **不完整数据的估计与填充** [173](#bookmark120)

8.1 引言 [173](#bookmark121)

8.2 基于统计关系学习的缺失数据估计与填充 [173](#bookmark122)

8.2.1 统计关系学习概述 [174](#bookmark123)

8.2.2 基于马尔可夫模型的缺失值估计方法 [178](#bookmark124)

8.2.3 基于关系马尔可夫模型的缺失值估计 [181](#bookmark125)

8.3 基于机器学习的缺失数据估计与填充 [192](#bookmark126)

8.3.1 基于k -近邻的填补算法 [192](#bookmark127)

8.3.2 局部敏感哈希技术 [193](#bookmark128)

8.3.3 LSH\_KNN 数据填补算法 [193](#bookmark129)

8.3.4 实验验证 [197](#bookmark130)

8.4 函数依赖一致性数据生成 [200](#bookmark131)

8.4.1 函数依赖一致性 [200](#bookmark132)

8.4.2 单函数依赖一致性数据生成算法 [201](#bookmark133)

8.4.3 基于有向无环图的多函数依赖一致性数据生成 [203](#bookmark134)

8.4.4 属性集划分和数据生成流水线 [206](#bookmark135)

8.5 本章小结 [209](#bookmark136)

参考文献 [209](#bookmark137)

**第** **9** **章** **条件函数依赖挖掘及其优化方法** [**211**](#bookmark138)

9.1 引言 [211](#bookmark139)

9.2 条件函数依赖挖掘及其常用算法 [211](#bookmark140)

9.2.1 条件函数依赖及其挖掘问题 [212](#bookmark141)

9.2.2 函数依赖挖掘 [215](#bookmark142)

9.2.3 CTANE算法 [217](#bookmark143)

9.2.4 CFDMiner 算法 [219](#bookmark144)

9.3 基于开项集剪枝的常量条件函数依赖挖掘算法 [221](#bookmark145)

9.3.1 剪枝与优化策略 [221](#bookmark146)

9.3.2 优化前后复杂度对比 [225](#bookmark147)

9.3.3 实验验证与结果分析 [226](#bookmark148)

9.4 本章小结 [228](#bookmark149)

参考文献 [229](#bookmark150)

**第10章** **基于规则的不一致数据检测与修复方法** [231](#bookmark151)

10.1 引 言 [231](#bookmark152)

10.2 基于Fellegi- Holt 方法的不一致数据检测 [232](#bookmark153)

10.2.1 Fellegi- Holt 方法 [232](#bookmark154)

10.2.2 检测流程及策略 [236](#bookmark155)

10.2.3 实验及分析 [238](#bookmark156)

10.3 基于Evidence - Rules 模型的不一致数据修复 [242](#bookmark157)

10.3.1 确定问题记录中待修改属性集 [243](#bookmark158)

10.3.2 基于函数依赖规则的属性值修复 [244](#bookmark159)

10.3.3 Evidence - Rules 模型与问题数据修复 [246](#bookmark160)

10.3.4 实验及分析 [253](#bookmark161)

10.4 本章小结 [256](#bookmark162)

参考文献 [257](#bookmark163)

**第11章** **数据质量工具** [259](#bookmark164)

11.1 引 言 [259](#bookmark165)

11.2 数据质量工具发展概况 [259](#bookmark166)

11.2.1 Gartner 分析报告 [259](#bookmark167)

11.2.2 数据质量管理工具分析 [261](#bookmark168)

11.3 基于表达式树的数据质量工具设计 [265](#bookmark169)

11.3.1 数据质量规则的分类与表达 [265](#bookmark170)

11.3.2 数据质量规则的存储与识别 [271](#bookmark171)

11.4 基于流程的数据质量工具设计 [276](#bookmark172)

11.4.1 数据模型 [277](#bookmark173)

11.4.2 作业模型 [278](#bookmark174)

11.4.3 执行方案模型 [280](#bookmark175)

11.5 本章小结 [281](#bookmark176)

参考文献 [282](#bookmark177)

**第12章** **大数据与大数据质量问题** [283](#bookmark178)

12.1 引 言 [283](#bookmark179)

12.2 大数据时代的特征 [283](#bookmark180)

12.2.1 大数据的含义 [284](#bookmark181)

12.2.2 大数据的特征 [284](#bookmark182)

12.2.3 进入大数据时代的必要条件 [285](#bookmark183)

12.2.4 大数据时代的革命性转变 [287](#bookmark184)

12.2.5 大数据时代的核心任务 [288](#bookmark185)

12.3 大数据质量面临的挑战 [290](#bookmark186)

12.3.1 数据安全问题 [290](#bookmark187)

12.3.2 大数据的偏见和盲区 [291](#bookmark188)

12.3.3 非结构化数据的质量控制 [292](#bookmark189)

12.3.4 结构化数据内缺少结构性 [292](#bookmark190)

12.3.5 分布式数据清洗 [293](#bookmark191)

12.3.6 数据化程度不够 [293](#bookmark192)

12.3.7 数据稀缺 [294](#bookmark193)

12.3.8 数据冗余 [294](#bookmark194)

12.3.9 数据对实际需求的适用性 [294](#bookmark195)

12.3.10 人为选择导致的信息失真 [295](#bookmark196)

12.4 数据治理 [295](#bookmark197)

12.4.1 数据治理的出发点 [295](#bookmark198)

12.4.2 数据治理的一般流程 [296](#bookmark199)

12.4.3 数据治理的系统框架 [297](#bookmark200)

12.5 本章小结 [300](#bookmark201)

参考文献 [300](#bookmark202)

**基金资助目录** [302](#bookmark203)

**Contents**

**Chapter 1 Introduction to Data Engineering and Data Quality** [1](#bookmark204)

1.1 Introduction [1](#bookmark1)

1.2 An Overview on Data Engineering Construction [2](#bookmark205)

1.2.1 Development Status of Data Processing and Application [2](#bookmark3)

1.2.2 Data Engineering in Information System Construction [3](#bookmark206)

1.2.3 Problems in Domestic Data Engineering Construction [9](#bookmark207)

1.3 An Overview on Data Quality [10](#bookmark208)

1.3.1 Nature of Data Quality [10](#bookmark7)

1.3.2 Management of Data Lifecycle Quality [12](#bookmark209)

1.3.3 Origin of Data Quality Problems [13](#bookmark210)

1.3.4 Development Status of Data Quality Research [14](#bookmark211)

1.4 Contents Arrangement [16](#bookmark212)

References [18](#bookmark213)

**Chapter 2 Frameworks of Data Quality Research and Data**

**Cleaning System** [20](#bookmark14)

2.1 Introduction [20](#bookmark15)

2.2 Framework of Data Quality Research [20](#bookmark16)

2.2.1 Typical Data Quality Frameworks [20](#bookmark17)

2.2.2 Research Themes of Data Quality [25](#bookmark18)

2.2.3 Research Methods of Data Quality [30](#bookmark19)

2.3 Thoughts on Data Quality Management [33](#bookmark20)

2.3.1 Development Pattern of Data Quality Management [33](#bookmark21)

2.3.2 Analysis of Data Quality Management Problems [35](#bookmark22)

2.4 Typical Framework of Data Quality Control [38](#bookmark23)

2.4.1 Hierarchical Data Quality Control Framework [38](#bookmark24)

2.4.2 Critical Problems Related to Hierarchical Data Quality

Control Framework [40](#bookmark214)

2.4.3 Introduction to Data Cleaning Technologies [42](#bookmark27)

2.4.4 Concept Resolution of Data Cleaning [42](#bookmark215)

2.4.5 General System Framework of Data Cleaning [45](#bookmark216)

2.5 Summary [47](#bookmark217)

References [48](#bookmark218)

**Chapter 3 State of Art of Typical Data Cleaning Technologies** [51](#bookmark33)

3.1 Introduction [51](#bookmark219)

3.2 Development of Entity Resolution Technologies [51](#bookmark220)

3.2.1 Data Blocking Algorithm [52](#bookmark221)

3.2.2 Record Comparison Algorithm [54](#bookmark222)

3.2.3 Matching Decision Model [55](#bookmark223)

3.2.4 Relationship- based Entity Resolution [57](#bookmark224)

3.2.5 Data Sets for Training and Testing [59](#bookmark225)

3.2.6 Challenages [61](#bookmark226)

3.3 Developments of Incomplete Data Cleaning Technologies [63](#bookmark227)

3.3.1 Concept and Evaluation of Data Completeness [63](#bookmark228)

3.3.2 Incomplete Data Classification [64](#bookmark229)

3.3.3 Incomplete Data Cleaning Tecnologies [65](#bookmark230)

3.4 Developments of Inconsistent Data Cleaning Technologies [66](#bookmark231)

3.4.1 Data Dependency Theory for consistency [67](#bookmark232)

3.4.2 Typical Data Dependency Mining Methods [72](#bookmark233)

3.4.3 Data Consistency Guarantee Based on Data Dependency [75](#bookmark234)

3.5 Summary [79](#bookmark235)

References [80](#bookmark236)

**Chapter 4 Data Blocking Methods in Entity Resolution** [86](#bookmark52)

4.1 Introduction [86](#bookmark237)

4.2 Redundancy Based Data Blocking [86](#bookmark238)

4.3 Redundancy Record Pairs Elimination Based on Inverted Indexing [87](#bookmark239)

4.3.1 Data Blocks Sorted Indexing [88](#bookmark240)

4.3.2 Records Inverted Indexing [89](#bookmark241)

4.3.3 Redundant Record Pairs Recognition [89](#bookmark242)

4.3.4 Experiments [90](#bookmark243)

4.4 Data Blocks Reduction Based on Space Mapping [94](#bookmark244)

4.4.1 Data Blocks Mapping [95](#bookmark245)

4.4.2 Data Blocks Reduction [96](#bookmark246)

4.4.3 Experiments [98](#bookmark247)

4.5 Canopy Based Data Blocking [101](#bookmark248)

4.5.1 General Process [102](#bookmark249)

4.5.2 ID Assignment [102](#bookmark250)

4.5.3 BK Generation [103](#bookmark251)

4.5.4 Canopy Clustering [103](#bookmark252)

4.5.5 Candidates Obtaining [105](#bookmark69)

4.5.6 Complexity Analysis [105](#bookmark253)

4.5.7 Experiments [105](#bookmark254)

4.6 Summary [108](#bookmark255)

References [109](#bookmark256)

**Chapter 5 Similarity Computing Methods in Entity Resolution** [111](#bookmark74)

5.1 Introduction [111](#bookmark257)

5.2 Similarity Computation Based on Multiple Edit Distance Fusion [111](#bookmark258)

5.2.1 Similarity Feature Definition and Standardization [112](#bookmark259)

5.2.2 Edit Distance [113](#bookmark260)

5.2.3 Edit Distance of Chinese -English Characters [114](#bookmark261)

5.2.4 Fusion of Multiple Edit Distances [116](#bookmark262)

5.2.5 Experiments [117](#bookmark263)

5.3 Relationship between Attribute Similarity and Function Depedency [119](#bookmark264)

5.4 Attribute Similarity Adjustment Based on Function Dependency [122](#bookmark265)

5.4.1 Partition of Attribute Simialrity [122](#bookmark266)

5.4.2 Adjustment of Attribute Similarity [123](#bookmark267)

5.4.3 Algorithm Description [126](#bookmark268)

5.4.4 Experiments [127](#bookmark269)

5.5 Summary [133](#bookmark270)

References [134](#bookmark271)

**Chapter 6 Relationship- based Entity Resolution** [136](#bookmark272)

6.1 Introduction [136](#bookmark273)

6.2 Record Pairs Partition for Entity Resolution Based on Cloud Model [137](#bookmark274)

6.2.1 Introduction to Cloud Model [137](#bookmark275)

6.2.2 Distribution of Record Similarities [138](#bookmark276)

6.2.3 Cloud Model Representation of Record Similarities [139](#bookmark277)

6.2.4 Partition Method [140](#bookmark278)

6.2.5 Result Analysis [142](#bookmark279)

6.3 Record Pairs Partition for Entity Resolution Based on Neighborhood

Rough Set [143](#bookmark280)

6.3.1 Neighborhood Rough Set [144](#bookmark281)

6.3.2 Record Pairs Partition Based on Neighborhood Rough Set [145](#bookmark282)

6.3.3 Experiments [146](#bookmark283)

6.4 Adaptive Entity Resolution Based on Relationship Types [150](#bookmark284)

6.4.1 Path Weight [150](#bookmark285)

6.4.2 Path Probability [151](#bookmark286)

6.4.3 Connection Strength [152](#bookmark287)

6.4.4 Adatptive Weight Learning of Relationship Types [153](#bookmark288)

6.4.5 Experiments [154](#bookmark289)

6.5 Summary [159](#bookmark290)

References [159](#bookmark291)

**Chapter 7 Classification and Detection of Incomplete Data** [161](#bookmark292)

7.1 Introduction [161](#bookmark293)

7.2 Incomplete data Classification and Detection Based on Bit

Operation [162](#bookmark294)

7.2.1 Incomplete data and its Classification [162](#bookmark295)

7.2.2 Binary Representation of Record [164](#bookmark296)

7.2.3 Classifying Detection Method for Incomplete Records Based on

Bit Operation [164](#bookmark297)

7.2.4 Applications [166](#bookmark298)

7.3 Incomplete Data Classification Based on Statistical Relationship [167](#bookmark299)

7.3.1 Classification of Data Missing Patterns [167](#bookmark300)

7.3.2· Classification of Data Missing Mechanisms [169](#bookmark301)

7.4 Summary [171](#bookmark302)

References [171](#bookmark303)

**Chapter 8 Estimation and Imputation of Incomplete Data** [173](#bookmark304)

8.1 Introduction [173](#bookmark305)

8.2 Missing Data Estimation and Inputation Based on Statistical

Relational Learning [173](#bookmark306)

8.2.1 A Survey on Statistical Relational Learning [174](#bookmark307)

8.2.2 Missing Values Estimation Based on Markov Model [178](#bookmark308)

8.2.3 Missing Values Estimation Based on Relational Markov Model [181](#bookmark309)

8.3 Missing Data Estimation and Imputation Based on Machine

Learning [192](#bookmark310)

8.3.1 k-NN Based Imputation Algorithm [192](#bookmark311)

8.3.2 Locality Sensitive Hashing Technology [193](#bookmark312)

8.3.3 LSH\_KNN Data Imputation Algorithm [193](#bookmark313)

8.3.4 Experiments [197](#bookmark314)

8.4 Consistent Data Generation Using Function Dependencies [200](#bookmark315)

8.4.1 Function Dependency Consistency [200](#bookmark316)

8.4.2 Consistent Data Generation Based on Single Function

Dependency [201](#bookmark317)

8.4.3 Consistent Data Generation Using Multiple Function Dependencies

Based on Directed Acyclic Graph [203](#bookmark318)

8.4.4 Attribute Sets Partition and Data Generation Pipeline [206](#bookmark319)

8.5 Summary [209](#bookmark320)

References [209](#bookmark321)

**Chapter 9 Conditional Function Dependency Mining and Optimization** 211

9.1 Introduction [211](#bookmark322)

9.2 Mining of Conditional Function Dependencies and Its Algorithms [211](#bookmark323)

9.2.1 Conditional Function Dependencies [212](#bookmark324)

9.2.2 Mining of Function Dependency [215](#bookmark325)

9.2.3 CTANE Algorithm [217](#bookmark326)

9.2.4 CFDMiner Algorithm [219](#bookmark327)

9.3 Mining of Constant Conditional Function Dependencies Based

on Pruning Free Itemsets [221](#bookmark328)

9.3.1 Pruning and Optimization [221](#bookmark329)

9.3.2 Complexity Comparison [225](#bookmark330)

9.3.3 Experiments [226](#bookmark331)

9.4 Summary [228](#bookmark332)

References [229](#bookmark333)

**Chapter 10 Inconsistent Data Detection and Repairing Based on Rules** [231](#bookmark334)

10.1 Introduction [231](#bookmark335)

10.2 Inconsistent Data Detection Based on Fellegi - Holt [232](#bookmark336)

10.2.1 Fellegi- Holt Method [232](#bookmark337)

10.2.2 Detection Procedure and Strategy [236](#bookmark338)

10.2.3 Experiments [238](#bookmark339)

10.3 Inconsistent Data Repairing Based on Evidence - Rules Model [242](#bookmark340)

10.3.1 Detecting Candidate Attributes Set in the Wrong Records [243](#bookmark341)

10.3.2 Repairing Attribute Values Based on Function Dependencies [244](#bookmark342)

10.3.3 Evidence - Rules Model and Inconsistent Data Repairing [246](#bookmark343)

10.3.4 Experiments [253](#bookmark344)

10.4 Summary [256](#bookmark345)

References [257](#bookmark346)

Chapter 11 Data Quality Tools [259](#bookmark347)

11.1 Introduction [259](#bookmark348)

11.2 Development Status of Data Quality Tools [259](#bookmark349)

11.2.1 Gartner ' s Analysis Report [259](#bookmark350)

11.2.2 Review on Data Quality Management Tools [261](#bookmark351)

11.3 Design of Data Quality Tool Based on Expression Trees [265](#bookmark352)

11.3.1 Classification and Expression of Data Quality Rules [265](#bookmark353)

11.3.2 Storage and Resolution of Data Quality Rules [271](#bookmark354)

11.4 Design of Data Quality Tool Based on Processes [276](#bookmark355)

11.4.1 Data Model [277](#bookmark356)

11.4.2 Job Model [278](#bookmark357)

11.4.3 Execution Model [280](#bookmark358)

11.5 Summary [281](#bookmark359)

References [282](#bookmark360)

**Chapter 12 Big Data and Big Data Quality Issues** [283](#bookmark361)

12.1 Introduction [283](#bookmark362)

12.2 Characteristics of Big Data Era [283](#bookmark363)

12.2.1 Meaning of Big Data [284](#bookmark364)

12.2.2 Characteristics of Big Data [284](#bookmark365)

12.2.3 Essential Conditions of Entering Big Data Era [285](#bookmark366)

12.2.4 Revolutions of Big Data Era [287](#bookmark367)

12.2.5 Core Tasks in Big Data Era [288](#bookmark368)

12.3 Challenages of Big Data Quality [290](#bookmark369)

12.3.1 Data Security Problems [290](#bookmark370)

12.3.2 Prejudice and Blindness of Big Data [291](#bookmark371)

12.3.3 Quality Control of Nonstructural Data [292](#bookmark372)

12.3.4 Structure Lacks within Structural Data [292](#bookmark373)

12.3.5 Distributed Data Cleaning [293](#bookmark374)

12.3.6 Insufficient Data Degree [293](#bookmark375)

12.3.7 Data Sparsity [294](#bookmark376)

12.3.8 Data Redundancy [294](#bookmark377)

12.3.9 Data Suitability to Specific Situations [294](#bookmark378)

12.3.10 Data Artifacts due to Artifical Choice [295](#bookmark379)

12.4 Data Governance [295](#bookmark380)

12.4.1 Purpose of Data Governance [295](#bookmark381)

12.4.2 General Process of Data Governence [296](#bookmark382)

12.4.3 Framework of Data Governance [297](#bookmark383)

12.5 Summary [300](#bookmark384)

References [300](#bookmark385)

**Catalogue of Supported Funds** [302](#bookmark386)

**第1章** **绪** **论**

**1.1** **引言**

信息系统和基础数据中心逐渐成为组织持续运营必不可少的部分，并成为企 业(代表个人或任何独立的商业组织、科学组织、技术组织及其他组织)取得成功 的基础。数据资源是信息系统最重要的资源之一，更是大数据(Big Data)时代的 新型战略资源，建设和利用数据资源既是信息化的出发点，又是信息化的目标。现 实中，随着信息化进程的不断深入，各行各业积累了海量数据，但当用户需要使用 这些数据时，却往往会因各种数据质量/信息质量(Data Quality/Information Quali- ty,DQ/IQ) 问题(本书对二者不加区分)给使用者带来损失(McGilvray,2010;Mad- nick et al.,2009;刘飞，2008;韩京宇，等，2008;曹建军，等，2010;Chen et al., 2005)。根据普华永道会计师事务所在纽约的调查，75%的公司存在因数据质量 造成经济损失的现象，仅35%的公司对自己的数据质量充满信心(Eppler et al., 2003)。另据数据仓库研究所的研究报告，美国企业每年因数据质量造成的直接 经济损失达6110亿美元，约占美国GDP的 6 %(Sara,2002)。著名国际咨询机构 互联网数据中心(Internet Data Center,IDC)在我国的调查显示，我国企业数据集成 环境比美国更加复杂，72%的企业存在相似重复数据，60%的企业存在不完整数 据；数据质量问题是数据集成相关项目面临的主要困难之一，以致大部分项目难以 达到预期目标(刘飞，2008)。因此，数据质量已成为制约一个国家经济建设特别 是信息化建设的障碍。为了使信息化建设健康发展，必须全面加强数据质量研究， 切实提高数据质量。

数据工作已经成为我国信息化建设的“短板”,其中的数据质量问题可谓“短 中之短”,数据质量问题甚至在一些信息系统项目实施方案中鲜有提及。缺少数 据资源的信息系统是不会有“重大效益”的，至多可以算作“演示”系统，而数据质 量差更会使信息系统沦为“观赏性系统”,不可能真正推广应用。大数据具有分散 存储整合使用、分析处理复杂度高、数据整体与关联协同程度决定价值高低的三大 特征，导致数据质量问题更为复杂(Lee et al.,2015)。

因此，结合我国当前信息化中的实际数据需求和国内信息环境特点，开展数据 质量研究与实践具有重要现实意义，相关成果应用前景广阔。

2 )数据质量导论

**1.2** **数据工程建设概述**

随着数据应用需求的不断加大，尤其是数据量的迅速增长，数据工程(Data Engineering)的概念及其技术体系迅速发展完善，数据质量问题日益突显。

**1.2.1** **数据处理与应用的发展简况**

数据处理与应用是数据工作的核心，经历了一个漫长而复杂的发展过程，在这 一过程中不断引发新的需求，推动相关技术发展。

人类在各项活动中不断产生数据。在计算机出现以前，数据处理与应用主要 通过人工方式或机械方式来完成。最早出现的典型数据处理工具就是中国传统的 算盘。20世纪初以来，机械式计算机的出现，使得人们开始以半自动的方式进行 数据处理。在这个过程中，由于人工方式、机械方式的处理能力有限，使得数据处 理所涉及的数据量还不够大、数据类型还较单一，数据处理相对比较简单，基于数 据的各类应用还处于初级阶段。

随着20世纪中期电子计算机的出现与广泛应用，数据处理技术有了长足发 展。20世纪50年代开始，计算机就开始应用于商务数据处理、科学计算等领域， 进入了以自动方式进行数据处理的时期。20世纪80年代以来，随着网络技术的 进一步发展，数据处理技术也在不断提高。并随着信息技术的进步，数据的应用领 域也越来越广泛，在数据处理中涉及的数据量也越来越大，应用背景越来越复杂。 在这一阶段数据存储、数据管理、数据分析、数据质量、模拟数据生成等作为相对独 立的问题也逐渐出现，并分别得到研究和发展。同时，与数据工作相关的方法和技 术也日渐丰富，且得到快速发展。

以下以比较典型的科学研究活动中的科学数据和社会活动中的经济数据为例 作进一步说明。

科学实验、系统模拟、理论研究是科学研究的三大方法。其中，科学实验和系 统模拟都涉及大量数据的处理与应用，通过对科研数据的分析处理，研究自然科 学、社会科学的发展规律。随着研究过程的不断深入，现在的科学研究产生了大量 的数据。尽管针对数据处理的信息技术(Information Technology,IT)取得了一定突 破，无论是在数据处理能力方面，还是在海量存储管理及高速网络等方面均取得了 很大的进展，但是硬件水平的提高并未能完全解决大规模数据处理的问题。事实 上，目前的数据处理能力还远远不能满足科学研究的要求，依然是制约科学研究的 一个重要因素。

在人们的社会活动中会产生大量的社会实践数据。这些数据大到一个国家的

第 1 章 绪论

宏观经济数据，小到一个公司的生产数据、财务数据、销售数据等，这些数据的数据 量也非常庞大，并且对于经济活动具有非常重要的价值。遗憾的是，国内目前对这 部分数据的利用很大程度上停留在信息技术层面的研究上，仅仅通过数据挖掘、数 据仓库等技术手段来研究社会活动的局部特征。要有效利用这部分数据资源，除 了研究信息技术外，更需要研究数据自身，系统研究数据处理过程中的各项活动， 利用工程化的观点从全局角度研究数据的关联性，只有这样才能更有效地利用数 据资源，从中发掘出更大的社会价值。

根据 IDC的监测统计，2011年全球数据总量已经达到1.8ZB(1ZB=27°B), 而这个数量还在以每两年翻一番的速度增长，预计到2020年，全球将总共拥有 35ZB 的数据量，比2011年增长近20倍。换言之，近两年产生的数据总量相当 于人类有史以来所有数据量的总和(李国杰，2012a; 李国杰，等，2012b) 。 在以 上背景下，从公司战略到产业生态，从学术研究到生产实践，从城镇管理乃至国 家治理，都将发生质变。国家竞争力将部分体现为拥有数据的规模及运用数据 的能力。

大数据(Big Data)产业的两个特点非常有利于我国信息产业跨越式发展(王 元卓，等，2013)。 一是大数据技术以开源为主，迄今为止，尚未形成绝对技术垄 断，即便是 IBM、甲骨文等行业巨擘，也同样是集成了开源技术和该公司已有产品 而已。开源技术对任何一个国家都是开放的，国内各行业同样可以分享开源的蛋 糕，但是需要以更加开放的心态、更加开明的思想正确地对待开源社区。二是我国 的人口和经济规模决定了数据资产规模处于世界前列。这在客观上为大数据技术 的发展提供了演练场，也亟待政府、学术界、产业界、资本市场四方面能力合作，在 确保国家数据安全的前提下，最大程度地开放数据资产，促进数据关联应用，释放 大数据的巨大价值。大数据超越信息技术，使人们重新界定国家竞争的主战场，重 新审视政府治理水平，重新认识科学研究的新范式，重新审视产业变迁的驱动因 素，重新理解投资的决策依据，重新思考公司的战略和组织结构(程学旗，等， 2016)。

**1.2.2** **信息系统建设中的数据工程**

[**1.2.2.1**](1.2.2.1) **数据工程的内涵**

信息工程(Information Engineering)是20世纪80 年代产生并发展起来的建立 集成化信息系统的方法论，“以数据为中心构建企业信息系统”是这一方法论的基 本思想：首先，应用开发应面向数据而不是面向处理过程；其次，数据是稳定的而过 程是多变的；最后，最终(高层)用户必须真正参加开发工作。

1987年，IEEE学会的知识与数据工程会刊“IEEE Transactions on Knowledge

数据质量导论

and Data Engineering(IEEE TKDE)”第一次提出使用数据工程取代之前的数据库 工程。数据工程植根于数据库技术，是对数据库工程的扩展，并体现了该领域的发 展和变化。鉴于数据在信息工程中的中心地位，数据工程一词的使用频度也越来 越高。美国国防部(Department of Defense,DoD)于第二次世界大战之后着手国防 物资编目工作，并且于20世纪90年代启动了数据工程，实现了国防数据字典系统 (DoD Data Dictionary System)、共享数据工程(Shared Data Engineering)和联合公共 数据库(Joint Common Database)。这些成果为实现 C4ISR(Command,Control,Com- munications,Computers,Intelligence,Surveillance and Reconnaissance)系统之间的数 据重用和数据共享奠定了基础，也确保了美军在海湾战争、科索沃战争和伊拉克战 争中的信息优势。

随着信息化建设的不断深入，数据工程在国内也越来越受到重视，近年启动 了多个数据工程项目，并成立了专门的数据工程研究机构，如中国人民大学的 “数据工程与知识工程教育部重点实验室”,北京科技大学的“高性能计算与数 据工程实验室”,解放军理工大学与信息质量研究组合建的“数据与知识工程实 验室”等。

所谓工程是指将自然科学的原理应用到工农业生产部门中去而形成的各学科 的总称。如土木工程、水利工程、冶金工程、机电工程、化学工程、海洋工程、航天工 程、计算机工程、软件工程等。这些学科是应用数学、物理学、化学、生物学等基础 科学的原理，结合生产实验与科学实践积累的技术经验而发展出来的。主要内容 有：对工程基地的勘测、设计、施工，原材料的选择研究，设备和产品的设计制造，工 艺和施工方法的研究等。

作战数据工程培训教材编写组认为：数据工程是规范和支撑数据产生、维护、 服务、使用、存储、销毁全过程的一系列技术、建设和管理活动的总称，其主要目标 是加强数据的管理，强化数据的可用性、可达性和可信性，最大程度地提高数据的 使用价值。

本书将信息系统建设中的数据工程的内涵表述如下：数据工程是以满足信息 系统全生命周期的数据需求为目标，以数据全生命周期过程为主线，所进行的数据 活动(Data Action)的集合。数据工程不但要求数据活动规范化，还要求具体数据 活动适应数据全生命周期过程，要考虑不同数据活动之间的联系与相互影响，实现 数据活动的优化整合。

[**1.2.2.2**](1.2.2.2) **数据工程中的数据活动**

综合国内外数据工程发展动态，将数据工程所涉及的数据活动(研究内容)归 纳为五类(见表1-1):法规与政策、规划与设计、获取与处理、应用与维护，以及存 档与监管。

第 1 章 绪 论

表1- 1 数据工程中的数据活动

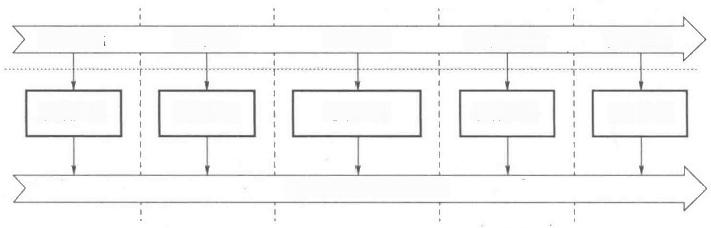
|  |  |
| --- | --- |
| **类别** | **数据活动** |
| 法规与政策 | 数据标准、数据规范、业务规则、数据字典、数据模型、指标体系、评价标准、组织流程、工 作手册、管理规定、激励措施 |
| 规划与设计 | 需求分析、方案设计、资源配置、数据格式、数据结构、元数据管理、数据建模、数据索引 |
| 获取与处理 | 数据采集、数据生成、数据集成、统计分析、数据挖掘、数据融合、不确定性、数据流、内部 关系、联机分析、数据耕耘、数据网格 |
| 应用与维护 | 数据网格、信息检索、趋势预测、决策支持、可视化、优化策略、用户接口、质量评估、系统 性能、业务运营、私密性、安全性、监测与清洗 |
| 存档与监管 | 数据存储、数据备份、数据容灾、数据销毁、数据注释、数据监管、世系管理、发布订购、云 存储、高效管理、分布式管理 |

如表1-1所列，随着数据在信息系统中重要性的提高，出现了种类繁多、功能 各异的数据活动，而对一个具体的数据工程项目，其任务便是选择合适的数据活动 并组织实施。

[**1.2.2.3**](1.2.2.3) **数据的全生命周期过程**

20世纪末，麻省理工学院(MIT) 数据质量研究项目得出了“将数据作为产品 进行管理”的研究结论。产品全生命周期管理(Product Lifecycle Management)以产 品整个生命周期的数据集成为基础，融合先进的管理理念，将协同和优化贯穿产品 的全生命周期。近年来，相关理论和技术受到了关注并得到快速发展。产品全生 命周期思想是从全生命周期的角度考虑产品的设计、生产、使用、管理、质量等 问题。

图1-1给出了产品形式的历史演变示意图。



计算机时代

电子产品

相关理论、方法和技术

数据产品

机械产品

软件产品

电子时代

电气时代

信息时代

蒸汽时代

电气产品

图1-1 产品形式的历史演变示意图

如图1-1所示，科学技术的发展造就了新产品的出现，并不断催生以新产品 为标志的新时代。因此，科学技术本身的进步是以产品为标志的时代出现为原动 力的，表现为科学技术发展的纵向力，也是主动力。当前正在进入信息时代，数据

6)数据质量导论

产品将是信息时代的重要标志，与其他时代的标志产品一样，数据产品将会带来社 会生产力的再次飞跃，与此同时，相关的理论、方法和技术也将得到进一步发展。 因此，新产品的出现也拓宽了相关理论和技术的应用范围和发展空间，这也是科学 技术发展的横向力，是被动力。科学技术发展的纵向力使数据作为新的产品形式 出现，数据产品出现的同时也要求生产、管理、保障等相关理论与技术适应数据产 品的新特点和要求，进而推动横向发展。

通常，数据工作是完整信息系统项目(Project) 的一部分，数据的生命周期蕴含 在信息系统的生命周期中，数据全生命周期管理必须结合项目管理方式和具体项 目方案。实现对一个具体信息系统的全生命周期管理所涉及的所有行动集合构成 一个相应的信息系统项目，通过在信息系统项目的生命周期内采取行动实现信息 系统的全生命周期管理。因此，明确定义信息系统项目的生命周期是信息系统建 设的第一步。

在运行一个项目时存在许多不同的方法。 一个项目生命周期定义了制订解决 方案的方法和项目的各个阶段。解决方案的实例包括建立一个新的应用，将数据 从现有应用迁移到新的应用或者改进流程中。项目生命周期为项目方案和项目团 队要承担的任务提供依据。

项目生命周期也可被看作是解决方案(Solution, 或软件(Software), 或系统 (System)) 开发生命周期(Solution Development Life Cycle,SDLC),这个方案既可以 是软件的也可以是系统的，尽管有人认为项目生命周期与 SDLC 并不一样。如果 将这一点应用于数据工程，几乎在任何项目中都存在数据活动的时机，这将会增加 项目成功的机会，同时防止一旦投入生产后出现不适当的数据活动。例如，几乎在 每个项目中，都存在与收集需求相关的某种类型的行动。可用不同的名称来称呼 这一阶段，然而，如果正在收集需求，应在项目规划阶段考虑某些数据活动。

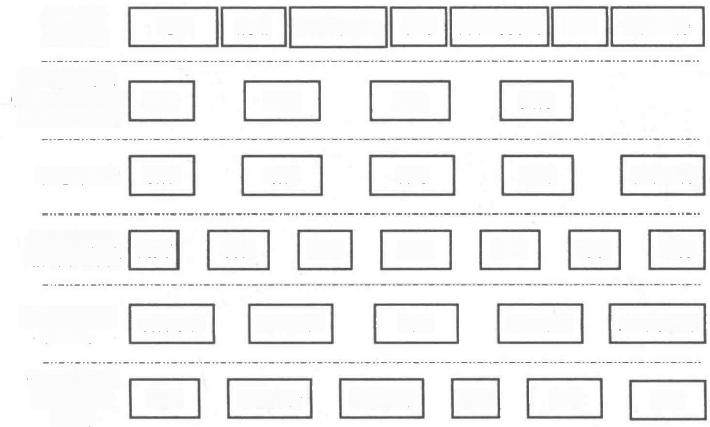
可采用不同的方式来构建一个项目生命周期，并且许多公司和供应商都有他 们自己特有的方式。图1-2呈现了一个典型项目生命周期的各个阶段，以及其他 五个项目生命周期变种的各个阶段(McGilvray,2010)。

无论选择使用图1-2中的哪种全生命周期过程模型，数据工程都应当将该模 型整合进项目方案。在项目初始阶段认真规划，并确保将适当的数据活动整合进 整个信息项目中。与信息系统项目的生命周期类似，数据的生命周期也有不同的 构建方式。数据产品的全生命周期过程是一个相对概念，其定义依赖于问题主体 所涉及数据产品的具体阶段。如为了解决数据的全面存储管理，企业存储研究小 组(Enterprise Storage Group)将数据的全生命周期过程划分为创建、修改、发布、利 用和删除/归档等几个阶段，在不同的阶段，这些数据的利用价值也会不同。因此， 需要对这些数据在不同阶段实施不同的数据存储管理，技术手段以分级存储管理 (Hierarchical Storage Management)为主。IBM公司的信息生命周期管理(Informa-

第 1 章 绪 论

tion Lifecycle Management)方案就是对不同的数据进行贯穿其整个生命周期的管 理，通过完整的信息生命周期管理解决方案，可以让不同类型的数据存放在合适的 存储设备上，利用适当的技术手段对这些数据进行处理和分析；另外，美国康孚 (CommVault)系统公司的 CommVault Simpana一体化信息管理平台也是依据数据

全生命周期过程的定义进行数据全生命周期存储管理的 。 不同数据的生命周期比 较如图1-3所示。



典型项目

生命周期

统一开发过程

(Rational Unified

Precess,RUP)

改进的RUP

定义 发现 设计 配置

快速SAP方法

(ASAP)

Oracle应用软

件集成方法

(AIM)

初始 细化 构造

初始 细化 构造

论证 规划 需求与分析 设计

Siebel电子线 路图实施方法

开发与测试 部署

验证 部署

上线与支持

企业蓝图

项目准备

最终准备

方案设计

后期支持

操作分析

日常支持

创建

生产

维持

交付

交付

交付

定义

实现

图 1 - 2 不同项目全生命周期过程模型对比图



规划 获取 存储和共享 维护 应用 报废

创建 修改 发布 应用 删除/归档

描述 获取 管理 应用

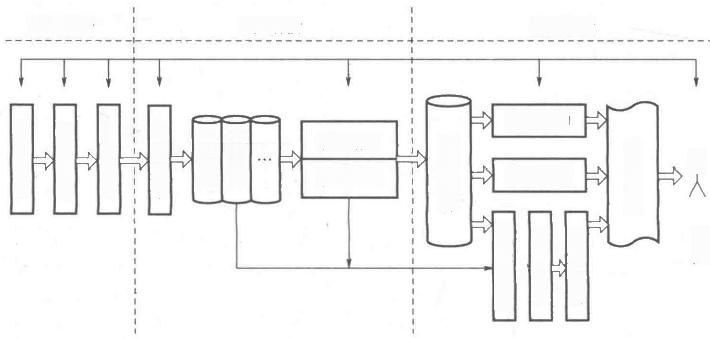
设计 生产 应用

图 1 - 3 数据生命周期对比图

尽 管 图 1 - 3 中 各 数 据 生 命 周 期 的 划 分 各 不 相 同 ， 但 各 种 划 分 都 包 括 “ 应 用 ” 阶段 。 只有当数据被应用时才能从信息中获取价值 。 数据生命周期的所有阶段都 需要花费成本，数据工程的最终效果受各阶段的影响。

)数据质量导论

数据管理以用户获取有用信息为目标，当前大多沿用的是传统数据存储管理 对数据全生命周期过程的定义，数据生命的起点一般从采集(生产)开始，没有涉 及设计阶段。而在产品设计阶段关注产品质量尤为重要，设计质量影响产品全生 命周期内产品成本的80%以上，为此，以用户获取有用信息为目标的数据生命周 期过程的起点应向前延伸。如图1-4所示，以满足信息系统用户对数据质量的需 求为目标，建立数据全生命周期过程模型。



子 用 户

有用信息

模式描述

数据挖掘

选择变换

数据仓库

数 据 库

数据库

数据采集

详细设计

概要设计

需 求 分 析

数据查询

数据分析

数据清洗

数据集成

使用阶段

生产阶段

设计阶段

H

图1-4 数据全生命周期过程模型

在图1-4中，以用户得到有用信息为最终目标来规划数据的设计、生产和 使用。

如前所述，将数据作为产品进行管理是20世纪末期提出的数据管理思想，而 产品全生命周期管理思想也是当前正在发展中的管理理念，数据全生命周期管理 具有如下特点：

(1)主体是数字化数据。作为信息时代标志的数据产品是指计算机可以直接 存储和处理的数字化数据，不包括数字化转化前的传统的纸质数据或其他形式的 数据。

(2)数据全生命周期的界定具有相对性。虽然有些数据对企业来说已经毫无 意义，存储管理全生命周期过程已经结束，但在很多情况下，它对竞争对手并未失 去价值，其生命周期也没有终止，所以数据的全生命周期过程是相对的，图1-4的 数据全生命周期模型是面向数据分析用户的。

(3)以数据的数据集成为基础。产品全生命周期管理是以产品的数据集成为 基础的，所以，数据产品全生命周期过程管理要以数据的数据集成为基础。

数据工程项目的基础是与其相关的生命周期定义，与数据工程项目相关的生 命周期之间的关系如图1-5所示。

第1章绪论 9

图1-5中的生命周期由外向内具有包含关系，但是内部生命周期对外部生命 周期又会产生直接影响。因此，从概念层面应由外向内明确各生命周期的定义以 及生命周期中各阶段与数据工程项目相关的职责、任务、角色；而从实施层面应从 内向外明确具体行动以及各行动与上一层生命周期的关系。



信息质量生命周期

数据生命周期

数据行动生命周期

数据角色生命周期

图1-5 数据工程涉及生命周期的嵌套关系示意图

**1.2.3** **我国数据工程建设面临的问题**

随着我国各行各业信息化建设不断向广度和深度推进，数据工程建设取得了 长足的进步，信息化建设的重点由基础设施和业务应用系统建设，正在转入信息资 源的开发和利用。当前，我国数据工程建设中仍然存在许多未解决的问题，并且不 断暴露出新问题。

首先，将数据看成信息系统的辅助部分的观点仍不在少数。以“数据”为核心 构建信息系统的思想还没有达成普遍共识，特别是没有得到决策层的一致认可，因 此，信息化项目的运作模式没有得到根本转变，信息系统中的数据工程行动难以有 效组织推行。这种观点直接影响信息系统项目的任务目标与具体实施方案，决定 数据工程在整个信息系统项目中所占的比重低、资源投入少，最终形成越到项目后 期数据问题越严重，直至进入信息系统大规模早期失效的怪圈(吴建明，2004)。

其次，分散建设遗留问题严重，统管工作滞后，企业级数据集成工作仍面临困 难。 一是多个部门按不同的业务渠道，分别进行数据建设，重点考虑各自的业务需 求，对数据共享或互操作性考虑不够，同时，数据采集内容有交叉，存在重复性采 集，最终导致一数多源，冲突严重。二是相关数据规范、标准建设存在重复，制定流 程不合理，造成权威性不强，甚至存在冲突，最终导致推行困难。三是目标不透明， 下级单位只知道按照要求完成任务，不理解实际需求，也看不到收益，直接影响工 作效率和工作效果。

再次，“数据量大”与“无数据可用”的矛盾突出。 一是现有数据与信息系统的

数据质量导论

真正数据需求存在偏差。有的大型信息系统投入使用后，经常会遇到数据的完整 性、准确性、可信度达不到要求，造成“有数据不可用，有数据不敢用”,导致信息系 统迟迟得不到应用，难以产生效益。二是现有数据质量控制手段落后，影响到数据 质量和数据管理效率。 一个典型的数据采集任务，数据汇总时，数据整理和清洗的 任务繁重，甚至占到整个项目工作量的80%以上，相当多的模式级数据质量问题， 有些实例级数据质量问题因脱离最初产生环境而不易解决，甚至不可能解决。三 是差的数据质量直接影响到信息系统的功能完善和深度开发应用。低劣的数据质 量不能保障数据分析方法的正常使用和分析结果的正确性，甚至严重扭曲分析结 果，进而间接影响到信息驱动方法的潜在效益。

从以上分析可以看出，当前我国数据工程建设所面临的问题绝大多数属于数 据质量的范畴。数据质量问题已被美军列为信息化建设面临的、正致力解决的首 要问题，而我国信息化建设阶段相对滞后且不平衡，多数尚未完成企业级数据集 成，甚至面临的首要问题是“业务需求不明确”,“数据质量问题”位列其次(刘飞， 2008),这也是数据质量得不到充分重视的客观原因。但是，数据质量问题客观上 正在阻碍我国信息化建设进程，将其作为数据工程的重要内容，加强数据质量研究 与实践，切实提高数据质量势在必行。

**1.3** **数据质量概述**

最近20余年，在麻省理工学院的全面数据质量管理计划(Total Data Quality Management Program,TDQM)的引领下，数据质量已成为一门发展迅速的新兴学 科。从宏观上说，数据质量研究的目标是确保正确的利益相关者在正确的地点和 时间，拥有正确格式的正确信息。无独有偶，云计算(Cloud Computing)、智慧的地 球(Smarter Planet)、感知中国(Sensing China)、美军全球信息栅格(Global Informa- tion Grid)等近年出现的新的信息化战略同样以此为目标。从另一个角度讲，“数 据质量”是实施当前各种信息化战略的重要内容。

**1.3.1** **数据质量的含义**

虽然有观点认为“数据质量”是一个面向目标的不可正式定义的术语(Lueeb- ber et al.,2003;Jeusfeld et al.,1998),但是，从具体研究需要出发，有关文献仍然从 不同角度对数据质量进行了定义。

**定义1** **-** **1(Aebi et al.,1993)** :数据质量是指模式和数据实例的一致性、正确 性、完整性和最小性四个指标在信息系统中满足的程度。

**定义1-** **2(Huang et al.,1998;Lee et al.,2003):**数据质量是数据适合使用的

第 1 章 绪 论 

程度。

**定义1** **-** **3(Kahn et al.,1998;Cappiello et al.,2004)**:数据质量是数据满足特 定用户期望的程度。

作者认为，数据质量与车辆质量、电器质量等属于同等的概念范畴，是可定义 的，著名质量管理专家朱兰博士(Joseph, 美国),从用户的角度将质量定义为(Jo- seph et al.,2003):质量是适用性，即产品使用过程中成功地满足用户目标的程度。 显然，定义1-2和定义1-3与此质量定义一致。

国际标准化组织(International Organization for Standardization,ISO)曾三次给出 质量定义：

(1)反映产品或服务满足明确或隐含需要能力的特征和特性的总和 (ISO8402:1986《质量术语》)。

(2)反映实体满足明确和隐含需要的能力的特性总和(ISO8402:1994《质量 管理和质量保证术语》)。

(3)质量是一组固有特性满足要求的程度(ISO9000:2000《质量管理体系基 础和术语》)。

ISO 对质量的定义进行了不断修订和完善，1994年将“产品或服务”扩展到 “实体”,即“可单独描述和研究的事物”,包括活动或过程、产品、组织、体系或人， 以及它们的任何组合，说明20世纪90年代质量工作领域比20世纪80 年代有所 拓宽，但“实体”的概念不易被人理解和接受，2000 年的定义更加简明。

由1994年ISO 质量定义：在某些文献中，质量是指“适用性”“适合目的”“顾 客满意或符合要求”。按照质量的定义，这些仅表示了质量的某些方面。所以，定 义1-2和定义1-3只反映了质量的某些方面，具有片面性。术语“质量”不应作 为一个单一的术语来表示在比较意义上的优良程度，也不应用于定量意义上的技 术评价。而定义1-1直接将质量定义为4个量化指标，具有明显的局限性。定义 1-1至定义1-3均不能全面反映数据质量的含义，对进一步的数据质量控制研 究有不利影响。

根据以上分析，参照ISO 9000:2000《质量管理体系基础和术语》中的质量定 义，给出数据质量的定义。

**定义1-4:**数据质量是数据的一组固有特性满足要求的程度。

数据质量具有如下特点：

(1)要求会随着时间变化，数据质量也会随时间发生变化。

(2)数据质量是数据的本质属性，要求不同表现也不同，但不影响其他要求下 质量的客观存在。

(3)数据质量可以借助信息系统来度量，但独立于信息系统而存在。

(4)数据质量之间可能存在矛盾，某一特性的提高，可能导致另一特性的

数据质量导论

降低。

(5)数据质量存在于数据的整个生命周期，随着数据的产生而产生，随着数据 的消失而消失。

**1.3.2** **数据全生命周期质量管理**

正如质量管理大师戴明(Deming) 博士所言：“不要对质量的修正流程抱有幻 想，在生产的第一个环节中提高质量才是最有效的方法。 ”(Joseph et al.,2003)因 此，数据全生命周期质量管理实施应分布在数据全生命周期的各阶段，是一项基于 全生命周期过程综合优化的质量管理工程。

[**1.3.2.1**](1.3.2.1) **数据设计阶段的质量管理**

当前在信息系统研发过程中，在数据设计阶段对数据质量还不够重视，在产品 设计阶段关注产品质量尤为重要，设计质量影响到普通产品全生命周期内成本的 80%以上(李成刚，等，1997;周康渠，等，2002)。首先列举几个典型的设计缺陷引 起的数据质量问题：

(1)局限性。如在有的个人情况表格中有年龄一项，但年龄本身是随时间变 化的，更合理的是填入出生时间，根据出生时间计算出年龄。

(2)针对性。如某调查公司进行的面向普通消费者的有关热水器的调查活动 中，所设计的问题过于专业，以至于被调查者对多数问题看不懂，只能凭感觉填写 问卷，因设计的问题针对性差，直接导致所获取的数据使用价值低。

(3)必要性。如在某有关博士后政策的调查中，调查对象为博士后管理人员、 博士后合作导师和博士后工作人员，其中有关博士后出站之后待遇的调查内容，因 受调查对象的限制，所得结果对决策参考价值不大。

(4)歧义性。有多对一和一对多两种。多对一的情况如某项目申请书中有 “申请资助金额”和“项目经费预算总金额”两个输入项，因缺乏必要说明，相当多 的填报者将二者均填写成了本次所申请项目的经费； 一对多的情况如身份证号码 输入项，对军人而言有军官证号码和居民身份证号码两种正确输入。

(5)适用性。用户对数据质量的要求是十分复杂的，在设计阶段制定的数据 标准、数据结构和采集内容等并不一定能满足数据分析用户的要求。

以上实例说明在数据设计阶段进行质量管理是很有必要的，相关理论和技术 的研究还较少，当前可行的方法是通过调研、咨询等措施，在设计阶段尽可能考虑 周全。

[**1.3.2.2**](1.3.2.2) **数据生产阶段的质量管理**

当前的数据质量管理工作主要集中在此阶段，所研究的问题不仅包括当前数 据生产过程的问题，还包括数据设计阶段积累下的问题(集中在模式级),主要技

第 1 章 绪 论 13

术手段包括数据集成(Data Integration)和数据清洗(Data Cleaning)(陈伟，2004)。 当前，这一阶段数据质量管理工作过于繁重，设计缺陷带来的问题经过生产阶段后 演变得更加复杂，并对生产过程产生直接影响，这一现状的缓解也依赖于将部分的 数据质量管理工作追溯到设计阶段解决。

[**1.3.2.3**](1.3.2.3) **数据使用阶段的质量管理**

在如图1-4所示的使用阶段中，如在相对较长的时间内不对数据库做修改操 作，则对数据分析用户而言数据是静态的。虽然数据的选择和变换对数据挖掘结 果有直接影响，但一般将数据的选择和变换作为数据挖掘的一部分来考虑，这种情 况下较少直接涉及数据质量管理问题，但可以通过应用发现新的数据质量问题以 改进数据质量管理方法。

对于动态数据的应用情况，由于数据在使用过程中的状态是不断变化的，极端 情况下在每一次添加、删除、替换后，数据的质量都可能发生改变，此时，及时对数 据质量进行状态监测是数据质量管理的前提(曹建军，等，2007)。

**1.3.3** **数据质量问题的来源**

结合本书的研究工作重点，主要讨论实例级数据质量问题的来源。数据在其 生命周期内，要经历人员交互、计算、传输等操作步骤，每一环节都可能引入错误， 产生数据质量问题，导致数据异常。

1)数据录入(更新)错误

首先，当录入人员从声音中提取信息，或依据书写、打印资料源键入数据时，由 于印刷错误或对原始数据资料的曲解，造成数据录入错误；其次，当录入人员不知 道正确值时，经常编造一个容易输入的默认值，或他们认为的典型值录入，通过引 入“脏数据”以达到所谓的伪完整性(Spurious Integrity),这样的数据通常可以通过 数据完整性约束设计和数据输入系统的初步数据完整性测试，而在数据库中 没有事实上无意义或异常的迹象，具有相当的隐蔽性；再次，软件缺陷也可造 成数据异常，如删除某条记录时，没有删除与其关联的记录，而又进行添加操 作引起的错误。

2)测量错误

不包括测量工具本身的问题，主要指两种人为引入的数据异常。 一是无意的 人为错误，例如方案问题(如不合适的调研和采集策略等),以及方案执行中的问 题(如测量工具误用等);二是有意的人为舞弊，即出于某种不良意图的造假，如虚 报训练成绩、夸大装备完好率等，这类数据可以直接导致信息系统决策错误，造成 严重后果。并不是所有的数据造假就一定对应数据质量下降，数据造假也可以有 正面作用(吴建明，2004)。

14)数据质量导论

3)简化错误

许多情况下，原始数据入库之前需要预处理和简化，这一过程中多种操作可以 导致信息损失。例如：减少原始数据复杂性或噪声；执行数据库管理员所不了解的 域统计分析；经常使用的一些以减少数据占用存储空间为目的的简化处理(如数 据稀疏化)。以上操作导致在入库的简化数据中，或与简化相关的最终分析中产 生错误风险，造成数据质量下降。

4)数据集成错误

在数据库建立过程中，将多个数据源中的数据并入一个数据库是常见的操作， 这种数据集成任务需要解决数据库之间的不一致或冲突问题，在模式级主要是命 名冲突和结构冲突(Joseph,2008;Rahm et al.,2000),在实例级主要指因集成而产 生的相似重复问题(Parent et al.,1998;Milo et al.,1997)(可以看成数据集成的一 部分，但解决此类问题属于数据清洗的范畴)。吴建明(2004)和 Rahm(2000)将各 数据源之间的不一致和冲突视为数据质量问题，其实，在没有事先约定的情况下 (这种情况是常见的),多数据源之间的冲突和不一致的存在是正常的，合并过程 中解决这些不一致和冲突是数据集成过程中必要的也是正常的操作步骤，模式级 的问题主要用数据抽取、转换和加载(Data Extraction,Transformation and Loading, 数据 ETL)工具(王新英，等，2004)来解决。问题是为解决多数据源之间的不一致 和冲突时，在基于多数据源的数据集成过程中可能导致数据异常，如不一致和冲突 的解决不彻底，甚至引人新的异常，再如冲突记录的误识别，因此，数据集成是数据 质量问题的一个来源，而数据集成本身(主要指模式级冲突的解决)应视为数据生 产过程的正常操作。

**1.3.4** **数据质量研究发展简况**

20世纪80年代末Wang和 Madnick 开始应用系统方法研究数据质量相关问 题(Wang et al.,1989),主要识别和解决从包含交叉重叠记录(Overlapping Re- cords)的多个数据源进行信息集成时，出现的实体分辨(Entity Resolution)问题。 相关问题研究者进行方法探索，用以确定不同的记录实际上是否对应于同一实体。 随后，Wang和 Madnick 提出了查询处理机(Query Processor)中考虑数据来源标记 的多源模型(Wang et al.,1990),该模型能够回答诸如“数据来自哪里?”“得到这 一数据用到了哪些数据源?”等问题。随后的研究包括在概念数据库设计阶段系 统捕获综合数据质量标准作为元数据的建模方法(即质量实体关联关系模型 (Quality Entity Relationship Model))(Wang et al.,1993;Storey et al.,1998),并应用 扩展的关系代数(Relation Algebra)实现查询处理机对层次数据质量元数据的处理 (Wang et al.,1995)。这一研究方向在诸如数据溯源(Data Provenance)与数据世

第 1 章 绪 论 **15**

系(Data Lineage)等现代数据库研究与设计(Buneman et al.,2001),以及针对数据 安全性(Data Security)与数据私密性(Data Privacy)管理的其他拓展应用中产生了 影响。更重要的是，这些早期的研究工作激发了研究者对数据质量问题的全方位 探索，特别是麻省理工学院的学者们敏锐地意识到这是一个具有发展潜力的新研 究领域。

20世纪90年代初，麻省理工学院的Wang首次使用了“信息质量”一词，并启 动了全面数据质量管理计划(Total Data Quality Management Program,TDQM)。该 计划的首要贡献便是得出了“将数据作为产品进行管理”的研究结论，并借鉴普通 产品的“全面质量管理(Total Quality Management)”思想，提出了“全面数据质量管 理”的思想。全面数据质量管理的思想很快被美军采纳，美国国防部也成立了称 为“全面数据质量管理”的研究机构。全面数据质量管理计划每年承担超过50项 的数据质量方面项目，其中不乏军事项目，如美国海军研究部项目“Data Cleansing Research Project N00014-99-1-0730”。

全面数据质量管理计划引领了数据质量领域20余年的发展，产生了广泛影 响。由TDQM 发起的信息质量国际会议(International Conference on Information Quality,ICIQ)自1996年起已召开21届(前13届均在 MIT 召开，ICIQ 2009 在德国 波茨坦的哈索·普拉特纳研究所(Hasso Plattner Instiute,HPI)召开，ICIQ 2010 在 美国阿肯色州的阿肯色大学召开，ICIQ 2011 在澳大利亚阿德莱德市南澳大学召 开，ICIQ2012 在法国巴黎的法国国立工艺学院召开，ICIQ 2013 在美国阿肯色州的 阿肯色大学召开，ICIQ2014 在中国西安交通大学召开，ICIQ 2015 在美国 MIT 召 开，ICIQ2016 在西班牙雷阿尔城卡斯蒂利亚拉曼查大学召开，ICIQ 2017将再次重 返阿肯色大学)。传统的VLBD、SIGMOD、ICDE等高级别数据工程领域会议也设 有数据质量专题研讨，如 SIGMOD 的信息系统中的信息质量专题研讨会(SIG- MOD Workshop on Information Quality in Information Systems) 、GAiSE 信息质量专 题研讨会(CAiSE Workshop on Information Quality)。面向应用的 MIT 信息质量 工业研讨会(MIT IQ Industry Symposium Program,IQIS)自2007年起已召开10 届。IQIS自2012年起更名为首席数据官与信息质量研讨会(Chief Data Officer and Information Quality Symposium,CDOIQ),这一变化说明了对信息质量工作中 专职人员的重视。

2001年，美国国会通过了《数据质量法》。2002年，美国国会决定由美国总统 事务办公厅管理和预算办公室发布政府信息质量指南，并要求每个联邦机构在 2002年10月1日前，结合本部门实际发布数据质量指南，建立管理机制使得数据 的使用者能够发现并得到联邦机构所维护与传播的正确信息，同时要定期向主管 报告。

2006年，在全面数据质量管理计划的帮助下，美国阿肯色大学设立了世界首

16)数据质量导论

个数据质量硕士和博士学位授予点，开始持续培养高层次数据质量人才。当前阿 肯色大学重点关注信息系统的实体完整性问题，在实体分辨(相似重复记录检测) 方面做了系统深入的研究。

2009年，全面数据质量管理计划创办了数据质量学术期刊——ACM Journal of Data and Information Quality(JDIQ),其中首篇论文系统讨论了数据质量的研究内 容，将数据质量分为4大类19个研究主题，从而形成数据质量的完整研究框架 体系。

越来越多的组织设置首席数据官(Chief Data Officer,CDO)或具有类似职责的 高级主管来监管数据生产过程并管理数据改进举措。 一些组织甚至开始使用信息 战略家(Information Strategist)头衔来表明数据质量对组织战略的重要性。2009 年，Wang由 MIT 调往白宫，任美军首席数据质量官(Chief Data Quality Officer, CDQO),专职负责数据质量工作。

国内方面，解放军军械工程学院、解放军国防科学技术大学、清华大学、西安 交通大学等科研机构在数据质量方面做了一些富有成效的工作，研究内容主要 涉及数据清洗、数据质量评估等。近年国内对数据质量的关注度有所提高，“数 据质量管理的基础理论与关键技术(F020204)”还被列为了国家自然科学基金 委信息科学部的2011年度重点项目研究领域。但当前国内数据质量整体上仍 然以零星分散研究为主，尚未形成合力，尤其缺乏针对我国信息环境特点的研究 与实践。

**1.4** **本书内容结构安排**

本书的相关工作先后得到国家自然科学基金面上项目“基于蚁群算法和云模 型的领域无关数据清洗(No.61371196)”,中国博士后科学基金特别资助项目和中 国博士后科学基金面上项目“××信息质量控制方法研究及应用(No.201003797, 20090461425)”,江苏省博士后科研资助计划项目(No.0901014B)“基于蚁群算法 的领域无关数据清洗方法研究”,解放军理工大学预研基金项目(No.20110604) “业务领域无关的数据清洗方法研究”,解放军理工大学预研基金项目 (No.41150301)“依赖模式挖掘及在缺失数据处理中的应用”等基金项目的支持。

本书的具体内容结构安排如下：

第1章至第3章是本书的总述部分。

第1章是绪论。从分析国内数据工程建设现状入手引出数据质量问题，介绍 了数据质量以及数据全生命周期质量管理的含义，分析了数据质量问题的来源并 归纳了其研究发展简况；给出了本书内容结构安排。

第2章分析构建了数据质量研究和数据清洗系统框架。重点从研究主题和研

第 1 章 绪论 

究方法两个维度，分析了数据质量研究的框架；分析了数据质量管理的传统发展模 式，引入了数据质量管理的并行发展模式，分析了制约数据质量管理发展的因素和 当前数据质量研究存在的误区，对数据质量管理提出了建议；构建了适用于国内数 据组织模式的层次结构数据质量控制框架，分析了其实现所涉及的关键问题，在进 一步辨析数据清洗概念的基础上，构建了数据清洗的一般性系统框架。本章工作 从实施框架层面对数据质量和数据清洗做了全面深入的分析，给后续章节工作提 供了清晰定位。

第3章综述了典型数据清洗技术的发展动态。系统归纳了实体分辨、不完整 数据、不一致数据三类实例层数据质量问题的数据清洗技术发展动态。

第4章至第10章是实体分辨、缺失数据和不一致数据三类数据清洗技术的研 究成果。

第4章研究了实体分辨中的数据分块方法。针对基于冗余的数据分块方法， 提出了一种具有线性复杂度的冗余记录对识别方法；提出了一种基于空间映射的 数据块约减方法；针对XML的数据分块，提出了一种基于Canopy 聚类的数据分块 方法。

第5章研究了实体分辨中的相似度算法。针对中西文混合字符串，提出一种 融合多种编辑距离的字符串相似度计算方法；为了提高属性相似度的准确性，提出 了基于函数依赖的属性相似度调整方法。

第6章研究了基于关系的实体分辨。为了解决基于特征(属性)相似度(Fea- ture Based Similarity,FBS)的方法到基于关系的数据清洗(Relationship-based Data Cleaning,RelDC) 切换的问题，提出基于云模型的方法和基于邻域粗糙集的方法来 选择 FBS所不能有效识别的记录对；为了解决基于关系的实体分辨在计算路径权 重过程中对训练数据的数量和必须满足的条件要求过于严格的问题，提出了一种 基于关系类型的自适应实体分辨方法。

第7章研究了不完整数据的分类与检测。提出了基于位运算的不完整记录分 类与检测方法，介绍了基于统计关系的不完整数据分类。

第8章研究了不完整数据的估计与填充。提出了基于统计关系学习的填充算 法和基于机器学习的填充算法；将数据生成当作缺失的一种极端情况(即所有属 性都缺失),提出了一种基于函数依赖一致性的数据生成方法。

第9章研究了条件函数依赖挖掘及其优化方法。给出了条件函数依赖的相关 基本概念，介绍了函数依赖挖掘和条件函数依赖挖掘的常用方法，提出了一种新的 基于开项集剪枝的常量条件函数依赖挖掘算法。

第10章研究了基于规则的不一致数据检测与修复方法。提出了基于 Fellegi- Holt方法的不一致数据检测策略，实现了基于已知规则自动推演出完备的最小规 则集，并基于该规则集进行不一致数据检测；提出了一种基于Evidence-Rules 模

)数据质量导论

型的不一致数据修复方法，基于上述检测结果，利用原有记录信息，有效提高了修 复的正确性和效率。

第11 章研究了数据质量工具的发展概况及设计方法。综述了数据质量工具 的发展概况；研究了基于表达式树的数据质量工具设计和基于流程的数据质量工 具设计方法。

第12章研究了大数据与大数据质量问题。在归纳大数据时代特征的基础上， 总结提出了大数据质量面临的挑战；研究了数据治理的动机及 一 般流程，提出了适 应我国信息环境特点的基于相对不变过程的数据治理系统框架。

**参考文献**

[1]王元卓，靳小龙，程学旗.2013.网络大数据：现状与展望[J]. 计算机学报，36(6):1125-1138. [2]王新英，陈语林.2004.数据抽取、转换、装载综述[J]. 企业技术开发，23(8):3-5.

[3]刘飞.2008.中国企业数据集成与数据质量市场白皮书[R]. 北京：IDC.

[4]李成刚，胡迪青.1997.设计领域发展的新趋势[J]. 中国机械工程，8(6):56-59.

[5]李国杰.2012a. 大数据研究的科学价值[J]. 中国计算机学会通讯，8(9):8-15.

[6]李国杰，程学旗.2012b. 大数据研究：未来科技及经济社会发展的重大战略领域[J]. 中国科学院院刊， 27:647-657.

[7]吴建明.2004.病态信息理论及其在装备保障中的应用[D]. 石家庄：军械工程学院.

[8]陈伟.2004.数据清理关键技术及其软件平台的研究与应用[D]. 南京：南京航空航天大学.

[9]周康渠，徐宗俊，郭钢.2002.制造业新的管理理念—产品全生命周期管理[J]. 中国机械工程，13(15); 1343 -1346.

[10]曹建军，张培林，石志勇，等.2007.大型移动多机电设备在线状态监控系统研究[J]. 兵工学报， 28(9):58-62.

[11]曹建军，刁兴春，汪挺，等.2010.领域无关数据清洗研究综述[J]. 计算机科学，37(5):26-29. [12]韩京宇，徐立臻，董逸生.2008.数据质量研究综述[J]. 计算机科学，35(2):1-5,12.

[13]程学旗，靳小龙，杨靖，等.2016.大数据技术进展与发展[J]. 科技导报，34(14):49-59.

[14]McGilvray D.2010. 数据质量工程实践[M]. 刁兴春，曹建军，张健美，等，译.北京：电子工业出版社. [15]Joseph M J,Blanton G A,Robert E H,et al.2003. 朱兰质量手册[M]. 焦叔斌，译.5版.北京：中国人民

大学出版社.

[16]Lee Y W,Pipino L L,Funk J D.2015. 数据质量征途[M]. 黄伟，王嘉寅，苏秦，等，编译.北京：高等教育出 版社.

[17]AebiD,Perrochon L.1993.Towards Improving Data Quality[C]//Proceedings of the International Conference on Information Systems and Management of Data.New Delhi:Towards Improving Data Quality,CISMOD:273-281.

[18]Buneman P,Khanna S,Tan W C.2001.Why and Where:A Characterization of Data Provenance[C].London, UK:Springer:316-330.

[19]Cappiello C,Francalanci C,Pernici B.2004.Data Quality Assessment from User's Perspective[C].Paris, France:International Workshop on Information Quality in Information System(IQIS):68 -73.

[20]Chen ZQ,Kalashnikov D V,Mehrotra S.2005.Exploiting Relationships for Object Consolidation[J]//Proceed-

第 1 章 绪 论 19

ings of the International Workshop on Information Quality in Information Systems(IQIS)Workshop at ACM SIG- MOD Conference.Balimore,Maryland,USA:International Conference on Management of Data(MD):1 -12.

[21]Eppler MJ,René A,Marcus D.2003.Quality Criteria of Content -Driven Websites and Their Influence on Customer Satisfaction and Loyally:an Empirical Test of an Information Quality Framework[C].Califormia,

USA:MIT:108 -120.

[22]Huang K T,Lee Y W,Wang R Y.1998.Quality Information and Knowledge Management[M].New Jersey:Prentice

Hall.

[23]Jeusfeld M A,Quix C,Jarke M.1998.Design and Analysis of Quality Information for Data Warehouses[M]// Proceedings of the 17th International Conference on the Entity Relationship Approach(ER'98),Singapore: Springer,1507:349 -362

[24]Joseph MH.2008.Quantitative Data Cleaning for Large Databases[J/OL].EECS Computer Science Divi-

sion.UC Berkeley,<http://db.cs.berkeley.edu/jmh/cleaning> -unece.pdf.

[25]Kahn B K,Strong D M.1998.Product and Service Performance Model for Information Quality:An Update[C]// Proceedings of the 1998 Conference on Information Quality.Cambridge,Massachusetts,USA:MTT,102-115.

[26]Lee Y W,Strong D M.2003.Knowing-Why About Data Processes and Data Quality[J].Journal of Manage- ment Information Systems,20(3):13-39.

[27]Lueebber D,Grimmer U.2003.Systematic Development of Data Mining Based Data Quality Tools[C]//Pro- ceedings of the 29th VLDB.Berlin,Germany:VLDB.

[28]Madnick S E,Wang R Y.2009.Overview and Framework for Data and Information Quality Research[J].ACM

Joumal of Data and Information Quality(JDIQ),1(1):1-22.

[29]Milo T,Zohar S.1977.Using Schema Matching to Simplify Heterogeneous Data Translation[C]//Proceedings of 24th International Conference on Very Large Databases.New York:Morgan Kaufmann,122-133.

[30]Parent C,Spaccapietra S.1998.Issues and Approaches of Database Integration[J].Communications of the ACM,41(5):166-178.

[31]Rahm E,Do HH.2000.Data Cleaning:Problems and Current Approaches[J].IEEE Data Engineer Bulletin,23 (4):3-13.

[32]Sara Cushman.2002.Billions Lost to Poor Data Quality[J/OL].<http://searchdatabase.techtarget.com>.

[33]Storey V,Wang R Y.1998.Modeling Quality Requirements in Conceptual Database Design[C]//Proceedings of the International Conference on Information Quality.Cambridge,Massachusetts,USA:MTT:64-87.

[34]Wang R Y,Madnick S E.1989.The Inter-database Instance Identification Problem in Integrating Autonomous Systems[C]//Proceedings of the 5th International Conference on Data Engineering Los Angeles,California,

USA:IEEE Computer Society:46-55.

[35]Wang RY,Madnick S E.1990.A Polygen Model for Heterogeneous Database Systems:The Source Tagging Per- spective[C]//Proceedings of the 16th VLDB Conference.Brisbane,Queensland,Australia:VLDB:519 -538.

[36]Wang R Y,Kon H B,Madnick S E.1993.Data Quality Requirements Analysis and Modeling[C]//Proceedings of the 9th International Conference of Data Engineering.Vienna,Austria:IEEE Computer Society,670-677.

[37]Wang R Y,Reddy M,Kon H.1995.Toward Quality Data;An Attribute -based Approach[J].Decision Support System(13):349-372.

**第2章** **数据质量研究和数据清洗系统框架**

**2.1** **引言**

最近20余年，在麻省理工学院的全面数据质量管理计划的引领下，数据质量 研究得到了迅速发展，发现和解决了许多数据质量问题。然而，数据质量作为一门 新兴学科，除了面临的具体理论、方法和技术难题外，还缺乏对其框架体系的统一 认识，尤其是国内，在数据质量研究的基本内容等问题的认识上还存在局限性，直 接影响到数据质量研究与实践的健康发展。另外，具体到数据质量控制，直接局限 于某项数据质量提高手段，只能解决部分局部性问题，不利于问题的最终全面解 决，甚至不利于明确地确立阶段性目标。

本章借鉴国内外研究成果，重点从主题(Topics)和方法(Methods)两个维度阐 述数据质量研究框架(Madnick et al.,2009),对数据质量的研究体系进行梳理的 同时，引用已有典型成果对这一框架的具体维度进行举例说明，试图清晰地描绘出 该领域的发展脉络和发展趋势。然后，结合国内数据工程建设环境的特点，研究数 据质量管理发展模式和数据质量控制框架，并进一步剖析数据清洗的任务目标，构 建数据清洗系统框架。

**2.2** **数据质量研究框架**

本节综述数据质量研究框架，重点介绍 Madnick 的数据质量研究框架，试图给 出数据质量研究范畴的完整视图。

**2.2.1** **典型的数据质量框架**

*Wang、Storey 和Firth 于1995年8月在IEEE Transactions on Knowledge and Da- ta* *Engineering上发表了题为“A* *Framework* *for* *Analysis* *of* *Data* *Quality* *Research”的*

论文，提出了数据质量研究的一个早期框架(Richard et al.,1995)。该框架改编自 ISO 9000,将有形产品(Physical Products)和信息产品(Information Products)进行类 比，是一个基于产品生产过程的框架，对超过100多项处理数据质量问题或与数据

第2章 数据质量研究和数据清洗系统框架 (21

质量管理问题密切相关的文献进行分类。该框架包括对信息质量产生影响的7个 要素，相关描述列于表2-1。

表2 - 1 Wang的数据质量研究的框架

|  |  |
| --- | --- |
| **要素** | **描述** |
| 1.管理职责 | ·共同数据质量政策的制定  数据质量系统的建立 |
| 2.运营和保障成本 | ·运营成本包括预防、评价与故障费用  ·保障成本和用户及管理者所需的质量证明与论证相关 |
| 3.研究和开发 | ·数据质量维度的定义以及它们的测量  ·数据产品质量特性的分析和设计  融入数据质量特性的数据生产系统设计 |
| 4.生产 | ·数据产品生产所需的原始数据获取、组成部分、汇集中的质量需求 ·原始数据、工作进展，以及最终数据产品的质量审核  ·不一致数据列项的识别和相关行动说明 |
| 5.发布 | 数据产品的存储、标识、包装、安装、交付、售后服务  数据产品的质量文档和记录 |
| 6.人员管理 | ·雇员对数据质量相关问题的了解  雇员生产高质量数据产品的动力  ·雇员数据质量成绩的评测 |
| 7.法律职能 | ·数据产品的安全性和法律责任 |

尽管表2-1所列的数据质量研究的框架对数据质量的影响要素描述是全面 的，但缺少直观的描述数据质量的关键词集合，描述粒度不够细，不易实际应用。

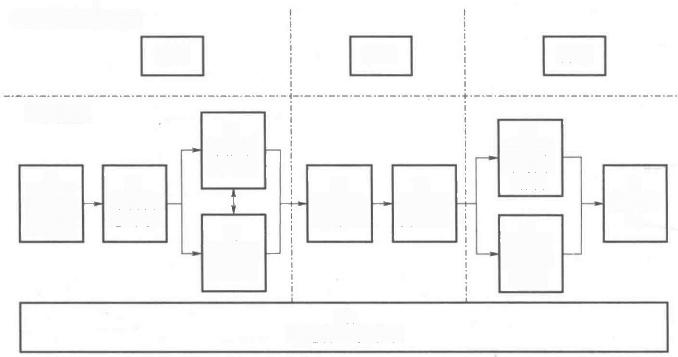
Danette 以数据质量工程实践为基点提出了数据质量提高周期，以及其实施的 10步流程(The Ten Steps Process)(McGilvray,2010),两者的示意图及其的对应关 系如图2-1所示。

如图2-1所示的数据质量提高周期中，评估(Assessment)(评判实际的环境 和数据，并将它们与需求和预期进行比较)是认知的关键，认知(Awareness)(了解 数据和信息的真实状态、对业务的影响及其根本原因)将引起行动(Action)(除纠 正当前数据错误外，还预防未来信息和数据质量问题),行动又由定期评估(Peri- odic Assessments)来核实。因此，信息与数据质量提高周期得以继续循环进行。

图2-1中的10步流程是关于评估、提高以及获得高质量数据的一种实施方 法框架。

有关10步流程的详细说明见表2-2。

数据质量导论



数据质量提高周期

评估 认知 行动

10步流程

7.

1.

定义业

务需求

和方法

纠正当 前数据 错误

10.

沟通行动和结果

预防未 来数据

错误

9.

实施控

制

6.

制订提 高方案

4.

评估业 务影响

2.

分析信 息环境

3.

评估数 据质量

5.

确定根 本原因

8.

图 2 - 1 Danette 的数据质量提高周期和10步流程及两者的对应关系

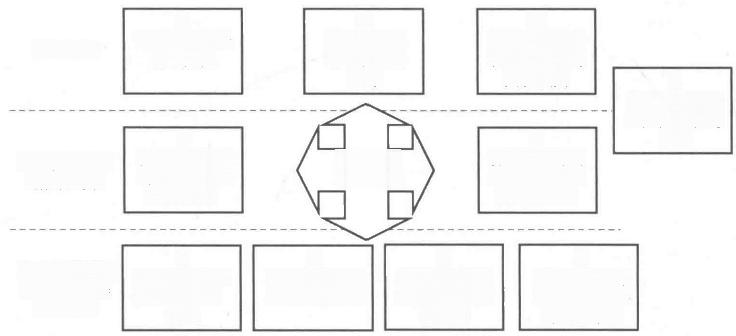
表 2 - 2 Danette 的1 0 步 流 程

|  |  |
| --- | --- |
| 步骤 | 描述 |
| 1.定义业务需求和方法 | 定义和商定问题、时机或目标，以指导整个项目期间的所有工作。为了把目 标放在所有行动的最前面，对这一步的参考应贯穿其他步骤 |
| 2.分析信息环境 | 收集、汇编、分析有关当前形势和信息环境的信息。归档和核实信息生命周 期。信息生命周期为未来步骤提供基础，确保相关数据得以评估，并帮助发 现根本原因。设计数据捕获和评估方案 |
| 3.评估数据质量 | 针对适用于这一问题的数据质量维度，评估数据质量。评估结果为未来步骤 提供基础，例如，确定根本原因、需要的改进和数据纠正 |
| 4.评估业务影响 | 使用各种技术来确定劣质数据对业务的影响。该步骤为建立改进业务案例、 获取信息质量支持、确定适当的信息资源投资提供输入数据 |
| 5.确定根本原因 | 确定引起数据质量问题的真实原因，并区分它们的优先次序，以及为解决这 些问题形成具体的建议 |
| 6.制订提高方案 | 最终确定关于行动的具体建议。基于这些建议制定并且执行提高方案 |
| 7.预防未来数据错误 | 实施解决引起数据质量问题根本原因的解决方案 |
| 8.纠正当前数据错误 | 实施进行适当数据纠正的步骤 |
| 9.实施控制 | 监测和核实所进行的改进。通过标准化、归档和对成功改进进行持续监测， 维护已改良的结果 |
| 10.沟通行动和结果 | 归档与沟通质量测试结果、所做的改进以及这些改进的结果。沟通非常重 要，可成为每一个步骤的一部分 |

第2章数据质量研究和数据清洗系统框架(23

数据质量提高周期是10步流程的基础，10步流程为对关键业务信息进行 持续评估、维护和提高描述了一个实施方法框架。对规划与实施数据质量提 高项目，10步流程是具体操作指南，并且是可定制的、逐步进行的、线性的一 系列事件，但数据质量提高流程是迭代的——项目团队可以返回到先前的步 骤以改善其工作，可以选择满足业务需求的步骤，甚至重复整个10步流程，以 支持持续数据质量提高。Danette 的数据质量工程实践框架在实际数据质量 项目中有较强的可操作性，注重实际效果，但并不过于关注具体的方法技术 细节。

Redman 认为数据质量管理不仅需要相关理论、技术，以及系统的支持，更为重 要的是，还需要一个完善的企业管理机制，为全方位开展数据质量管理工作提供保 证。如图2-2所示的是各类企业拥有高质量数据必须具有的10个要素，及其与 企业中角色之间的对应关系，它们一起构成企业的全方位数据质量管理框架 (Redman,2008)。



10.

促进重视数据

和数据质量的

文化氛围

3.

管理(内部的 和外部的)数 据“提供者”

1.

重点关注最重

要的(客户)

需求

5.

为了消除错误

在各个层面实

施控制

7.

自上而下地设

定并实现改进

目标

2.

管理产生数据

的流程(保证

它们正确)

6.

改进：发现并 消除产生错误 的根本原因

9.

开明、有见

地、敬业的

领导

所有接触数据

的人=四项基 本“任务”

4.

对照客户需求

测量质量水平

8.

明确界定数据

管理责任

中层管理人员

(指挥人员)

各项工作 紧密关联

高层领导

图 2 - 2 Redman 的全方位数据质量管理框架

如图2-2所示的全方位的数据质量管理框架，形成了各类企业获取高质量数 据的环境基础，也是最终实现全面数据质量管理的基础。首先，数据质量管理需要 高层领导、中层管理人员和所有接触数据人员的通力合作，但其前提是企业内部在 理念上应达成高度一致，并形成共同的数据质量目标；其次，数据质量管理工作不 仅涉及用户需求、数据质量测量、消除数据错误和产生错误的原因，要管理产生数 据的流程和数据“提供者”,还要有专门的领导、良好的文化环境以及明确的数据 质量改进目标；再次，图2-2中的10个数据质量管理要素是相互关联的，对一个 具体企业，要结合企业自身实际进行综合考虑。构建如图2-2所示的数据质量管 理框架，需要企业管理高层进行推进。

 数据质量导论

2009年，在JDIQ的第一篇论文中(Madnick,2009),Madnick 、Wang 、Lee 等 提出了一个更易于应用的研究框架，框架包括两个维度：主题和方法。这种维 度划分方式比较简明：任何数据质量研究都是应用某种(几种)研究方法来讨 论某个(几个)主题。对每一维度，选择意义简明的关键词集进行描述，列于 表2 - 3。

表2 - 3 Madnick 的数据质量研究框架

|  |  |
| --- | --- |
| **主题** | **方法** |
| **1.信息质量影响**  **1.1** **应用领域**  **1.2性能、费效比与业务运营**  **1.3** **IT管理**  **1.4** **企业过程及其变化**  **1.5策略与原则**  **2.关系数据库信息质量的技术解决方案**  **2.1数据集成与数据仓库**  **2.2企业架构与概念建模**  **2.3** **实体分辨与企业内部管理**  **2.4** **监测与清洗**  **2.5溯源、世系与来源标识**  **2** **.** **6不确定性**  **3.计算机科学和IT中的信息质量**  **3.1测量与评估**  **3** **.** **2信息系统**  **3** **.** **3** **网** **络**  **3.4协议与标准**  **3** **.** **5私密性**  **3** **.** **6安全性**  **4.监管中的信息质量**  **4.1** **监管的标准与原则**  **4.2** **监管的技术解决方案** | **1.行动研究**  **2.人工智能**  **3.个案研究**  **4.数据挖掘**  **5.设计科学**  **6.经济计量学**  **7.经验方法**  **8.实验方法**  **9.数学建模**  **10.定性方法**  **11.定量方法**  **12.统计分析**  **13.系统设计及实施**  **14.理论和形式化证明** |

表2-3中的主题和方法构成了信息质量的层次结构研究框架，研究内容涉及 4类共19个研究主题，本节以下内容综合 Madnick(2009)和作者的研究积累对这 一框架展开进一步描述。

第2章 数据质量研究和数据清洗系统框架

**2.2.2** **数据质量的研究主题**

信息质量是一个多学科交叉研究领域，目前，研究者还多见于管理信息系统 (Management Information Systems,MIS)和计算机科学(Computer Science)两个学科 领域，应当鼓励更多其他领域研究者参与信息质量研究。多学科交叉性导致信息 质量研究包含更广的主题范围，结合当前及可预见的未来研究主题展开讨论。

[**2.2.2.1**](2.2.2.1) **信息质量影响**

该类主题研究信息质量的影响，提出评价这些影响的方法，并提出使积极影响 最大化和消极影响最小化的设计试验机制。

1)应用领域

研究与信息系统具体应用领域相关的信息质量问题。如用户关系管理(Cus- tomer Relationship Management,CRM)、知识管理(Knowledge Management,KM)、供 应链管理(Supply Chain Management,SCM)和企业资源规划(Enterprise Resource Planning,ERP)等。例如，患者记录中常常包含不准确的属性值，导致医疗人员难 以得到详实的患者信息，给诊断和治疗造成影响。Xu 等人(2002)提出了一个从 正在实施的 ERP系统中识别信息质量问题的框架。

2)性能、费效比与业务运营

研究信息质量对部门(含个人)职能、数据质量行动的费用和效益、业务运营 和决策的影响。差的信息质量可能破坏企业战略和战术的效力，乃至引起更严重 问题发生。Jung 研究了信息质量对决策效能的影响，Sheng初步研究了信息质量 对企业职能的影响，Y.Lee(2004) 等人研究了工作角色认知或求知的模式，以及具 备知识的不同对工作效能的影响。最近研究证明，信息质量和企业成果的关系是 可以系统测量的，通过测量信息质量可以预测企业成果(Slone,2006) 。 这一方向 的研究仍需加强。

3)IT 管理

研究信息质量与IT 管理之间的相互影响。例如，IT投资、首席信息官(Chief Information Officer,CIO)领导能力和IT 治理(IT Governance)。适用性信息质量观 点，决定了在支持企业业务运营和竞争能力的IT 应用中信息质量的重要性，当前， 企业管理信息质量的方式已开始从被动管理(Reactive Management)向主动管理 (Proactive Management)转变，因此，需要更多的经验研究(Empirical Research)度量 主动管理的有效性，并探讨它们的影响。

4)企业过程及其变化

研究信息质量和企业变化过程的相互关系。信息质量提高的要求在企业行动 中常常发生变化，Y.W.Lee(2004) 研究了大规模生产公司的信息质量提高方法，

)数据质量导论

提出了响应业务过程和需求变化的基于迭代自适应技术的数据完整性规则； Y.W.Lee(2004) 的纵向研究建立了一个信息质量问题求解模型，通过对五个机构 的五年的行动研究(Action Research),分析了实践者对行动中信息质量的描述，研 究表明，有经验的人员可以通过上下文(包括隐含其间的)知识来解决信息质量问 题，文中还详细说明了五个关键的信息质量要素：角色(Role)、范例(Paradigm)、时 间(Time)、目的(Goal) 和地点(Place)。

5)策略与原则

研究组织和制度层面信息质量管理和提高的各种策略和原则。Kerr(2006) 研 究了新西兰卫生保健部门所采纳的策略和原则，表明当部门间或企业内部通过信 息质量提高计划进行改革时，所采用的信息质量评估框架(Data Quality Evaluation Framework)和当前国家数据质量提高战略(Data Quality Improvement Strategy)给它 们提供了宏观指导路线。

[**2.2.2.2**](2.2.2.2) **关系数据库信息质量的技术解决方案**

该类主题提出评价、提高和管理信息质量的数据库技术，并提出能够生产出高 质量信息的论证和系统设计技术。

1)数据集成与数据仓库

为了完成分析或决策任务，需要从内部和外部数据源(贸易伙伴、数据供应 商、因特网等)搜集和集成数据。因此，信息系统处理的数据通常是分布式的和异 构的。数据集成(Data Integration)通过改善一致性(Consistency)、完整性(Com- pleteness)、可访问性(Accessibility)及其他信息质量维度来提高数据的可用性。通 过即问即答式多源数据查询应答(Query Answering)系统或者为已知或预期用户而 建立的数据仓库(Data Warehouse),能够实现数据集成。此方向的研究已经持续 20 余年，当前仍然异常活跃。

Madnick 等(2006)提出了一个柔性查询应答系统，称为上下文交互(COntext INterchange,COIN),该系统使用了知识表示(Knowledge Representation)、结合约束 求解(Constraint Solving)的反绎推理(Abductive Reasoning)和查询优化(Query Op- timization)技术，允许用户在多个数据源中查询，而不必关心大多数的句法和语义 的差异。许多所谓的“信息质量”问题实际上是“数据曲解(Data Misinterpreta- tion)”问题，COIN 试图通过数据源和数据用户的上下文解决数据曲解问题，必要 时，将数据转化成用户乐见和便于理解的格式。

实体分辨和模式匹配(Schema Matching)是数据集成解决的另两个问题，实体 分辨将在下文讨论。模式匹配研究如何自动或半自动地进行数据模式匹配，利用 模式匹配结果，数据仓库可以构建全局模式，查询应答系统可以实现从全局模式到 与其匹配的局部模式的查询重写。

通常，通过数据 ETL过程建立数据仓库，并进行快速数据查询，以及获取多维

第2章数据质量研究和数据清洗系统框架 

视图(如季节销售、地区销售等)。Ballou等(1999)提出了提高数据仓库中信息质 量的框架；为了提高信息质量，数据仓库质量(Data Warehouse Quality)项 目(Jarke et al.,1999)提出了一系列描述和管理数据 ETL过程的建模工具。

2)企业架构和概念建模

企业架构(Enterprise Architecture)是认识IT 基础结构和相关业务及管理过程 的框架，旨在提出存货管理(Inventory)、可视化展现、分析和最优化技术，并将它们 的功能与业务需求相关联。企业架构可根据业务目标设计并调整信息系统，这一 调整通常通过查阅书面材料、观察分析系统之间的关系和组织需求完成。企业架 构方法应用广泛，如美国的联邦机构(Federal Agency)需要通过一系列 IT 运营、规 划和预算相关的联邦企业架构(Federal Enterprise Architecture)方案。

概念建模(Conceptual Modeling)允许依据一系列具体业务需求进行数据库设 计，最常用的概念建模技术是实体-关系模型(Entity-Relationship Model)及其扩 展。将信息质量特征加入实体-关系模型是重要的扩展应用之一，如 Storey 把获 取的数据质量需求作为单元级(Cell Level)元数据。另外，扩展后的查询应答系统 可以有效处理信息质量元数据(Wang et al.,1995)。满足信息质量元数据要求的 扩展建模和查询应答机制是当前的研究热点，如质量度量(Quality Metrics)、私密 性、安全性和数据世系研究等。

3)实体分辨和企业内部关系管理

同一实体(如人员、组织)在不同系统中甚至同一系统中常常有不同的表示。 实体分辨也叫记录链接(Record Linkage)或对象识别(Object Identification),提供 确认表示同一实体记录的技术手段，在数据集成过程中，常用这类技术提高信息完 整性、解决不一致性并消除冗余。

一个实体经常包括具有复杂结构和错综关系的多个子实体，对这些子实体的 结构和关系可以有不同的理解，例如，问题“2008年IBM的全部收益是什么?”的 答案依赖于问这一问题的“目的”(如信用风险评估或提交给监管机构的文件), “目的”将决定是否包括子公司、分部及合资企业的收益，类似问题称为企业内部 关系(Corporate Householding)问题，某些情况下可以用集结异构问题(Aggregation Heterogeneity Problem)建模(Madnick et al.,2006);供企业行动参考的知识和它们 的内外关系被认为是企业内部关系知识(Madnick et al.,2005)。企业内部关系管 理旨在解决获取、分析、理解、定义、管理和有效使用企业内部知识的问题。Mad- nick 给出了应用上下文仲裁的企业内部关系管理的初步结果。

4)监测和清洗

在数据录入时，可以通过在线或适时批处理检测和纠正某些信息质量问题，此 方向主要研究如何自动化地完成这些任务。Madnick 等(2004)提出了一种从大数 据集中检测相似重复记录的方法；Ajax清洗框架拥有详细说明数据清洗操作的声

**28**)数据质量导论

明型语言(Galahardsh et al.,2001),这一声明方式允许数据清洗所需数据转换的 逻辑表达和物理实现分离，Ajax 框架适用于生物学数据库。

5)溯源、世系与来源标记

数据溯源和数据世系信息(如用于追溯数据的来源和过程相关的知识)十分 重要，数据用户需要评价数据质量时这些信息可以得到合适利用。该主题早期提 出了用来源标记(Source Tagging)标记每一数据元素的数据模型，并给出了处理来 源标记的关系代数方法，Buneman 等(2001)提出了一个更一般的模型，不但能用 于关系数据库也可以用于层次结构数据(如 XML)。在应用方面，斯坦福大学提出 了一个处理数据世系信息和数据不确定性的数据库管理系统(Data Base Manage- ment System,DBMS);Widom 提出了一种基于数据追源的数据可信度评估方法。

6)不确定性

从概率论的角度，每一个数据元素都具有某种程度的不确定性。另外， 一个属 性也存在有多个值的可能性，而属性值是数字值时，还存在精度问题。该主题主要 研究不确定数据的存储、处理和推理技术，Dalvi 提出了一种可同时处理不确定性 和世系信息的扩展的关系模型。然而， 一些任务需要高精度、低不确定性的数据， 而另一些任务用低精度、高不确定性的数据便可完成。因此，需要对具有不同程度 精度和不确定性的数据有效利用，以满足多种应用需求。Benjelloun 描述了一个可 以动态地选择不同图像数据源的系统，系统根据不同用户约束和参数选择产生信 息；在其他一些应用中，需要平衡确定性或精度，以及信息质量的其他维度，Ballou 等(1995)给出了一种信息系统设计的准确性和时效性优化机制。

[**2.2.2.3**](2.2.2.3) **计算机科学和信息技术中的信息质量**

该类主题研究除已有关系数据库技术外的管理、保护和提高信息质量的技术 和方法。

1)测量和评估

为了管理信息质量，首先需要评价现有系统和过程中的信息质量，如果信息系 统和信息生产过程复杂，则实现精确且低成本的信息质量评估将面临诸多困难。 该主题的研究旨在提出对企业或特定应用中信息质量的测量和评估技术，并且可 能要求此类测量能够定期、持续地进行下去。Lee 等(2002)提出了一种信息质量 评估和提高的方法，包括一个测量信息质量的调查表和一种解释信息质量测量的 差异分析技术。

2)信息系统

在信息系统领域，信息质量研究是有关确认企业中的信息质量问题、调查提高 (削弱)信息质量的实践活动，并提出具体背景下的信息质量管理技术和解决方 案。例如，采用将信息看成产品的观点，Shankaranarayang等(2003)提出了一种称 为IPMap的建模技术来描述信息产品的生产过程；应用类似的建模技术，Ballou 等

第2章 数据质量研究和数据清洗系统框架 (**29**

(1998)阐明了如何对一个信息产品制造过程建模，并提出了一个确定系统的信息 质量维度的方法。

3)网络

网络包括物理通信网、不同系统间的逻辑和语义链接、系统和用户间的连接， 以及用户之间的连接(如社会网)。对这些网络进行研究可深入了解数据是如何 被使用的以及数据质量从一个节点到另一个节点是如何变化的，研究结果可用于 最优化网络拓扑和设计分析、管理网络的工具，如 Marco 等(2003)研究了密集无 线传感器网的传输能力和数据的可压缩性。

4)协议和标准

协议和标准可能影响信息质量。数据协议或标准可在一致性、可理解性、准确 性等维度上提高信息质量，但是，当协议或标准太繁琐时，用户可能避开这些协议 或标准而引入违背协议或标准的数据。因此，需要研究协议和标准是如何影响信 息质量的，以及如何促使用户遵从这些协议和标准。另外，也应对协议或标准自身 的质量进行评估，如 Bovee 等(2002)讨论了可扩展的业务报告语言(eXtensible Business Reporting Language,XBRL)标准的词汇表。

5)私密性

某些系统包含个人私密信息(如用户、雇员和病人的信息),这些信息的管理 应确保只有授权用户才能查看。私密性信息质量研究是关于如何处理私密数据， 才能将不同权限的私密规则强制加入不同需求的研究。数据的既定私密遭到破坏 应视为一种数据质量问题，尽管已有商业工具能创建私密规则并用规则进行在线 审核，但在规则表达和规则有效执行机制方面仍面临挑战。近期的研究还讨论了 当必须透露某些私密数据时的私密保护问题(即应采取措施，以防从透露的数据 中推断出其他私密信息),其目的是提出防止下游用户推断出私密信息的数据操 作算法(Li et al.,2006)。

私密涉及到许多需求，例如，私密的一个特征为不受干涉权利，又叫个体自主 (Autonomy), 美国的“请勿拨打”名单就是私密个体自主需求的一个法律保护实 例。随着与用户通信模式的发展，未来的研究需要提出描述不同私密需求的解决 方案，并设计满足这些需求的系统。有些私密问题更加复杂，如私密性可能受到数 据溯源设计的影响。

6)安全性

信息的安全性越来越受到重视，此主题的研究旨在提出安全信息访问解决方 案、调查影响安全性的因素、提出评价组织内和组织间的全面信息安全性度量。最 近的研究从三个层面拓展了信息安全性含义： 一是场景，超出了一个企业的范围， 包括了合作伙伴；二是角色，不仅包括信息管理员，还包括用户和管理者；三是对 策，应从技术维度和管理维度考虑对策问题。当前的研究基于这一拓展定义提出

**30)**数据质量导论

了评价信息安全性的技术手段。

[**2.2.2.4**](2.2.2.4) **监管中的信息质量**

数字监管(Digital Curation)是来源于信息科学和图书馆领域的新兴研究方向， 包括为了满足当前和未来的信息应用，以更加容易发现和恢复的方式选择、保存和 管理数字信息。数字监管应同时考虑当前和将来的信息质量问题，如信息质量的 可访问性维度，因很难找到配有兼容软驱的计算机，所以，如果数据保存在8英寸 和5英寸(1英寸=2.54cm)的软盘上几乎是不可访问的；数字监管还应在考虑技 术问题的同时考虑非技术问题，例如，今天已知的、含蓄的上下文信息(常常是不 言自明的)可能发生变化，必须对那些未来可能变成未知的信息进行注释，确保所 监管数据将来使用时的可理解性，使将来的使用者可以获取明确信息。

1)监管的标准和原则

应提出改进数据监管过程、监管策略的标准和原则。目前有关监管标准的汇 总详见由英国数字监管中心(Digital Curation Centre) 维护的网站：http://

www.dcc.ac.uk/diffuse/。

2)监管的技术解决方案

除了关系数据库的有关问题，监管过程中还会有其他问题，如在注释中添加 “说明”,会对数据追源设计提出挑战。Buneman等人提出了一项适于用户手动从 多个数据源汇总进一个监管数据库时的数据溯源技术，得到的溯源信息可用来查 询监管数据的由来和形成过程。

**2.2.3** **数据质量的研究方法**

数据质量包括丰富的研究主题，同样有多种方法适用于数据质量研究，在此讨 论14类方法。

1)行动研究

行动研究是被研究者和实践者协作用来改进组织实践和提高学科理论的一种 经验和解释方法。行动研究对理论和实践的贡献既与咨询不同，又与个案研究 (Case Research)不同：除对目标进行观察外，更重要的是进行干预。Lee(2004)给 出了这种研究方法的一个实例，研究了当某全球制造企业建立全球数据仓库时如 何提高信息质量。

2)人工智能

人工智能(Artificial Intelligence)研究已经持续了半个多世纪。人工智能的多 种方法可用于信息质量研究，如知识表示(Knowledge Representation)和自动推理 技术(Automatic Reasoning Techniques)可用于实现异构系统的语义相容性，如 Mad- nick 等(2006)使用此类技术对信息质量的可理解性和一致性两个维度进行了改

第2章 数据质量研究和数据清洗系统框架

进。使用人工智能中改进的代理技术(Agent Techniques)可以将许多任务自动化， 如资源选择(Source Selection)、数据转换(Data Conversion)、增进系统性能，以及用 户经验的预测搜索和输入。

3)个案研究

个案研究是一种用于分析数据上下文中某种现象的经验方法。这种分析 结合了单一实例和事件的定性和定量证据，不但可以更深入地理解事件发生 的原因和过程，还可以通过个案研究提出并验证有用假设。信息质量研究广 泛使用了此方法，如Madnick 的纵向个案研究，讨论了如何开发信息产品管理 分析与规划系统(Management Analysis and Planning System,MAPS)并将其用于 提高信息质量。

4)数据挖掘

数据挖掘(Data Mining)起源于人工智能的机器学习和统计学的统计学习，是 从大数据集中抽取隐含的、先前未知的潜在有用信息的技术。数据挖掘方法如数 据异常(Data Anomaly)检测算法可以用于信息质量监测、数据清洗和入侵检测 (Instrusion Detection)。数据挖掘还可用于模式匹配，发现一对一匹配和复杂匹配 关系。尽管大多数据挖掘算法是健壮的，但当对有某种信息质量问题的数据挖掘 时，往往需要特别处理(Zhu et al.,2007)。

5)设计科学

在经历了IT项目的失败和低质量信息的负面影响后，企业对信息系统良好设 计的需求也日益增强。因此，需要对信息系统的设计技术进行系统研究。根据以 交付件为中心(Artifact-centric) 的设计科学(Design Science)观点，Hevner 等 (2004)提出了一个框架和一系列指导原则用于理解、执行和评估这一新兴领域的 研究。与其他交付件一样，如质量实体关系(Quality Entity -Relationship)和 IPMap (Shankaranarayang et al.,2003)均针对信息质量管理中的具体问题，需要用合适的 框架(如 Hevner等(2004)的框架)对它们进行评估。

6)经济计量学

经济学领域的经济计量学(Econometrics) 使用统计方法，研究和阐明经济学原 理。当前，虽然尚未提出信息质量的广义经济学理论，但人们对低质量信息代价的 认识正在增强，涌现出了大量的研发经济学、信息质量经济学等经济计量学成果。 随着继续积累经验数据，经济计量学研究也将不断深入，目标是提出信息质量经济 学理论并全面了解信息质量实践。

7)经验方法

经验方法是对从观测证据得出结论的研究方法的总称，如前面讨论的行动研 究、个案研究、统计分析和经济计量学均属于此类方法。此外，基于用户调查的研 究也是经验方法，Wang等(1996)采用用户调查方法确定信息质量维度并对这些

32)数据质量导论

维度分组；Slone(2006)用事后统计分析研究了不同信息角色的知识模式与信息质 量性能的关系。

8)实验方法

实验方法(Experimental Method)用来研究自然系统(如物理学)、人员和组织 (如实验心理学)、交付件(如不同方法的性能评估)的特性，如Jung 等(2005)用主 观实验检测上下文信息质量和任务复杂度对决策性能的影响。

9)数学建模

通常用数学模型描述系统行为，Ballou 等(1998)用一个数学模型描述信息质 量维度(如时效性和准确性)在信息系统中的变化。为了处理诸如软件质量安全 和开发等问题，系统动力学(System Dynamics)已用于对多种复杂系统的建模，该 建模技术源于系统和控制理论。

10)定性研究方法

定性研究(Qualitative Research)是对用于了解人类行为的探索性研究方法的 总称。行动研究、个案研究和民族志方法是适用于人类行为信息质量定性研究的 方法。前面已讨论了行动研究和个案研究，民族志学方法是指研究者深入课题环 境，通过直接观察和访谈搜集资料的研究方法，Kerr(2006) 用此方法研究了新西兰 保健部门的信息质量实施情况。

11)定量研究方法

定量研究(Quantitative Research)是对用于分析某现象的可量化属性以及它们 之间关系的研究方法的总称。统计计量学和数学建模是适用于信息质量研究的典 型定量方法。

12)统计分析

信息质量研究广泛使用了统计分析(Statistical Analysis)方法，如 Wang 等 (1996)用因子分析确定调查数据的信息质量维度，另外，统计分析方法也是其他 定量方法(如数据挖掘和经济计量学)的数学基础。

13)系统设计及实现

针对特定信息质量解决方案，可以利用软件工程的设计方法论、数据库设计与 数据建模、系统结构等进行系统设计。利用这些方法，可以通过在特征空间中对市 场进行系统评估使选择目标最优化。这些方法可用于设计和实现概念证明 (Proof-of-concept) 系统，如 Goh 等(1999)基于这些方法提出的 COIN 系统。

14)理论和形式化证明

这一方法广泛应用于理论计算机科学研究，如提出新的逻辑形式和计算复杂 性属性证明。信息质量理论研究可以使用该类方法，Shankaranarayang等(2003)应 用图论证明了IPMap 的一些属性，Fagin 等(2005)对数据交换问题进行了形式化 描述，提出了该环境下的查询应答计算复杂度理论。

第 2 章 数据质量研究和数据清洗系统框架 

**2.3** **对数据质量管理的思考**

随着从机械化建设向信息化建设转变，越来越多的信息系统被研发并投入使 用，而现实中因数据质量引发的问题越来越多，甚至延缓了信息化的进程(吴建 明，2004)。实践经验表明：信息化的核心是实施全面的“信息驱动”,而不是简单 的计算机化和网络化。数据是信息的载体和来源，虽然我国已具有了开发大型复 杂信息系统的能力和经验，但在数据质量管理环节上还很薄弱。

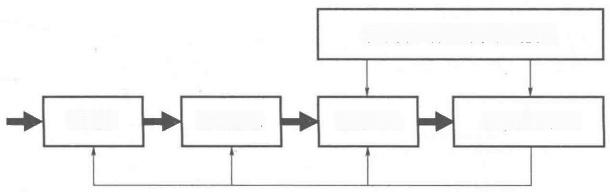
**2.3.1** **数据质量管理的发展模式**

[**2.3.1.1**](2.3.1.1) **数据质量管理的传统发展模式**

数据质量研究经过了20多年的发展，对其认识也从最初的“准确性”,进一步 发展到“特定用户的满意程度”,20世纪90年代初，麻省理工学院数据质量研究项 目得出了“将数据作为产品进行管理”的重要研究结论，这一结论改变了将数据作 为系统副产品的观念，是借用管理普通产品质量的方法管理数据质量的前提或 基础。

在相当长的时期，对数据质量的认识和研究，与对其他普通产品质量的研究是 相互独立的，随着对数据质量研究的深入和质量概念的逐步扩展，才将二者结合起 来。随后，利用信息系统环境和其他有形产品制造环境的相似性，建立了数据产品 和其他有形产品的联系，全面质量管理(Total Quality Management,TQM)的理论和 方法也被引入到数据质量管理，全面数据质量管理(Total Data Quality Management, TDQM)继承了完善的质量管理体系2000版ISO 9000 族标准(包括 ISO 9000、ISO 9001、ISO 9004)的原则、要求和指南，形成了先进的数据质量管理框架(Richard et al.,1995)。

传统的数据质量管理发展模式(非严格)如图2-3所示。

数据质量引发的事故或损失

数据库 信息系统 数据质量 数据质量管理

图2-3 传统数据质量管理发展模式示意图

数据质量导论

如图2-3所示，传统数据质量管理发展模式的主体是串行的，和其他信息化 要素发展之间呈现出时间上有序的特点，这种串行有序的模式符合“实践-认 识-再实践-再认识”的事物一般发展规律，呈现出明显的螺旋式前进趋势。传 统的发展模式，通过对数据质量的研究，包括问题来源、提高途径和技术等，最终实 现对数据质量的控制，数据质量管理又会对信息化其他要素提出新的具体要求。

因数据质量问题引发的事故是数据质量研究阶段性发展的直接推动力，往往 这种推动力的作用是巨大的，并且事故造成的损失越大，对数据质量发展的推动作 用也越大，这是螺旋式前进的具体表现。根据数据仓库研究所(The Data Ware- house Institute,TDWI)2002 年的一项研究表明，美国企业每年因数据质量造成 6110亿美元损失，如果包括邮寄、打印和人工费用等，数据质量造成的损失更高 (Sara,2002)。而几乎同步，美国国会决定由美国总统事务办公厅管理和预算办公 室发布政府信息质量指南，并要求每个联邦机构在2002年10月1日前，结合本部 门实际发布信息质量指南，建立管理机制，使得数据的使用者能够发现并得到联邦 机构所维护与传播的正确信息，同时要定期向主管报告。另一典型的案例便是 “三鹿牌婴幼儿奶粉事件”,该事件一方面推动了食品质量管理的重大进展，如食 品不得实施免检、加强不定期抽检、改进检测手段等，并进行了立法；但另一方面， 受害消费者付出了生命代价，政府为此立即启动国家重大食品安全I 级响应， 一批 企业高层、政府官员受到了处理，从企业到奶农蒙受了巨大损失，消费者对奶制品 严重失去信心，对直接相关企业的打击甚至是致命的。数据质量管理和其他产品 质量管理的发展遵循相同的螺旋式前进规律，质量问题引起的事故推动质量管理 的发展，这种发展的目的又是通过对质量问题本身和控制方法的研究，防止同类或 类似事件的发生。

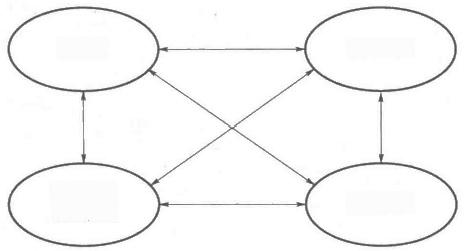
技术先行、政府参与是传统产品质量管理发展的两个基本规律。通常政府参 与产品质量管理的时机是在技术研究发展到一定程度时介入，但这种参与的力度 决定于对该产品质量问题的整体认识，在发生质量事故时政府参与力度往往会 加强。

[**2.3.1.2**](2.3.1.2) **数据质量管理的并行发展模式**

具体到数据质量管理建设，目前，信息化其他要素仍在快速发展完善之中，尚 未进入平稳发展时期，如数据库的结构仍在不断调整完善，数据采集的内容也在变 化，信息系统功能要求也不断改变，对数据质量的重视程度和研究水平还比较低。 综合当前现状，提出数据质量管理建设的并行发展模式，如图2-4所示。

数据质量管理建设要与其他信息化要素发展同步考虑、协同发展。如图2-4 所示的并行发展模式也是管理信息化质量效益型发展的具体表现形式，数据质量 管理建设与信息化其他要素关系密切，互为目的和前提。并行发展模式便于顶层 设计的实施，使各要素发展交叉进行，通过各要素之间不间断的交流与协作，及早

第2章数据质量研究和数据清洗系统框架(

信息系统

数据库

全面数据

数据质量

质量管理

图2 -4 数据质量管理并行发展模式示意图

发现并校正发展偏差，避免了串行发展模式中为适应后者，前者需要做较大代价修 改的情况。并行发展模式可以充分借鉴先进管理理论和技术，是一种自上而下的 高效的发展模式。

引入全面数据质量管理思想。全面质量管理是以质量为核心，建立在全员参 与基础上的一种管理的系统方法，是当前产品质量管理的先进理论，在各行各业得 到了广泛推广应用，直接将其引入到数据质量管理，使数据质量管理工作有一个高 的切入点和收益回报率。

技术研究和政府参与同步。数据质量管理建设，涉及人才队伍建设、编制调 整、标准与规范的推行等，因此，数据质量管理建设需要高层管理部门的尽早直接 参与，在并行发展模式下，建立起良好的发展机制。

减少以数据质量事故为推动力的发展方式。数据质量问题引起的后果是不言 而喻的，在并行发展模式下，首先对国内外以往发生的数据质量事故，特别是其发 生之后采取的相应措施进行引进和消化，以一个较高的认识水平为建设起点，避免 同类数据质量问题的产生和同类事故的发生。

**2.3.2** **数据质量管理问题剖析**

近年，随着数据工作日渐受到重视，数据质量管理工作也逐渐受到关注，但距 对数据质量问题的全面认识和数据质量管理建设的全面规划要求仍有较大差距。 以下对数据质量管理面临的问题进行剖析。

[**2.3.2.1**](2.3.2.1) **制约数据质量管理发展的因素**

1)认识不一致

首先，数据质量的定义不一致，目前对数据质量的理解主要以“满足特定用户 的期望程度”或类似表述为主，但也有认为数据质量不可定义的观点和其他定义 形式，而对数据质量的理解是对数据质量问题研究的根本基点；其次，对数据质量

)数据质量导论

问题的属性认识不一致，有观点认为数据质量问题是一个政治思想问题，不是一个 学术和技术问题，还有观点认为数据质量问题是一个主观问题，不可客观检测；再 次，对数据质量的研究内容认识不一致，既有将数据ETL 划分为数据质量研究内 容的观点，也有将数据质量问题简单地认为是记录相似重复问题，或是数据录入错 误的观点。最后，具体数据质量主题的研究边界不够清晰，如记录链接问题涉及了 多个研究主题(见2.2.2节数据质量的研究主题)。

2)重视程度不够

当前多数信息化项目并没有在计划中体现数据质量管理的内容，重点关心 的仍然是信息系统本身的功能，仍将数据作为信息系统的附属产品对待。在采 集环节，虽然信息系统主研部门根据需要下发数据采集软件和采集规范，并进行 培训，但执行单位并不能充分重视，大多临时安排人员，作为一项阶段性工作 对待。

3)专业人才缺乏

经过多年发展，我国已拥有一支数据库应用和软件开发专业人才队伍，他们在 我国信息化工作中做出了卓越贡献，对数据质量重要性和数据质量管理建设的迫 切性认识也最为直接，并且从某些具体问题出发在技术层面上做了一些工作，但这 种由下向上的研究只能解决某些具体问题，而根据2.3.1节的并行发展模式，数据 质量管理建设首先需要进行顶层设计，同时数据质量管理需要更多的数据处理知 识和领域知识，现有人才队伍中达到要求的专业人才还较少。

4)研究与安全存在矛盾

数据具有较高的保密要求，通常信息系统开发只需要数据库接口和系统功能 说明，只在调试阶段接触真实数据，时间较短、人员有限，容易管控。而数据质量管 理的研究对象便是数据，需要对数据进行较长时间的分析、研究与实验，安全工作 是首先要考虑的问题。安全问题客观上会增加研究成本或限制研究进展，当前还 缺乏相关合理机制。

[**2.3.2.2**](2.3.2.2) **当前数据质量研究存在的误区**

当前国内数据质量研究工作还比较零散，主要集中于对数据质量评估和清洗 技术的探讨，由于缺乏顶层设计以及受所接触问题的局限等原因，当前研究中存在 一些误区。

1)过于追求完美

许多工作都有两面性，数据清洗也不例外，其完整表述应该是：在尽量保留有 用信息的前提下，尽可能多地去除脏数据。但当前只注重了数据清洗中去除脏数 据的功能，以得到尽可能干净的数据，甚至绝对干净的数据。事实上，不当的数据 清洗工作对有用信息也可能有所损坏，另外，原有的数据质量问题解决了，往往还 会发现有新的质量问题，即某些症状会随着另外一些症状的解决而显现，再就是随

第2章数据质量研究和数据清洗系统框架 

着时间的推移和数据的演化，会有新的数据质量问题产生。所以，数据清洗工作应 以满足应用为目标，在数据生命周期内是一项经常性工作。

2) 过于追求通用性

当前研究往往强调数据清洗的通用性(领域无关数据清洗),而数据清洗的任 务是对数据进行维修，不同领域的数据产品具有不同的特点，与普通产品一样，不 同品牌的手机、汽车、家用电器有着各自完善的保修体系，通用往往意味着低端，专 修、精修才是专业和高端的体现，数据清洗应树立起专修、精修的理念(领域相关 数据清洗),重点研究适于各种数据特点的、领域无关与领域相关相结合的清洗 技术。

3)过于依赖事后清洗

数据质量管理的目的是将数据质量控制在满足要求的范围内，其中的数据质 量提高途径有数据录入接口设计、组织管理、数据集成、事后清洗，等等，在当前缺 乏顶层设计的情况下，事后清洗最受关注。其实，数据质量提高是一项综合性工 作，要合理地利用不同途径，对数据异常应有预防和早发现、早处理的意识，事后清 洗由于脱离了错误的原发环境，客观上给数据质量的提高增加了难度，在不同环节 应用不同的数据质量提高手段，才能达到数据质量管理的目的。

4)过于强调自动化

数据清洗对象往往是海量数据，纯手工清洗是不现实的，所以很自然地提出了 实施自动检测、自动分析、自动修正等自动化处理和控制要求。在故障诊断理论体 系中，后两者分别对应于自诊断和自修复，待清洗的数据往往存在多个(种)异常， 并且它们之间相互影响，数据清洗要处理的是典型的复合故障，到目前为止，复合 故障的自动分析仍是一个未很好解决的命题。因此，当前对数据清洗的自动要求 过于超前，由人员参与决策、自动化完成规律性工作的交互式数据清洗模式才是符 合实际的近期努力方向。

[**2.3.2.3**](2.3.2.3) **对数据质量管理的建议**

针对数据质量现状及存在的问题，对数据质量管理建设提出以下建议。 1)加强专业人才的培养

数据质量管理建设首先要解决人才问题，人才培养也是立足长远发展的战略 性举措，有目的地逐步培养数据质量建设专业管理和技术人才，逐步扩大数据质量 管理人才队伍并增强科研实力，是数据质量管理建设的根本途径，同时，也为将来 的数据质量管理和技术岗位需求作好人才储备。

2)重视基础理论研究

当前在数据质量的定义、数据质量问题的来源、数据质量研究的范围，以及全 面数据质量管理等基础理论方面还有许多不完善的地方，给数据质量管理的深入 研究与实践工作带来了障碍。引入普通产品的质量管理和维修保障的先进理论和

数据质量导论

方法，进一步完善数据质量管理和数据清洗的基础理论体系，对数据质量管理建设 的长远发展具有重要意义。

3)加快标准与规范的制定

标准与规范是数据质量管理建设顶层设计的重要组成部分和具体体现，也是 数据质量管理的最初目标和依据。数据质量管理相关标准与规范的尽快制定与执 行，是推动数据质量建设正规化、科学化、规范化发展的必要条件。

4)坚持沿途“下蛋”的发展策略

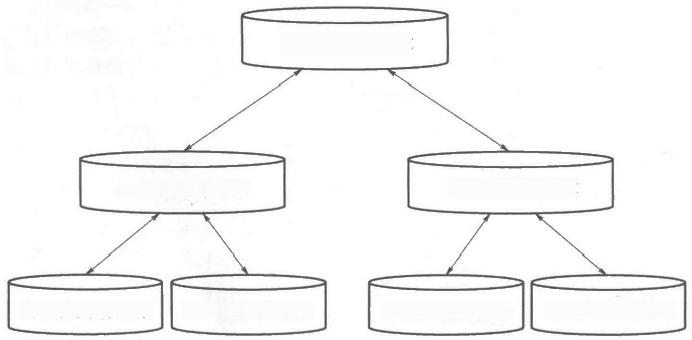
数据质量管理建设不是一次性的建设，而是一项需要持续发展的任务，在一次 建设的成果基础上，边应用边研究，继续进行适应新发展需求的建设。数据质量管 理建设应保持与需求持续发展相适应的建设模式。

**2.4** **典型数据质量控制的框架**

数据质量管理的直接目标是实现对数据质量控制(Data Quality Control),英文 中的“Control”一词同时有控制和管理的含义，而中文中管理是比控制更高层面的概 念。数据质量控制是针对某个应用以明确的测量标准对数据质量进行控制的过程。

**2.4.1** **层次结构数据质量控制框架**

由层次组织结构所决定， 一个典型信息系统的数据组织方式如图2-5所示。



国家级系统数据库

省部级系统数据库 省部级系统数据库

地区级系统数据库 地区级系统数据库 地区级系统数据库 地区级系统数据库

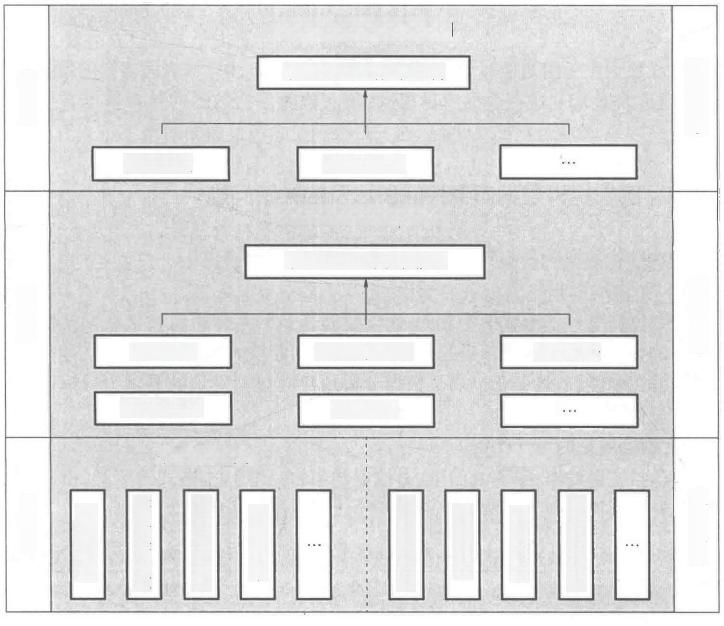
图2-5 一个典型信息系统的数据组织方式

图2-5中的这种数据组织方式具有典型的层次结构，该结构决定了数据组织 具有跨地区(空间)、跨编制单位的特点，数据质量控制要考虑这种层次组织结构

第2章数据质量研究和数据清洗系统框架 

所涉及的人员、技术、环境等因素。

层次结构数据质量控制框架如图2-6所示。



数 据 上 报 数据录入

上 下 文 约 束

函 數 依 賴

逻辑约束

定义域约束

参照约束

必填项约束

字典表约束

主键约束

集成层 清 洗 层 约 束 层

数据不一致

实例级数据质量问题控制

数据检测、分析和修正

拼写错误

不完整记录

模式级数据质量问题控制

数据抽取、转换和加载

相似重复记录

数据设计约束

数据汇总

模式冲突

逻辑错误

规则约束

异常数据

图2-6 层次结构数据质量控制框架

图2-6的层次结构数据质量控制框架分为三层，分别在数据生命周期的不同 阶段用不同的控制手段进行数据质量控制，同样，不同层面的数据质量控制手段针 对不同的数据质量问题。

1)约束层

数据质量控制系统的约束层位于数据录入阶段，主要功能为通过数据完整性 约束(Integrity Constraint,IC)进行数据质量控制，主要针对的是记录内部字段级、 字段之间的逻辑关系，以及一定范围内上下文的记录级的数据质量问题。此部分 功能在数据录入时被激活，通过将错误信息直接反馈给数据录入人员，根据错误信 息进行及时处理。

2)清洗层

清洗层主要功能是通过数据清洗(Data Cleaning,DC)解决实例级数据质量问

40)数据质量导论

题，对象是一个相对完整的数据集，通过数据检测、分析和修正进行数据质量控制， 解决的实例级数据质量问题包括拼写错误、相似重复记录、异常数据、不完整数据、 逻辑错误等。在层次数据获取流程中，数据上报前是数据清洗的最佳时机。

3)集成层

集成层主要功能是通过数据集成(Data Integration,DI)解决模式级的数据质量 问题，通过数据ETL进行数据质量提高，重点解决多源数据向单源数据转化时面 临的模式冲突、数据不一致等问题，这一过程中包含部分数据清洗工作。

**2.4.2** **层次结构数据质量控制所涉及的关键问题**

层次结构数据质量控制的实现，需要解决以下关键问题。

1)数据质量评估

将数据质量评估分成两个层面：面向管理人员的数据质量评估和面向数据分 析用户的数据质量评估。前者是指以数据质量问题检测、统计打分为主，包括评价 和辅助修正功能；后者是指以满足数据分析用户对数据质量的需求为目标进行的 数据质量评估。

2)数据质量的设计约束

其包括主键约束、字典表约束、必填项约束、参照约束、格式约束等，由数据库 或采集软件共同实现。这些约束所对应的数据质量问题如表2-4所列。

表2 - 4 设计约束与数据质量问题的对应表

|  |  |
| --- | --- |
| **约束名称** | **针对问题** |
| 主键约束 | 重复记录 |
| 字典表约束 | 枚举型属性异常值 |
| 必填项约束 | 不完整记录 |
| 参照约束 | 数据冲突 |
| 格式约束 | 格式不一致 |

研究表2-4中所列的数据质量设计约束与在线规则约束、数据清洗之间的关 联关系，给出适度约束方案。

3)数据质量的在线规则约束及其优化

包括定义域约束、逻辑约束、函数依赖约束、上下文约束等，此部分由单独的软 件模块实现，但与数据采集软件整合，与数据质量的设计约束一起在数据录入时被 激活。重点解决对检测算法的优化问题，提高规则约束的实时性；在线规则约束针 对的同样是实例级数据质量问题，所关注数据质量问题范围的界定是重点研究 内容。

第2章 数据质量研究和数据清洗系统框架(

4)数据清洗方法及实现

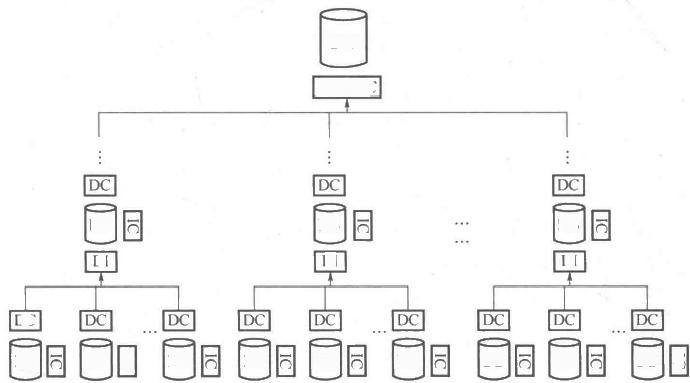
包括拼写错误、相似重复记录、异常数据、不完整数据、逻辑错误等实例级数据 质量问题的检测与修正，应开发适应具体需求的数据清洗工具。

5)数据集成方法及实现

数据集成解决的是数据异构性、分布性和自治性，虽然数据采集在统一的组织 和规范下进行，但由于所涉及单位和人员众多，不确定因素多而复杂，数据集成在 数据采集中十分重要。另外，各级数据正常汇总和数据清洗也需要数据集成。数 据集成主要依据元数据和具体需求进行，应开发适应具体集成需求的数据抽取、转 换和加载工具。

6)分布式数据质量控制系统构建

这包括数据全生命周期过程中设计阶段、生产阶段、使用阶段的全面数据质量 管理任务和实施时机；针对数据库建设模式，构建分布式数据质量控制系统，实现 实际信息系统的数据质量控制(典型的层次结构数据质量控制系统的部署如 图2-7所示)。同时，数据质量控制系统应着重考虑与信息系统其他部分的良好 整合。



DW

DI&DC

DB

DI

DC

DB|

DB 司 DB DB

DB DB 同

DB

DB

DB

DB

DB

DI

DI

图 2 - 7 层次结构数据质量控制系统的部署图

图2- 7展现了创建数据库(DataBase,DB), 最终构建数据仓库(Data Ware- hourse,DW) 的分布式过程中数据质量的控制手段及时机。图2-7中的分布式数 据质量控制系统，通过综合应用完整性约束、数据清洗和数据集成三种具体数据质 量提高技术手段实现数据质量控制，并且数据质量评估工作贯穿始终，与具体的数 据质量控制手段形成多个控制闭环。

数据质量导论

**2.4.3** **数据清洗技术简介**

如前所述，经过前期信息化建设，很多行业已经积累了大量数据，这些数据中 存在大量实例级数据质量问题，同时，当前我国信息环境复杂，大量数据已脱离数 据获取环境，不可能重新获取或返回修改，要提高这些数据的质量，很大程度上依 赖数据清洗技术。数据清洗是通过检测和消除实例级数据质量问题来提高数据质 量的技术手段，是数据质量的19个研究主题之一(Madnick et al.,2009;Erhard et al.,2000) 。 实例级数据质量问题包括相似重复记录(Madnick et al.,2009;Cao et al.,2010) 、不完整记录、逻辑错误、异常数据等。数据是信息时代的标志性产品形 式，数据清洗的地位和作用等同于传统设备(产品)的故障诊断与维修(Cao et al., 2009),相关研究对改善我国当前数据质量现状，增加与数据相关的收益具有现实 意义。

数据清洗可分为“特定领域(Domain-specific) 数据清洗”和“领域无关(Do- main-independent) 数据清洗”两类。“特定领域数据清洗”需用到相关领域知识， 并要求参与清洗过程的人员掌握相关领域知识(Chaudhuri et al.,2005);“领域无 关数据清洗”面向普通数据库用户，适用于不同的业务领域，更方便与传统数据库 管理系统相整合，且具有依赖数据集规模的合理复杂度(Kalashnikov et al.,2006)。

实现数据清洗的典型技术是数据检测、分析和修正(Data detection,Analysis and Modification,DAM),即发现和定位错误、对错误进行分析，以及对错误进行修 正的相关技术，此处用了修正一词，原因是虽然数据清洗最终目的是消除实例级数 据质量问题，但事实上因不确定性等因素的存在，仅从技术层面上将脏数据消除是 非常困难的，可操作的解决方法往往是对脏数据进行一定程度的修正，达到提高数 据质量的实际效果(陈伟，2004;戴剑伟，等，2010;Lee et al.,2000)。数据清洗是 一个代价高昂的过程，需要花费大量的资源，相当的工作需要人工交互地完成，如 何提高数据清洗效率、增加数据清洗的自动化水平是数据清洗技术的研究重点 之一。

**2.4.4** **数据清洗的概念辨析**

[**2.4.4.1**](2.4.4.1) **数据清洗的基本概念**

数据清洗主要有数据仓库(Data Warehouse)、数据挖掘(Data Mining)、全面数 据质量管理三个应用领域，其定义因应用领域不同而不同。广义上讲，数据清洗是 将数据库精简以除去重复记录(De-duplication), 并使剩余部分转换成符合标准 的数据的过程；狭义上的数据清洗特指在构建数据仓库和实现数据挖掘前对数据

第2章 数据质量研究和数据清洗系统框架(

源进行处理，使数据实现准确性(Accuracy) 、完整性(Completeness)、一致性(Con- sistency)、适时性(Timeliness)、有效性(Validity) 以适应后续操作的过程；从提高数 据质量的角度，凡是有助于提高数据质量的处理过程，都可以认为是数据清洗。 Muller 等(2003)认为数据清洗是对数据进行处理以保证数据具有较好质量的过 程，即得到干净数据的过程。Galhardas 等(2000b) 认为数据清洗是一个消除数据 中的错误和不一致，解决实体分辨问题的过程。Hernandez 等(1998)则把数据清 洗定义为重复记录的合并/消除(Merge/Purge) 问题。Erhard 等(2000b) 认为数据 清洗通过检测和消除数据中的错误或不一致来提高数据质量。Madnick 等(2009) 认为数据清洗研究如何自动地通过数据录入时在线或适时批处理的方式检测和纠 正某些信息质量问题。

以上对数据清洗定义的研究讨论因出发点不同等原因，存在不明确、不完善或 覆盖范围过宽的问题。本书综合以上对数据清洗的描述，从与其他产品形式故障 诊断与维修相类比的角度给出数据清洗的定义：数据清洗是通过检测和消除实例 层数据质量问题(脏数据)来提高数据质量的数据处理过程。

[**2.4.4.2**](2.4.4.2) **数据清洗的进一步讨论**

数据清洗研究虽然持续了半个多世纪，但因数据产品的特殊性，对数据清洗的 理解至今仍存在诸多偏差(如<2.4.4.1>节所述),以下对“数据清洗”做一全面 说明：

(1)数据清洗洗掉的是数据错误(Data Error)而不是错误数据(Error Data), 数据清洗的目的是要解决脏数据的问题(罗雪山，等，2010),所以容易产生将脏数 据洗掉的误解，正确理解应该是将脏数据洗干净。

(2)数据清洗只针对实例层数据质量问题，对一个给定的数据集，实例层数据 质量问题是有限的、可检测的和可隔离的。

(3)数据清洗不能完全解决所有的实例层数据质量问题，即通过数据清洗提 高数据质量的程度是有限度的。如对缺失值(Missing Value)的估计有不确定性； 实现全面数据质量管理还要结合数据质量的其他研究主题，如通过概念建模(Con- ceptual Modeling)减少实例层问题的产生。

(4)在全面数据质量管理中，数据清洗是不可替代的，实例层数据质量问题产 生的原因异常复杂，基于概念建模的数据库设计等其他手段只能减少、不能完全避 免实例层数据质量问题，有些实例层问题必须通过数据清洗来处理。

(5)数据清洗需在数据全生命周期中定期进行， 一次清洗后，随着新数据的产 生、进入或数据更改， 一定时间后又会产生实例层数据质量问题，还需要再一次的 数据清洗。

(6)数据清洗是一个代价高昂的数据处理过程(陈伟，2004;Hernandez et al., 1998),因此，数据清洗的目标应该是明确且有限的，应结合具体的应用需求而适

44))数据质量导论

当确定，不应一味追求完美(吴建明，2004)。

(7)数据清洗是有风险的，数据清洗的完整表述应该是“在尽可能不破坏有 用信息的前提下，尽可能多地去除数据错误”,数据清洗可能损失有用信息，也可 能产生新的数据质量问题。

(8)数据清洗过程应是柔性的， 一个良好的数据清洗系统框架允许根据用户 要求完成相应工作，如对某种数据质量问题的检测、对检测结果的统计等，即数据 清洗过程是多样的，甚至可以不包括“修正”。

(9)不同的实例层数据质量问题特性有明显差异，所需的数据清洗方法、技术 也不相同(韩京宇，等，2008;陈伟，2004);对给定数据清洗任务，往往含有若干种 实例层数据质量问题，需要结合特定领域数据清洗与领域无关数据清洗，并且要共 同应用自动处理与手动处理(Hernandez et al.,1998;罗雪山，等，2010;程开明， 2010)。因此，数据清洗的实施过程通常是松耦合的。

(10)数据清洗虽然是事后行为，但可以在数据录入时在线实现(McGilvray, 2010;Madnick et al.,2009;Kuang et al.,2011),虽然同为解决实例层数据质量问 题，但在线方式是受资源限制的增量数据处理过程，其功能受到相应限制，实现技 术与批处理方式也有很大不同。在线方式和批处理方式是不可以相互取代的 (Kuang et al.,2011)。不加说明情况下，数据清洗通常指的是批处理方式。

[**2.4.4.3**](2.4.4.3) **数据清洗与数据集成**

尽管数据集成和数据清洗分属于两个相对独立的数据质量研究主题(Madnick et al.,2009),但二者是当前数据质量领域最不易区分的概念。现实中，数据集成 的应用需求较数据清洗更为直接。事实上，数据清洗研究源自数据集成，最初用来 解决多数据源合并及查询、大型数据仓库创建中的记录链接、实体分辨、对象合并 等问题(Madnick et al.,2009)。一般而言，数据集成项目有如下数据清洗需求： 一 是数据集成对数据源的质量有要求，数据集成之前需要对各数据源进行数据清洗； 二是数据集成产生的实例级数据质量问题需要数据清洗。当前数据集成和数据清 洗是相互补充的两个数据质量研究主题。

数据清洗、数据集成和数据 ETL及它们之间的关系在近年文献中有多种表 述：罗雪山等(2010)认为数据 ETL是数据清洗工具；有的学者认为数据 ETL在理 论界和数据清洗等同，在工程界和数据集成等同，甚至认为数据清洗和数据集成是 数据 ETL要解决的两个问题；还有的学者将数据集成列为了数据清洗的一个研究 内容。

在数据质量领域，特别是在全面数据质量管理中，数据清洗和数据集成是同等 的概念范畴。与数据清洗不同，数据集成是将互相关联的分布式异构数据源集成 到一起，使用户能够以透明的方式访问这些数据源(Maurizio,2003; 陈跃国，等， 2004)。尽管数据清洗和数据集成联系密切，往往相互交织、互相渗透，但二者从

第2章数据质量研究和数据清洗系统框架 

任务目标到典型技术的侧重点都有明显区别。表2-5对数据清洗和数据集成进 行了比较。

如表2-5所列，首先，数据清洗和数据集成针对的问题不同；其次，数据清洗 主要解决的是实例层的问题，而数据集成解决的是模式层的问题；再次，数据集成 主要依据已知数据信息进行，如元数据、数据说明文件等，而数据清洗首先需要通 过检测发现和定位脏数据，并进行分析，作为最终数据修正的依据；第四，实现数据 清洗的典型技术是数据DAM,而实现数据集成的典型技术是数据 ETL,尽管实际 中有所交叉，但二者有本质不同；最后，数据清洗的难点是对错误数据如何更好地 进行检测、修正，而数据集成的难点是数据源的异构性。

表2- 5 数据清洗和数据集成比较表

|  |  |  |
| --- | --- | --- |
| **项目** | **数据清洗** | **数据集成** |
| 针对问题 | 脏数据 | 数据异构 |
| 问题层面 | 实例层 | 模式层 |
| 实施依据 | 检测信息 | 已知信息 |
| 典型技术 | 数据DAM | 数据ETL |
| 技术难点 | 数据检测、修正 | 数据异构性 |

**2.4.5** **数据清洗的一般性系统框架**

在已有的数据清洗方法和工具中， 一部分数据清洗工具只提供了有限的清洗 功能(如一些数据 ETL工具),另一部分则专门针对数据清洗。

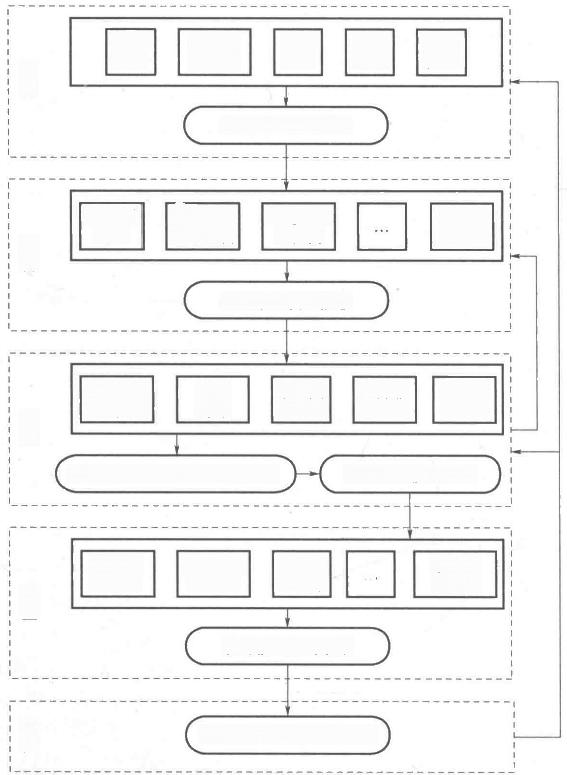
Galhardas 等(2001)提出了一个数据清洗框架，试图清晰地分离逻辑规范层和 物理实现层；提出了一种描述性语言，该描述性语言可以在逻辑层上指定数据清洗 过程所需采取的数据转换操作，并指定何时可以弹出例外，要求用户的交互。Gal- hardas 等(2000)基于该框架实现了一个可扩展的数据清洗工具AJAX(Galhardas et al.,2000a),其实验结果证明了该框架的价值。Raman 等(2001)提出了数据清 洗的一个交互式系统框架，它紧密地集成数据转换和差异检测(Discrepancy Detec- tion), 具有良好的交互性。

陈伟(2004)分别研究了相似重复记录、不完整记录、异常数据、逻辑错误的数 据清洗流程。王芳潇等(2010)提出了一个粗粒度的、紧耦合的自动化数据清洗 框架。

在以上讨论的基础上，与有形产品、软件产品的故障诊断与维修框架(范兆 军，等，2006;单锦辉，等，2011)相类比，提出一个数据清洗的一般性系统框架，该 框架由准备(Preparation)、检测(Detection)、定位(Location)、修正(Modification)、

 数据质量导论

验证(Validation)5个阶段组成，简称为 PDLMV,如图2-8所示。



信息环

境分析

获得数据清洗方案

不完整记

录检测

获得数据质量信息

数据质量

问题定位

确定数据质量问题性质及位置 确定数据修正方案

不可用数

据删除

获得修正后的数据

验证与任务定义的符合性

相似重复 记录检测

根本原 因分析

数据世 系管理

业务影 响分析

问题数 据标记

重复记 录合并

数据质 量评估

数据预 处理

检测结 果统计

需求 分析

基本 配置

方法 定义

任务 定义

数据 追踪

准备

定位

修正

检测

验 证

图 2 - 8 数据清洗的一般性系统框架 PDLMV

图2-8中的一般性系统框架 PDLMV,允许在不同位置停止以完成不同的数 据清洗任务，是一个柔性、可扩展、交互性好、松耦合的框架。下面对5个阶段分别 说明：

(1)准备。包括需求分析、信息环境分析、任务定义、方法定义、基本配置，以 及基于以上工作获得数据清洗方案等。通过需求分析明确信息系统的数据清洗需 求，通过信息环境分析明确数据所处的信息环境特点，任务定义要明确具体的数据 清洗任务目标，方法定义确定合适的数据清洗方法，基本配置完成数据接口等的配

第2章数据质量研究和数据清洗系统框架 

置，最终要形成完整的数据清洗方案，并整理归档。

(2)检测。包括检测必需的数据预处理，并进行相似记录、不完整记录、逻辑 错误、异常数据等数据质量问题的检测，对检测结果进行统计，以获得全面的数据 质量信息，并将相关信息整理归档。

(3)定位。包括数据质量问题定位、数据追踪分析，并根据检测结果对数据质 量进行评估，分析问题数据及修正的业务影响，分析产生数据质量问题的根本原 因；进而确定数据质量问题性质及位置，给出数据修正方案，并将相关信息归档。 根据定位分析结果，来确定是否可能需要返回检测阶段。

(4)修正。在定位分析的基础上，对检测出的实例层数据质量问题进行修正， 具体包括问题数据标记、不可用数据删除、重复记录合并、缺失数据估计与填充等， 并对数据修正过程进行数据世系管理。

(5)验证。验证修正后的数据与任务定义的符合性(用到部分检测操作),如 果结果与任务目标不符合，则做进一步定位分析并修正，甚至返回准备阶段来调整 相应准备工作。

**2.5** **本章小结**

本章首先对数据质量研究的框架进行了综述，较完整地描述了典型的数据质 量研究框架体系。早期 Wang的数据质量研究框架是通过将信息产品与普通有形 产品相类比，参考ISO 9000提出的一个较宏观的框架，该框架对数据质量的影响 要素描述是全面的，在概念层面具有较好的指导作用，但描述粒度较粗，不易实际 应用。面向数据质量工程实践的 Danette 的10步流程具有较强的可操作性，注重 实际效果，但并不关注具体的方法技术细节，适于指导实践而不适于指导理论研 究。Redman的全方位数据质量管理框架，列出了要拥有高质量数据必须具有的 10个要素，强调了企业各阶层共同协作及持续改进。Madnick 的数据质量研究的 框架由研究主题和研究方法两个维度构成，包括了数据质量涉及或可能涉及的研 究内容，对开展数据质量研究具有指导作用。但该框架仍面临以下局限性： 一是研 究主题之间的界线不够清晰，如实体分辨涉及了数据集成与数据仓库、实体分辨与 企业内部关系管理、监测和清洗三个研究主题；二是该框架试图用数据质量将传统 独立的研究主题统一起来，但事实上在短期内难以从根本上改变分别独立发展的 状况；三是研究主题的数据质量要素与非数据质量要素区别不清，有些主题与非数 据质量要素密切相关。因此，Madnick 的数据质量研究的框架尚需进一步发展 完善。

然后，研究了数据质量管理工作的现状及发展。在分析数据质量管理传统发 展模式的基础上，提出了适合我国信息化进程的数据质量管理的并行发展模式；深

48)数据质量导论

入剖析了影响数据质量管理发展的因素，以及当前数据质量研究存在的误区，并对 数据质量管理建设提出了具体建议。

最后，构建了适用于国内层次结构数据组织模式的装备数据质量控制框架，并 对相关问题进行了分析。对层次结构数据质量控制框架进行了描述，并分析了所 涉及的关键问题；研究了数据清洗与数据集成的区别，在全面数据质量管理领域， 二者是同等的数据质量提高技术手段；进一步分析了数据清洗的任务特点，纠正了 对数据清洗理解的偏差，对数据清洗给出了10点说明，构建了数据清洗的一般性 系 统 框 架 PDLMV。

本 章 工 作 为 后 续 的 具 体 数 据 清 洗 技 术 研 究 提 供 了 清 晰 定 位 。

**参考文献**

[1]王芳潇，曹建军，汪挺，等.2010.一种通用数据清洗框架的研究与应用[J]. 现代军事通信，18(1):60-63. [2]吴建明.2004.病态信息理论及其在装备保障中的应用[D]. 石家庄：军械工程学院.

[3]陈伟.2004.数据清理关键技术及其软件平台的研究与应用[D]. 南京：南京航空航天大学. [4]陈跃国，王京春.2004.数据集成综述[J]. 计算机科学，31(5):48-51.

[5]范兆军，郑海起，戚洪海.2006.基于信息融合技术的机械系统故障诊断框架研究[J]. 科学技术与工

程，6(23):4709-4713.

[6]罗雪山，罗爱民，张耀鸿，等.2010.军事系统体系结构技术[M]. 北京：国防工业出版社，184-198. [7]单锦辉，徐克俊，王戟.2011.一种软件故障诊断过程框架[J]. 计算机学报，34(2):371-382.

[8]曹建军，刁兴春，杜鹚，等.2010a. 基于蚁群特征选择的相似重复记录分类检测[J]. 兵工学报，31(9):

1222-1227.

[9]曹建军，刁兴春，吴建明，等.2010b.基于位运算的不完整记录分类检测方法[J]. 系统工程与电子技 术，32(11):2489-2492.

[10]曹建军，刁兴春，汪挺，等.2010c. 领域无关数据清洗研究综述[J]. 计算机科学，37(5):26-29. [11]韩京宇，徐立臻，董逸生.2008.数据质量研究综述[J]. 计算机科学，35(2):1-5,12.

[12]程开明.2010.统计数据质量诊断与管理研究[M]. 浙江：浙江工商大学出版社，12:31-113.

[13]McGilvray D.2010. 数据质量工程实践[M]. 刁兴春，曹建军，张健美，等，译，北京：电子工业出版社. [14]Ballou D,Pazer H.1995.Design Information System to Optimize Accuracy-timliness Trade off[J].Information

System Research,6(1):51 -72.

[15]Ballou D,Tayi G K.1999.Enhancing Data Quality in Data Warehouse Environments[J].Communications of the ACM,42(1):73-78.

[16]Ballou D,Wang R Y,Pazer H,et al 1998.Modeling Information Manufacturing Systems to Determine Informa- tion Product Quality[J].Management Science,44(4):462-484.

[17]Bovee M,Ettredge M L,Srivastava R P,et al.2002.Does the Year 2000 XBRL Taxonomy Accommodate Cur- rent Business Financial -Reporting Practice?[J].Joumal of Information Systems,16(2):165 -182.

[18]Buneman P,Khanna S,Tan W C.2001.Why and Where:A Characterization of Data Provenance.In Internation-

al Conference on Database Theory[J].Lecture Notes in Computer Science,Springer(1973):316 -330.

[19]CaoJJ,Diao X C,Wang T,et al.2009.Some Innovative Viewpoints for Improving Data Quality[C]//The 8th

第2章数据质量研究和数据清洗系统框架 **49**

International Symposium on Test and Measurement.Chongqing,China:[S.n]:6:3563-3566.

[20]CaoJJ,Diao X C,Tan M C,et al.2010.Classifying Detection using Classifier and Feature Selection by Ant

Colony Optimization:A Methodology for Identifying Approximately Duplicate Records[C]//The 15th Interna-

tional Conference on Information Quality(ICIQ 2010).Litle Rock,Arkansas USA:UALR.

[21]Chaudhuri S,Ganjam K,Ganti V,et al.2005.Data Cleaning in Microsoft SQL Server 2005[C]//Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data.Baltimore:International Conference on Management of Data(MD):918 -920.

[22]Erhard R,Hong H D.2000.Data Cleaning:Problems and Current Approaches[J].IEEE Data Engineering Bul- letin,3(4):3-13.

[23]Fagin R,Kolaitis P G,Miller R,et al.2005.Data Exchange:Semantics and Query Answering[J].Theoretical Computer Science,336(1):89-12.

[24]Galahardsh,Flreslud,Shasha D,et al.2001.Declarative Data Cleaning:Language Model and Algorithms[C]//

In Proceedings of 27th VLDB Conference.Roma,Italy:Morgan Kaufmann:371-380.

[25]Galhardas H,Florescu D,Shasha D,et al.2000a.AJAX:an Extensible Data Cleaning Tool[J].SIGMOD Re- cord:590-590.

[26]Galhardas H,Florescu D,Shasha D,et al.2000b.An Extensible Framework for Data Cleaning[C]//The 16th International Conference on Data Engineering.San Diego,California:International Conference on Data Engi- neering(ICDE):312-312.

[27]Galhardas H,Florescu D,Shasha D,et al.2001.Declarative Data Cleaning:Language,Model and Algorithms [C]//Proceedings of the 27th International Conference on Very Large Data Bases.Roma:Morgan Kaufmann:

371-380.

[28]Goh C H,Bressan S,Madnick S E,et al.1999.Context Interchange:New Features and Formalisms for the Intel-

ligent Integration of Information[J].ACM Transactions on Information Systems,17(3):270-293.

[29]Hernandez M A,Stolfo S J.1998.Real -world Data is Dirty:Data Cleansing and the Merge/Purge Problem[J].

Data Mining and Knowledge Discovery,2(1):9-37.

[30]Hevner A T,March S T,Park J,et al.2004.Design Science in Information Systems Research[J].MIS Quart,28 (1):75-105.

[31]Jarke M,Jeusfeld M A,Quix C,et al.1999.Architecture and Quality in Data Warehouse:An Extended Reposi- tory Approach[J].Information Systems,24(3):229 -253.

[32]Jung W,Olfman L,Ryant,et al.2005.An Experimental Study of the Effects of Contextual Data Quality and Task Compexity on Decision Performance[C]//In proceedings of the IEEE International Conference Informa- tion Reuse and Integration.[S.I.]IEEE,149 -154.

[33]Kalashnikov D V,Mehrotra S.2006.Domain -Independent Data Cleaning Via Analysis of Entity -Relationship Graph[J].ACM Transactions on Database Systems,31(2):716-767.

[34]Kerr K.2006.The Institutionalization of Data Quality in the New Zealand Health Sector[D].New Zealand:

Auckland University.

[35]Kuang C,Harr C,Neil C,et al.2011.USHER:Improving Data Quality with Dynamic Forms[J].IEEE Transac- tion on Knowledge and Data Engineering,23(8):1138 -1153.

[36]Lee M L,Ling T W,Low W L.2000.IntelliClean:a Knowledge-based Intelligent Data Cleaner[C].Boston:

ACM Press:290-294.

[37]Lee Y,Srong D.2004.Knowing-Why about Data Processes and Data Quality[J].Jourmal of Management In-

**50】**数据质量导论

formation Systems,20(3):13-39.

[38]Lee Y,Strong D,Kahn B,et al.2002.AIMQ:A Methodology for Information Qualiy Assessment[J].Infe Man- agement,40(2):133-146.

[39]Lee Y W.2004.Crafting rules:Context -reflective Data Quality Problem Solving[J].Journal.Management In- formation Systems,20(3):93-119.

[40]Li X B,Sarkar S.2006.Privacy Protection in Data Mining:A Perturbation Approach For Categorical Data[J].

Information System Research,17(3):254-270.

[41]Madnick S E,Wang R Y.2009.Overview and Framework for Data and Information Quality Research[J].ACM Journal of Data and Information Quality(JDIQ),1(1):1-22.

[42]Madnick S E,Zhu H.2006.Improving Data Quality with Effective Use of Data Semantics[J].Data and Knowl- edge Engineering,59(2):460-475.

[43]Madnick S E,Wang RY,Xian X.2004.The Design and Implementation of a Corporate Householding Knowl- edge Processor to Improve Data Quality[J].Journal Management Information Systems,20(3):41 -69.

[44]Madnick S E,Wang R Y,Krishna C,et al.2005.Exemplifying Business Opportunities for Improving Data Qual- ity from Corporate Household Research[M].New York,USA:MIT,Sloan school of Management.181 -196.

[45]Marco D,Duate-Melo E,Liu M,et al.2003.On the Many-to-one Transport Capacity of a Dense Wireless Sensor Network and the Compressibility of its Data[J].Lecture Notes in Computer Science,Springer Berlin (2634):556.

[46]Maurizio L.2003.Data Integration:a Theoretical Perspective[C]//ACM SIGMOD Conference.San Diego,Cali- fornia,USA:ACM:233-246.

[47]Muller H,Freytag J.2003.Problems,Methods and Challenges in Comprehensive Data Cleansing[J].http:// www.dbis.informatik.hu-berlin.de/fileadin/research/papers/techreports/2003-hub\_ib\_ 164-mueller.pdf.

[48]Raman V,Hellerstein J.2001.Potter's Wheel:an Interactive Data Cleaning System[C]//Proceedings of the

27th International Conference on Very Large Data Bases.Roma:Morgan Kaufmann:381 -390.

[49]Redman T C.2008.Data Driven;Profiting from Your Most Important Business Asset[M].Boston ,USA:Har- vard Business Press.

[50]Richard Y W,Veda C S,Christopher P F.1995.A Framework for Analysis of Data Quality Research[J].IEEE Transactions on Knowledge and Data Engineering,7(4):623 -640.

[51]Sara C.2002.Billions Lost to Poor Data Quality[J/OL].<http://searchdatabase.techtarget.com>.

[52]Shankaranarayang,Ziadm,Wang R.Y.2003.Managing Data Quality in Dynamic Decision Enviorment:An In- formation Product Approach[J].Jourmal of Database Management,14(4):14 -32.

[53]Slone J P.2006.Information Quality Strategy:An Empirical Investigation of the Relationship Between Informa- tion Quality Improvements and Organizational Outcomes[D].Minneapolis;Capella University.

[54]Wang RY,Reddy M,Kon,H.1995.Toward Quality Data:An Attribute-based Approach[J].Decision Support System,13(3-4):349-372.

[55]Wang R Y,Strong D M.1996.Beyond Accuracy:What Data Quality Means to Data Consumers[J].Journal of Management Information Systems,12(4):5-34.

[56]Xu H,NordJH,Brown N,et al.2002.Data Quality Issues in Implementing an ERP[J].Industrial Manage- ment.Data System,102(1):47-58.

[57]Zhu X,Khoshgoftaar T,Davidson I,et al.2007.Editorial:Special Issue on Mining Low -quality Data[J]. Knowledge and Information Systems,11(2):131 -136.

**第** **3** **章** **典型数据清洗技术的发展动态**

**3.1** **引言**

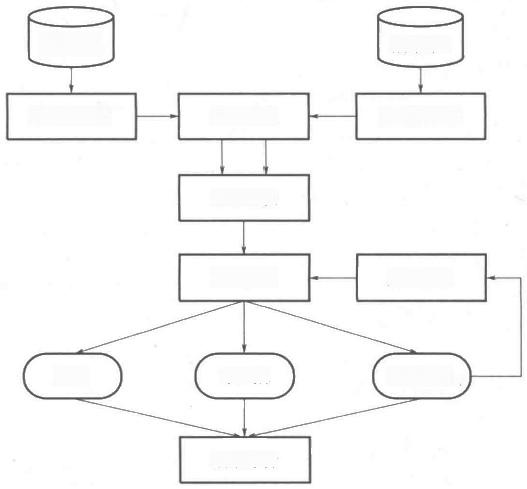
本章综述实体分辨、不完整数据、不一致数据三类实例层数据质量问题的数据

清洗技术发展动态。

**3.2** **实体分辨技术的发展动态**

典型的实体分辨过程以数据库中的记录作为输入，以记录是否描述相同实体

的决策结果作为输出，依据数据处理的先后顺序，通常划分为5个步骤(Churches et al.,2002;Christen et al.,2002):数据预处理、数据分块、记录比较、匹配决策和 效果评价。图3 - 1给出对两个数据库进行实体分辨的过程示意图。



数据库A

数据预处理

记录比较

匹配决策

可能匹配

效果评价

数据预处理

数据分块

人工匹配

数据库B

不匹配

匹配

图3 - 1 实体分辨过程示意图

数据质量导论

图3-1中，第一个步骤是数据预处理，该步骤用来确保参与实体分辨的数据 具有相同的格式；第二个步骤是数据分块，在一些文献中也被称为数据索引(Data Indexing)(Christen,2012b;Christen,2012a),目的是利用计算复杂度较低的算法尽 可能去除不可能匹配的记录对，从而减少需要比较的记录对的数量；第三个步骤是 记录比较，该步骤一般会用到各种属性和记录比较函数对数据分块步骤中产生的 候选记录对进行比较，计算相似度，这些比较函数通常具有较高的计算复杂度；第 四个步骤是匹配决策，该步骤依据记录比较步骤得到的相似度，判断两条记录的匹 配状态，将记录对划分到匹配、不匹配和可能匹配三个类别中，如果记录对被划分 到可能匹配类别中，则可能需要领域专家以人工方法对记录对的匹配状态进行进

一步判断；第五个步骤是效果评价，该步骤对实体分辨结果的质量、完整性进行评 价，并对实体分辨过程的时间效率进行评价。以上的步骤划分对于只有一个数据 库参与的实体分辨依然有效，只不过预处理步骤相对简单，记录比较步骤中不是对 两个数据库中的记录进行比较，而是将一个数据库中的每条记录与剩余的其他记 录进行比较。

在这五个步骤中，数据预处理和效果评价一般被认为是实体分辨任务开始前 和完成后的工作，在实体分辨的研究过程中通常只关注数据分块、记录比较和匹配 决策三个步骤。

**3.2.1** **数据分块算法**

传统的实体分辨方法需要对数据库中的记录进行两两比较，而对于任意一条 记录而言，数据库中与其匹配的记录数目远远少于不匹配的记录数目，因此，绝大 部分记录比较是在非匹配记录对之间进行的，而实体分辨的最终目的是发现匹配 记录对，当需要处理的数据规模较大时，这就会造成大量计算能力的消耗。数据分 块的目的就是在尽量不影响实体分辨准确性和完整性的同时尽量减少参与比较的 记录对数目，从而提高效率。数据分块算法通常依据一定的分块主键(或排序主 键)对参与实体分辨的记录进行划分，将每条记录划分到一个或多个数据块中(或 对数据库进行记录排序操作),使得可能匹配的记录尽量出现在同一数据块中，在 记录比较时只对同一数据块(或窗口)中的记录进行两两比较，从而在减少候选记 录对数目的同时尽量保留可能匹配的记录对(Christen,2012a)。按照记录聚集方 式的不同，实体分辨中的记录分块算法可以分为标准分块算法、基于近邻排序的分 块算法和基于q-gram 的分块算法三类。

[**3.2.1.1**](3.2.1.1) **标准分块算法**

标准分块算法为每条记录产生一个分块键值，将具有相同分块键值的记录划 分到同一数据块中，候选记录对由同一数据块中的记录两两组合产生，如果某一键

第3章典型数据清洗技术的发展动态(**53)**

值只有一条记录与之对应则去除该键值和与之对应的记录(Fellegi et al.,1969)。 标准分块算法只能将一条记录划分到一个数据块中，如果分块主键定义不合理，两 条匹配记录可能会划分到两个不同的数据块中，从而遗漏匹配的记录对。为了减 少遗漏，Winkler 等(2010)通过定义不同的分块主键，并按照不同的分块主键对记 录进行多次划分，使得某一分块主键下被划分到两个数据块中的匹配记录在其他 分块主键下可能被划分到同一分块中，从而增加匹配记录对被检测到的机会，但这 样也增加了需要比较的记录对的总体数目。

[**3.2.1.2**](3.2.1.2) **基于近邻排序的分块算法**

基于近邻排序的分块算法依据排序主键对所有记录进行排序，使得相似的记 录能够彼此靠近，然后利用滑动窗口在排序后的记录上滑动，每滑动一步，将窗口 内的第一条记录与其他记录进行两两比较，直到窗口滑过所有记录(Winkler, 2006)。传统的近邻排序算法使用固定大小的窗口，窗口大小设置不合理会导致 匹配记录的丢失或增加候选记录对的数量。Yan 等(2007)提出可变窗口的IA-

SNM 和 AA-SNM 算法，IA-SNM 算法递增地增加窗口大小，直到窗口中第一条 记录与最后一条记录之间的距离大于某一阈值，AA-SNM 算法在IA-SNM 算法 的基础上保留前一窗口的部分记录，从而可以把多个相邻的窗口进行合并。理论 上讲，可变窗口可以获得更好的性能，然而Uwe Draisbach 等人在实验验证时并没 能得出AA-SNM 算法和IA-SNM 算法的性能比传统的近邻排序算法更好的结 论(Draisbach et al.,2012) 。Christen(2012a)为排序主键建立倒排索引，窗口在索 引值上滑动，使得具有相同索引值的记录都可以划分到相同的窗口中，提高了匹配 记录对出现在同一窗口内的概率。

[**3.2.1.3**](3.2.1.3) **基于q-gram 的分块算法**

当分块主键(或排序主键)中存在错误或变体时，标准分块算法和基于近邻排 序的分块算法可能无法将两条匹配的记录划分到同一个数据块中。Baxter 等 (2003)和邱越峰等(2001)提出基于q-gram 的分块算法，将分块键值分解为长度 为q(通常为2或3)的子串，并依据得到的子串生成该分块键值的多种变体，再将 具有相同键值变体的记录划分到同一个数据块中，这样一条记录可能被划分到多 个不同的数据块中，从而提高了对错误和变体的容忍能力。基于类似的思想，De 等(2011)提出利用后缀数组生成分块键值变体的方法，并通过合并相似变体的方 法降低产生候选记录对的数量。由于一条记录可能被划分到多个数据块中，增加 了要比较的记录对的数目，Whang 等(2009)提出了迭代分块模型，在实体分辨过 程中将匹配记录合并为新的记录，再将合并后的记录与其他记录一起迭代地进行 数据分块，在某一个数据块中检测到的匹配信息会传递到其他数据块中，有助于发 现更多的匹配记录并减少记录比较次数。

分块主键的定义是数据分块算法的关键，直接影响到候选记录对的数量和实

)数据质量导论

体分辨结果的完整性(Christen,2012a) 。Christen 等(2007)指出定义分块主键时 通常要考虑属性值的缺失情况、属性值的概率分布和数据块大小等问题。Michel- son等(2006)提出利用机器学习技术定义分块主键的方法，选择具有最优覆盖率 和准确性的分块主键，然而此类方法需要的训练数据在实际应用中难以得到满足。 目前大部分数据分块算法中的分块主键还是由领域专家进行定义。

**3.2.2** **记录比较算法**

记录比较是实体分辨的核心步骤，主要任务是利用各种相似度算法计算两条 记录各个属性的相似度，得到该记录对的比较向量 V=<s₁,S₂,…,s>, 其 中s₁, i=1,2,…,n 为各个属性的相似度。这些比较向量将作为决策模型判断两条记录 是否描述同一个实体的依据。因此，记录比较算法得到的相似度向量的准确性一 定程度上决定着实体分辨结果的准确性。现有的属性相似度算法按照字符串划分 粒度的不同，通常可以分为基于字符的相似度算法，基于token 的相似度算法和基 于读音的相似度算法三类(Elmagarmid et al.,2007)。

[**3.2.2.1**](3.2.2.1) **基于字符的相似度算法**

基于字符的相似度算法以单个字符为比较单位，度量两个字符串在字符级别 上的差异。如 Levenshtein 距离利用两个字符串相互转换所需字符的增、删、改操 作的数目计算两个字符串之间的距离(Koudas et al.,2006)。Smith -Waterman 距 离在Levenshtein 距离基础上引入了字符交换操作(Christen,2012b) 。LCS 算法通 过迭代地查找和移除两个字符串的最大公共子串计算两个字符串的相似度，可以 处理字符串中的单词位置变换产生的错误(Elmagarmid et al.,2007)。基于 q- gram的相似度算法利用滑动窗口分别将两个字符串划分为长度为q 的子串，然后 利用两个字符串对应的子串集合的Jaccard 系数计算相似度(Cohen et al.,2003)。 Jaro-Winkler 距离结合了Levenshtein 和q-gram 算法，利用两个字符串的公共子 串和字符顺序计算相似度(Snae,2007) 。 基于字符的相似度算法可以很好地处理 字符拼写错误，但不能很好地捕捉单词交换、缩写等问题。

[**3.2.2.2**](3.2.2.2) **基于** **token** **的相似度算法**

基于 token的相似度算法通过分词技术将字符串分解为多个 token, 以 token 为单位计算两个字符串的相似程度，并且可以通过不同的 token 比较方法处理单 词缩写、拼写错误等问题。Atomic Strings 算法以标点符号为边界将字符串分解为 多个 token,如果两个token 相同或其中一个为另外一个的前缀，则认为两个 token 匹配，进而计算整个字符串的相似度(Elmagarmid et al.,2007);WHIRL算法在对 字符串进行分词后，利用信息检索领域中的 TF-IDF 和余弦相似度计算两个字符 串的相似程度，但WHIRL算法不能处理拼写错误(Elmagarmid et al.,2007) 。Soft

第3章 典型数据清洗技术的发展动态 

TF-IDF 算法和q-gram TF-IDF算法对WHIRL进行了改进，Soft TF-IDF算法 在计算余弦相似度时考虑了相似的 token, 而 q-gram TF-IDF 算法则用q-gram 替代单词作为 token,解决了单词拼写、插入、删除等问题(Bilenko et al.,2003a; Gravano et al.,2003)。

<3.2.2.3> 基于读音的相似度算法

在信息检索等领域，需要计算字符串在读音上的相似程度，基于读音的相似度 算法根据单个或多个字符的读音对字符串进行编码，利用得到的编码比较两个字 符串读音上的相似程度。如Soundex算法根据英文字符的美式发音，将任何一个 单词映射为由1个字符和3个数字组成的编码，使得读音接近的单词能够得到相 同的编码(Christen,2012b);NYSIS 算法不使用数字编码，而是使用发音相同的字 母对原始字符串中的字符进行替换，替换后的单词依然可以作为单词进行处理 (Elmagarmid et al.,2007);ONCA 算法则结合了前两者，先对字符串应用NYSIS 进行替换，再利用Soundex进行编码，提高了Soundex 编码的准确性(Gill,1997) 。 目前大多数基于读音的相似度算法都是针对英文发音，而对于中文字符读音相似 度的计算通常是将中文字符转换为对应的拼音，再利用已有的方法进行编码(刘 兵，2010;曹犟，等，2009)。

大量相似度算法的提出为记录相似度的计算提供了更多的选择，但同时也带 来了问题。Cohen 等(2003)和Snae(2007) 通过对多种相似度算法的实验比较，指 出没有一种相似度算法能够适用于所有数据。基于此，Wang 等(2011)对相似度 算法选择问题进行了形式化，证明了该问题是 NP 难问题，并提出基于训练数据选 择最优相似度算法和最优匹配阈值的启发式算法 SiFi-Greedy、SiFi-Gradient 和 SiFi-Hill, 然而在实际应用中由于难以获得合适的训练数据，仍然难以应用。

3.2.3 匹配决策模型

通过记录比较得到记录对的比较向量后，匹配决策步骤依据各种决策模型对 记录的匹配状态进行判断。根据记录关联方式的不同，现有的决策模型可分为以 下三类：两两比较模型、聚类模型和基于关系的决策模型。

<3.2.3.1> 两两比较模型

两两比较模型将实体分辨看作分类问题，依据记录对的比较向量将记录对划 分到匹配、不匹配和可能匹配的类别中，然后通过传递闭包获得描述同一客观实体 的所有记录。现有的两两比较决策模型主要包括概率决策模型和基于规则的决策

et al.,1969)在具有训练数据和属性条件独立的假设

模型。概率决策模型(Fellegi

下，计算条件概率的比值

,其中γ和厂分别为比较向量和比较

**56**)数据质量导论

空间，为当前记录对，M 和 U分别为匹配和不匹配记录对集，通过将R 与上下两 个阈值进行比较判断记录对r的匹配状态。由于实际应用中常常缺少计算条件概 率所需的训练数据，Winkler(1991) 和 Herzog 等(2007)利用期望最大算法对条件 概率进行估计。Verykios 等(2003)在概率决策模型的基础上考虑了错判匹配状态 的不同代价，提出了基于最小代价的概率决策模型。Naumann 等(2010)提出基于 规则的决策模型，匹配规则由一组布尔表达式及其对应的匹配状态组成，形式为 P=(term₁₁Vterm,₂V…)A…A(term,IVterm,₂V…)→C, 其中 term,,是对属性 i的第j 个相似度S 的一个判断(如，若s>0.8, 则 term=1, 否则 term=0)。

如果某个比较向量中的属性相似度满足布尔表达式P,那么该记录对的匹配状态 就为该表达式对应的匹配状态C。 匹配规则通常由领域专家制定， 一般难以保证 规则集合的完备性，Fan等(2009)为了降低规则过少对实体分辨准确性的影响，将 决策规则描述为匹配依赖，并给出了匹配依赖的推理规则，可以根据少量已知规 则通过推理获得蕴含的规则集合，从而提高决策准确性。为了减少领域专家的 参与程度，Cochinwala 等(2001)、Christen(2008b) 、Christen(2008a) 和 Arasu 等 (2010)分别提出了利用CART决策树、ID3 决策树、支持向量机和主动学习技术 从训练数据中学习决策规则的方法。由于决策规则会随着人们对数据、模式和 应用的认识程度的加深而发生改变，Whang等(2010)对规则的演化进行了形式 化，提出了规则单调和上下文无关两个约束，指出满足这两个约束的规则可以使 用增量方式进行处理，在当前的决策结果的基础上获得新规则的决策结果，从而 减少计算复杂度。

[**3.2.3.2**](3.2.3.2) **聚类模型**

聚类模型将实体分辨过程看作聚类问题，每一个类别对应着一个客观实体。 Monge(2000)和 Hassanzadeh 等(2009)利用多种相似度测度和优先队列对数据库 中的记录进行聚类，为每一个实体选出一个代表性的记录保存在优先队列中，每条 记录与优先队列中的代表记录进行比较，依据不同的相似度值将记录划分到一个 已有的实体中或建立一个新的实体。Hernández等(1995)利用聚类方法对两两比 较模型的结果中可能匹配的记录对进行后续处理，在匹配记录对的传递闭包的基 础上建立记录之间的匹配关系图，然后通过基于阈值的子图划分将描述同一实体 的记录对聚集到一起，解决因传递闭包带来的冲突和误判，该方法的不足之处是 用记录两两比较的结果确定记录之间匹配关系图的结构，无法检测比较结果中 没有的匹配记录。Chaudhuri 等(2005)为了解决 Hernández等(1995)的问题，在 整个比较空间上构建记录匹配关系图，利用紧密集和稀疏近邻的概念对记录对 进行聚类。

[**3.2.3.3**](3.2.3.3) **基于关系的决策模型**

基于关系的决策模型利用记录中的关联关系构建记录之间的关系图，利用图

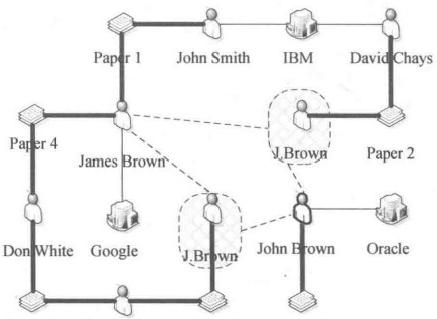
第3章 典型数据清洗技术的发展动态(

论知识计算连接强度，进而判断记录对是否匹配。基于关系的决策模型主要涉及 两个问题：①如何构建关系图；②如何计算关系图中连接的概率或权重。Kalashni- kov 等(2006)构建了数据库中不同实体间的关联关系图，每个节点代表数据库中 的一个实体，每条边表示两个实体间存在关联关系，利用两个节点之间的路径长度 估计两个节点的关联程度。Dong等(2005)构建了数据库中的依赖关系图，每个节 点表示一对实体间的相似度，节点间的边表示两个实体相似度之间存在依赖关系， 利用相似度间的依赖关系迭代地计算两条记录的相似度。Bhattacharya 等(2007) 构建了记录间的关联关系超图，以每条记录作为节点，用超边连接多个存在关联关 系的节点，每条边的权重由属性相似度和关联相似度的加权获得。Kalashnikov 等(2006)、Dong等(2005)和 Bhattacharya等(2007)的实验结果表明基于关系的 决策模型的准确性整体上高于基于规则和基于聚类的决策模型，但同时也大大 增加了计算复杂度，Rastogi 等(2011)为提高此类决策模型的效率，提出划分独 立关系子图的方法，切断连接强度小于一定阈值的边，将整个关系图划分为多个 关系子图，独立地在每个子图中判断记录的匹配状态，最后合并所有子图的决策 结果。

**3.2.4** **基于关系的实体分辨**

基于关系的实体分辨通常是在基于特征相似度(Feature Based Similarity,FBS) 方法的实体分辨结果的基础上，利用图模型对数据集进行建模，对 FBS 方法不能 正确识别的实体描述进行处理。将数据集看作一个实体关系图 G(V,E), 其中 V 表示数据集中的实体集合，E 表示实体之间的各种关系。V中包含两种类型的结

点： 一种是普通结点，对应的 是利用 FBS 方法可以识别出 的具体实体； 一种是歧义结 点，对应的是 FBS 方法无法 确定究竟对应哪个具体实体 的实体描述。歧义结点通过 选择边与多个普通结点连 接，表示该歧义结点描述的 是与它连接的普通结点对应 的实体中的某一个，即这些 普通结点为该歧义结点的候 选结点。图3-2给出了一 个文献数据集(表3-1)的



Paper 5 Dustin Lange Paper 6 Paper 3

图3-2 文献数据集的实体关系图示例

数据质量导论 实体关系图示例。

表3 - 1 文献数据集示例

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **序号** | **文献** | **作者1** | **作者1的单位** | **作者2** | **作者2的单位** |
| 1 | Paper 1 | James Brown | Google | John Smith | IBM |
| 2 | Paper 2 | David Chays | IBM | J.Brown |  |
| 3 | Paper 3 | John Brown | Oracle |  |  |
| 4 | Paper 4 | James Brown | Google | Don White |  |
| 5 | Paper 5 | Dustin Lange |  | Don White |  |
| 6 | Paper 6 | Dustin Lange |  | J.Brown |  |

如图3-2所示，数据集中的6篇文献、作者和作者单位分别对应实体关系 图中的各个节点，作者和文献之间的“写作”关系以及作者和单位之间的“隶属” 关系对应于图中的边(实线),边的粗细区分了实体间两种不同的关系，粗线表 示“写作”关系，细线表示“隶属”关系。图中加虚线框的两个“J.Brown” 结点表 示这两个作者无法通过基于FBS的方法进行识别，它们分别通过选择边(虚线) 连接到“James Brown” 和“John Brown” 两个节点，表示它们可能是两个中的 一个。

为了讨论方便，将数据集D 中的任一实体描述记为d,d 的候选实体集可以利 用基于FBS的方法选择相似度大于一定阈值的方法获得，记为O₄={om,o,…,

oa},o₄ 表示d 的第i个候选实体，d 真正描述的实体记为o,εO₄ 。d 到 og的连接 强度记为cs(d,oa) 。 基于关系的实体识别的目的就是通过分析d 与候选实体间 的连接强度，从O₄ 中选出正确的o₄, 使得o₄=0₂o

由于并不知道d 真正描述的实体，因此，需要依据某种原则为d 选择对应 的实体。上下文吸引原则(Context Attraction Principle,CAP)由 Kalashnikov 等 人(2005)针对基于关系的实体识别问题提出，在之前的诸多研究中已经证实 了该原则的准确性和有效性，因此，这里依然将 CAP 原则作为实体分辨的基 本依据。

CAP原则：对于数据集中的任一实体描述d, 它的候选实体集为O₄={oa,

0m,…,0a}, 如果d 真正描述的实体为og, 那么d 与 o 之间的连接强度应明显 大于d 与其他候选实体之间的连接强度，即cs(d,ou)>cs(d,om),k=1,2,…,

n, 且 k≠j.

表3-2中的伪代码给出了依据CAP原则进行实体识别的具体算法。

第3章典型数据清洗技术的发展动态(**59**

表 3 -2 CAP\_ER 算法

|  |
| --- |
| Name:CAP ER(D,G) |
| **Input:**数据集D及其关系图G  **Output**:实体识别结果 |
| 1.For each de D  2. Ifd可以通过FBS方法识别  3. **Break;**  4. **Else**  5.0₄=Get\_All\_Option(d)  6. **EndIf**  7. **For** each oeO  8. R[d][o]=Find\_All\_Path(G,L,d,o)  9. **For** each reR[d][o]  10. cs[r]=Get\_One\_Path\_CS(r);  11. **EndFor**  12. CS[d][O₄ i]=Get \_Two\_Node\_CS(cs);  13. E₄=Select\_MaxCS\_Entity(CS[d]);  14. **EndFor**  15.End |

CAP\_ER 算法首先寻找每个FBS 方法不能识别的实体描述d 的候选实体(1~ 5行),然后在关系图中查找d 与其每个候选实体之间的所有L-short 路径(7 行),之后计算每条路径的连接强度，以及d 与其每个候选实体之间的连接强度 (8～10行),最后选择与d 连接强度最大的候选实体作为d 实际描述的实体。其 中，查找L\_short 路径和计算连接强度是该方法的关键，由于这里主要关注连接强 度的计算，而对于L\_short 路径的查找问题，则借鉴Chen 等(2007)、Nuray-Turan 等(2013)和FanX 等(2011)中采用的基于动态规划的路径查找算法，本书不对该 问题进行更深入研究。

**3.2.5** **实体分辨中的训练和测试数据集**

各种实体分辨方法的训练和测试需要用到大量的已知记录匹配状态的数据， 然而在实际应用中，人们难以获得一个数据集中任意两条记录的匹配状态，尤其是 高校和科研院所中的研究人员，甚至连获得真实的数据都十分困难。目前，在实体 分辨研究当中使用的数据主要有两类： 一类是公开的测试数据；另一类是利用工具 生成的人造数据。

)**数据质量导论**

[**3.2.5.1**](3.2.5.1) **公开测试数据集**

公开测试数据集一般由一些数据公司或组织在互联网上公开发布，任何用户 皆可访问和使用，如 DBLP(Ley,2002) 、IMDB(Andrei et al.,2012),以及由 SIGKDD 组织的 KDD-CUP 竞赛数据集(Tavallaee et al.,2009),等等；另外一些开源或商 业的实体分辨系统和软件也会在其软件发行的同时附带一些公开的测试数据集， 如 SecondString工具箱(Cohen,2003) 以及 FEBRL系统(Christen et al.,2004;Chris- ten,2008)。

目前在实体分辨研究中应用较多的公开测试数据集主要包括以下几种。

Cora数据集：该数据集是一个文献数据集，包含1295条机器学习领域的书籍 和文献记录，这些记录对应了189篇(部)现实书籍和文献，每条记录包含文献标 题、发表时间、作者姓名、期刊或会议名称等属性，数据集中存在着姓名缩写、录入 错误等问题。该数据集提供了所有记录对的真实匹配状态，在所有可能的837865 对记录中有17184对记录是描述相同的书籍或文献。

UCD人员数据集：该数据集包含的是部分都柏林大学(University College in Dublin,UCD)工作人员的数据信息。每个人员的信息表示为一个长字符串，包含 了该员工的姓名、职务、职称等信息。该数据集是一个干净的数据集，每一个人员 的信息都分配有唯一的标识。

DBLP 数据集：该数据集是一个开放的计算机科学专业文献数据库，在数据挖 掘、知识发现和实体分辨领域均有应用。数据集中的每条记录包含文献标题、期刊 名称、发表时间，以及作者名称等信息。在该数据集中， 一篇文献可能存在多个不 同的标题， 一个期刊(或者会议)可能存在不同的名称。该数据集不提供记录对的 真实匹配状态。

IMDB 数据集：该数据集是互联网上的一个影视数据库，其中包含了多种不同 种类的客观实体，如人员(演员、制片人、导演等)、影片、影视公司以及影片排行等 信息。在该数据集中，同一个人可能有不同的姓名(存在缩写、昵称等问题),同一 部电影也可能具有不同的名字，此外，相同的名字也可能指的是不同的人，但该数 据集也不提供记录对的真实匹配状态。

CDDB数据集：该数据集包含了从 FreeCD 数据库中随机抽取的9763 条 CD 记录，每条记录包含了CD的名称、类型、艺术家、发行时间等信息。该数据集提 供了所有记录对的真实匹配状态，对同一张 CD 的不同记录分配了相同的标 识符。

这些公开的测试数据具有真实、不涉及隐私和版权问题等特点，但这些数据集 通常规模较小，虽在测试实体分辨准确性上具有较好的应用效果，但在测试方法可 扩展性时具有一定的局限性。另外，不同的数据集中包含的数据质量问题不尽相 同，因此在一个数据集上得到的结果的通用性也有待考究。

第3章 典型数据清洗技术的发展动态 

[**3.2.5.2**](3.2.5.2) **人造数据**

鉴于公开测试数据在数量和内容方面的限制，研究人员在实体分辨的研究过 程中也使用了大量的人造数据。这些数据由各种数据生成工具生成，没有数据量 方面的限制，而且通过各种参数的控制，得到的数据具有与真实数据相同或相似的 特征，如可以使属性值具有与真实数据接近的概率分布，且具有与真实数据中类似 的数据质量问题。

目前，在实体分辨研究领域常用的数据生成工具有UIS DBGen(Bilenko et al., 2003b) 、FEBRL(Christen,2005;Christen et al.,2009) 、SOG(Talburt et al.,2009) 等，这些数据生成工具使用了多种不同的数据生成方法，主要可以分为两类。

第一类方法是利用算法生成每个属性值，即所有的属性值是利用算法生成的。 生成的数据与真实数据不存在任何关联关系，因此不存在任何隐私和版权问题。 在生成属性值时通常是依据属性的特点构造生成规则或建立查询表。生成规则一 般用来生成一些结构化的属性值，如身份证号、电话号码等；而查询表一般用于生 成包含大量不同属性值的属性，如生成姓名时需要建立姓氏表和名字表，生成地址 时则需要建立省市表、街道表等。在现实世界中，不同的属性值具有不同的分布概 率，而且属性之间存在着各种各样的依赖关系，因此在生成数据时需要对这些约束 关系进行建模，需要设置的参数越多，算法的复杂度也就越高。这种方法通常用来 生成干净的数据集，然后利用工具或通过人工的方式向其中引入不确定数据，因此 可以确定每对记录的匹配状态。

第二类方法是从真实数据中抽取需要的属性值。这种方法通常是在拥有多个 数据集的情况下，从不同的数据集中抽取不同属性的属性值，组成一个新的数据 集。由于属性值来自多个不同的数据集，因此通常也会避开隐私和版权问题， 一般 适用于拥有大量数据集的企业和研究部门。这些来自不同数据集的属性值保留了 真实数据的分布特征、依赖关系以及存在的数据质量问题，能较好地反映数据的真 实特征。当采用的数据集是干净数据时，这种方法生成的是干净数据，也需要向生 成的数据中引入不确定数据。

**3.2.6** **实体分辨面临的挑战**

尽管各个领域的研究人员对实体分辨问题进行了很多研究，但仍然面临着诸 多挑战。

1)缺少准确的实体标识

参与实体分辨的数据库中通常不包含唯一的实体标识或主键，如居民身份证 号、商品的货号等。如果数据库中包含正确的实体标识，那么实体分辨问题就变成 了数据库中的join 操作，可以直接通过 SQL语言来实现。不过，即使数据库中包

数据质量导论

含了实体标识，仍然需要确认这些标识是否准确、完整，是否随时间发生了变化等， 否则也不可能得到完全正确的结果。而且，在数据库中经常存在具有不同标识的 重复记录，这也为实体分辨任务增加了难度。

在数据库中不包含唯一实体标识的情况下，实体分辨需要用到数据库中的一 些普通属性。如在人员信息数据库中，这些属性可能是姓名、地址、出生日期等，这 些属性不是唯一标识，只能对记录进行部分区分，此外，这些普通属性通常存在着 各种各样的数据质量问题，如录入错误、信息不完整或随时间发生变化等。

2)缺少训练和测试数据

在大多数实体分辨应用中，并不知道两条记录是否真正描述相同实体(匹 配),实体分辨结果只是对两条记录是否描述相同实体的一个判断。这与模式识 别、机器学习等其他应用不同，模式识别等应用中一般都有可用的训练数据以及标 准的结果。在实体分辨应用中，如果不依据一些额外的信息，如咨询当事人等，无 从知道实体分辨结果正确与否。当数据库中包含大量记录时，如人口数据库或销 售记录数据库，基本上无法获得实体分辨的全部标准结果。因此，如何构建合适的 训练和测试数据对分辨结果进行评估是实体分辨面临的一个挑战。

3)计算复杂度高

在实体分辨过程中，如果参与分辨的是两个数据库，那么每个数据库中的每条 记录都需要与另外一个数据库中的所有记录进行比较，来确定他们是否描述相同 客观实体。如果是在单个数据库中进行实体分辨，那么每条记录都要与该数据库 中其他所有记录进行比较。因此，实体分辨的计算复杂度以数据库大小的平方递 增，而数据库中真正匹配的记录数目则呈线性递增，而且实体分辨中的记录比较算 法通常也具有较高的时间复杂度。显然，大量的计算资源花费在比较不可能匹配 的记录上面。

4)相似度函数的准确性受适用范围影响

在诸多实体分辨相关的研究中，大多数相似度函数是针对英文文本提出的，国 内在最近几年才逐渐意识到实体分辨对于数据应用的重要性，大多只是将针对英 文文本提出的相似度函数套用在中文文本上。但由于中文文本和英文文本在存储 和录入上存在的差异，针对英文文本的相似度函数并不一定完全适用于中文文本。 另外，即使具有了适用不同情况的多种相似度函数，在具体的实体分辨任务中，为 不同的属性选择合适的相似度函数也是一个挑战，如果选择的相似度函数不合适， 依然不能得到准确的结果。

5)传统的决策模型不再适应现有的数据环境

传统的实体分辨通常只涉及到单个类型的客观实体，如人员、商品、文献等，针 对不同客观实体的实体分辨通常是相互独立的，即对人员的实体分辨结果不会影 响到对商品或文献的分辨结果。然而随着信息技术的发展，尤其是互联网、电子商

第3章典型数据清洗技术的发展动态 

务和大数据技术的快速发展，数据库中可能同时包含多种类型记录，如在电子商务 网站的销售记录中可能既包含人员的基本信息也包含了商品的一些信息。因此， 对各个类型的记录进行实体分辨的结果将不再相互独立，它们通过各种关系相互 关联、相互影响，如何针对这些新出现的情况提高分辨的准确性也是实体分辨需要 面对的一个挑战。

**3.3** **不完整数据清洗技术的发展动态**

完整性是首先要考虑的数据质量维度，其他数据质量维度均以数据完整性为 基础。本节介绍不完整数据的相关概念及清洗技术。

**3.3.1** **数据完整性及其评价方法**

本节简要分析不完整数据的产生原因，并介绍数据完整性的评价方法。

[**3.3.1.1**](3.3.1.1) **数据不完整的原因**

Wand 等(1996)在其研究中提出了真实世界系统到信息系统状态映射的四个 缺陷中数据完整性是数据质量的重要维度之一，并且在很多情况下完整性会直接 影响到其他维度。在数据挖掘、数据仓库和数据质量管理中，数据缺失是影响数据 完整性问题的主要原因，不完整数据的处理是数据清洗等研究领域重要内容(曹 建军，等，2010)。

现实中由于人为或系统因素造成的数据不完整十分常见，刘永楠(2013)归纳 了造成不完整数据的主要原因：

(1)数据采集过程的不完整。例如在物理仪器采集时，由于仪器故障等原因 可能丢失部分数据；在问卷调查中，受调查者拒绝回答一些问题会导致数据的 缺失。

(2)数据传播过程中的不完整。例如信息在网络中传播时，由于带宽、时延等 原因，造成数据发生丢失。

(3)异构数据转换过程中的不完整。不同结构的数据由于模式的不同，相互 转换时会导致信息的丢失。

(4)数据更新不及时导致的不完整。多个用户共同维护一个数据库时，更新 不及时导致的不完整十分常见。例如某公司管理部门为了统计信息建立的一些关 系，下属人员没有及时进行录入导致属性值的缺失。

[**3.3.1.2**](3.3.1.2) **数据完整性评价方法**

在数据完整性度量的研究中，Nauman 等(2004)提出了在多信息源集成环境 下，数据集成的完整性计算问题。他将完整性分为两个维度：覆盖度和密度。覆盖

数据质量导论

度是指数据库中存了多少数据，密度指其中有多少是非空的。在指示不同数据源 的完整性时，给出了在不同精度下计算完整性的方法。

Motro 等(1996)使用了抽样技术来分析样本的完整性，然后根据完整性将数据进 行划分，将完整性相当的数据划分到同一视图中。从而可以对不同视图上的查询结果 完整性进行估计。刘永楠(2013)认为现有数据完整性评价方法中存在以下不足：

(1)计算当前数据的完整性时，需要额外的信息，而这些信息经常是难以获取的。

(2)现有评价完整性方法是评价非空值占整体数据的比例，但是通过使用函 数依赖等数据库本身特性可以填充部分空值，因此这种评价方法并不能真实反映 数据库本身包含的信息完整性。

因此，刘永楠(2013)提出了一种可以适用于不同应用的数据完整性计算模 型，这个模型由属性值完整性、元组完整性、关系完整性来刻画不同粒度的完整性。

**3.3.2** **不完整数据的分类**

从数据清洗的角度，不完整数据处理包括缺失数据的检测、分类和估计填充等 步骤，分别对应数据清洗的检测、分析和修正三类典型技术。在实施缺失数据估计 填充前，首先要完成缺失数据的检测和分类，以针对不同的缺失数据类型选择合适 的处理方法。

[**3.3.2.1**](3.3.2.1) **不完整数据的层次分类**

根据缺失发生的层次，可以将数据的不完整分为三个层次：表层次、记录层次 以及字段层次。

表层次的不完整有三种情况：表缺失是指该表没有创建；空表是指没有填记录 的表；表不完整指数据集从表的层面来说是“完整”的，也就是该表被正确的建立 且非空表，但是该表存在记录层次的不完整。

记录层次的不完整有三种情况：记录缺失是指表中缺少某些记录；空记录是指 该记录只有主键等创建记录时生成的字段值，其余字段为空字段；记录不完整表示 该记录在记录层面来说是“完整”的，该记录被正确地建立并且是非空记录，但是 存在字段层次的不完整。

字段层次的不完整有三种情况：空字段是指字段值为空值的字段；字段不完整 是指字段内容不完整，如中文姓名只填写一个中文字符等；隐性缺失(Disguised Missing)是 由Hua 等(2007)提出的概念，指表面看起来是完整但实际上属于系统 默认值或者用户随意填写的字段，由于其值并不能反映任何真实的信息，因此认为 这是一种字段不完整的情况。

[**3.3.2.2**](3.3.2.2) **不完整数据缺失机制分类**

导致数据不完整的原因和过程不同，对应着不同的缺失处理方法。Rubin

第3章 典型数据清洗技术的发展动态(

(1976)提出使用数据缺失机制来描述数据发生缺失的过程，并按照缺失机制将不 完整数据分为完全随机缺失、随机缺失以及非随机缺失三种类型。

当属性缺失的概率同属性自身的取值和其他属性取值都无关时，属于完全随 机缺失中；当属性缺失的概率同属性自身的取值无关而同其他未缺失属性相关时， 属于随机缺失；当属性缺失的概率同自身的取值相关时，属于非随机缺失。

完全随机缺失和随机缺失属于可忽略缺失，处理较为简单；而非随机缺失属于 不可忽略缺失，处理难度最大。

**3.3.3** **不完整数据清洗技术**

在应用统计学中为了消除或降低不完整数据的不利影响，许多研究者提出了 各种不完整数据的处理方法，Little 和Rubin(2002) 将常见的不完整数据处理方法 归纳为以下几类：

(1)基于完整实例的方法；

(2)基于加权方法；

(3)基于填充的方法；

(4)基于模型的方法。

其中基于模型的方法过于复杂， 一般用于特殊应用中，而基于完整实例、基于 加权和基于填充的方法经常被用在不完整数据的清洗技术中。

[**3.3.3.1**](3.3.3.1) **基于完整实例和基于加权完整实例的方法**

基于完整实例的方法(简称完整实例法)只关注所有属性都完整的记录，而删 除存在缺失的记录。这一方法的优点有： 一是简单性，删除存在缺失的记录之后可 以使用标准的数据统计分析方法，而不需要修正；二是一元统计量的可比较性，因 为这些统计量都是基于相同样本计算的。

完整实例法的缺点是抛弃了不完整的个体，存在信息损失。这一信息损失有 两个方面：精度损失以及当缺失机制不是完全随机缺失时出现的偏离。当精度损 失和偏离很小时，完整实例法从简单性来讲是合理的，因为从不完整记录中可以获 取的信息是微乎其微的，特别是当完整实例的比例较高时似乎更是如此。然而很 难找到一种经验或者法则来对此进行判断，因为精度和偏离的程度不仅依赖于完 整实例的比例和缺失数据的模式，还依赖于完整个体和不完整个体的差异程度以 及感兴趣的参数。

为了降低完整实例分析中的偏离，加权是个简单有效的措施：在清洗之后的分 析中对这些记录加权进行使用。在抽样调查中这是最常用的策略，常用于处理单 元不响应问题。在那些被删除的不完整样本中其实蕴含有一些可用的信息，例如 他们的地理位置，利用这些信息可以对响应者指定权重，这样至少可以部分地调整

 数据质量导论

缺失数据所带来的偏离。

[**3.3.3.2**](3.3.3.2) **基于填充的方法**

数据填充方法总体上可以分为两类：单一填充和多重填充。

1)单一填充

在这种方法中填充值是缺失值分布的一个平均值或者抽样值，也就是说要求 以观测到的数据为基础，建立一个模型来估计缺失值。Little 和 Rubin 论述了关于 缺失值填充建模的方法。

明确建模的方法有：①均值填充(Mean Imputation),以变量观测到的部分的均 值来填充缺失值；②回归填充(Regression Imputation),以有缺失值的变量对完整观 测的变量做回归，用缺失变量的完整部分的数据估计回归模型的参数，再用此模型 来预测缺失部分的值；③随机回归填充(Random Regression Imputation),用回归填 充值加上一个随机项来预测缺失值。

模糊建模的方法有：①热卡填充(Hot Deck),将缺失项的值用类似样本中点的 对应值代替；②冷卡填充(Cold Deck),用一个从其他来源的，如以往调查中的完整 个体值代替缺失值；③替代法，主要在抽样阶段使用此方法，是指当前个体属性有 缺失时使用另外个体进行调查。

2)多重填充

多重填充是指对每项缺失都用一个填充向量(维数M 不小于2)来代替，对于 填充向量中的每一个填充值可以用单一填充中的方法来得到。这样可以构造M 个完整的数据集。最终的参数估计值就是由M 个完整数据集得到的估计值的某 种综合(如 Rubin 法则)。

上述算法大部分都是利用已有数据集中的所有完整数据作为参照来建立填充 模型，但是这样的方法效率较低而且并不一定具有更高的准确性。Bhekisipho 等 (2010)提出一种基于集成学习的缺失数据填充方法，该方法是一种有监督机器学 习方法，通过决策树将训练数据分片，然后在不同的数据分片上使用不同的填充模 型来估计缺失值 ， 最后将多个结果综合起来作为最终的填充结果 。 Anagnostopoulos 等(2014)为了提高填充的精度和效率提出的分布式填充策略，将 数据分散到多个节点上，并在每个节点上训练一个填充模型。对一条待填充数据 只选择其中部分节点进行填充，然后综合多个填充结果作为最终填充值。

本书第8章中对不完整数据的填充技术进行了详细论述，并提出了多种填充 方法对不完整数据进行填充。

**3.4** **不一致数据清洗技术的发展动态**

不一致数据是另一大类实例层数据质量问题，本节重点介绍基于数据依赖理

第3章 典型数据清洗技术的发展动态(

论的不一致数据清洗技术。

**3.4.1** **针对一致性的数据依赖理论**

数据依赖(Data Dependency)定义了关系数据库属性间的语义约束，是现实世 界数据间内在联系在数据库中的抽象反映(胡艳丽，等，2009)。本节将以关系数 据的一致性为背景介绍各类依赖理论，说明数据依赖的定义形式。

假设关系R 建立在一组属性{A₁,A₂,…,Am} 上，记作Attr(R)={A₁,A₂,…, Am}。I为 R 上的一个实例集，记作IεR,t为I 上的元组，即t∈l。

Dom(A): 属性A的值域；

t[A]: 元组t 在属性A∈Attr(R)上的取值。类似地，t[X] 表示元组t 在属性集 XCAttr(R) 上的取值。

[**3.4.1.1**](3.4.1.1) **统计依赖**

统计依赖应用于提高数据质量是一个新兴的课题，主要以概率论为基础提高 数据一致性。典型的统计依赖理论包括：概率依赖(Probabilistic Dependencies, PDs)、贝叶斯网络(Bayesian Networks,BNs)、马尔可夫链(Markov Chains,MCs)、隐 马尔可夫模型(Hidden Markov Models,HMMs)、动态贝叶斯网络(Dynamic Bayesian Networks)等。统计依赖弥补了传统的函数依赖在数据一致性研究领域的局限性， 为确定性数据“基本一致”的描述需求和不确定性数据的一致性度量提供了理论 依据(周傲英，等，2009)。

定义3-1(概率依赖):若任意元组t 在属性A上的取值，以稳定概率依赖于t 在属性集X 上的取值，称为属性A与属性集X 存在概率依赖关系，其概率记作P(t [A]lt[X])=P(AIX)。

可见，关系数据的概率依赖是使用概率论中的条件概率表达的，根据概率的归 一性，若属性A 与属性集X存在概率依赖关系，则 

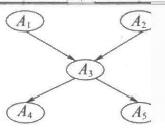
若已知属性(字段)之间的概率依赖关系与相应的概率阈值区间，则可以检测 在区间外的数据，为判断和提高数据一致性和准确性提供参考。

**定义3-2(贝叶斯网络(Nir et al.,1997;Korb et al. ,2004)**):关系R 上，属性 的贝叶斯网络用二元组表示：B=(S,P) 。 其中：

S表示反映R 上属性(A₁,A₂,…,Am) 依赖关系的有向无环图，存在依赖关系的 属性之间根据因果关系用有向弧连接；

P 表示存在依赖关系的属性之间的概率表集合，每个概率表对应一个属性。

图3-3给出了关系R 上属性(A₁,A₂,…,A 。)的贝叶斯网络示意图。可见，实 际上是基于贝叶斯网络理论构建关系R 上属性的概率依赖S, 每个属性对应的条

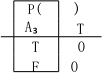
)数据质量导论

 A₁  A₂  T  F     0.001 0.999

件概率表构成P。 例如，当A₁ 和A₂ 属性分别取T 和F 时 ，A₃ 属性取T 值的概率为 0.94,即P(A₃=TIA₁=T,A₂=F)=0.94。

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | P(A₁) |  |  |  | P(A₂) |  |  |
|  | T | F |  |  | T | F |  |
|  | 0.0010.999 |  |  |  | 0.002|0.998 |  |  |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |
|  | A₁ | A₂ | A₃ | A₄ | As |  |
| tl | all | a21 | a31 | a41 | a51 |  |
| t2 | a12 | a22 | a32 | a42 | a52 |  |
| t3 | al3 | a23 | a33 | a43 | a53 |  |

 t4  al4  a24  a34  a44  a54 A₄A₃

|  |  |  |  |
| --- | --- | --- | --- |
| P(As | A₃) |  |  |
| A₃ | T | F |  |
| T | 0.7 | 0.3 |  |
| F | 0.01 | 0.99 |  |

图3-3 关系R上的贝叶斯网络示意图

不难看出，贝叶斯网络是概率依赖在整个关系R 上的表达，P 中的概率表的每 一行的和都为1。使用贝叶斯网络描述数据之间的依赖关系，实际上是将基于图 模型的概率理论应用到数据一致性的控制中。

定义3-3(马尔可夫链(Levin et al.,2008;Resnick,2002)):若任意元组t;在 属性A 上的取值，以稳定概率依赖于t;-1,t--2,…,t₁- 在属性A 上的取值，称为属性

A 在元组t 序列上存在马尔可夫性，其概率记作P(t₁[A]It-₁[A],-2[A],…,r-k [A]) 。对于一阶马尔可夫假设，有P(t₁[A]lt-₁[A],t₁-2[A],…,t;-k[A])=P(t [A]lt;-₁[A])。

马尔可夫依赖关系一般适用于物理信息系统中经传感器采集的各种物理对象 和物理过程的一维测量数据，如实时监控、实时检测数据等，主要用于生产调度、过 程控制、现场指挥、环境保护中的数据一致性的保证(中国计算机学会大数据专家 委员会，2013)。马尔可夫链也被看作是最简单的时间贝叶斯网络(Ghahramani, 2001)、

定义3-4(隐马尔可夫模型(Eddy,1998 )):若任意元组t;在属性A₁ ,A₂ 上的 取值，存在下列依赖关系：

t; 在属性A₁上的取值以稳定的概率依赖于元组t-, 在属性A, 上的取值，即马 尔可夫假设：P(t₁[A₁]lt₁-₁[A₁],t-2[A₁],…,t₁[A₁])=P(t₁[A₁]lt₁-₁[A₁])。

t, 在属性A₂上的取值以稳定的概率依赖于t, 在属性A₁上的取值，即独立性假 设：P(t₁[A₂]lt-₁[A₂],t₁[A₁])=P(t₁[A₂]lt₁[A₁])。

隐马尔可夫模型将马尔可夫理论扩展到两个字段上，是马尔可夫理论在关系 数据上的应用。类似于马尔可夫链，该模型适用于解决时间序列测量数据的一致 性问题，如气候数据中天气属性和气温属性、病历表中的血压和BMI值等(Ling et

第3章典型数据清洗技术的发展动态 

al.,2013)。

一方面，虽然较马尔可夫链有了进一步的扩展，但是其不能反映多于两个字段 的关系，对于全面反映时间关系数据的依赖存在局限；另一方面，模型的独立性假 设也可能在一定程度上导致丢失数据中的其他依赖信息，如事实上t; 在属性A₂ 上 的取值不仅依赖于t, 在属性A₁ 上的取值，而且依赖于t;-; 在属性A₂ 上的取值。

因此，相比隐马尔可夫模型，动态贝叶斯网络理论在反映记录上下文之间的依 赖关系和属性之间的依赖关系更加全面和具有一般性。

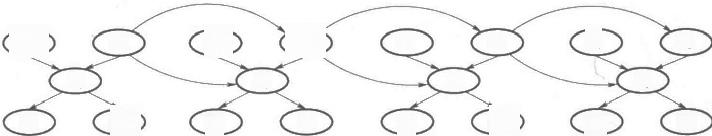
定义3-5(动态贝叶斯网络(Murphy,2002 )):关系 R 上，包含记录上下文和 属性之间的依赖关系可以用动态贝叶斯网络表示：(B₁,B)。 其中，B, 表示第1 条记录属性之间的依赖关系，B\_ 则表示相邻两条记录上下文之间的属性依赖 关系：

B 定义了相邻记录之间的条件依赖关系，即 Pa(t₁[A₁])), 其中Pa(t₁[A₄]) 表示t;[A₄] 的父节点，N为属性的数目。

而对于整个记录表的联合概率，则有： ),

其中M 为记录的条数。

以图3-3的关系R 为例，假设关系表中存储的是时间序列数据， 一个可能的 动态贝叶斯网络如图3-4所示。图3-4中第2条以上的记录的A₃ 属性，同时依 赖于当前记录的A₂ 属性和上一条记录的A₂ 属性。可以看出，隐马尔可夫模型是 动态贝叶斯网络的特例，利用动态贝叶斯网络对天气数据进行预测和修正是常用 的方法，如(Matthew,2008)。



*A₁ A₂* A A₂

*A₃*

A

*A₂* A

*A₃*

 *A₂*



*A₃*

*A₄*



As



*A₂*

*A₅*

*A₃*

图 3 - 4 关系R 上的动态贝叶斯网络示意图

<3.4.1.2> 函数依赖

函数依赖(Functional Dependencies,FDs)被认为是研究数据一致性的主要理 论，因此在数据质量研究中，函数依赖及其扩展的条件函数依赖(Conditional Func- tional Dependencies,CFDs)、条件谓词依赖(CFDs with Built-in Predicates,CFDps) 理论更加丰富和成熟。函数依赖对数据一致性的要求是“完全一致”,因此其更加 适用于已知确定依赖关系的数据约束。

定义3-6(函数依赖(Alechina,2000) ):满足实例I 的一个函数依赖，记作： X→Y。其中，属性子集X,YCAttr(R),“→” 表示“决定”关系，即当任意的元组t;,

70) )数据质量导论

t;∈l,t₁[X]=t;[X] 时，都有t₁[Y]=t[Y] 。 实例1满足φ:X→Y记作Ⅱ=φ。

**定义3-7(条件函数依赖(Fan et al.,2012;Fan et al.,2011)** ):满足实例I 的 一个条件函数依赖，记作：(X→Y,T,[XIIY]),“Ⅱ”为 LHS 和 RHS 的分隔符(LHS: Left Hand Side,RHS:Right Hand Side)。其中：

属性子集X,YCAttr(R),X→Y 是标准函数依赖；

T,[XUY] 为 XUY 上的某个具体的取值模式，即对于A∈XUY,T,[A] 为变量 值或者 Dom(A)中某个特定的常量值，“决定模式”X 与“被决定模式”Y一般称为 CFD的 LHS 和 RHS,即 LHS(Y)=X,RHS(X)=Y。

按照T,[XUY] “全部为常量值”和“部分为常量值”,条件函数依赖可以分为 常量条件函数依赖(Constant CFDs,CCFDs)和变量条件函数依赖(Variable CFDs, VCFDs)。

从定义可以发现，函数依赖将数据一致性约束在模式层，是对整个实例I 的约 束。但是在很多情况下，并非整个实例上的所有元组，都必须满足同样的函数依 赖。例如在A 国，邮编能唯一地决定乡镇，而在B 国，多个相邻的乡镇却可能公用 一个邮编。当这两国的数据出现在同一张数据表中时，显然不能用传统的函数依 赖Zip→County 表达实例I 的一致性。此时，扩展的条件函数依赖通过加入“条件 属性”,将这样的约束关系具体到部分的实例子集，更加客观灵活地反映实例I 的 一致关系。

函数依赖和条件函数依赖又被看作是“等号”依赖，但是如果要在一致性中表 达诸如“当任意的元组t,,t;εl,t₁[X]>a 时，都有t₁[Y]<b” 这样的约束时，则需要 用到条件谓词依赖。

**定义3-8(条件谓词依赖(Chen** **et** **al.,2009))** :满足实例I 的一个条件谓词 依赖，记作：(X→Y,T[X||Y]) 。 其中：

属性子集X,YCAttr(R),X→Y 是标准函数依赖；

T,[XUY] 为XUY 上的某个具体的谓词模式，即对于A∈XUY,T,[A] 为变量 值或者关于aεDom(A) 的逻辑表达式“opa”,op 为逻辑运算符“=,≠,>,<,≥, ≤”。

类似于条件函数依赖，按照T,[XUY] 的模式中“全部为常量值”和“部分为常 量值”,条件谓词依赖可以分为常量条件谓词依赖和变量条件谓词依赖。

条件谓词依赖实际上是将(条件)函数依赖的等号约束扩展到谓词约束，有一 类只针对数值属性的谓词依赖，即数值函数依赖(Numeric Functional Dependen- cies,NFDs)(Grace et al.,2014)。

区别于统计依赖的概率范围，上述定义中，无论是“等号依赖”还是“谓词”依 赖，对数据的约束都是确定的，即“对于任意的元组t,t;∈l,t₁[X]opa 时，都有“t; [Y]op'b" 。 因此，函数一般被看作是“确定性依赖”。

第3章典型数据清洗技术的发展动态 

<3.4.1.3> 过渡依赖

上述统计依赖和函数依赖不仅在对一致性的约束强度和灵活性上存在很大的 不同，其定义式也有不小的差异。而过渡依赖则同时借鉴了上述两种数据依赖的 相关理论，其拥有接近函数依赖的形式，并且通过附加的指标设置约束的概率。使 用最为广泛的过渡依赖是关联规则(Association Rules,ARs)。

定义3-9(关联规则):满足实例I 的一个关联规则，记作：(X→Y,t[X l Y]) 。其中：

属性子集X,YCAttr(R),X→Y 表示 X 与 Y在某个具体的常量模式上有一定 程度的关联，“→”表示“关联”关系；

t[XUY] 为 XUY 上的某个具体的取值模式，即对于A∈XUY,t[A] 为 Dom (A)中某个特定的常量值，t[XUY] 表示构成该AR的频繁项，“Ⅱ”表示将t[XU Y]分隔为该AR的 LHS和 RHS。

AR 有两个基本的度量：

(1)构成该AR的频繁项t[XUY] 的支持度 supp,即满足t[XUY] 取值的元 组在整个I 上的占比：supp(t[XUY])=N(t^[XUY])/N(t[XUY])。

(2)该AR的 LHS与 RHS的关联程度用置信度 conf 表示，即满足t[XUY] 的元组在满足t[X] 的元组中的占比：conf(t[XUY])=supp(t[XUY])/supp(t

[X])。

关联规则通过支持度和置信度两个指标，放宽了对标准函数依赖的约束强度。 传统的关联规则应用于关系数据的一致性检查，实际上是对关系数据的某个单一 属性在记录中是否一致进行检查，如(Buys₁→Buys₂,[Beer|Diaper]), 即上述定义 式中的X 和 Y实际上是单一属性值的不同组成部分。这种关联规则也称为1维 规则(Zhao et al.,2003)。而对于数据中的多个属性检查而言，需要将传统的关联 规则扩展到多维，即定义式中的X 和 Y 代表多个属性，如((Age,occupation)→ Buys,[20-29,Student ||Laptop])。

除了关联规则，概率函数依赖、近似函数依赖、模糊函数依赖理论本质上也属 于过渡依赖的范畴，只是所定义的附加指标的形式不同(Wang et al.,2009;Chris et al.,2002;Chris et al.,2004;Liu et al.,2003)。如近似函数依赖则是通过定义误差 阈值e(O≤e≤1) 来设置表示式“X→A”的约束范围，其含义为“不满足X决定A的 元组比例不超过误差阈值ε”。这里不再举例赘述。

[**3.4.1.4**](3.4.1.4) **基于约束强度的数据依赖的划分**

从上文的定义中可以看出，关联规则、函数依赖、条件函数依赖之间存在非常 紧密的联系。表3-3举例说明这几种依赖理论在数据一致性表达上的特点和 异同。

72)数据质量导论

表3 - 3 关联规则与函数依赖示例与含义

|  |  |
| --- | --- |
| **表达式** | **含义** |
| φ1:((CC,AC)→CT,[01,212]||NY) supp=0.5,conf=0.9 | AR:国家属性为“美国”(CC=01),地区码属性为“212”,且城  市名属性为纽约(CT=NY)的记录在实例I中的比例为50%;  对于国家属性为美国，地区码属性为“212”的记录集合，其城市 名属性为纽约的记录占比为90%。 |
| φ2:((CC,AC)→CT,[01,212]ⅡNY) | CCFD:国家为美国，地区码为“212”,则城市应为纽约 |
| φ3:((CC,AC)=CT,[01,908]||MH) | CCFD:国家为美国，地区码为“908”,则城市应为默里山 |
| 44:((CC,AC)→CT,[01,\_]|l\_) | VCFD:国家为美国，则地区码能唯一地决定城市 |
| ps:(CC,AC)→CT | FD:国家和地区码都一致地记录，则其城市也应一致 |
| 备注：CC(Country Code)、AC(Area Code)、CT(City)、NY(New York)、MH(Murray Hill) | |

由此看出，关联规则以支持度和置信度作为宽约束，是介于统计依赖(不确定 依赖)和函数依赖(确定依赖)的过渡形式。函数依赖是在模式层对属性的约束， 是对整个实例I 中的所有记录一致性的约束。而条件函数依赖是在实例层对属性 值的约束，是对实例I 中的部分记录的一致性的约束，是函数依赖的特例。而常量 条件函数依赖同时也是关联规则的特例(置信度为100%的关联规则为常量条件 函数依赖)。因此，相关研究证明，关联规则、函数依赖、条件函数依赖本质上是一 样的，能够使用统一的层次框架进行描述，并且这样的层次框架可以扩展到近似依 赖和近似关联规则中(Raoul et al.,2009)。

需要说明的是，过渡依赖虽然放宽了对传统函数依赖对数据“完全一致”的苛 刻要求，但是一般其置信度还是更接近于100%,即更多的关心的是“强关联规 则”,而非“弱关联规则”。而统计依赖则没有这样的要求，所以通常情况下，数据 一致性的研究更偏爱“过渡依赖”和“函数依赖”。

因此，按照约束强度从弱到强，可以将数据依赖划分为：不确定性依赖(统计 依赖)、过渡依赖、确定性依赖(函数依赖)。

**3.4.2** **典型数据依赖(规则)挖掘方法**

无论是使用统计依赖、函数依赖还是过渡依赖描述数据的一致性，首先要给定 数据属性之间的依赖关系或者字段之间的依赖规则。通常，有三种途径给定针对 某实例I的数据依赖： 一是由业内专家根据先验知识指定依赖关系或者规则，主要 适用于数据属性较少、关系简单或专家经验可信度较高的场合；二是从数据中自动 挖掘出依赖规则集，主要适用于数据关系复杂且专家不可得的场合；三是结合专家

第3章典型数据清洗技术的发展动态 

领域知识，给定先验挖掘准则或部分依赖规则，从数据中挖掘出依赖规则集后，再 由专家对规则集进行检查和筛选。

显然，对于大型关系数据实例而言，依靠专家制订的依赖规则集的可靠性难以 保证，而单纯依赖自动挖掘算法从数据中挖掘相应的依赖规则集不加选择，则难免 搜索出很多对实际需求无意义的规则(如弱规则)。通常情况下，需要同时结合专 家知识和自动挖掘算法。

这里主要介绍典型数据依赖自动挖掘算法的研究现状。

[**3.4.2.1**](3.4.2.1) **贝叶斯网络构建算法**

在统计依赖中，概率依赖、马尔可夫链比较简单，只需要使用全概率公式计算 相应的条件概率或者检测每个字段是否具有马尔可夫性即可。而隐马尔可夫模型 只针对特定的两个字段，通常情况下在领域知识中是已知的，只需要学习模型的参 数(转移矩阵和生成矩阵)。但是，对于能够全面反映数据属性之间的依赖关系的 贝叶斯网络和动态贝叶斯网络，则需要从数据中自动构建出反映属性之间的依赖 性和条件独立性的网络结构。

贝叶斯网络学习算法分为两大类(Cheng et al.,2001;Margaritis,2003):基于 评分的启发式搜索算法和基于依赖分析的搜索算法。第一类方法基于给定的评分 标准，通过比较搜索出的新网络和旧网络确定网络解，因此，评分标准的选择很大 程度上决定了计算出的网络与真实解的匹配程度，常用的评分标准包括贝叶斯评 分方法、基于熵的评分方法和最小描述长度评分方法等。第二类算法主要通过条 件独立性测试计算变量之间的依赖程度，以确定网络的连线。

动态贝叶斯网络可以看作是展开的贝叶斯网络，同样可以使用上述贝叶斯网 络学习算法。

[**3.4.2.2**](3.4.2.2) **函数依赖挖掘**

函数依赖挖掘算法已经有相当丰富的研究，因为条件函数依赖包含了传统函 数依赖，首先以条件函数依赖为例，说明函数依赖挖掘的几个重要概念。

**定义3-10(不平凡、非冗余、k- 频繁的** **CFD** **)**:对于某一个实例I 上的 CFD: φ=(X→A,t;[X]l|a)。

若A≠X, 则φ为不平凡的CFD;反之，φ为平凡的CFD;

若 6YCX,使得φ':(Y→A,t;[Y]l|a) 成立，(即6BeX, 使得t;[B] 升级为变 量之后的φ依然成立),则φ为非冗余的 CFD;反之，φ为冗余的 CFD;

将实例I 上所有匹配φ的元组的集合记作 supp(φ,1), 若此集合的元组数目 lsupp(φ,I)l≥k, 则φ为 k- 频繁的CFD。

定义3-11(最小CFDs 集):若∑为CFDs集合，Vφe2 都有(X→A,t;[X] Ⅱ a) 的形式，且是不平凡、非冗余的CFD,则三是最小的CFDs。

定义3- 12(CFDs 正则覆盖集):若三是最小的 CFDs集，且三包含了实例I

**74)**数据质量导论

上所有的k- 频繁的 CFDs, 则 为 I 的 CFDs正则覆盖集。

**定义3-** **13(CFDs 挖掘问题)**:挖掘一个实例I 的 CFDs正则覆盖集，称为该实 例I 的 CFDs 挖掘问题。

上述定义说明，对于函数依赖的挖掘，只关心不平凡、非冗余、k- 频繁的 CFD 极小函数依赖集，也称为最小函数依赖集，或最小覆盖(Canonical Cover)。

传统的标准函数的挖掘已经趋于成熟，包括自顶向下的算法 TANE、FUN、FD\_ Mine和自底向上的算法 Dep-Miner、FastFDs 已经在数据库系统中得到了广泛的 应用(Huhtala et al.,1999;Wyss et al.,2001;Hong et al.,2008)。而对于条件函数 依赖的挖掘，Fan WF等人在提出此依赖的形式化定义的同时，也给出了相应的挖 掘算法，包括常量条件函数依赖挖掘算法 CFDMiner、变量条件函数依赖挖掘算法 CTANE 和深度挖掘算法 FastCFD(Fan et al.,2011)。这些算法成为条件函数依赖 的主流挖掘算法，并扩展后(approxCFDMiner,approxCTANE 和 approxFastCFD)被 用来挖掘近似条件函数依赖(Hiroki et al.,2013)。

CFDMiner算法基于开项集和闭项集进行二次挖掘，效率较高，而 Li等(2013) 对常量条件函数依赖挖掘进行了深入研究和剪枝优化，进一步提高了常量条件函 数依赖挖掘的效率。

对于常量条件函数依赖挖掘，根据上文，置信度为100%的关联规则即为常量 条件函数依赖。 一个自然的思路是将各类关联规则挖掘算法经过扩展，用于常量 条件函数依赖挖掘(Thierno et al.,2012)。如将 FP-tree 、Apriori 等扩展后的用于 挖掘常量条件函数依赖的算法 CCFD-FPGrowth 、CCFD-AprioriClose 和 CCFD- ZartMNR(Devi,2013) 。 显然，这些算法特点类似于其原型算法。

对于条件谓词依赖的挖掘尚不多见，相关的研究可以参考(Antonella, 等， 2014)和(Grace, 等，2014),限于篇幅，不再赘述。

[**3.4.2.3**](3.4.2.3) **关联规则挖掘**

作为数据挖掘的代表性研究，关联规则的挖掘算法已经相当丰富。关联规则 挖掘的主要目标就是搜索出所有满足给定最小支持度和置信度的关联规则。通常 关联规则挖掘会被分解为两个子任务：即挖掘满足最小支持度的频繁项集(也称 为大项集)和基于频繁项集挖掘出满足最小置信度的关联规则。

Zhao 等(2003)综述了各类关联规则的挖掘方法，包括早期的关联规则算法一 AIS(Agrawal,Imielinski,Swami)算法(只产生形如X,Y→Z 的关联规则，而不产生 形如X→Y,Z 的关联规则),最为著名的Apriori 算法(通过高效的剪枝策略，避免 了多次遍历数据库以及产生很多无效的候选项集，使得效率得到很大提升),基于 树结构的FP-tree 算法(只需对数据库进行两次遍历，将数据压缩为 FP-Tree, 从 树中挖掘关联规则),以及快速关联规则挖掘算法 RARM(通过使用SOTrieIT 结构 只遍历一次数据库，只产生1-项集和2-项集)等。这些算法各有优势和局限，如

第3章典型数据清洗技术的发展动态7

FP-tree 和 RARM在效率上高于 Apriori,但是在交互式系统中的应用却遇到了困 难，也无法进行增量挖掘，因为一旦用户改变支持度阈值或者关系数据库加入了新 的数据，FP-tree 和 RARM 则需要重新启动整个挖掘过程。

针对数据一致性检查的不同维度，相关的研究也扩展了多维关联规则挖掘算 法，如果数据的属性均为枚举型， 一种直接的方法是将关系数据的多个属性进行合 并，看作一个属性，直接使用上述传统的挖掘算法。而对于连续性属性而言，需要 解决的首要问题是对属性进行离散化处理，这一过程对于是否能够挖掘出强关联 规则至关重要。按照离散化的方法，可以将多维关联规则挖掘算法分为三类：静态 离散方法(按照与预先定义的距离将属性值离散化为适合的枚举值)、动态离散方 法(在挖掘过程中动态调整离散间距，以适应支持度和置信度)和基于距离的离散 方法(对距离相近的值进行聚类以离散化)。相关的研究成果已经由Nir 等 (1997)进行了综述，这里不再赘述。

对于近似依赖挖掘等其他过渡依赖的挖掘，Huhtala 等(1999)、Hiroki 等 (2013) 、Huhtala等(1998)和 Jalal 等(2009)也进行了相关的研究。

**3.4.3 基于数据依赖的数据一致性保证**

本节主要讨论利用数据依赖理论提高数据一致性的方法和步骤，以及基 于数据依赖减少或消除不一致数据研究中的重点和难点内容 依赖规则集 的推理与数据修复。

[**3.4.3.1**](3.4.3.1) **数据获取与维护中的一致性保证**

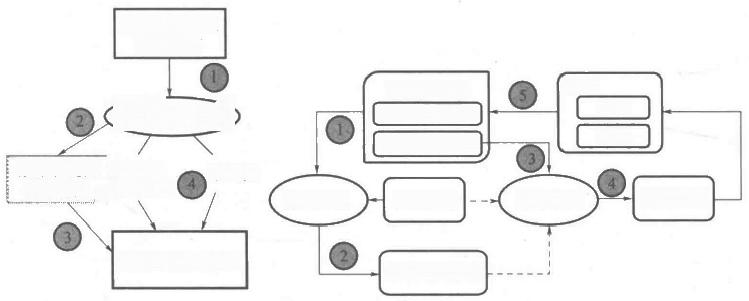
数据一致性保证的主要任务是根据给定的数据属性之间的依赖关系提高数据 的一致性。除了数据维护阶段，信息生命周期的数据获取阶段对于数据质量的保 证也是非常重要的时机(Danette,2010;Kuang et al.,2012)。在不同阶段提高数据 一致性的差异是：在数据获取阶段，主要利用依赖关系或规则对数据获取进行约束 和限制，尽可能地避免不一致数据的产生；而在数据维护阶段，则主要是利用依赖 规则集对数据进行检测，发现不一致的数据并进行修复，尽可能消除数据的不 一致。

过去利用统计依赖提高数据一致性的研究很少。然而，考虑到企业对数据 “基本一致”(而非“完全一致”)的度量需求，近年来也出现了使用贝叶斯和马尔 可夫理论控制数据质量的研究(Ling et al.,2013;Kuang et al.,2012)。

Kuang等(2012)首次提出了一个端对端的，在数据获取中进行数据质量控制 的数据收集系统 USHER,如图3 - 5中(a) 所示为 USHER 数据收集系统的组件和 数据流。需要进行数据获取的企业或组织在 USHER 客户端按要求提交格式化的 表格规范说明文件(包括字段名、数据类型、控件样式和基本约束等明细)和训练

)数据质量导论

数据，USHER 服务器生成相应的表格，并根据训练数据学习 一 个概率模型。在用 户录入过程中，系统将用户的录入值实时提交给该模型，模型对用户的录入值进行 预测或给出错误概率以提示用户。



数据表字段及

数据样本

标签

正常 存疑

概率模型 阈值设置

数据库服务器/UI

数据表字段 基本信息字段] 病症信息字段

历史数据

数据表规范 预测值、 说明文件 F错误似然(

贝叶斯网络模型

相似病患分组

分类模型

录入值

(a) (b)

图 3 - 5 两种基于统计依赖理论提高数据质量的框架

Ling 等(2013)则提出了一种录入错误检测与标记流程，对电子病例系统中满 足一般统计学一致性却违背医学一致性的数据进行了检测。例如， 一个肥胖的、 BMI>35 的糖尿病患者，很可能同时患有心血管疾病(表现为血压偏高、血脂异 常)。如果病例记录显示其血压低(但在统计值范围内),那么这个属性值被定义 为 Population Inconsistent。而 一个男性患者的怀孕属性如果为是，那么这个属性值 被定义为Personal Inconsistent。在该流程中，每一个病人被抽象为一系列数据属性 (包括病人的人口统计学信息和临床症状),表格被用于收集一些特殊类型的信 息，如血压、体重等。系统将病人的数据属性分为两类：在错误检测中保持不变的 值称为独立属性；需进行收集并被标记的属性称为依赖属性。

错误检测流程分为五步：①根据当前病人的独立属性和历史记录，使用分类模 型将病人记录分组；②在分组中搜索到与当前病人的相似病人；③在每一个分组 中，根据临床症状和历史记录学习一个概率模型；④对录入值设置一个阈值；⑤根 据概率模型和阈值对录入值添加“正常”或“异常”标签。这样给不一致数据库的 查询提供了可信度的参考，与吴爱华等(2012)的设计思路不谋而合。

在数据清洗中，数据检查的下一步工作，通常是利用依赖集合修复不一致的数 据，使其与记录的上下文“完全一致”或“基本一致”。使用数据依赖规则消除数据 中不 一 致通常有如下几种方法：

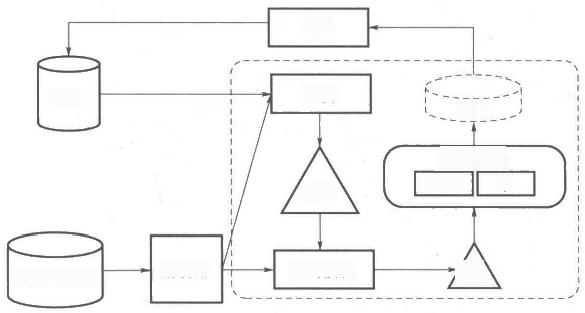
(1)直接删除违反依赖规则集的不一致数据；

(2)标记出违反依赖规则集的不 一 致数据，交由用户处理；

第3章典型数据清洗技术的发展动态 

(3)遵循所设置的准则(如最小操作准则),根据依赖规则集，使用自动修复 算法修复不一致数据。

如图3-6所示的是一种基于数据依赖的不一致数据消除与修复框架(Xu et al.,2013) 。 在该图中，数据依赖根据检测和修复需求被定义为“否定约束”的形式 (即函数依赖的求反，用于直接匹配不一致数据)。框架的输入为关系数据库和否 定约束集(Denial Constraints,DCs)。解析器从否定约束集中解析出约束规则，通 过检测模块筛选出不一致数据的可能组合，以超图(Confict Hypergraph,CH)的形 式表示。LookUp模块结合解析出的约束规则和超图，计算可能的修复方案(Re- pair Context,RC)。所有修复方案提交给决策模块，通过其中两个核心组件——修 复代价函数(Repair Cost Function,RCF)和二次规划(Quadratic Programming,QP)工 具选择一个方案进行修复。将修复后的数据更新到数据库中，重启整个流程，直到 超图为空。



应用

检测

更新

决策模块 VFM QP

否定约束集

查询模块

解析器

数据

超图

Rc

图 3 - 6 基于数据依赖的不一致数据检测与修复框架

[**3.4.3.2**](3.4.3.2) **依赖规则集的推理与数据修复**

使用函数依赖对数据一致性进行检查，首先需要对给定的依赖集合本身 的一致性进行检查，即保证依赖集中不存在矛盾的依赖规则，在很多情况下， 还需要对集合的最小性进行验证(Fan et al.,2012;Fan et al.,2008;Wang et al.,2014) 。 如果直接使用挖掘算法对数据中的依赖规则进行挖掘，则可以事 先给定挖掘的准则(例如上文中不平凡、非冗余、k- 频繁的 CFD的定义),那 么输出的结果必然满足一致性和最小性法则。但是如果由人工指定依赖规则 集或者在自动挖掘的结果中加入专家指定的规则，则必须检查依赖规则集的 一致性。

检查依赖集的一致性，有时比检查数据的一致性更加困难。

函数依赖集的一致性，通常包含可满足性(Satisfiability)和蕴含性(Implica-

**78** ))数据质量导论

始方女电的能可

tion)两个方面(Fan et al.,2012)。可满足性是指是否存在一个数据实例能够满足 给定的依赖规则集。显然，如果依赖规则集本身存在相互矛盾的规则，那么不存在 能够满足此函数依赖集的数据实例。而蕴含性则是指对于给定的函数依赖集 和函数依赖φ,是否有任意的数据库实例D, 使得D1=φ。 可以看出，如果 和φ 对于任意的数据库都有此关系，则表示φ对于 是冗余依赖规则。

FDs的“可满足性”和“蕴含性”检查已经被证明是复杂度为0(1)和O(IZ1) 的问题，而对于CFDs, 则 复 杂 度 分 别 是NP 完 全 的 和coNP 完全的。因此条件函数 依赖集的一致性保证是研究的难点。

在实际应用中，过多的不一致数据会给数据分析带来影响，因此，不一致数据 的修复对提高数据的可用性很重要。过去一般都是使用传统的函数依赖等对数据 进行约束和修复(George et al.,2010),随着条件依赖理论的深入和发展，约束规则 和修复规则的研究也进行了相关的拓展(George et al.,2014)。

从上文中可以看出，根据依赖规则集检测出的不一致数据组合经常并不唯一， 即便对于同一组合的不一致数据，也存在多种修复，如图3-7所示。所以，没有一 种普适性的最优修复方案，通常情况下会依据置信度或给定的属性的权重选择修 复方案。

4

B

3

3]

B

2

t

t₂

g

A t

t₂

*t₃*

A B

tt 2 t₂ 3

3

Z=(A→B)

t A

B

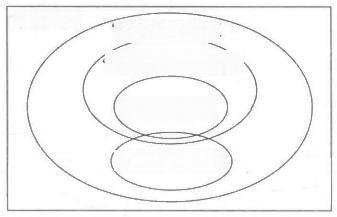
12

s [1](#bookmark387)

满足“基最小”准则；

满足“基-集合最小”准则；

满足“集合最小”准则

修复空间

不满足“基最小”准则；

满足“基-集合最小”准则；

满足“集合最小”准则

不满足“基最小”准则；

不满足“基-集合最小”准则；

满足“集合最小”准则

“集合最小”修复

“基-集合最小”修复

“基最小”修复

“代价最小”修复

A *I₄*t

B

3

tz

ts₃[ 1

3

不满足“基最小”准则；

不满足“基-集合最小”准则；

不满足“集合最小”准则

图 3 - 7 不一致消除的几种最小修复

修复方案的集合称为修复空间(George et al.,2014)。图 3 - 7 给 出 了 几 种 修 复方案的示例以及不同修复之间的包含关系。George 等(2014)介绍了这些修复

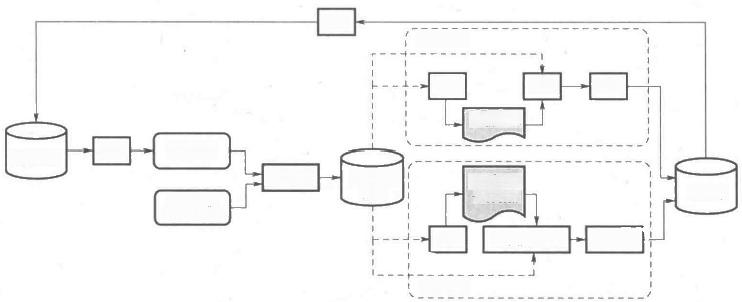
的形式化定义，并对它们之间的包含关系给出了详细证明。

[**3.4.3.3**](3.4.3.3) **提高数据** **一** **致性框架**

结合基于数据依赖的数据 一 致性保证的关键步骤和大多数研究成果，本书给

第3章 典型数据清洗技术的发展动态(79

出从数据获取阶段到数据维护阶段提高数据一致性的框架，如图3-8所示。从图 3-8中可以看出，无论是在数据获取阶段或数据维护阶段，首先需要根据已有的 训练数据学习出数据依赖模型(数据依赖关系或规则集),在有专家指定规则的前 提下，还需要对规则的可满足性和蕴含性进行冲突检测，以保证数据质量规则是一 致的极小集合。



更新

数据获取

预测 校验

当前录入

数据依赖

可能的不 一致数据

检测 搜索修复方案

数据维护

数据依赖 (先验知识)

数据依赖 模型

确认修义

冲突检测

历史数据

数据库

确认

训练

图3-8 基于数据依赖的数据一致性保证框架

在数据获取阶段，主要利用数据依赖模型提供的属性依赖关系或依赖规 则集对用户的录入值进行预测，辅助用户输入正确的值，并对已录入的值进行 校验，经过确认后提交。在数据维护阶段，主要检测和修复不一致数据。在大 多数一致性保证的模型和框架中，都需要提供多个候选的预测或清洗方案，并 经过交互式的用户确认过程(Hui et al.,2013;Helena et al.,2011)。

此框架包含了大部分成熟的研究成果，例如从训练数据中学习出贝叶斯网络 或最小函数集等。但是仍有许多值得研究的问题：

(1)如何提高基于训练数据挖掘依赖规则的效率和模型的准确性；

(2)在数据收集过程中，如何通过不断增长的训练数据实现依赖规则的增量 挖掘；

(3)如何优化函数依赖集的冲突检测算法；

(4)如何确定多种不一致数据修复方案的标准；

(5)如何在自动的挖掘和修复中加入人工干预，使得数据一致性保证更加符 合实际需求。

**3.5** **本章小结**

本章较系统地综述了实体分辨技术、不完整数据清洗技术和不一致数据清洗

)数据质量导论

技术，不难发现，从技术方法层面，以上三类实例级数据质量问题的数据清洗技术 是相互独立的。尽管如此，从对具体数据集的数据质量整体影响的角度，其三者又 是相互关联的。接下来，将对以上数据清洗技术分别具体研究。

**参考文献**

[1]中国计算机学会大数据专家委员会.2013.中国大数据技术与产业发展白皮书[R]. 北京：中国计算机 学会大数据专家委员会.

[2]刘兵.2010.基于拼音输入法的中文字符串近似匹配技术研究[D]. 沈阳：东北大学. [3]刘永楠.2013.数据完整性模型及评估算法的研究[D]. 哈尔滨：哈尔滨工业大学.

[4]吴爱华，谈子敬，汪卫.2012.不一致数据库上带信任标记的查询结果[J]. 软件学报，23(5):1167-1182. [5]邱越峰，田增平，季文赟，等.2001.一种高效的检测相似重复记录的方法[J]. 计算机学报，24(1):69-77. [6]周傲英，金澈清，王国仁，等(译).2009.不确定性数据管理技术研究综述.计算机学报，32(1):1-16. [7]胡艳丽，张维明，罗旭辉，等.2009.基于数据依赖的数据修复研究进展[J]. 计算机科学，36,10:11-15. [8]曹犟，邬晓钩，夏云庆，等.2009.基于拼音索引的中文模糊匹配算法[J]. 清华大学学报：自然科学版，

49(S1):1328-1332.

[9]曹建军，刁兴春，吴建明，等.2010.基于位运算的不完整记录分类检测方法[J]. 系统工程与电子技术， 32(11):2489-2492.

[10]Danette MeGilvray.2010. 数据质量工程实践[M]. 刁兴春，曹建军，张建美，译.北京：电子工业出版社.

[11]Alechina N.2000.Functional Dependencies Between Variables[J].Studia Logica,66(2):273 -283.

[12]Anagnostopoulos C,Triantafillou P.2014.Scaling Out Big Data Missing Value Imputations:Pythia vs.Godzilla [C]//Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Min- ing.New York:ACM,651-660.

[13]Andrei O,Mathias B,Manos T,et al.2012.Predicting IMDB Movie Ratings Using Social Media[M].Springer Berlin Heidelberg:Advances in Information Retrieval,503 -507.

[14]Antonella Z,Alberto T.2014.Discovering non-constant Conditional Functional Dependencies with Built -in Predicates[C]//International Conference on Database and Expert Systems Applications.Munich,Germany: Springer:35-49.

[15]Arasu A,Gotz M,Kaushik R.2010.On Active Learning of Record Matching Packages[C]//Proceedings of the International Conference on Management of Data.Indiana,USA:ACM:783-794.

[16]Baxter R,Christen P,Churches T.2003.A Comparison of Fast Blocking Methods for Record Linkage[J].Kdd Workshops,3:25-27.

[17]Bhattacharya I,Getoor L.2007.Collective Entity Resolution in Relational Data[J].ACM Transactions on Knowledge Discovery from Data(TKDD),1(1):5.

[18]Bhekisipho T,Michelle C.2010.Ensemble Missing Data Techniques for Software Effort Prediction[J].Intelli- gent Data Analysis,14:299-331.

[19]Bilenko M,Mooney R J.2003a.Adaptive Duplicate Detection Using Leanable String Similarity Measures [J]. ACM:39-48.

[20]Bilenko M,Mooney R,Cohen W,et al.2003b.Adaptive Name Matching in Information Integration[J].Inteli- gent Systems,IEEE,18(5):16-23.

第 3 章 典型数据清洗技术的发展动态 **81**

[21]Chaudhuri S,Ganti V,Motwani R.2005.Robust Identification of Fuzzy Duplicates[C]//Proceedings of 21st International Conference on Data Engineering.Tokyo,Japan:IEEE:865-876.

[22]Chen Z,Kalashnikov D V,Mehrotra S.2007.Adaptive Graphical Approach to Entity Resolution[C]//Proceed- ings of the 7th ACM/IEEE-CS Joint Conference on Digital Libraries.Vancouver,Canada;ACM:204-213.

[23]Chen W,Fan W,Ma S.2009.Analyses and Validation of Conditional Dependencies with Built -in Predicates [J].Springer Berlin Heidelberg.Database and Expert Systems Applications(DEXA),5690:576-591.

[24]Cheng J,Greiner R.2001.Learning Bayesian Belief Network Classifiers:Algorithms and System[M].Springer Berlin Heidelberg:Advances in Artificial Intelligence:141-151.

[25]Chris G,Edward R.2004.On Approximation Measures for Functional Dependencies[J].Information Systems, 29:483-507.

[26]Chris M,Mehmet M.2002.Improving Query Evaluation with Approximate Functional Dependency Based De- compositions[J].Lecture Notes in Computer Science,26-41.

[27]Christen P.2005.Probabilistic Data Generation for Deduplication and Data Linkage[M].Springer Berlin Hei-

delberg:Intelligent Data Engineering and Automated Learning -IDEAL 2005:109-116.

[28]Christen P.2008a.Automatic Training Example Selection for Scalable Unsupervised Record Linkage [M]. Springer Berlin Heidelberg:Advances in Knowledge Discovery and Data Mining:511-518.

[29]Christen P.20086.Awtomatic Record Linkage Using Seeded Nearest Neighbour and Support Vector Machine classification[C]//Proceeding of the 14th ACM SIGKDD Ihternational conference on Knowledge Discorery and Data Mining ACM,151-159.

[30]Christen P.2012a A Survey of Indexing Techniques for Scalable Record Linkage and Deduplication[J].IEEE

Transactions on Knowledge and Data Engineering,24(9):1537-1555.

[31]Christen P.2012b.Data Matching:Concepts and Techniques for Record Linkage,Entity Resolution,and Dupli-

cate Detection [M].New York,Springer Science &Business Media.

[32]Christen P,Churches T.2002.Febrl-Freely Extensible Biomedical Record Linkage [M].Canberra,Austrlia:

Australian National University.

[33]Christen P,Goiser K.2007.Quality and Complexity Measures for Data Linkage and Deduplication[M].Quality

Measures in Data Mining.[S.l.]:Springer Berlin Heidelberg:127-151.

[34]Christen P,Pudjjono A.2009.Accurate Synthetic Generation of Realistic Personal Information[M].Heidel-

berg,Germany:Springer-Verlag:507-514.

[35]Christen P,Churches T,Hegland M.2004.Febrl:A Parallel Open Source Data Linkage System[J].Advances in Knowledge Discovery &Data Mining:638-647.

[36]Churches T,Christen P,Lim K,et al.2002.Preparation of Name and Address Data for Record Linkage using Hidden Markov Models[J].BMC Medical Informatics and Decision Making,2(1):9.

[37]Cochinwala M,Kurien V,Lalk G,et al.2001.Efficient Data Reconciliation[J].Information Sciences,137 (1):1-15

[38]Cohen W W,Ravikumar P,Fienberg S E.2003.A Comparison of String Distance Metrics for Name -matching Tasks[C]//Proceedings of the IJCAI-2003 Workshop on Information Integration on the Web(IIWeb-03). Acapulco,Mexico:73-78.

[39]De V T,Ke H,Chawla S,et al.2011.Robust Record Linkage Blocking using Suffix Arays and Bloom Filers

[J].ACM Transactions on Knowledge Discovery from Data(TKDD),5(2):9.

[40]Devi KD.2013.Mining Constant Condional Functional Dependencies for Improving Data Quality[J].Inter-

 数据质量导论

national Journal of Computer Applications,74(15):12-20.

[41]Dong X,Halevy A,Madhavan J.2005.Reference Reconciliation in Complex Information Spaces[C]//Proceed- ings of the 2005 ACM SIGMOD International Conference on Management of data.Baltimore,Maryland,USA:

ACM:85-96.

[42]Draisbach U,Naumann F,Szott S,et al.2012.Adaptive Windows for Duplicate Detection[C]//IEEE 28th In- ternational Conference on Data Engineering(ICDE).Washington,DC,USA:IEEE:1073-1083.

[43]Eddy S R.1998.Profile Hidden Markov Models[J].Bioinformatics,14(9):755-763.

[44]Elmagarmid A K,Ipeirotis P G,Verykios V S.2007.Duplicate Record Detection:A Survey[J].Knowledge and Data Engineering,IEEE Transactions on,19(1):1 -16.

[45]Fan W,Jia X,Li J,et al.2009.Reasoning about Record Matching Rules [J].Proceedings of the VLDB Endow- ment,2(1):407-418.

[46]Fan X,WangJ,Pu X,et al.2011.On Graph-Based Name Disambiguation[J].Journal of Data and Information Quality(JDIQ),2(2):10.

[47]Fan W F,Geerts F.2012.Foundations of Data Quality Management[M].[S.I.]:Morgan &Claypool.

[48]FanWF,GeertsF,Jia XB.2008.Conditional Functional Dependencies for Capturing Data Inconsistencies[J]. ACM Journal of Data and Information Quality,33(2):1-6,48.

[49]Fan W F,Geerts F,Li J Z,et al.2011.Discovering Conditional Functional Dependencies[J].IEEE Transac-

tions on Knowledge &Data Engineering,23(5):683-698.

[50]Fellegi IP,Sunter AB.1969.A Theory for Record Linkage [J].Journal of the American Statistical Associa- tion,64(328):1183-1210.

[51]George B,Lhab F,Lukasz G.2010.Sampling the Repairs of Functional Dependency Violations under Hard Con- strains[J].Proceedings of the VLDB Endowment,3(1):197 -207.

[52]George B,Lhab F,Lukasz G,et al.2014.Sampling the Repairs of Conditional Functional Dependency Viola- tions.VLDB Joumal,23(1):103-128.

[53]Ghahramani Z.2001.An Introduction to Hidden Markov Models and Bayesian Networks.International Journal

Of Pattern Recognition and Artificial Intelligence,15(1):9-42.

[54]Gill L 1997.OX-LiNK:The Oxford Medical Record Linkage System[C]//Proceedings Int'I Record Linkage

Workshop and Exposition.Oxford,UK:Record Linkage Techniques:15-33.

[55]Grace F,Fan W F,Floris G.2014.Detecting Errors in Numeric Attributes[C]//International Conference on Web-Age Information Management.Macau,China:Springer:125-137.

[56]Gravano L,Ipeirotis P G,Koudas N,et al.2003.Text Joins in an RDBMS for Web Data Integration[C]//Pro-

ceedings of the 12th International Conference on World Wide Web.Budapest,Hungary:ACM:90-101.

[57]Hassanzadeh O,Miller R J.2009.Creating Probabilistic Databases from Duplicated Data [J].The International Journal on Very Large Data Bases,18(5):1141 -1166.

[58]Helena G,Antonia L,Emanuel S.2011.Support for User Involvement in Data Cleaning[J].DaWaK 2011,Lec- ture Notes in Computer Science:136-151.

[59]Hemández M A,Stolfo S J.1995.The Merge/Purge Problem for Large Databases[C]//ACM SICMOD Re- cord.San Jose,California:ACM,24(2):127-138.

[60]Herzog T N,Scheuren F J,Winkler W E.2007.Data Quality and Record Linkage Techniques [M].[S.1.]: Springer.

[61]Hiroki H,Ayako H,Chihiro I,et al.2013.Formalization and Discovery of Approximate Conditional Functional

第3章典型数据清洗技术的发展动态 

Dependencies[J].Lecture Notes in Computer Science,1:118 -128.

[62]Hong Y,Howard J.2008.Mining Functional Dependencies from Data[J].Data Mining Knowledge Discovery, 16:197-219.

[63]Hua M,Pei J.2007.Cleaning Disguised Missing Data:A Heuristic Approach[C]//ACM KDD'07.San Jose,

Califomia,USA:ACM:950-958.

[64]Huhtala Y,Karkkainen J,Porkka P,et al.1999.TANE:An Efficient Algorithm for Discovering Functional and Approximate Dependencies[J].The Computer Journal,42(2):100-111.

[65]Huhtala Y,Kinen J,Porkka P,et al.1998.Efficient Discovery of Functional and approximate dependencies using partitions[C]//International Conference on Data Engineering Orlando,Florida,USA:IEEE:392-401.

[66]Hui X,Wang H Z,Li JZ,et al.2013.A Data Cleaning Framework Based on User Feedback[J].WAIM

2013.Lecture Notes in Computer Science;514 -520.

[67]Jalal A.2009.Mining Approximate Functional Dependencies from Databases Based on Minimal Cover and EquivalentClasses[J].European Journal of Scientific Research,33(2):338-346.

[68]Kalashnikov D V,Mehrotra S.2006.Domain-independent Data Cleaning via Analysis of Entity -relationship

Graph[J].ACM Transactions on Database Systems(TODS),31(2):716-767.

[69]Kalashnikov D V,Mehrotra S,Chen Z.2005.Exploiting Relationships for Domain -Independent Data Cleaning [C]//Proceedings of the 2005 SIAM International Conference on Data Mining.Newport Beach,Clafornia, USA:SIAM International Conference on Data Mining(SDM):262 -273.

[70]Korb K,Nicholson A E.2004.Bayesian Artificial Intelligence [J].Chapman &Hall/cre Boca Raton FI: xxiv,364.

[71]Koudas N,SarawagiS,Srivastava D.2006.Record Linkage:Similarity Measures and Algorithms[C]//Proceed- ings of the 2006 ACM SIGMOD International Conference on Management of data.Chicago,Illinois,USA:ACM, 802 -803.

[72]Kuang C,Harr C,Neil,et al.2012.Usher:Improving Data Quality with Dynamic Forms.IEEE Transactions on

Knowledge &Data Engineering,23(8):1138-1153.

[73]Levin D A,Peres Y,Wilmer E L.2008.Markov Chains and Mixing Times[J].American Mathematical Society Providence Ri,3(4):xvi+,371.

[74]Ley M.2002.The DBLP Computer Science Bibliography:Evolution,Research lssues,Perspectives[C]//String Processing and Information Retrieval.Berlin Heidelberg:Springer:1 -10.

[75]LiJY,LiuJX,Hannu T,et al.2013.Effective Pruning for the Discovery of Conditional Functional Dependencies [J].The Computer Joumal,56(3):378-392.

[76]Ling Y,An Y,Liu M W,et al.2013.An Error Detecting and Tagging Framework for Reducing Data Entry Errors in Electronic Medical Records(EMR)System[C]//International Conference on Bioinformatics and Biomedi-

cine.Shanghai,China:IEEE Computer Society:249-254.

[77]LitleR,Rubin DB.2002.Statistical Analysis with Missing Data [M].New Jersey:John Wiley &Sons:19-20.

[78]Liu WY,Song N.2003.Fuzzy Functional Dependencies and Bayesian Networks[J].Journal of Computer Science & Technology,18:56-66.

[79]Margaritis D.2003.Learning Bayesian Network Model Structure From Data[D].Pittsburgh,Pennsylvania, USA:Carnegie Mellon Universiy.

[80]Matthew D K.2008.Weather Forecasting Using Dynamic Bayesian Networks[J].Department of Computer Sci- ence,University of Cape Town.

**数据质量导论**

[81]Michelson M,Knoblock C A.2006.Learning Blocking Schemes for Record Linkage[C].Menlo Park,Califor-

nia;Cambridge,Massachusetts;London:AAAI Press;MIT Press,21(1):440.

[82]Monge A E.2000.Matching Algorithms Within a Duplicate Detection System[J].IEEE Data Engineering Bul- letin,23(4):14-20.

[83]Motro A,Rakov I.1996.Estimating the Qcuality of Data in Relational Databases[C]//Proceedings of the 1996 Conference On Information Quality.Massachusetts,USA:MIT:94-106.

[84]Murphy K P.2002.Dynamic Bayesian Networks:Representation,Inference,and Learning[D].Berkeley:Uni-

versity of California.

[85]Naumann F,Herschel M.2010.An Introduction to Duplicate Detection[J].Synthesis Lectures on Data Man- agement,2(1):1-87.

[86]Naumann F,Freytag J C,Leser U.2004.Completeness of Integrated Information sources[J].Information Sys- tems,29(7):583-615.

[87]Nir F,Dan G,Moises G.1997.Bayesian Network Classifiers[J].Machine Learning,29:131-163.

[88]Nuray-Turan R,Kalashnikov D V,Mehrotra S.2013.Adaptive Connection Strength Models for Relationship- Based Entity Resolution [J].Journal of Data and Information Quality(JDIQ),4(2):8.

[89]Raoul M,Lhouari N.2009.A Unified Hierarchy for Functional Dependencies,Conditional Functional Depend- encies,and Association Rules [J].Lecture Notes in Computer Science:98 -113.

[90]Rastogi V,Dalvi N,Garofalakis M.2011.Large -scale Collective Entity Matching [J].Proceedings of the

VLDB Endowment,4(4):208-218.

[91]Resnick S I.2002.Markov Chains[J].Adventures in Stochastic Processes,140(1):107-116. [92]Rubin D B.1976.Inference and Missing Data [J].Biometrika,63:581 -592.

[93]Snae C.2007.A Comparison and Analysis of Name Matching Algorithms [J].International Journal of Applied Science,Engineering and Technology,4(1):252-257.

[94]Talburt J R,Zhou Y,Shivaiah S Y.2009.SOG:A Synthetic Occupancy Generator to Support Entity Resolution Instruction and Research [C ]//Proceedings of the 14th International Conference on Information Quality (ICIQ).Potsdam,Germany:9:91-105.

[95]Tavallaee M,Bagheri E,Lu W,et al.2009.A Detailed Analysis of the KDD CUP 99 Data Set [C]//Proceed- ings of the Second IEEE International Conference on Computational.Ottawa,Canada:Computational Intelli- gence for Security and Defense Applications(CISDA):1 -6.

[96]Thiermo D,Noel N,Jean -Marc P.2012.Discovering(Frequent)Constant Conditional Functional Dependen-

cies[J].International Journal of Data Mining,Modelling and Management,4(3):205-223.

[97]Verykios V S,Moustakides G V,Elfeky M G.2003.A Bayesian Decision Model for Cost Optimal Record Matching [J].The VLDB Journal,12(1):28-40.

[98]Wand Y,Wang R.1996.Anchoring Data Quality Dimensions in Ontological Foundations [J].Communications of the ACM,36(11):86-95.

[99]Wang J,Li G,YuJX,et al.2011.Entity Matching:How Similar is Similar [J].Proceedings of the VLDB En- dowment,4(10):622-633.

[100]Wang D Z,Dong X L,Sarma A D,et al.2009.Functional Dependency Generation and Applications in Pay-as - you -go Data Integration Systems[C]//Proceedings of the 12th International Workshop on the Web and Data- bases.Rhode Island,USA:WebDB.

[101]WangJN,Tang N.2014.Towards Dependable Data Repairing with Fixing Rules[C]//SIGMOD.Snowbird,

第3章 典型数据清洗技术的发展动态 

USA:ACM,457-468.

[102]Whang S E,Garcia -Molina H.2010.Entity Resolution with Evolving Rules [J].Proceedings of the VLDB

Endowment,3(1-2):1326-1337.

[103]Whang S E,Menestrina D,Koutrika G,et al.2009.Entity Resolution with Iterative Blocking [C]//Proceedings

of the 35th SIGMOD International Conference on Management of Data.Rhode Island,USA:ACM:219-232.

[104]Winkler W E.2006.Overview of Record Linkage and Current Research Directions [C]//Statistical Research

Division U.S.Census Bureau Washington D C.Washington DC,USA:Statistical Research Division;1-44.

[105]Winkler W E,Thibaudeau Y.1991.An Application of the Fellegi -Sunter Model of Record Linkage to the

1990 US Decennial Census [J].US Bureau of the Census:1 -22.

[106]Winkler W E,Yancey W E,Porter E H.2010.Fast Record Linkage of Very Large Files in Support of Decenni- al and Administrative Records Projects [C]//Proceedings of the Section on Survey Research Methods,Bos- ton,Massachusetts USA:American Statistical Association JSM:2120-2030.

[107]Wyss C,Giannella C,Edward R.2001.FastFDs:A Heuristic -Driven,Depth -first Algorithm for Mining Functional Dependencies from Relation Instances[C]//International Conference on Data Warehousing and Knowledge Discovery.Munich,Germany:Springer:101 -110.

[108]Xu C,Ihab F,Paolo P.2013.Holistic Data Cleaning:Putting Violations into Context[J].International Confer-

ence on Data Engineering:458-469.

[109]Yan S,Lee D,Kan M Y,et al.2007.Adaptive Sorted Neighborhood Methods for Efficient Record Linkage

[C]//Proceedings of the 7th ACM/IEEE -CS Joint Conference on Digital libraries.Vancouver,Canada:

ACM:185-194.

[110]Zhao Q K,Sourav S.2003.Association Rule Mining:A Survey[J].Nanyang Technogical University,3(3):

157-159.

**第4章** **实体分辨中的数据分块方法**

**4.1** **引言**

为了提高实体分辨效率，研究人员提出了各种各样的数据分块方法(Winkler et al.,2010;Draisbach et al.,2012;De et al.,2011;Kenig et al.,2013;Liang et al., 2014;Ramadan et al.,2014),然而随着大数据和社交媒体的兴起，互联网上半结构 化和非结构化数据不断增加，传统的数据分块技术导致漏检的风险逐渐加大，基于 冗余的数据分块方法得到了研究人员的关注。基于冗余的数据分块方法的核心思 想是允许同一条记录可以划分到不同的数据块中，使得描述相同实体的两个记录 有更大可能划分到某一个数据块中，从而减少漏检，提高查全率(De et al.,2011; Whang et al.,2009;Papadakis et al.,2011;Papadakis et al.,2013;Shu et al.,2012; Papadakis et al.,2012)。然而，这种基于冗余的数据分块方法会造成数据块之间 的大量重叠，相同的记录对可能会在多个不同的数据块中被多次比较，从而造成了 计算资源的浪费。

因此，本章首先针对基于冗余的数据分块方法，利用类似倒排索引的数据结构 保存记录的分块信息，提出了一种具有线性复杂度的冗余记录对识别方法，并通过 空间映射提高冗余记录对识别的效率。通过识别数据块中冗余的记录对，可以尽 可能地避免相同的记录对在多个数据块中被多次比较，从而提高实体分辨的效率， 且不会因此降低实体分辨的效果。然后为了减少需要识别的记录对，提出基于空 间映射的数据块约减方法。最后针对XML 的数据分块，提出一种基于 Canopy 聚 类的数据分块方法，提高分块的可扩展性以及效率。

**4.2** **基于冗余的数据分块**

为了描述方便，首先给出基于冗余的数据分块及其相关概念的定义。对给定 数据集D 中的任意一条记录r 均为一个二元组<ID,A,>, 其中ID 为该记录的全 局唯一标识，A,为该记录中所有属性的集合。每一个属性a,∈A, 也是一个二元组 <n₄,v₄>,n₁ 为该属性的属性名称，v, 为其属性值。基于冗余的数据分块的定义 如下。

第4章实体分辨中的数据分块方法 (87

定义4-1(基于冗余的数据分块):对于给定的数据集D,r;,r;为 D 中任意两 条记录，i,j=1,2,3,…,P₁,P, 为记录r,r, 的分块键值，(p,P₁) 为一组映射函数， 基于冗余的数据分块指的是将D中所有广(P,P) 的值为true 的两条记录划分到 同一个数据块b₂中的过程，使得b₁CDAV r,₁∈D:f(Pi,P)=true→r,';∈b₄。 所有数据块的集合称为数据集 D 的一个数据分块方案，记为B。

值得注意的是，同一条记录可能与不同的记录组合满足不同的映射函数(如 f(P₁,P)=true 且f(p₁,Pm)=true), 而且同样的两条记录也可能满足不同的映射 函数(如产(P₁,P₁)=true且f(p₁,P₁)=true), 因此，相同的记录可能会被划分到两 个不同的数据块中，不同的数据块可能存在记录重叠部分，从而产生冗余的记 录对。

参与实体分辨的数据可能来自于同一个数据集，也可能来自于两个不同的数 据集，当数据来自于两个不同数据集时，通常假设每一个数据集内部不存在描述相 同实体的记录(如果存在则首先对每个数据集单独进行实体分辨)。下面对两种 数据来源情况下的数据分块进行定义。

定义4-2(单源数据分块):单源数据分块指的是参与实体分辨的只有一个数 据集D, 数据分块方案B 中任意一个数据块b₄中的记录全部来自于同一个数据集， 即 Vb₄ ∈B,b₄={r₁,r₂,r₃,…,rn}→r₁ ∈D,i=1,2,3,…,n。

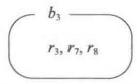
定义4-3(双源数据分块):双源数据分块指的是参与实体分辨的有两个数据 集 D₁和D₂, 数据分块方案 B'中任意一个数据块b^中的记录均来自于两个不同的 数据集D₁和D₂, 即 Vb^∈B',b={{rn,Tz,rig,…,rin},{rz₁,rz,Tza,…,Tzm}}→

r₁ ∈D₁Ary∈D₂,i=1,2,3,…,n;j=1,2,3,…,m 。 其 中ba={rn,riz,rg,…, ria},b}₂={rz₁,Tzz,Tz₃,…,fzm} 称为b?的内部分块。

实体分辨算法对每个数据块中的记录进行两两比较(双源数据分块只在两个 内部分块间进行比较),由于数据块之间存在重叠，因此，相同的记录对可能会在 不同的数据块中进行多次比较。记录对的比较算法通常具有较高的计算复杂度， 而且相同记录对的重复比较并不会提高实体分辨的查全率和查准率，因此，冗余记 录对的存在消耗了大量的计算资源，且对实体分辨效果没有任何提高。

4.3 基于倒排索引消除冗余记录对

实体分辨算法对同一数据块中的数据进行两两比较，计算过程中并不保存记 录对的比较状态，因此无法判断两条记录是否在其他的数据块中已经进行过比较。 通过对数据块进行排序索引，并利用哈希表保存每条记录的倒排索引，可以在不保 存记录比较状态的情况下，快速判断两条记录是否已经进行过比较，避免冗余 计算。

88))数据质量导论

4.3.1 数 据 块 排 序 索 引

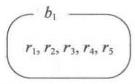
为了对冗余记录对进行识别，首先对划分好的数据块进行排序索引。对于数 据分块方案B 中的所有数据块，为每个数据块分配一个唯一的大于0的整数作为 索引值。对数据块进行赋值的依据多种多样，通常是根据具体应用需要为每个数 据块分配不同数值，不同的分配方案对于冗余记录对的识别没有影响(见4.3.3 节),这里为了满足下文中数据块约减(见4.4节)的应用需求，依据每个数据块中 包含的记录(对)数目为每个数据块分配索引值。

(1)对于单源数据分块方案 B,按照每个数据块b₄包含的记录数目Ib₄ l由大 到小的顺序对所有数据块排序，并按照排列顺序，从最大的数据块开始以逐个递增 的方式依次为每个数据块b 分配一个大于0的整数 index₁作为索引值，使得Vb₄, b₁∈B,lb₄l<lb₁l→index₄>index₁。

(2)对于双源数据分块方案B', 按照每个数据块b?包含的记录对数目1bkl× 1bz| 由大到小的顺序对所有数据块进行排序，并按照排列顺序，从最大的数据块 开始以逐个递增的方式依次为每个数据块b{分配一个大于0的整数 index,作为索 引值，使得Vb},b/∈B',Ib²l×1b₂ l<1b₁ l×1b₂ l→index>index₁。

图4-1给出了对单源数据分块方案进行排序索引和记录倒排索引的一个示 例。双源数据分块方案与此基本类似，不同之处只是每个数据块中的记录被分成 了两个内部分块。

|  |
| --- |
| r |
| F2 |
| r |
| F4 |
| rs |
| 76 |
| r |
| F8 |



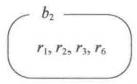
index₁=1

r:<1,2>

rz:<1,2>

:<1,2.3>

r4:<1>



index₃=3

index₂=2

r²<l>

s<2>

:<3>

Fg∵<3>

图4- 1 数据块排序索引、记录倒排索引示例

如图4- 1所示，b₁,b₂,b₃ 为记录r₁~rg 的一个单源数据分块方案，按照每个数 据块包含的记录数目由大到小的顺序进行排序，并按照排序结果，从包含记录数目 最多的数据块b,开始，逐个递增为每个数据块分配索引值，得到b₁,b₂,b₃ 的索引值 分别为index₁=1,index₂=2,index₃=3。

对数据块进行排序算法的时间复杂度为O(IBl·log|BI), 即使对于包含大量

第4章 实体分辨中的数据分块方法 

数据分块的划分方案也具有较好的可扩展性。

**4.3.2** **记录倒排索引**

在为每个数据块分配索引值后，采用类似倒排索引的结构为每条记录保存包 含此记录的所有数据块的索引值，并将其保存在哈希表中。哈希表中的 key 是每 条记录的唯一标识，对应的 value 是包含该记录的所有数据块的索引值。

如图4-1所示，在对b₁,b₂,b₃ 进行排序索引后，每个数据块对应一个唯一的 索引值，对于任意一条记录而言，都可以得到包含该记录的所有数据块的索引值。 如对于记录r₁ 而言，包含r₁ 的数据块分别是b₁和b₂, 对应的索引值分别是1和2, 因此可以在哈希表中创建一条记录r₁:<1,2 >保存该信息。

采用这种数据结构保存每条记录的倒排索引，可以快速地获得包含每条记录 的所有数据块信息，以便进一步识别冗余记录对。

**4.3.3** **冗余记录对识别**

在实体分辨过程中，记录比较算法对每个数据块中的记录进行两两比较，通常 不考虑处理数据块的先后顺序。为了识别冗余记录对，规定按照数据块索引值从 小到大的顺序逐个数据块进行处理，即首先处理包含记录(对)数目最多的数据 块，然后依次处理记录(对)数目较少的数据块。

在对两条记录进行比较之前，通过表4-1中的算法判断两条记录是否在之前 的数据块中已被处理。

表4- 1 IsRedundant 算法

|  |  |
| --- | --- |
| **Name:IsRedundant(**r₁,rz,curldx) | |
| **Input:**需要识别的两条记录r₁,₂,以及当前数据块的索引值curldx  **Output:**如果r₁,r₂是冗余记录对输出true,否则false | |
| 1.list1 =getldxList(r₁.ID);  2.list2=getldxList(r₂.ID);  3.ldx\_comm[]=getCommonldx(list1,list2); 4.If curldx =min(ldx\_comm)  5. **Return** false  6.Else  7. **Return** **true**  **8.End** **if** | //获取包含r₁ 的数据块索引列表  //获取包含r₂ 的数据块索引列表  //获得所有包含r₁和r₂ 的公共数据块  //如果当前数据块是索引值最小的公共数据块， //那么这两条记录是第一次进行处理  //否则这两条记录在其他数据块中已经被处理过 |

数据质量导论

表4- 1中的getldxList() 方法首先根据记录的唯一标识从哈希表中分别获得 包含两条记录的数据块列表(第1、2行),然后 getCommonldx()方法获得两个列表 中的所有公共部分，即同时包含两条记录的数据块(第3行),并判断当前数据块 是否是索引值最小的公共数据块(第4~7行)。由于实体分辨算法按照数据块索 引值由小到大的顺序，对相同数据块中的记录进行两两比较，如果当前数据块是索 引值最小的公共数据块，说明两条记录在之前没有进行过比较，否则则说明这两条 记录在之前的数据块中已经进行过比较。表4-1中的算法在判断两条记录是否 已经进行过比较时，只应用到了两条记录的数据块索引列表，而该列表在建立后不 会发生改变，该算法的复杂度为0(n),n 为记录对数目。

对于图4-1中的例子而言，按照b₁,b₂,b₃ 的顺序对每个数据块中的记录进行 两两比较，对于记录r₁ 和r₂ 而言，在处理数据块b₁时，IsRedundant 算法首先获得两 条记录的数据块列表，分别是r₁ :<1,2> 和 r₂ :<1,2 >,两者最小公共数据块 索引值为1,等于数据块b₁的索引值，说明r₁ 和 r₂ 是第一次进行比较，则正常 对这两条记录进行比较；而当处理数据块b₂ 时，虽然b₂ 同样包含记录r₁ 和 r₂, 但当前数据块b₂ 的索引值为2,不等于记录r₁ 和r₂ 的最小公共数据块索引值， 说明这两条记录已经在之前的数据块中进行过比较，从而不再对这两条记录 进行处理。

表4-1中的算法可以在实体分辨算法执行之前识别出所有冗余的记录对， 避免了对冗余记录的重复处理。这种方法要对所有的记录对进行判断，对于记 录数目为n 的一个单源数据块，需要n·(n-1)/2 次判断，而对于内部数据块 大小分别为m,n 的双源数据块，需要mn 次判断。基于冗余的数据分块通常会 产生大量的数据块，为了减少判断次数，进一步提高效率，通过空间映射对数据 块进行约减。

**4.3.4** **实验分析**

为了验证提出的方法，将提出的方法应用到基于冗余的数据分块算法中，统计 识别出的冗余记录对的数目，以及识别冗余记录对带来的计算代价。

[**4.3.4.1**](4.3.4.1) **实验设置**

1)数据集

本实验在两个真实数据集上进行验证， 一个是 Cora 数据集(Christen, 2012),用于验证单源数据分块的情况， 一个是 DBPedia 数据集(Liang et al., 2014),用于验证双源数据分块的情况。两个数据集的具体细节如表4-2和 表4- 3所列。

第4章实体分辨中的数据分块方法 

表4 - 2 Cora 数据集

|  |  |
| --- | --- |
| **Cora数据集** | |
| 文献数目 | 189 |
| 记录数目 | 1295 |
| 单源数据块数目 | 238 |
| 记录对数目 | 3570 |
| 冗余记录对数目 | 893 |
| 实体覆盖比 | 0.9530 |

表 4 - 3 DBpedia 数据集

|  |  |  |
| --- | --- | --- |
| **DBPedia数据集** | | |
| 2007 | 文章数目 | 396011 |
| 名-值对数目 | 5817838 |
| 2009 | 文章数目 | 721352 |
| 名-值对数目 | 12217795 |
| 双源数据块数目 | | 810262 |
| 记录对数目 | | 1.98×10° |
| 冗余记录对数目 | | 1.75×10° |
| 实体覆盖比 | | 0.9720 |

Cora数据集中包含1295条文献记录，但只对应着189篇现实文献，平均每篇 文献有接近7条存在各种差异的记录。选取每条记录的文献标题、发表时间、作者 姓名、期刊/会议名称等4个属性，将“作者姓名”和“期刊/会议名称”作为分块主 键，分别将“作者姓名”或“期刊/会议名称”的 Soundex 值相同的记录划分到相同 的数据块中，作为实验所需的单源数据分块方案。

DBPedia 数据集中包含了2007年和2009年从维基百科中抽取的每篇文章对 应的Inforbox中的名-值对，由于2007年到2009年之间维基百科 Inforbox 进行了 很大的调整，即使完全相同的文章也只有25%左右的名-值对是完全相同的，因 此分辨两个版本中描述相同文章的Inforbox 名-值对是一个严峻的挑战。在进行 数据分块时，采用基于属性不可知的数据分块方法(De,2011) 来产生实验所需的 双源数据分块方案。

2)评价指标

本节的目的是在不降低实体分辨效果的前提下，识别冗余的记录对，避免实体 分辨过程对相同记录对的重复比较，并通过数据块约减尽量减少数据分块方案产 生的记录对的数量，因此对以下指标进行计算：

 数据质量导论

(1)冗余记录对减少比RR:



(2)记录对减少比RO:



(3)实体覆盖比 RE:



<4.3.4.2> 冗余记录对识别验证

(4-1)

(4-2)

(4-3)

在两个数据集上，通过数据块排序索引和记录倒排索引，为每条记录保存包含 该记录的数据块列表，按照数据块索引递增顺序依次对每个数据块中的记录对进 行识别，计算冗余记录对减少比和识别后的实体覆盖比。结果如表4-4所列。

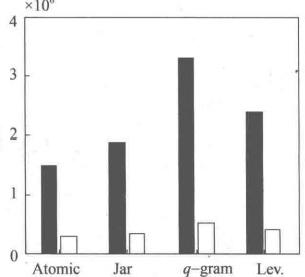
表4 - 4 冗余记录对识别结果

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **数据集** | **所有冗余**  **记录对数目** | **准确识别冗余**  **记录对数目** | **冗余记录对**  **减少比(RR)** | **识别后实体** **覆盖比(RE)** |
| Cora | 893 | 893 | 1 | 0.9530 |
| DBPedia | 1.75×10° | 1.75×10° | 1 | 0.9720 |

表4-4中的结果与前文中的分析相一致，IsRedundant 算法能够识别出数据 分块方案中的冗余记录对，且并不会降低实体覆盖比，因此能够在不降低实体分辨 效果的情况下，减少记录对比较次数。

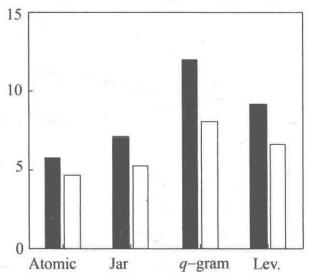
由于识别冗余记录对也要消耗计算资源，尤其是要对数据块和记录进行排序， 如果因识别冗余记录对消耗的时间大于对这些冗余记录对进行重复比较消耗的时 间，那么识别冗余记录对的工作是得不偿失和没有必要的，为此，本实验对比进行 冗余记录对识别和不进行冗余记录对识别情况下实体分辨消耗的时间。由于不同 的实体分辨方法的效率也不相同，为了实验更具有说服力，采用一种简单高效的方 法进行实体分辨，即计算属性平均相似度。分别利用四种常用的相似度算法 (Atomic String(Elmagarmid et al.,2007) 、Jaro(Snae,2007) 、q-gram(Bilenko et al., 2003;Gravano et al.,2003)和 Levenshtein距 离(Levenshtein,1966)) 计算各个属性 的相似度，然后用所有属性的平均相似度与指定阈值进行比较，判断两条记录是否 描述相同实体。这里并不关心实体分辨的准确性，只关注消耗的时间，因此阈值的 设定对于实验结果没有影响。实验结果如图4-2所示。

从图4-2中可以看出，无论是 Cora数据集还是 DBPedia 数据集，在采用冗余 记录对识别的情况下实体分辨消耗的时间都小于不进行冗余记录对识别直接进行

第4章实体分辨中的数据分块方法 

时 间 / s

时间/s



实体分辨方法 (a)Cora数据集

|  |
| --- |
| 不识别冗余记录对 识别冗余记录对 |

实体分辨方法

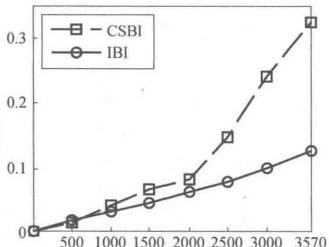
(b)DBPedia数据集

图4-2 冗余记录对识别对实体分辨消耗时间的影响

实体分辨消耗的时间，这说明通过避免对冗余记录对的重复处理所节省的时间要 大于冗余记录对识别所消耗的时间。由于实验中采用的是一种效率较高的实体分 辨方法，因此可以认为对于大多数实体分辨方法而言，利用基于倒排索引的冗余记 录对识别方法都可以提高实体分辨的效率。对比图4-2的两个子图可以发现，冗 余记录对识别在Cora数据集中节省的时间要远小于在 DBPedia 数据集中节省的 时间。这是由于Cora 数据集中存在的冗余记录对所占的比例远小于 DBPedia 数 据集中冗余记录对所占的比例。Cora数据集中冗余记录对所占比例约为1/4,而 DBPedia数据集中约为7/8。对于四种相似度算法而言，从图4-2中还可以看出， 相似度算法效率越高消耗的时间越少，冗余记录对识别的效果越不明显，相反，相 似度算法效率越低，冗余记录对识别所能节省的时间占整个实体分辨时间的比例 就越大。在实体分辨任务中，准确性较高的相似度算法通常伴随着较低的效率。 因此，利用冗余记录对识别提高实体分辨效率可以为使用准确性较高的实体分辨 方法提供保证。

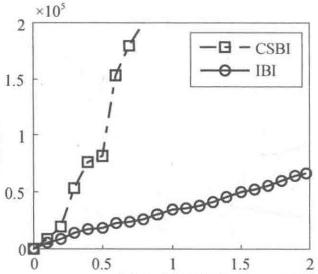
为了验证基于数据块倒排索引的识别算法(IBI) 的可扩展性，对基于倒排索引的 冗余记录对识别算法与保存记录对比较状态算法(CSBI) 进行比较。保存记录对比较 状态的算法在实体分辨算法之外保存所有记录对的比较状态，识别某一记录对是否已 被处理时，在保存的状态列表中进行查找。比较两种方法对两个数据集中的不同数量 的记录对进行识别(不进行实体分辨)消耗的时间，比较结果如图4-3所示。

如图4-3所示，无论在 Cora数据集还是 DBPedia 数据集上，基于倒排索引的 方法随着记录对数目的增长基本呈线性增长，而保存记录比较状态的CSBI算法则 随着记录对的增长呈幂函数的形式增长。这主要是由于CSBI算法每次识别记录

数据质量导论

时间/s

时间/s



|  |  |
| --- | --- |
| 记录对数目/对 | 记录对数目/对 ×10° |
| (a)Cora 数据集 | (b)DBPedia 数据集 |

图4-3 冗余记录对识别的可扩展性

对时都要到保存的记录对状态列表中查找，而记录对状态列表的大小随着识别记 录对数目的增加也在增加。基于倒排索引的方法只需要两条记录对应的数据块列 表，该列表在建立之后不会发生变化，而且数量远小于记录对数目。

**4.4** **基于空间映射的数据块约减**

为了减少需要识别的记录对数目，对三类数据块进行约减，分别是重复数据 块、完全子集数据块和重叠数据块。首先给出这三类数据块的具体定义如下：

**定义4-4(重复数据块)**:对于单源数据分块方案B 中的任一数据块b, 如果 在B 中存在一个数据块b,, 使得Vr₁∈b₄→r;∈b,且 Vr;∈b→r;∈b, 则 b₂和b₁为重 复数据块，记为b₄⇔b₁。对于双源数据分块方案B'中的任一数据块b{, 如果在B' 中存在一个数据块b},使得b₁→b} 且bi₂⇔b₂, 则b}和b}为重复数据块。

定义4-5(完全子集数据块):对于单源数据分块方案B 中的任一数据块b, 如果在B 中存在一个数据块b,, 使得lb₄I<Ib₂l 且 Vr;∈b₄→r;∈b, 则 b₄为b₂的完 全子集数据块，记为b₄Cb₁ 。对于双源数据分块方案B'中的任一数据块b', 如果 在B'中存在一个数据块b},使得 bCb′₁Ab{₂Cb₂, 或者b⇔b}Ab′₂Cb?₂, 或者 b₁₁Cb′₁Ab'₂→b₂, 则b{为b;的完全子集数据块。

**定义4-6(重叠数据块)**:对于单源数据分块方案B 中的任一数据块b, 如果在 B中存在一个数据块b,,b₁₂和b,既不是重复数据块也不是完全子集数据块，且b₄∩b₁≠〇, 则b 和b₁为重叠数据块，记为b₄ ×b。 对于双源数据分块方案B'中的任一数据块

b, 如果在B'中存在一个数据块b},使得 b;b′₁Ab}₂ ∩b²₂≠② 或ba∩b≠ A

b}₂×b₂, 则b{和b'为重叠数据块。

第 4 章 实体分辨中的数据分块方法 95

对于重复数据块和完全子集数据块可以直接将重复或被包含的数据块去除， 而对于重叠数据块则需要判断重叠的部分所占比例，对于重叠比例超出设定阈值 的数据块进行合并，从而减少记录对数目。判断数据块是否是重复数据块、完全子 集数据块或重叠数据块需要对两个数据块进行比对，为尽量减少对数据块进行的 详细比对，将数据块映射到欧氏空间，利用空间特性对三类数据块进行约减。

**4.4.1** **数据块映射**

数据块映射是通过为每条记录分配一个坐标值，从而将数据块映射到空间坐 标系中的某个区域上。将单源数据分块方案映射到一维空间，双源数据分块方案 映射到二维空间。为了尽可能使较大的数据块映射到连续的空间上，根据4.3.1 节中的排序结果，按照数据块索引值从小到大的顺序(即数据块从大到小),依次 为每个数据块中的每条记录分配坐标值。在同一个数据块中，按照每条记录出现 的频率(包含此记录的数据块的数目)由小到大的顺序分配连续坐标值，不同数据 块中的相同记录只分配一次坐标。表4-5中的伪代码描述了如何为单源数据分 块方案分配坐标的过程。

表4 - 5 BlockMapping算法

|  |  |
| --- | --- |
| **Name:BlockMapping(B)** | |
| **Input:**数据分块方案B  **Output:**所有记录及其分配的坐标C=|<ri,c>| | |
| 1.B\_ord=blockOrderBySizeDescending(B);  2.c=0;  3.C={|;  **4.For** **each** b,e B\_ord  5. R\_order =recorderOrderByFreqAscending(b); 6.**For** **each** re R\_order  7. **If**!C.has(r)  8. C=CU|<r,c>};  9. c++;  10. **End** **if**  11.End for  12.End for  13.Return C; | //对数据块进行排序  //当前坐标值  //对每个数据块中的记录排序 |

按照表4-5中的算法，首先 blcokOrderBySizeDescending()方法按照由大到小 的顺序对所有数据块进行排序(第1行),然后按照顺序，对于每个数据块中的记 录，recorderOrderByFreqAscending()方法按照每条记录出现频率由小到大的顺序

96))数据质量导论

进行排序(第5行),并依次分配坐标(第6~9行),这样每条记录都对应一个坐标 值，每个数据块都被映射到一维空间上。如图4-4中的单源数据分块方案b₁=|rj,

rz,'₃,T₄,⁷ s},b₂={r₁,r₂,'₃y'₆},b₃={r₃,'₇,'g},BlockMapping算法首先为包含记录最

多的数据块b₁中的记录分配坐标，然后是b₂、b₃,分配坐标结果为 C={<r₁,2>,

<r₂,3>,<r₃,4>,<r₄,0>,<rs,1>,<r₆,5>,<r₇,6>,<rg,7>}, 图 4 - 4 给 出了三个数据块在一维空间上所占的区域。

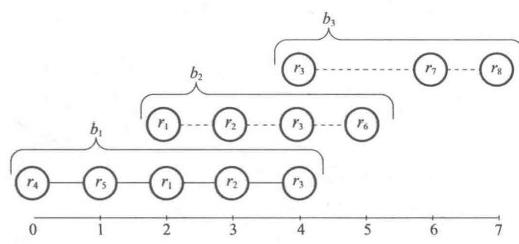


图4-4 单源数据分块方案空间映射

值得注意的是，并不是所有的数据块都能映射到一个连续区域上(如上例中 的b₃), 因此， 一个数据块映射后占有的区域的大小总是会等于或大于数据块中记 录的数目。两者之间的差别越小，则说明被映射到连续区域的数据块越多。 BlockMapping 按照数据块大小和记录出现频率进行分配坐标，目的是使有着大量 重叠记录的数据块更可能映射到两个连续的空间上，为下一步数据块约减提供便 利。对于双源数据分块方案进行数据块映射的方法与单源数据块类似，只需要独 立地对每个双源数据块中的内部分块应用 BlockMapping 算法，两个内部分块分别 映射到两个不同的维度上，因此，每个双源数据块都被映射到二维空间上的一个区 域上。由于 BlockMapping算法需要对数据块和记录进行排序，其时间复杂度为0 (IBl·log|Bl+|Dl·log|Dl)。

**4.4.2** **数据块约减**

将数据块映射到欧氏空间后，利用空间位置特性和集合运算对重复数据块、完 全子集数据块和重叠数据块等三类数据块进行约减，减少数据块和记录对的数量。

[**4.4.2.1**](4.4.2.1) **重复数据块约减**

对于两个重复的数据块，去除其中任何一个都不会对实体分辨结果造成影响。 在映射到欧氏空间后，通过以下四个步骤判断两个数据块是否重复：

(1)判断两个数据块的大小是否相同，如果相同进行步骤(2);

第4章实体分辨中的数据分块方法 

(2)判断两个数据块对应的空间区域是否重合，如果重合进行步骤(3);

(3)判断两个数据块中所有记录的坐标值之和是否相等，如果相等进行步骤 (4);

(4)对于满足条件(1)、(2)和(3)的数据块进行比对，判断两者的坐标是否完 全相同。

其中步骤(1)、(2)和(3)是为了减少需要比对的可能重复的数据块数目，对 于步骤(4)可以通过集合操作高效的完成。

<4.4.2.2> 完全子集数据块约减

对于两个完全子集数据块，由于较小数据块中的记录都包含在较大数据块中， 因此去除较小的数据块不会对实体分辨结果造成影响。通过以下三个步骤判断两 个数据块是否是完全子集数据块：

(1)判断两个数据块的大小是否相同，如果不相同进行步骤(2)。

(2)判断较小数据块对应的空间区域是否包含在另一个数据块中，如果被包 含进行步骤(3)。

(3)对满足条件(1)和(2)的数据块进行比对，判断其中较小的数据块中的记 录是否完全包含在较大数据块中。

<4.4.2.3> 重叠数据块合并

基于冗余的数据分块会产生大量的或多或少存在重叠的数据块，合并两个大 部分记录都重叠的数据块，能够有效减少记录对的数量。首先在欧氏空间上判断 两个数据块是否存在重叠，对于任意两个存在重叠的数据块，判断其重叠部分的大 小是否满足合并条件，并对满足条件的数据块进行合并。

合并两个数据块的目的是减少实际产生的记录对数量，如果两个数据块 合并后产生的记录对数目少于合并前两个数据块产生记录对数目之和，那么 就可以将这两个数据块进行合并。通过以下方法判断两个数据块是否可以 合并：

(1)对于单源数据块而言， 一个包含n 条记录的数据块能够产生n·(n-1)/2 个记录对。任意两个重叠数据块b, 和b, 其重叠部分记为b。,两个数据块合并前

产生的记录对总数 ,合并后产生的记录对数 目为- 合并数据块的目的是为了

减少产生记录对的数目，因此希望合并后的数据块能够满足式(4-4)中的条件：



(4-4)

数据质量导论

化简后得

Ib,l²-[2(1bI+1b₄l)-1]·Ib₀I+2Ib;I·1bI<0 (4-5)

基于此，对于任意两个存在重叠的单源数据块，首先计算两个数据块重叠部分 的记录数目，然后判断是否满足式(4-5)的约束，如果满足则合并两个数据块形 成新的数据块。

对于双源数据块而言，如果两个内部分块的大小分别为n 和 m,则该数据块能 够产生n×m 个记录对。任意两个重叠的双源数据块b’和bx,b’包含的内部分块 记为b 和b}₂,b{包含的内部分块记为 b 和 b₂, 其中b’₁和b 重叠的部分记为 b₀,b'₂ 和bk₂重叠的部分记为b₂, 合并时b, 和b/₁合并，b’₂和bx₂合并。两个数据 块合并前产生的记录对数目为1b₇l·IbI+|b}l·Ib}₂1, 合并后产生的记录对数 目为(1b}l+|bl-Ib1)·(1b'₂ l+Ib₂ l-1b₂ 1), 与单源数据块类似，希望合 并后数据块能够满足式(4-6)中的条件：

1bI·1b₂l+1bl·1b²₂¹>(1b{1+1bl-1b%)·(1bl+1bl-1b₂)

(4-6)

基于此，对于任意两个存在重叠的双源数据块，首先分别计算两个数据块的内 部分块重叠部分的记录数目，然后判断其是否满足式(4-6)的约束，如果满足则 分别合并两个数据块对应的内部分块形成新的数据块。

4.4.3 实验分析

<4.4.3.1> 数据块映射效果验证

数据块映射的目的是尽量将数据块内的记录映射到连续的空间上，为了验证 本章映射方法的有效性，将第4.4.1节中的 BlockMapping 算法与随机映射算法 RandomMapping 进行对比，统计映射到欧氏空间上的数据块所占的空间与数据块 原始大小之间的差别，差别越小说明映射为连续空间的数据块越多，效果越好。

为了计算方便，对于单源数据分块，数据块原始大小记为Ib,l-1,映射后所占的 空间为其对应线段长度。对于双源数据分块，数据块原始大小记为(1bl-1) ·

(lb₂ l-1), 映射后所占的空间为其对应区域的面积。为了准确起见，对随机映射 算法进行了10次实验，取所有结果的平均值，验证结果如表4-6所列。

表4 - 6 数据块映射结果

|  |  |  |
| --- | --- | --- |
| **算法** | **映射空间与数据块大小差别(之和)** | |
| **Cora** | **DBPedia** |
| BlockMapping | 2.13×10⁵ | 0.81×10⁸ |
| RandomMapping | 2.96×10⁵ | 1.27×10⁸ |

第4章实体分辨中的数据分块方法((**99**

如表4-6中的结果所列，BlockMapping算法在两个数据集上的效果都要优于 RandomMapping算法，其中在Cora 数据集上，BlockMapping算法的映射空间与数据 块大小差别比 RandomMapping 算法小超过28%,而在 DBPedia 数据集上则小超过 36%。这一方面是由于DBPedia数据集的数据分块方案中的包含的重叠的数据块 更多，且重叠部分也更大；另一方面是因为映射算法对双源数据分块方案的影响大 于对单源数据分块方案的影响。

[**4.4.3.2**](4.4.3.2) **数据块约减效果验证**

为验证数据块约减的效果，分别对两个数据集依次进行重复数据块约减、完全 子集数据块约减和重叠数据块合并，统计每个步骤完成后减少的数据块数目，以及 相应的记录对减少比RO和冗余记录对减少比 RR。验证结果如表4-7所列。

表4 - 7 数据块约减效果

|  |  |  |  |
| --- | --- | --- | --- |
| **数据集**  **约减步骤** | | Cora | **DBPedia** |
| 重复数据块约减 | 减少数据块数目 | 11 | 30406 |
| RO | 0.0030 | 0.0005 |
| RR | 0.0123 | 0.0006 |
| 完全子集数据块约减 | 减少数据块数目 | 24 | 131649 |
| RO | 0.0070 | 0.0353 |
| RR | 0.0279 | 0.0413 |
| 重叠数据块合并 | 减少数据块数目 | 8 | 16125 |
| RO | 0.0327 | 0.4141 |
| RR | 0.1758 | 0.5430 |
| 合计 | 减少数据块数目 | 43 | 178180 |
| RO | 0.0427 | 0.4494 |
| RR | 0.2160 | 0.5849 |

从表4-7中的结果可以看出，对于两个数据集，通过数据块约减均可有效减 少数据块数目以及产生的记录对和冗余记录对数目。三个步骤中，重复数据块约 减对于减少记录对和冗余记录对的效果最不明显(RO 和RR 的值最小)。这主要 是因为完全重复的数据块一般是较小的数据块，越大的数据块完全重复的可能性 就越小。因此，重复数据块约减的数据块数目虽然比较多，但对应的记录对和冗余 记录对却比较少。对于完全子集数据块约减也是如此，被其他数据块完全包含的 数据块一般都是较小的数据块，所以即使减少的数据块数目较多，但减少的记录对 和冗余记录对数量并不是很多。而对于重叠数据块合并而言，被合并的多是有着

100) 数据质量导论

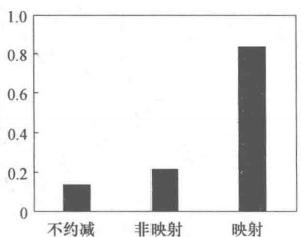
时 间 / s

时间/s

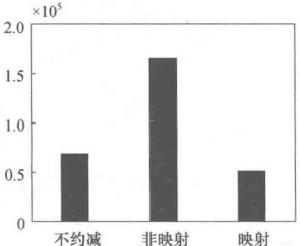
大量重叠记录的数据块，包含的记录数目较多，因此，减少的记录对和冗余记录对 的数目也就最多。通过对比在 Cora 和 DBPedia 数据集上的结果，可以发现，在 Cora 数据集上的约减效果明显低于DBPedia 数据集。这是因为两个数据集采用数 据分块方法不同，DBPedia 数据集采用基于属性不可知的数据分块方法，能够产生 更多的相互重叠的数据块和冗余记录对。

值得注意的是，对于重复数据块约减和完全子集数据块约减，去除的数据块产 生的记录对全部是冗余记录对，因此不会降低实体覆盖比，也不会降低后续的实体 分辨结果。而重叠数据块的合并在减少冗余记录对的同时也会引入原有数据分块 方案中没有的记录对，会对实体覆盖率产生影响。但由于减少的只是冗余记录对， 因此只可能增加而不会降低实体覆盖率。

虽然基于空间映射的数据块约减可以有效减少需要识别的记录对数量，但由 于其也需要消耗时间，如果消耗的时间大于其减少的冗余记录对识别时间，那么数 据块约减将是得不偿失的。为此，本实验对比在不利用数据块约减的情况下和在 利用基于空间的数据块约减的情况下识别冗余记录对消耗的总时间(数据块约减 时间+冗余记录对识别时间)。另外，为了验证空间映射能够提高数据块约减效 率，也将不采用空间映射的数据块约减方法(直接通过集合比较)加入对比。对比 结果如图4-5所示。



(a)Cora 数据集



(b)DBPedia 数据集

图4-5 基于空间映射的数据块约减的有效性

图4-5中，“不约减”表示的是不进行数据块约减直接进行冗余记录对识别 消耗的时间，“非映射”表示直接采用集合比较进行数据块约减，然后进行冗余记 录对识别消耗的时间，“映射”表示采用基于空间映射的数据块约减方法，然后进 行冗余记录对识别消耗的时间。从图4-5中可以看出，在两个数据集上，基 于空间映射的数据块约减方法得到了几乎相反的结果。图4-5(a) 中的结果 显示，对于Cora数据集上的数据分块方案，虽然数据块约减可以减少需要识 别的记录对数目，但消耗的时间反而增加，几乎是不进行数据块约减(“不约

第4章 实体分辨中的数据分块方法 (101

减”)情况下的6倍。而图4-5(b) 中的结果则显示，对于 DBPedia数据集上 的数据分块方案，采用数据块约减方法大约可以减少23%的识别时间。将 图4-5与表4-4和表4-7进行对比，可以发现，Cora 数据集中的数据块和 记录对数目都比较少，数据块约减方法能够减少的记录对数目有限 (4.27%),因此，基于空间映射的数据块约减减少的记录识别时间远不如其 自身消耗的时间，造成图4-5(a) 中数据块约减反而增加了冗余记录对识别 的总体时间的结果。而 DBPedia 数据集包含了大量的数据块和冗余记录，数 据块约减方法可以大量减少需要识别的记录对数目(44.94%),虽然其本身 也消耗了一定的计算资源，但仍能减少23%的时间消耗。

对于直接采用集合比较进行数据块约减的方法(非映射),在 Cora 数据集上的 效果要好于基于空间映射的数据块约减方法，但在 DBPedia 数据集上则相反。但 与不进行数据块约减的方法进行比较，可以发现，直接采用集合比较进行数据块约 减的方法在两个数据集上的效果都不如不进行数据块约减的方法，也就是说对于 两个数据集而言，与其采用集合比较的方法进行数据块约减，还不如不进行数据块 约减。

从本实验得到的结果可以得出，基于空间映射的数据块约减具有一定的局限 性，对于数据块和冗余记录对数目较少的数据分块方案，基于空间映射的数据块约 减不一定能够减少时间消耗，甚至会增加时间消耗，而对于包含大量数据块和冗余 记录对的数据分块方案可以有效减少记录对识别时间，提高效率。

**4.5** **基于Canopy 聚类的数据分块**

相比关系型数据而言，XML数据对于重复问题更敏感，因为文本信息和结 构信息都可能引起重复。XML数据的实体分辨得到了较为广泛的研究，Puhl- mann 等将关系型数据中的排序邻居方法扩展到XML数据，提出了排序XML邻 居算法 SXNM(Puhlmann et al.,2006) 。SXNM假设XML数据都有相同的模式， 并运用一个有关候选对、对象描述、分块键定义和分块键模式的配置文件为SX- NM生成分块键，排序后，用自底向上的方法比较两个 XML元素。该方法效率 较高。

Leitao 等提出了一种自动分块方法，它包括两个策略： 一个是成组(List - Wise)优化，类似于关系型数据实体分辨的分块，目的是聚合可能重复的记录； 一 个是成对优化，它关注成对对象比较，即分辨方法本身(Leitao et al.,2013a)。

本节关注成组优化，与目前的关注结构比较的分辨方法不同，由于文本信息相 比结构信息能提供更强的分辨能力(Ribeiro et al.,2009),因此，假设相似 XML对 象一定具有较高的文本相似度。另外，相比结构比较，文本相似度的计算效率更

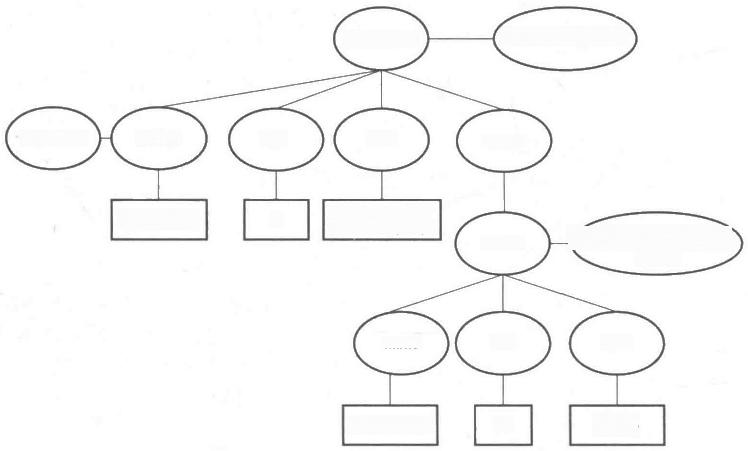
)数据质量导论

高。因此，采用文本信息生成分块键(Blocking Key,BK),当数据规模较大时，需要 采用高效和可扩展的分块以将可能重复的对象聚合到一个较小的块，实现了一个 MapReduce 版本的 Canopy 聚类(McCallum,2000)以聚合分块键。

在成组优化中，Leitao 等运用 Canopy 聚类算法对分块键进行聚类，他们将从 分块键的特征向量到对应参数的映射函数看作一个回归问题，并利用支持向量回 归进行求解。然而，由于XML结构非常灵活， 一些属性可能不完整，甚至缺失，导 致分块键不完整甚至缺失，影响分块效果。

**4.5.1** **整体流程**

一个典型的XML对象如图4-6所示，该对象抽取自Cora数据集(Leitao et al.,2013b) 。 参照文献(Puhlmann et al.,2006),图4-6中，椭圆、矩形和虚线椭圆 分别表示 XML元素、文本节点以及属性。

publication ID="ahlskog1994a

ID="199" author uid title venue

M.Ahlskog

0

0.Inganas and

M.R.Andersson

venue (publD="ahlskog₁994a"

id="1"

name vol date

JAppl.Phys 76 (1994)

图 4 - 6 典型的XML对 象

如图4-6所示，分块的步骤：首先，为每个元素指定一个唯一的ID; 其次，抽 取所有的文本节点，再用 Canopy聚类聚合可能重复的对象；最后，利用Xpath 语句 从各簇中抽取对应的元素，并生成候选对象对。

**4.5.2** **ID 指定**

和关系型数据不同，XML对象通常并不包含一个全局唯一的主键，使得难以

第4章实体分辨中的数据分块方法(103)

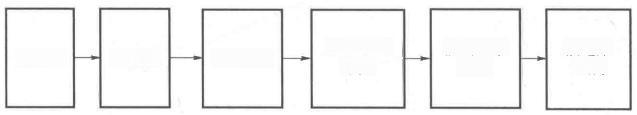
唯一定位某个元素。各对象的属性或文本信息都不能唯一定位某个元素，因为由 于结构的存在，两个文本信息完全相同的元素也可能不相同。因此，为每个元素指 定 一 个如图4 - 6所示的唯 一 的ID。

**4.5.3** **BK** **生成**

Leitao 等利用模拟退火算法选择部分文本节点以生成分块键，由于XML 结 构 的灵活性，使得生成的分块键不完整，甚至缺失。另外，文本节点的选择是一个组 合优化问题，时间复杂性较高。

另一方面，Canopy 聚类中常用的Jaccard 相似度效率较高，因此，直接抽取所 有文本节点，将其组合，生成分块键。

分块键的形式如下：ID, 分隔符以及文本节点的组合。例如，图4-7中的对象 生成的对应分块键如下：0:M.Ahlskog;0.Inganas and M.R.Andersson;J Appl. Phys.;76;(1994) 其中，0是ID,“:” 是分隔符，M.Ahlskog;0.Inganas and M.R. Andersson;JAppl.Phys.;76;(1994)是所有文本节点的组合，并用“;”分隔。

数据源 ID指派 BK 生成

候选集 生成

CANOPY 聚类

候选对象 获取

图 4 - 7 流程图

**4.5.4** **Canopy 聚** **类**

自 McCallum 提出以来，Canopy 算法广泛应用于各种聚类。Cohen 将 Canopy 算法应用于关系型数据的分块(Cohen,2002),Leitao 将 Canopy 应用于 XML 的数 据分块(Ribeiro et al.,2009),说明了利用 Canopy 聚类进行分块的有效性。

Mahout 项目组已经实现了一个 MapReduce 版本的 Canopy,但是由于 Mahout 定义了一个向量以表示聚类中的一个对象，而该向量仅支持数值型数据，使得现有 的 Canopy 聚类算法仅对数值型数据有效。对文本数据，通常的方法是将它转换为 词频向量，并对词频向量进行聚类，该方法显然不适用于本场景。

参考该实现的过程，将现有的基于 Mahout 框架的聚类扩展为仅基于 MapRe- duce 框架，以支持字符型聚类，并更适用于分块。

实现过程中，每一个 Canopy类都包括两个成员变量， 一个是簇中心，对应簇中 心的XML元素的文本节点，另一个是簇的成员，对应簇中所有成员的ID。

Mapper: 输入数据为一个存储在HDFS 上的文本文件， 一行为一个分块键，用

)数据质量导论

整个分块键作为Mapper 的 value, 并将key 置为空。文本文件被分到各Mapper上， 对每个map 任务，首先构建一个称作 canopies 的 Canopy 簇的全局集合。给定 BK 和两个阈值，mapper用方法 Addpointtocanopies 增量地将每个分块键加入各个 Can- opy。Addpointtocanopies 算法见表4-8。

表4 - 8 Addpointtocanopies 算法

|  |
| --- |
| **Name:Addpointtocanopies**(value,canopies) |
| **Input:**value,每一个value为一个BK,canopies是Canopy簇的集合 |
| 1.将value分为ID和文本；  2.For each canopy  3.If distance(center,text)<toose//text是各元素的所有文本节点的组合，distance是一个距离函数 4.将ID加入Canopy类中；  **5.End** **if**  6.Flag=distance(center,text)>trigh  7.End for  8.If Flag  9.将一个新的类canopy(text,ID)加入到canopies中  10.End if |

显然，Addpointtocanopies 是增量实现，且对各分块键，是单遍实现。各对象仅 需被执行一次，在大多数情况下，仅需要存储各对象的ID, 在少数情况下，也只需 要存储ID 和文本信息。存储和计算开销都很小。

重写mapper 方法如表4-9所列。将所有对象的 key 都设为相同，value 是 Canopy 中心的组合。

表4 - 9 Map 算法

|  |
| --- |
| Name:Map(values,canopies) |
| **Input:**每个value是一个生成的key,canopies是Canopy簇的集合  **Output:**context |
| 1.Addpointtocanopies(value,canopies);  2.For each canopy  3.组合类的成员、分隔符以及类中心以生成一个字符串mc;  4.将key和mc写入到context中； //所有的key都一样；  5.End for |

在 Mahout 的 Canopy 实现中，簇中心是各簇成员的均值，与原对象格式一致。 此处修改输出格式，也是为了使得输出的格式与原始对象一致。这样 Addpoint- tocanopies 也能用于 Reduce 中。

第4章 实体分辨中的数据分块方法(

Reducer:在 Reducer 阶段，簇中心相似度小于阈值的簇被进一步合并，并输出 所有簇的成员(表4-10)。

表4-10 Reduce 算法

|  |
| --- |
| **Name:Reduce**(value,canopies) |
| **Input:**每个value都是Canopy成员、分隔符以及簇中心的组合，canopies是所有Canopy簇的集合 **Output:**context |
| 1.Addpointtocanopies(value,canopies);  2.For each canopy  3.If size>2 //size是簇中的成员个数  4.将每个canopy中的成员写入context;  5.End if  6.End for |

**4.5.5** **候选对象获取**

由于每一个XML元素都要一个唯一的ID, 并且聚类的结果是 XML 元素的 ID,因此，读取每一行，将它划分成 ID 的组合，并应用XPath 获取各元素。针对 图4-6中的元素，所用的Xpath 是//publication[ulD=0]。

**4.5.6** **复杂性分析**

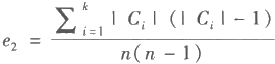
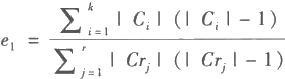
ID指定、分块键生成和候选对获取的时间复杂性均为线性，Canopy 分块 的复杂性已在(Christen,2011) 中得到分析，它生成的记录对在[(n/2) · (n/b-1),(nv/2)·(nv/b-1)] 区间之内，其中n 是对象的数目，b 是生成 的簇的数目。当各元素仅加入1个簇时，生成的候选对个数最小；当各元素 加入v 个簇时，生成的候选对个数最大。MapReduce 实现并不改变复杂性， 只是使其并行化。

**4.5.7** **实验分析**

[**4.5.7.1**](4.5.7.1) **评估标准**

采用(Hassanzadeh et al.,2009)提出的查准率(Precision)、查全率(Recall) 和 F1 等指标，令G={G₁,G₂,…,G₁} 表示数据R 的真实匹配结果，C={C₁,C₂,…,

C,}表示方法分辨出的匹配结果，每个簇G;或 C, 都包括一组匹配实体，f 是一个从

106)数据质量导论

G到C的映射函数，使得每一个簇G;都被映射到一个簇C;=f(G₁), 且满足G;和

C, 之间有最多的公共元素，查准率Pr、查全率Re 和 (4-1 1)所示，











F1 的定义如式(4-7)至式

(4-7) (4-8) (4-9) (4-10)

(4-11)

该标准在有真实值存在的情况下更适合于评估聚类，因为它不仅评估了每个 簇的查准率、查全率，还评估了加权查准率、查全率。

由于查准率、查全率关注准确性，定义两个度量指标来度量分块的有效性：

(4-12)

(4-13)

e₁ 评价分块生成的实体对与真匹配实体对之比，e₂ 评价分块生成的实体对与 所有实体对之比。

<4.5.7.2> 数据集以及实验配置

采用 Cora(Leitao et al.,2013b,CDDB10000 以及 CDDB1000 数据集作为测试 数 据 。Cora是一个参考文献集合(Cohen,2002), 它有1878个元素，具有相同出版 ID 的对象为重复对象。CDDB是一个抽取自FreeDB 的 CD 数据集合，CDDB1000 有500个干净的元素以及对应的人造的重复元素，CDDB10000 有10000条包含重 复对象的记录，我们人工找出重复对象。

由于Canopy 仅引入两个参数，不需要复杂的参数设置，为进一步简化考虑，将 t.…设置为与t 相同，并将它从0.3～0.6按0.05 的步长进行变化。

将本方法与 SXNM(Puhlmann et al.,2006)进行对比，滑动窗口的大小固定为 31和101,即每个簇包括31或101个元素，即1个元素以及它的后30或100个邻 居。SXNM生成的候选对的数目为(n-k/2)·(k-1), 其中，n 为对象个数，k 为 窗口大小。

第4章实体分辨中的数据分块方法(**107**

实验环境包括3个节点，每个节点配置为2.5Ghz Xeon CPU,64G 内存，Cen- tOS10.0 操作系统。

[**4.5.7.3**](4.5.7.3) **实验结果和分析**

Cora 的分块结果如表4-11所列。

表4-11 Cora 的分块结果

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **方法** | **参数** | **查准率** | **查全率** | **F1** | **e₁** | **e2** |
| Canopy | 0.3 | 0.9484 | 0.7127 | 0.8139 | 0.8026 | 0.0333 |
| 0.35 | 0.9414 | 0.7799 | 0.8531 | 0.9679 | 0.0401 |
| 0.4 | 0.9240 | 0.8052 | 0.8605 | 0.9646 | 0.0400 |
| 0.45 | 0.9066 | 0.8601 | 0.8827 | 1.0064 | 0.0417 |
| 0.5 | 0.9021 | 0.8696 | 0.8855 | 1.1055 | 0.0458 |
| 0.55 | 0.8636 | 0.9191 | 0.8905 | 1.2049 | 0.0499 |
| 0.6 | 0.8291 | 0.9433 | 0.8791 | 1.3588 | 0.0563 |
| SXNM | 30 | 0.5900 | 0.4540 | 0.5132 | 0.7405 | 0.0307 |
| 100 | 0.4049 | 0.6219 | 0.4905 | 2.4780 | 0.1027 |

CDDB10000 的分块结果如表4-12所列。

表4- 12 CDDB10000 的分块结果

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 方法 | 参数 | 查准率 | 查全率 | F1 | e1 | e₂ |
| Canopy | 0.3 | 0.4791 | 0.4312 | 0.4539 | 0.2075 | 1.6788×10-6 |
| 0.35 | 0.5607 | 0.5229 | 0.5411 | 0.2672 | 2.1614×10-6 |
| 0.4 | 0.6075 | 0.5749 | 0.5908 | 0.3320 | 2.6861×10~6 |
| 0.45 | 0.6616 | 0.6269 | 0.6438 | 0.4955 | 4.0081×10-6 |
| 0.5 | 0.7487 | 0.7095 | 0.7286 | 0.7160 | 5.7918×10~6 |
| 0.55 | 0.7806 | 0.7615 | 0.7709 | 2.5110 | 2.0313×10- |
| 0.6 | 0.9400 | 0.9755 | 0.9574 | 33.0117 | 2.6705×10-4 |
| SXNM | 30 | 0.0401 | 0.5657 | 0.0748 | 734.1383 | 0.0059 |
| 100 | 0.0127 | 0.5780 | 0.0249 | 2487.9 | 0.0202 |

CDDB1000 的分块结果如表4-13所列。

108)数据质量导论

表4- 13 CDDB1000 的分块结果

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **方法** | **参数** | **查准率** | **查全率** | **F1** | **e₁** | **e₂** |
| Canopy | 0.3 | 0.7160 | 0.7160 | 0.7160 | 0.3580 | 7.1672×10-4 |
| 0.35 | 0.8420 | 0.8430 | 0.8425 | 0.4230 | 8.4685×10-4 |
| 0.4 | 0.8780 | 0.8800 | 0.8790 | 0.4430 | 8.8689×10-4 |
| 0.45 | 0.9140 | 0.9160 | 0.9150 | 0.4610 | 9.2292×10-4 |
| 0.5 | 0.9320 | 0.9350 | 0.9335 | 0.4730 | 9.4695×10-4 |
| 0.55 | 0.9540 | 0.9590 | 0.9565 | 0.4890 | 9.7898x10- |
| 0.6 | 0.9590 | 0.9660 | 0.9625 | 0.4970 | 9.9499×10-4 |
| SXNM | 30 | 0.0594 | 0.9210 | 0.1116 | 28.6118 | 0.0573 |
| 100 | 0.0185 | 0.9340 | 0.0363 | 94.1145 | 0.1884 |

从上述三个表可知，随着 增加，查全率增加，这是因为随着t. 增加，距离 小于t 的元素将更多，加入簇中的元素也更多，从而得到更多的候选对象对。

e₁和e₂ 度量和查全率的趋势相同，原因也一样。由e₁ 和e₂ 可知，分块后，所需 比较的元素对数目显著下降。查全率的提高伴随着更多的候选对，将导致分块效 果变差。

三个表中的查准率的趋势不完全相同，这是因为在 Cora 中，选出的候选对的 数目接近甚至大于真实值，然而在CDDB数据集中，选出的真实值小于重复值，由 Hassanzadeh 等(2009)的定义可知，更小的t 将导致每簇的查准率提高，但是由 于生成的候选对的数目较小，导致总的加权查准率降低。

由于 SXNM 的滑动窗口大小固定为31和101,查全率可能因为一个小窗口而 降低，然而， 一个大的窗口也将导致e₁ 和 e₂ 指示的分块效果的下降。在 CDDB 上 的查准率较低，是因为分母是窗口大小，当真实簇只有几个成员而窗口大小较大 时，结果较差。

SXNM只需要一次排序，因此效率较高，而 Canopy的复杂性为0(nb) 。 然而， 滑动窗口后生成的候选对数目较大，而 Canopy 分块后生成的候选对数目较小。因 此 Canopy分块效果更好。总体来说，SXNM 效率更高，而 Canopy 分块效果更好。

**4.6** **本章小结**

本章针对数据分块中存在的对实体分辨效果没有任何提高而消耗大量计算资 源的冗余记录对问题，提出了一种基于倒排索引和空间映射的冗余记录对识别方 法。基于空间映射的数据块约减有效减少了参与实体分辨的数据块和记录对数 目，且不会因此降低实体分辨的效果。针对XML 的数据分块，提出基于Canopy聚

第4章 实体分辨中的数据分块方法 

类的数据分块，设计实现了MapReduce 版本的 Canopy 聚类算法。

在实体分辨任务中，数据分块是保证实体分辨效率的主要步骤，后续的记录比 较和决策模型步骤为了提高实体分辨的准确性，通常具有较高的计算复杂度，在数 据分块步骤中节省的计算资源越多，后续步骤就有越多的余地使用更精确和复杂 的算法提高实体分辨的准确性。因此，本章的工作为后续的记录比较和匹配决策 步骤提供可靠的效率保证。

**参考文献**

[1]Bilenko M,Mooney R,Cohen W,et al.2003.Adaptive Name Matching in Information Integration[J].Intelligent Systems.IEEE,18(5):16-23.

[2]Christen P.2011.A Survey of Indexing Techniques for Scalable Record Linkage and Deduplication[J].IEEE Transactions on Knowledge and Data Engineering,24(9):1537 -1555.

[3]Christen P.2012.Data Matching:Concepts and Techniques for Record Linkage,Entity Resolution,and Duplicate Detection[M].New York:Springer Science &Business Media.

[4]Cohen WW,Richman J.2002.Learming to Match and Cluster Large High Dimensional Data Sets for Data Inte- gration[C]//SICKDD.Edmonton,Alberta,Canada;Knowledge Discovery &Data Mining:475-480.

[5]De Vries T,Ke H,Chawla S,et al.2011.Robust Record Linkage Blocking Using Suffix Arays and Bloom Filters

[J].ACM Transactions on Knowledge Discovery from Data(TKDD),5(2):9.

[6]Draisbach U,Naumann F,Szott S,et al.2012.Adaptive Windows for Duplicate Detection[C]//IEEE 28th Inter- national Conference on Data Engineering(ICDE).Washington,DC,USA:IEEE:1073-1083.

[7]Elmagarmid A K,Ipeirotis P G,Verykios V S.2007.Duplicate Record Detection:A Survey[J].IEEE Transac-

tions on Knowledge and Data Engineering,19(1):1 -16.

[8]Gravano L,Ipeirotis P G,Koudas N,et al.2003.Text Joins in an RDBMS for Web Data Integration[C]//Pro- ceedings of the 12th International Conference on World Wide Web,Budapest,Hangary:ACM:90-101.

[9]Hassanzadeh O,Chiang F,Lee H C,et al.2009.Framework for Evaluating Clustering Algorithms in Duplicate Detection[C]//The VLDB,Lyon.France:VLDB,2(1):1282-1293.

[10]KenigB,GalA.2013.Mfiblocks:An Effective Blocking Algorithm for Entity Resolution [J].InformationSys- tems,38(6):908-926.

[11]Leitao L,Calado P.2013a.An Automatic Blocking Strategy for XML Duplicate Detection[J].Applied Compu- ting Review,13(2):42-53.

[12]Leitao L,Calado P,Herschel M.2013b.Efficient and Effective Duplication Detection in Hierarchical Data[J]. IEEE Transactions on Knowledge and Date Engineering,25(5):1028-1041.

[13]Levenshtein V.1966.Binary Codes Capable of Correcting Deletions,Insertion,and Reversals[J].Soviet physics Doklady,10(8):707-710.

[14]Liang H,Wang Y,Christen P,et al.2014.Noise-tolerant Approximate Blocking for Dynamic Real -time Entity Resolution[M].Tainan:Advances in Knowledge Discovery and Data Mining Springer International Publishing: 449-460.

[15]MeCallum A,Nigam K,Ungar L H.2000.Efficient Clustering of High -Dimension Data Sets with Application to Reference Matching[C]//ACM Knowledge Discovery and Data Mining(KDD).Boston,USA:ACM 2000.

110))数据质量导论

[16]Papadakis G,loannou E,Niederee C,et al.2011.Efficient Entity Resolution for Large Heterogeneous Informa- tion Spaces[C]//Proceedings of the 4th ACM Intermational Conference on Web Search and Data Mining Hong Kong,China:ACM:535-544.

[17]Papadakis G,loannou E,Niederee C,et al.2012.Beyond 100 Million Entities:Large -scale Blocking -based Resolution for Heterogeneous Data[C]//Proceedings of the 15th ACM International Conference on Web Search and Data Mining.Seattle,USA:ACM:53 -62.

[18]Papadakis G,oannou E,Palpanas T,et al.2013.A Blocking Framework for Entity Resolution in Highly Heteroge- neous Information Spaces[J].Knowledge and Data Engineering,IEEE Transactions on,25(12):2665-2682.

[19]Puhlmann S,Weis M,Naumann F.2006.XML Duplicate Detection Using Sorted Neighborhoods[C]//Confer- ence on Extending Database Technology,Munich,Germany:Springer,Lecture Notes in Computer Science 3896.

[20]Ramadan B,Christen P.2014.Forest -based Dynamic Sorted Neighborhood Indexing for Real -time Entiy Resolution[C]//Proceedings of the 23th ACM International Conference on Conference on Information and Knowledge Management Shanghai,China;ACM:1787-1790.

[21]Ribeiro L A,Harder T,Pimenta F S.2009.A Cluster -Based Approach to XML Similarity Joins[C]//Interna- tional Database Engineering and Application Symposium.Cetraro,Calabria,Italy:ACM:182-193.

[22]Shu L,Lin C,Meng W,et al.2012.A Framework for Entity Resolution with Efficient Blocking[C]//IEEE 13th

International Conference on Information Reuse and Integration(IRI).Las Vegas,USA:IEEE:431-440.

[23]SnaeC.2007.A comparison and Analysis of Name Matching Algorithms[J].International Jourmal of Appplied

Science,Engineering and Technology,4(1):252-257.

[24]Weis M,Naumann F.2005.DogmatiX Tracks down Duplicates in XML[C]//ACM's Special Interest Group on Management Of Data(SIGMOD).Baltimore,USA;ACM:431-442.

[25]Whang S E,Menestrina D,Koutrika G,et al.2009.Entity Resolution with Iterative Blocking [C]//Proceedings of the 35th SIGMOD International Conference on Management of Data.Rhode Island,USA:ACM:219-232.

[26]Winkler W E,Yancey W E,Porter E H.2010.Fast Record Linkage of Very Large Files in Support of Decennial and Administrative Records Projects[C]//Proceedings of the Section on Survey Research Methods,American Statistical Association.Boston,USA:American Statistical Association JSM:2120 -2130.

**第** **5** **章** **实体分辨中的相似度计算方法**

**5.1** **引言**

属性相似度的计算是保证各种实体分辨方法准确性的基础，无论是基于特征 的、基于上下文的还是基于关系的实体分辨方法，属性相似度计算的准确与否直接 影响着实体分辨结果的准确性。在实际应用中，字符串型属性占据了整个数据集 的绝大部分比例，因此，如何准确计算字符串的相似度一直是实体分辨领域研究的 热点之一(Christen,2012;Koudas et al.,2006;Cohen et al.,2003;Snae,2007;Bacht- eler et al.,2013;Gomaa et al.,2013)。编辑距离是字符串相似度计算中最为常用 的方法，但由于中文与西文在存储、录入等方面的区别，针对西文提出的编辑距离 在计算包含中文的字符串时可能会导致各种误差的产生。大多数属性相似度算法 通常只利用单个属性的属性值信息计算相似度，每种属性相似度算法一般都有一 定的适用范围，如果不能确定选择的属性相似度算法是否最优，那么得到的属性相 似度值就可能存在一定的误差。而且参与实体分辨的数据往往包含多个属性，这 些属性并不一定相互独立，而利用单个属性值信息计算属性相似度的方法默认记 录的各个属性相互独立，忽视了属性间存在的关联和依赖关系(Fellegi et al., 1969;Winkler et al.,1991) 。Bhattacharya 等(2007)、Kalashnikov 等(2005)、Chen

等(2005)、Chen 等(2007)、Nuray-Turan 等(2013)和 Bhattacharya 等(2005)虽然 考虑了数据中存在的关系，但这种关系是存在于实体之间的关系，而对于单个实体 的多个属性之间的依赖关系并没有考虑。

本章针对中西文混合字符串，提出一种融合多种编辑距离的字符串相似度计 算方法。另外为了提高属性相似度的准确性，本章通过分析属性相似度与函数依 赖的关系，进而研究了基于函数依赖的属性相似度调整。

**5.2** **基于多编辑距离融合的相似度计算**

本节在讨论属性值相似特征的基础上，研究中西文字符串的编辑距离及其融 合方法，并对方法进行验证。

112)数据质量导论

5.2.1 相似特征定义及其标准化

这里对三种常见的属性类型定义其相似特征，经验表明，在分类问题中，特征 标准化利于节省训练时间，提高训练成功率，提升分类效率和查准率(曹建军， 2008;Hong et al.,2005),因此相似特征定义中考虑相似特征值的标准化。

1)字符串型

字符串型属性值的相似特征值为

*S=D(α,β)=D(β,α)* (5-1) 式(5-1)中，D(α,β)为字符串α和字符串β的编辑距离(Edit Distance 或 Leven- shtein Distance)(见5.2.2节)(Levenshtein,1966)。记 l,lg 为α,β的字符数，根据 编辑距离的定义，S 的取值满足式(5-2):

Se[0,max{l₀,l₈}] (5-2)

根据式(5-2),用max{l₄,lg} 对式(5-1)进行归一化：

 (5-3)

注意，此处更改操作将中文字符和西文字符视为同等字符基本单位，原因是中 文的更改操作不可能以半字节进行，更符合实际。

2)枚举型

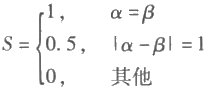
枚举型属性如性别、省份、学历、职称等，假定枚举型属性值用连续非负整数表 示。将枚举型属性分为两类， 一类是独立属性，即属性值之间是独立的，如性别、省 份；另一类是关联属性，即相邻属性值之间具有一定相似性，如学历、职称，此种情 况，可将属性值映射为有序整数表示。

对第一类枚举型属性值，由式(5-4)计算相似特征值：

(5-4)

对第二类枚举型属性值，为了描述相邻属性之间的相似性，由式(5-5)计算

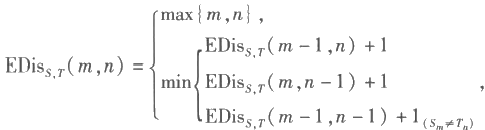
相似特征值：

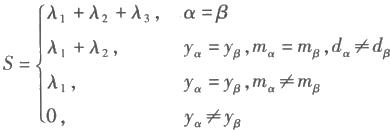


(5-5)

3)日期型

典型的日期型属性值一般由年、月、日三部分组成，分别记日期型属性α,β的 年、月、日三部分属性值为y,m 和d,则定义其相似特征为

第5章实体分辨中的相似度计算方法 

 (5-6)

式(5-6)中：

λ₁+λ₂+λ₃=1 且λ₁>λ₂>λ₃ (5-7)

式(5-6)和式(5-7)表明，对日期型属性值，年、月、日对相似特征的影响依 次减弱，并且后者影响以前者为条件。

5.2.2 编 辑 距 离

编辑距离是由Levenshtein(1966) 提出的用于测量两个字符串之间差别的量 化测度，因此也称作 Levenshtein 距离。对于两个字符串而言，它指的是从一个字 符串转换为另一个字符串所需的单字符编辑操作的数目，这些编辑操作包括插入 一个字符、删除一个字符和替换一个字符。

对于长度分别为m 和 n 的两个字符串S 和 T, 字符串S 与 T 之间的编辑距离 可以用递归的方式给出：

min{m,n}=0

其他

(5-8)

式(5-8)中，EDiss,(i,j),O≤i≤m,O≤j≤n,表示S 的前i个字符与T 的 前j个字 符之间的编辑距离；S;和T;分别表示字符串S 的第i个字符和 T 的第j 个字符； 1(s≠r))为指示函数，即

(5-9)

由于递归方法在计算过程中会重复计算相同子字符串的编辑距离，效率 较低，目前常用的方法是Wagner 等(1974)提出的利用动态规划方法计算 编辑距离的算法。表5- 1中为采用动态规划方法计算编辑距离算法的伪 代码。

114)数据质量导论

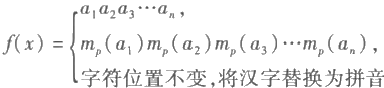
表 5 - 1 编辑距离计算的动态规划算法

|  |  |
| --- | --- |
| Name:EDis\_DP(S,T) | |
| Input:字符串S和T  Output:字符串S和T间的编辑距离 | |
| 1.d=int[m,n]  2.Fori=1to m  3.d[i,0]=  4.End for  5.Forj=1 ton  6. d[0,j]=j  7.End for  8.For j=1 to n  9.Fori=1 to m  10.If S[i]==T[j  11. d[i,j]=d[i-1,j-1]  12.Else  13. d[i,j]:=min(d[i-1,j]+1, d[i,j-1]+1,  d[i-1,j-1]+1  14. End if  15. End for  16.End for  17.Return d[m,n] | //字符相同，无操作  //删除字符  //插入字符  //替换字符 |

虽然采用动态规划方法需要保存各个子字符串的编辑距离，需要的计算内存 约为递归方法的mxn 倍，但这种方法时间效率明显高于递归方法，而且现代计算 机内存的大小完全能够满足该方法的需求。

**5.2.3** **中西文混合字符串的编辑距离**

编辑距离最初是针对西文字符提出的距离度量算法，其最小的计算单位是字 符，由于中文字符与西文字符在存储和录入等方面存在着较大差别，在利用编辑距 离计算包含中文的字符串之间的差别时存在诸多不足。最简单的方法是依据字符 串在计算机中的存储，直接将一个中文看作两个西文字符进行编辑距离计算，但是 这样可能得到半个字符距离的结果；俞荣华等(2002)将字符串分割成中文和西文 两部分，然后分别计算编辑距离。然而，中文是典型的非字母语言，在计算机录入 过程中必须使用各种输入法进行编辑，每个汉字都是多个字母按键的组合，因此， 把任意两个汉字的差别都算成同一个值，不能体现汉字之间真正的编辑差异。曹

第5章 实体分辨中的相似度计算方法(115)

犟等(2009)从汉字读音的角度提出了基于拼音改良的编辑距离，将汉字拼音分为 声母、韵母和音调三个部分，按不同的代价计算拼音的编辑距离，虽然较好地解决 了中文字符的读音相似问题，但没有考虑形近字带来相似重复的情况。

本节参考西文基于字符的编辑距离度量方式，以及目前使用最广泛的汉字输 入法，提出三种中西文混合字符串编辑距离的度量方法，即基于字符的编辑距离、 基于拼音编码的编辑距离和基于五笔编码的编辑距离。

1)基于字符的编辑距离

将单个汉字作为与西文字符等价的距离度量单位，即两个中西文混合字符串 A,B的基于字符的编辑距离(简称字符距离)等于将A转换为B 需要插入、删除、 替换的中文字符和英文字符的最少个数，记作D.(A,B)。

2)基于拼音编码的编辑距离

Sogou 实验室提供的用户日志分析结果显示，由于拼音输入法的广泛使用，大 多数用户的输入错误表现为同音字或近音字的替换误用(曹犟，等，2009)。如果 一个数据源中有类似的输入错误，那么当多个数据源进行集成时，便有可能产生相 似重复记录。拼音编码很大程度上反映了采用拼音输入法时用户真正的编辑操 作，基于此，提出基于拼音编码的编辑距离(简称拼音距离),从拼音输入的角度衡 量两个字符串的差异。

由于Sogou 输入法的易用性和用户的普及性，采用 Sogou 拼音输入法 3.6.0.1653版本的字码表进行汉字到拼音编码的映射。设x=a₁a₂a₃…a 。 为一中西 文混合字符串，a₁,a₂,a₃,…,a, 分别表示字符串中的中文或西文字符，1,2,3,…,n 标 记各个字符在字符串中的位置，m。(a₁),i=1,2,…,n 为单个汉字到拼音编码的映 射函数，则字符串x 到拼音编码的映射函数：

x 只包含西文字符 包今中文字符(

*2* *只* *包* *含* *中* *文* *子* *付* *(3-10)*

对式(5-10),当x 包含中西文字符时，只将对应中文字符位置用拼音编码， 而西文字符对应位置不变，如a₁,m 。(a₂),a₃,…,m,(an)。

若A,B 为两个中西文混合字符串，A'=f(A ),B'=f(B )分别为A,B 对应 的拼音编码，则字符串A,B 基于拼音编码的编辑距离D, 为 ：D,(A,B)=

D.(A',B')=D.(f(A),f(B))。

由于多音字的存在，在字码表中一个汉字可能对应多个拼音编码，但绝大多数 汉字的输入错误是由同音字和近音字引起的，多音字的影响可以忽略不计，因此， 在字码表中只为每个汉字保留一个最常用的拼音编码。

3)基于五笔编码的编辑距离

基于形码的五笔输入法凭着录入速度快、重码率低的特点，在专业录入人员中

116)数据质量导论

拥有大量用户。然而，由于五笔字根拆分的复杂性，输入过程中难免会产生形近字 的替换误用，基于类似拼音输入的考虑，提出基于五笔编码的编辑距离(简称五笔 距离),从五笔输入的角度衡量两个字符串的差异。

采用使用较广的陈桥五笔5.5版本的字码表进行汉字到五笔编码的映射。设 A,B 为两个中西文混合字符串，g(x) 为字符串到五笔编码的映射函数，A\*= g(A),B\*=g(B) 分别为A,B 对应的五笔编码，则字符串A,B 基于五笔编码的编 辑距离D 为：D(A,B)=D(A\*,B\*)=D.(g(A),g(B))。

这三种编辑距离分别从三个侧面描述了两个中西文混合字符串之间的差异程 度，字符编辑距离描述的是两个字符串字符存储上的差异程度，而拼音距离和五笔 距离则描述了两个字符串在采用两种输入方法进行录入过程中的差异。在计算两 个字符串的相似度时，综合考虑这三种编辑距离比单独考虑一种更能准确描述字 符串的相似程度。

**5.2.4** **多编辑距离字符串相似度融合**

编辑距离反映了两个字符串的绝对差异，而相似度以一个[0,1]之间的数值 反映两个字符串的相似程度，数值越大表示相似程度越高。

设α,β为任意两个字符串，l,l₉ 分别为两个字符串的长度，即字符串包含字符 (中文字符或英文字符)的个数，D(a,β)e[0,max}l₄,lg}] 为字符串α,β的编辑 距离，则字符串α,β的相似度S 用式(5-3)计算。

分别利用字符距离 D。、拼音距离 D,和五笔距离 D。代入式(5-3)进行相似度 计算，便得到字符串的字符相似度、拼音相似度和五笔相似度，分别记为S、S,和 S 。 三种相似度分别从不同侧面反映了两个字符串的相似程度，但三者并不是相 互独立的。对于使用拼音输入法，由同音字或近音字引起的相似重复，两个字符串 的S, 较大而S。较小；对于使用五笔输入法，由形近字引起的相似重复，两个字符串 的S, 较小而S。较大；而如果两个字符串差异很小(如完全相似重复),那么S, 和S。将都 比较大，而且接近；只有当两个字符串不相似重复时，S,和S,才会都比较小。基于此，可 以认为两个字符串无论是S, 还是S。较大，都反映了两个字符串具有较高的相似程度。

由于在计算字符串编辑距离时并不能确定其录入方式，单独使用拼音距离或 五笔距离都会带来一定的误差，因此，保留原字符距离，并将拼音距离和五笔距离 与字符距离进行融合。

设A、B 为两个中西文混合字符串，则融合相似度使用式(5-11)计算：

Sbmd =λ₁Sc+λ₂·max{S,S} (5-11)

式(5-11)中，λ₁,λ₂为各部分所占权重，O≤λ₁,λ₂≤1 且λ₁+λ₂=1 。

第5章实体分辨中的相似度计算方法 (17

**5.2.5** **实验分析**

为了验证本节提出的基于汉字编码的字符串相似度计算方法，设计两个实 验。实验的主要目的在于：①比较利用提出的相似度计算方法与其他距离度量 方法得到的相似度值，考察本方法是否能够更真实、准确地反映字符串的相似程 度；②比较各种相似度计算方法对实体分辨效果的影响，考察本方法是否能提高 实体分辨效果。

[**5.2. 5.1**](5.2.5.1) **相似度准确性验证**

1)实验设计和评价方法

使用5对已知相似重复类型和相似重复程度的样本字符串，分别利用基于字 符距离、拼音距离、五笔距离、对中西文进行分割计算编辑距离的方法，以及本节提 出的相似度计算方法计算每一对样本的相似度，然后将计算的结果进行对比。通 过比较各种方法对同一样本得到的相似度数值，考察各种方法对字符串相似程度反 映的真实程度。如对于完全相似重复的字符串，相似度数值越接近1,则认为越能反 映真实的相似程度，而对于不相似重复的字符串则正好相反。为了均衡字符相似度 和拼音、五笔相似度对结果的影响，实验中对式(5-11)中的参数取λ₁=λ₂=0.5。 在实际应用中可以依据实际情况灵活设置或利用样本数据获得最优的参数组合。

2)实验结果和分析

所用的5对字符串样本分属不同的相似重复类型，如表5-2所列。

表 5 - 2 字符串样本

|  |  |  |
| --- | --- | --- |
| **No** | **类型** | **样本** |
| 1 | 完全相似 | 栖霞区仙林街道学海路1号 |
| 栖霞区仙林街道学海路1号 |
| 2 | 不相似 | 栖霞区仙林街道学海路1号 |
| 玄武区钟灵街道9号 |
| 3 | 拼音编码相似 | 栖霞区仙林街道学海路1号 |
| 旗虾区闲林街道学海路1号 |
| 4 | 五笔编码相似 | 栖霞区仙林街道学海路1号 |
| 楂假区仙村街道学洛路1号 |
| 5 | 中英文交叉 | Java(5.0)入门教程 |
| Java入门教程(第5.0版) |

分别利用各种距离度量方法计算相似度，得到结果如表5-3所列。

118)数据质量导论

表5 - 3 五种编辑距离得到的相似度

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **No.** | **字符相似度** | **拼音相似度** | **五笔相似度** | **融合相似度** | **中西文分割** |
| 1 | 1 | 1 | 1 | 1 | —— |
| 2 | 0.3 | 0.3571 | 0.3103 | 0.3286 |  |
| 3 | 0.7 | 1 | 0.6897 | 0.85 |  |
| 4 | 0.6667 | 0.7273 | 0.8485 | 0.7576 | — |
| 5 | 0.2371 | 0.4629 | 0.5135 | 0.3753 | 0.8462 |

从表5-3可以看出，对于完全相似重复和不相似重复两种类型的样本1和样 本2,各种方法得到的相似度数值相差不大，均反映了样本真实的相似程度。对于 拼音编码和五笔编码引起的相似重复的样本3和样本4,基于拼音距离和五笔距 离的方法对于各自编码引起的相似重复反映得最好，但对其他类型的样本的适用 度较差，而基于多种汉字编码编辑距离融合的方法对两种类型的样本的相似程度 的反映虽然不是最好，却很接近最好值。更重要的是，该方法对于两个类型的样本 的相似程度都能较好的反映，可以看出融合公式(5-11)考虑了两种编码距离及 其相互关系的效果。从样本5的实验结果可以看出，中西文分割的方法忽略了单 个字符在字符串中的位置，夸大了字符串的相似程度。综合以上结果，可以认为基 于多种汉字编码的相似度融合的方法适用于各种相似类型的中西文混合字符串， 能够更加真实、准确反映字符串的相似程度。

<5.2.5.2> 实体分辨准确性验证

1)实验设计和评价方法

为了定量评测各种相似度计算方法对实体分辨效果的影响，首先构造了测试 数据生成器，生成123条测试记录，其中包含14对不同类型的相似重复记录，然后 分别利用<5.2.5.1>节中的5种方法进行实体分辨，最后使用查全率(Recall) 和查 准率(Precision) 两个指标对各自的检测结果进行评价。查全率 Re 和查准率 Pr 分 别用式(5-12)和式(5-13)计算(Lee et al.,2000):

(5-12)

(5-13)

2)实验结果和分析

通过对检测结果的统计，得到利用各种距离度量方法进行实体分辨的查全率 和查准率如图5-1所示。

从图5-1可以看出，基于字符距离的方法虽然得到了很高的查准率，但其查 全率却是最低的，这主要是因为基于字符的编辑距离将汉字作为一个字符进行处

第5章实体分辨中的相似度计算方法 

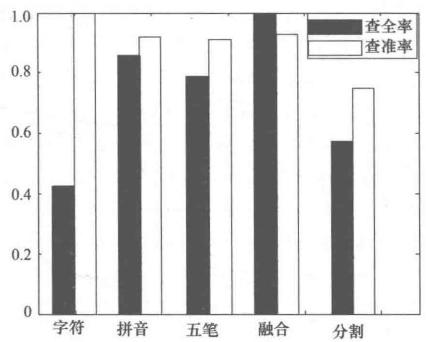


图5-1 五种编辑距离的准确性

理，无法处理由于汉字输入法引起的相似重复。基于拼音距离和五笔距离的方法 分别能够处理由于拼音录入和五笔录入引起的相似重复情况，查全率较基于字符 距离的方法有了较大的提升，但查准率有了一定的下降。基于中西文分割方法的 查全率较基于字符距离的方法也有一定提升，但查准率指标下降较大，这主要是因 为该方法会忽略单个字符在字符串中的位置，对字符串的相似度造成了一定的度 量误差。基于多种编辑距离融合的方法获得了所有方法中最高的查全率，而且在 查准率指标上虽然低于基于字符距离的方法，但仍然高于其他三种方法，而且相对 于查全率的提高程度，其查准率下降的程度是可以接受的。

为了进一步比较各种算法的综合效果，对每种算法得到的查全率和查准率进 行平均，得到结果如表5-4所列，可以清楚地看出，基于多编辑距离融合的方法的 综合检测效果在5种方法中是最好的。

表 5 - 4 综合检测效果

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **距离算法** | **字符距离** | **拼音距离** | **五笔距离** | **融合相似度** | **中西文分割** |
| Re | 0.429 | 0.857 | 0.786 | 1 | 0.571 |
| Pr | 1 | 0.917 | 0.909 | 0.929 | 0.750 |
| (Re+Pr)/2 | 0.715 | 0.887 | 0.848 | 0.965 | 0.661 |

**5.3** **属性相似度与函数依赖的关系**

为了讨论方便，将关系模式R 的任一实例记为I={t₁,t₂,…,tm},m≥1, 其中 t,j=1,2,…,m 为 I 中的任意一条记录，R 的属性集记为attr(R)={A₁,A₂,…,

120)数据质量导论

A,},n=1,2,… 。t, 中任一属性A;对应的属性值记为 t;[A,], 属性A,对应的相似度算 法记为sim;, 参与比较的两条记录(t,,₅) 在属性A;上的相似度记为s=sim,(t,[A;],

t₇ [A₁]),O≤s^"≤1,相似度向量记为V(t,,t₆)=<s,sF°,…,s)。

在实体分辨过程中，希望匹配记录对的相似度尽可能的大，不匹配记录对的相 似度尽可能的小，从而能够更准确地识别出描述同一实体的记录，并减少人工判断 的工作量，但实际情况并非如此。表5-5中给出了从多个数据源集成而来的部分客 户信息，包含姓名(name)、地址(addr)、邮编(postcard)、电话(tel) 、邮箱(email) 五个 属性，该实例中包含3条记录，其中记录t₁ 和t₂描述的是同一客户的信息。

表5 - 5 客户信息数据实例

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| ID | name | addr | **postcard** | tel | email |
| t | Mark Clifford | 10 Oak Street,MH,NJ | 07974 | 908-1111111 | mc@gmail.com |
| t₂ | M.Clivord | NJ | 07974 | 908-1111111 | mc@gmail.com |
| t₃ | Martin Crotons | 620 Elm Street,MH,NJ | 07976 | 908-2222222 | mc@hotmail.com |

使用Jaro(见<4.3.4.2>节)距离计算各个属性的相似度，得到所有记录对的相 似度向量如表5-6所列，其中最后一列为每个记录对的平均属性相似度。

表 5 - 6 客户信息数据中记录对相似度向量

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 相似度 | name | addr | **postcard** | tel | email | avgsim |
| V(t₁,t₂) | 0.68 | 0 | 1 | 1 | 1 | 0.73 |
| V(t₁,t₃) | 0.56 | 0.89 | 0.87 | 0.58 | 0.81 | 0.74 |
| V(t₂,t₃) | 0.57 | 0 | 0.87 | 0.58 | 0.81 | 0.57 |

由表5-6可以看出，记录t₁ 和t₂ 虽然描述的是同一客户的信息，但记录对(t, t₂) 的 addr 属性相似度却为0.0,而且平均相似度低于记录对(t₁,t₃) 的平均相似 度。如果决策模型依据记录对的平均相似度判断两条记录是否匹配，将无法识别 出记录t₁ 和t₂ 是描述同一客户信息的两条记录，或者将t₁ 和t₃ 错误地识别为同一客 户的信息。由于没有一种相似度算法适用于所有数据，这种情况在实体分辨过程 中是难认避免的。虽然可以通过分别为每个属性选择更合适的相似度函数来提高 属性相似度的准确性，但已有研究指出最优相似度函数选择问题是 NP 难问题 (Wang et al.,2011),而且实际应用中通常缺少用来进行相似度函数选择的训练样 本。因此，如果能够在不确定相似度函数是否最优的情况下对属性相似度进行调整， 从而提高属性相似度的准确性，对于提高实体分辨结果的准确性具有重要的意义。

在该例中，计算记录对各个属性的相似度时仅仅使用了单个属性的属性值信 息，例如计算记录对(t₁,tz) 中 name 属性的相似度时，用到的信息只有“Mark Clif-

第5章实体分辨中的相似度计算方法((121)

ford”和“M.Clivord” 两个字符串。这相当于默认各个属性之间相互独立，但实际 上各个属性之间存在着各种依赖关系，函数依赖就是最常见也是最重要的一种数 据依赖关系。在该例中存在以下3条函数依赖：

FD₁=tel→addr

FD,=email→name

FD₃=addr→postcode

FD₁是指如果两条记录的 tel 属性值相同，那么它们的 addr 属性值必然相同。 然而由于数据质量问题的普遍存在，这种严格的约束一般不会完全满足。对于实 体分辨而言，两个属性值是否匹配是用相似度的大小来判断的，因此FD₁的含义可 以理解为：如果两条记录的 tel 属性的相似度足够大。此外，FD₁ 还蕴涵了另外 一层含义，即如果两条记录的 addr 属性值不同，那么它们的 tel 属性值必然不 同。对于实体分辨而言可以理解为：如果两条记录的addr 属性相似度足够小， 那么它函数依赖的 tel 属性的相似度也应该比较小。同理 FD₂和 FD₃也具有类 似的含义。

对于任意一个函数依赖，其含义可以用以下命题及其逆否命题来描述。

**命题5-1:** 对于关系模式R,FD=X→Y 为R 上存在的任一函数依赖，t, 和t, 为 R的任一满足FD 约束的实例1中的任意两条记录，则有Vt,,t,(t,[X]=t,[X]→

t,[Y]=t₇[Y]) 成立。

命题5-1的逆否命题为：在命题5-1相同的条件下，有Vt,t(t,[Y]≠

t,[Y]=t,[X]≠t₁[X]) 成立。

对实体分辨而言， 一般通过相似度阈值判断两个属性是否匹配，如果两个属性 的相似度大于匹配阈值Th, 则认为两个属性值匹配，如果小于不匹配阈值Th~~, 则认为两个属性不匹配，如果介于两者之间，则不能直接判断两个属性值是否匹 配。根据命题5-1及其逆否命题，在实体分辨过程中，理想情况下属性相似度算 法应满足以下约束：

(1)Vt₁,t₂(sim(t₁[X],t₂[X])≥Thx→sim(t₁[Y],t₂[Y])≥Th")。

(2)Vt₁,t₂(sim(t₁[Y],t₂[Y])≤Ths\*=sim(t₁[X],t₂[X])≤Thx\*)。

然而在真实数据中，由于数据质量问题和相似度算法的局限性，约束(1)和 (2)可能会被违反。如表5-2中的V(t₁,t₂), 记录t₁ 和t₂ 在tel 属性上的相似度为 1.0,依据函数依赖FD₁,t₁ 和t₂ 在 addr 属性上的相似度应该也接近于1.0,而实际 上addr 属性的相似度却为0.0。因此，在计算某一属性相似度时，如果能够依据与 其具有函数依赖关系的属性的相似度进行调整(如依据记录t₁ 和t₂ 在 tel 属性上的 相似度提高在addr 属性相似度),使得最终得到的属性相似度向量尽量满足(1)和 (2)的约束，将能够很大程度上提高实体分辨的准确性。

数据质量导论

**5.4** **基于函数依赖的属性相似度调整**

基于上述发现和分析，本节提出了一种基于函数依赖的属性相似度调整算法 (Similarity Adjusting with Functional Dependency,SAWFD),该算法的主要思想是首 先依据属性间的函数依赖关系，把与某一属性存在函数依赖关系的属性集合划分 为函数依赖闭包和函数决定闭包，将属性相似度分为原始相似度、父相似度和子相 似度，然后根据属性相似度调整原则、函数依赖的传递性质和属性相似度调整代价 对原始相似度进行调整。

**5.4.1 属性相似度划分**

依据属性间是函数依赖关系还是函数决定关系，将与某一属性存在函数依赖 关系的属性集合划分为函数依赖闭包和函数决定闭包，它们分别对应该属性的父 相似度和子相似度。

**定义5-1(函数依赖闭包):**对于关系模式R,∑为 R 上存在的函数依赖集，R 中任一属性A,关于函数依赖集之的函数依赖闭包指的是能够由之根据 Armstrong 公理系统推理规则2(增广律)和推理规则3(传递律)推导出的属性A,完全函数依 赖的所有属性组成的集合，记为A₂。

定义5-2(函数决定闭包):对于关系模式R, 为 R 上存在的函数依赖集，R 中任一属性A,关于函数依赖集之的函数决定闭包指的是能够由三根据 Armstrong 公理系统推理规则2(增广律)和推理规则3(传递律)推导出的属性A;完全函数决 定的所有属性组成的集合，记为A。

由于某一属性的函数依赖闭包和函数决定闭包都是属性集，为了计算父相似 度和子相似度，首先给出属性集相似度定义如下。

**定义5-3(属性集相似度):**对于关系模式 R, 属性集attr(R)={A₁,A₂,…, A,},n≥1,I 为R的一个实例，I 的任意两条记录(t,t₄) 组成一个记录对，V(t,,)=

<s,s,…,sm〉 为该记录对的相似度向量，UCattr(R)为 R 上的任意属性集，则记 录对(t,,t) 在属性集U上的相似度为

Sg =min{s\*lA₁∈U} (5-14)

式(5-14)表明，任一记录对在某一属性集上的相似度由该属性集中相似度 最小的属性决定。计算集合相似度的方法有很多，如最大相似度、最小相似度、均 值相似度、Jaccard 系数等，而函数依赖的含义是如果任一记录对在该函数依赖的 所有左部属性上的属性值相同，那么它们在所有右部属性上的属性值必然相同，属 性集中任何一个属性值的不同都会违反该函数依赖的约束，因此选择属性集中相

第5章实体分辨中的相似度计算方法(

似度最小的属性代表整个属性集合是与函数依赖的含义相一致的。

**定义5-4(原始相似度、父相似度和子相似度):**对于关系模式R, 属性集 attr(R)={A₁,A₂,…,A,},n≥1,2 为 R 上存在的函数依赖集，I 为 R 的一个实例，I

的任意两条记录(t,,t₄) 组成一个记录对，V(t,,t)=<s,s{,…,sm〉 为该记录对的

相似度向量，对于任意属性A,eattr(R), 称 s” 为属性A;的原始相似度，其函数依 赖闭包A壹的属性集相似度称为属性A,的父相似度，其函数决定闭包A₂ 的属性集 相似度称为属性A,的子相似度，分别记为osim(A₁),psim(A₁) 和 csim(A₁), 在 不引起混淆的情况下简记为osim(A₁),psim(A₁) 和 csim(A;)。

如果某一属性的函数依赖闭包为空，说明该属性的相似度不受其函数依赖闭 包中属性相似度的影响；类似地，如果某一属性的函数决定闭包为空，则说明该属 性的相似度不受其函数决定闭包中属性相似度的影响。根据5.3节的分析，对于 任意两条记录，如果某一属性的函数依赖闭包中的属性值不同或函数决定闭包中的 属性值相同，该属性的相似度不会受到其函数依赖或函数决定的属性的影响。因此， 对于函数依赖闭包和函数决定闭包为空情况，规定若A₂= ,则 psim(A₁)=0, 若 A₂= ,则 csim(A₁)=1。

**5.4.2** **属性相似度调整**

[**5.4.2.1**](5.4.2.1) **相似度调整原则**

根据5.3节对属性相似度与函数依赖关系的分析和5.4.1节给出的定义可 知，若属性间存在函数依赖关系，在计算两条记录的属性相似度时，可依据某一属 性的函数依赖闭包或函数决定闭包中属性的相似度对当前属性的相似度值进行调 整，以降低由于相似度算法不合适带来的误差。基于以上分析，给出依据父相似度 和子相似度对原始相似度进行调整的基本原则如下。

相似度调整原则：对于关系模式R,FD=X→Y 为 R 上存在的任一函数依赖，t。 和t 为R 实例1中的任意两条记录，V(tp,t) 为t, 和t, 的属性相似度向量，那么:

原则1:如果记录对(t,,t) 在属性X 上的相似度大于等于匹配阈值，则提高 (t,) 在属性Y上的相似度。

原则2:如果记录对(t,,t₄) 在属性Y上的相似度小于等于不匹配阈值，则降低 (t,t) 在属性X 上的相似度。

原则1指的是如果一个属性的父相似度足够大，则提高该属性的相似度，原则 2指的是如果一个属性的子相似度足够小，则降低该属性的相似度。该调整原则 虽然简单且易于实现，但在应用中需要考虑以下三个问题： 一是如何控制相似度调 整的幅度；二是由于函数依赖的传递性，调整后的属性相似度可能会影响其他与该

124)数据质量导论

属性具有函数依赖关系的属性；三是当一个属性同时具有足够大的父相似度和足 够小的子相似度时，如何调整该属性相似度。

<5.4.2.2> 属性相似度非传递调整

首先考虑不存在函数依赖传递的情况，此时对某一属性相似度的调整不会影 响其他属性，依据<5.4.2.1>节中的相似度调整原则1和2对原始相似度osim(A₁) 进行调整，调整方法如式(5-15)所示：

SIM(A₁)=osim(A₁)+α·match(psim(A₁))-β·(1-unmatch(csim(A;)))

其中：





(5-15)

(5-16)

(5-17)

式(5-15)中α,β为最大相似度调整幅度，即各个属性相似度可以提高或降 低的最大值。相似度调整幅度的设置可以有多种策略，为保证调整后的相似度值 依然介于[0,1]之间，这里取α=1-osim(A₁),β=osim(A₁)。 式(5-15)表示如果 属性 A;的父相似度大于匹配阈值，则提高原始相似度，且提高的数值与父相似度 的大小成正比；如果属性A;的子相似度小于匹配阈值，则降低原始相似度，且降低 的数值与子相似度的大小成正比。式(5-16)和式(5-17)用来判断两个属性值 是否匹配，Th"、Th\*\*分别为匹配阈值和不匹配阈值。

<5.4.2.3> 属性相似度传递调整

当存在函数依赖传递时，属性A;的函数依赖闭包中的属性可能函数依赖于其 他属性，属性A;的函数依赖闭包中的属性的父相似度越高就越能提升属性A;的父 相似度；同理，属性A;的函数决定闭包中的属性也可能函数决定其他属性，属性A;

的函数决定闭包中的属性的子相似度越低就越能降低属性A;的子相似度，这种关 系按照函数依赖关系一直传递下去。据此，采用递归调整算法对属性的父相似度 和子相似度的计算方法进行修正，为了区别于前文中的父相似度和子相似度，分别 用psim²(A₁) 和 csim¹(A₁) 表示经过传递调整的父相似度和子相似度，计算公式如 式(5-18)和(5-19)所示：

psim²(A,)=min{osim(A,)+α·match(psim(A,))IA,∈A|(5-18) csim²(A₁)=min{osim(A)-β·(1-unmatch(csim²(A₄)))IA₄∈Ag}

(5-19) 由定义5-4容易证明，在不存在函数依赖传递的情况下psim²(A₁)=psim(A,),

csim¹(A₁)=csim(A₁) 。 因此，式(5-15)可以改写为式(5-20),不再区分是否存

第5章 实体分辨中的相似度计算方法(125)

在函数依赖传递。

SIM(A₁)=osim(A,)+α·match(psim'(A,))-β·(1-unmatch(csim²(A,)))

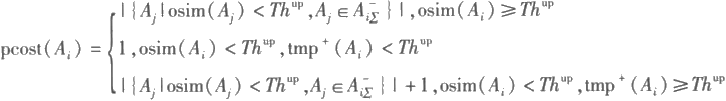
(5-20)

属性相似度传递调整以任一记录对的原始相似度向量为起点，按照式(5- 18)至式(5-20)对各个属性的原始相似度进行调整，得到第一次调整后的相似度 向量，然后再以调整后的相似度向量为起点再次进行调整，经过有限次执行后，若 每一个属性的父相似度和子相似度都趋于稳定(调整前后相似度差值小于w) 或 执行次数大于N, 则相似度传递调整结束，得到最终调整后的相似度向量。

<5.4.2.4> 属性相似度调整代价

根据函数依赖的传递性，统计依据父相似度和子相似度分别进行调整时受调 整结果影响的属性数目，受影响属性数目的多少反映了调整结果与其他属性的不 一致程度。依据群决策理论，在属性相似度调整时应当倾向于影响属性数目较少 的调整结果。基于此，给出正相似度调整代价和负相似度调整代价定义如下。

定义5-5(正相似度调整代价):对于关系模式R 的任一实例1,2为R 上存在 的函数依赖集，(t,,) 为任意两条记录组成的记录对，V(t,,t₄) 为该记录对的原始 相似度向量，记tmp\*(A₁)=osim(A₁)+α·match(psim¹(A₁)) 为任一属性A;的正



(5-21)

为属性A;的正相似度调整代价。

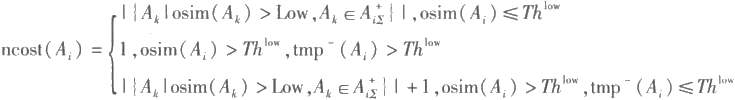
定义5-5指的是当根据父相似度提高某一属性A,的相似度时，如果该属性原 始相似度osim(A₁) 已经大于等于匹配阈值Th, 那么属性相似度的增加不会改变 该属性匹配状态，调整代价为属性A;的决定闭包中相似度小于Th” 的属性的个数； 如果原始相似度osim(A₁) 小于Th”, 且依据<5.4.2.1>节中的相似度调整原则1提 高相似度后的相似度值tmp\*(A₁) 仍然小于匹配阈值Th”,按照相似度调整原则， 调整结果对属性A,的决定闭包中的属性没有影响，调整代价为1(该属性自身);如 果原始相似度osim(A₁) 小于Th”, 且 tmp\*(A₁) 大于匹配阈值Th”, 按照相似度调 整原则，属性A,的决定闭包中相似度小于Th” 的属性的匹配状态可能会因相似度 的增加而发生改变，因此调整代价为属性A;的决定闭包中属性相似度小于Th” 的 属性的数目加该属性自身。

类似地，给出负相似度调整代价。

定义5-6(负相似度调整代价):对于关系模式R 的任一实例1, 为 R 上存在

126)数据质量导论

的函数依赖集，(t,,) 为任意两条记录组成的记录对，V(t,,t₄) 为该记录对的原始 相似度向量，记tmp(A₁)=osim(A₁)-β·(1-unmatch(csim(A₁))) 为任一属性 A;的负相似度调整值，则称



(5-22)

为属性A;的负相似度调整代价。

与正相似度调整代价类似，负相似度调整代价指的是当依据某一属性的子相 似度降低该属性相似度时，受其影响的属性数目。

在得到相似度调整代价后，依据正相似度调整代价和负相似度调整代价对式(5- 20)中的父相似度和子相似度进行加权，最终得到相似度调整公式如式(5-23)所示：





(5-23)

5.4.3 算法描述

综上所述，SAWFD算法将某一属性相似度分为原始相似度、父相似度和子相 似度，依据函数依赖的传递性质和调整代价对原始相似度进行传递调整，经过多次 调整达到平衡或超出最大执行次数后结束。具体算法如表5-7所列。

表 5 - 7 SAWFD 算法

|  |  |
| --- | --- |
| Name:SAWFD(2,V) | |
| Input:函数依赖集合2,相似度向量V  Output:调整后的相似度向量V' | |
| 1.sum=SUM(V);  2.n=0;  3.While n++<N Do  4.For each V[A;]in V 5.P=psim′(A;);  6.C=csim¹(A₁) | //计算各个属性相似度之和  //最多执行N次  //计算父相似度  //计算子相似度 |

第5章实体分辨中的相似度计算方法(127)

(续)

|  |
| --- |
| 8.End For  9.End While  10.If lsum-SUM(V')l<w//若相似度值趋于稳定则算法停止  11.BreakWhile;  12.Else  13.sum=SUM(V');  14.V=V';  15. End If  16.ReturnV"; |

从表5 - 7中的第3行和第8行可知SAWFD 算法一定收敛，最坏情况下执行 N次调整，按照最坏情况考虑，算法对某一相似度向量中的每个属性相似度进行调 整，时间复杂度为O(N·IVI), 其 中IVI 为相似度向量中包含的属性个数。由于第 7行中相似度传递调整和调整代价的计算都需要遍历相似度向量中的元素，因此 完成对某一相似度向量调整的最终时间复杂度为O(N·21VI²) 。 当实体分辨任 务中存在m 条相似度向量时，SAWFD 算法逐条对每条相似度向量进行调整，需要 的时间与相似度向量的数目成正比，即O(m·N·21VI²) 。 通常情况下，相似度向 量的数目要远远大于一个相似度向量包含的属性个数，而利用SAWFD 算法进行 调整需要的时间与相似度向量数目成线性递增关系。

5. 4.4 实 验 分 析

SAWFD 算法的目的是提高属性相似度的准确性，尽量区分匹配记录和不匹配 记录，进而提高实体分辨结果的准确性。因此，本实验从两个方面对该算法进行验 证：①区分能力验证，统计不同相似度算法在使用SAWFD 算法前后得到的记录对 平均属性相似度的分布，验证 SAWFD 算法能否增加匹配记录对与不匹配记录对 之间的相似度差别；②准确性验证，对相同的数据集采用不同相似度算法进行实体 分辨，验证使用SAWFD 算法能否获得更高的查全率 Re 和查准率 Pr 和 F¹ 值

(Pamadies et al.2012),其 中

最后对 SAWFD算法的效率和可扩

。

展性进行了验证。

**128)**数据质量导论

[**5.4.4.1**](5.4.4.1) **实验设置**

本实验使用的实验数据来源于MondialDB 数据集(Weis et al.,2004),包含 260个不同国家的信息，每个国家的信息由 name,continent,religion,government, language,car\_code,ethnicgroup,city 等8个属性组成，其上存在函数依赖：

FD₁=name→continent,city,government,ethnicgroup

FD₂=ethnicgroup→religion,language

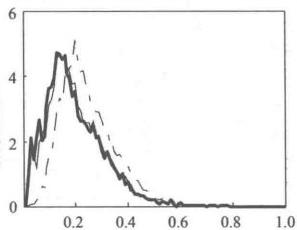
与 Weis 等(2004)类似，使用 DirtyXMLGenerator 工具向数据集中引入冗余记录， 其中字符交换、字符添加、字符删除和属性值删除所占比例分别为30%、30%、30%和 10%,得到的实验数据(记为MixedDB)包含不匹配记录6490对，匹配记录681对。

[**5.4.4.2**](5.4.4.2) **区分能力验证**

为了验证 SAWFD算法能否更好地区分匹配记录和不匹配记录，分别选择基 于字符和基于token的相似度算法中最为常用的Jaro 距离和 Atomic String 算法进 行对比验证。之所以没有选择基于读音的相似度算法是因为此类算法只能判断两 个字符串的读音是否相似，并不能作为判断两个字符串匹配的直接依据。除了 <5.4.4.1>节中的实验数据之外，本节额外构造了两组数据集， 一组数据集中不包含 匹配记录对(任意两条记录都不匹配),另外一组数据集中只包含匹配记录对(任 意两条记录都匹配),分别记为CleanDB和 DuplicateDB。首先，分别用Jaro 距离和 Atomic String算法计算每组实验数据中所有记录对的属性相似度，然后应用SAW- FD 算法对属性相似度进行调整，最后统计调整前后得到的匹配记录和不匹配记录 的相似度分布。匹配记录的相似度越大，不匹配记录的相似度越小。匹配记录与 不匹配记录之间相似度差别越大，说明属性相似度向量对记录匹配情况的描述越 准确，相似度算法区分匹配记录和不匹配记录的能力越强。

Snae(2007) 和Paradies 等(2012)中的研究，对于Jaro距离，取各个属性的匹配 阈值Th=0.9, 不匹配阈值Th~~=0.7, 对于Atomic String 算法，取各个属性匹配 阈值Th=0.75, 不匹配阈值 Th~=0.55 。 在实验中发现当SAWFD算法最大执 行次数N 的取值大于5时，相似度调整结果的变化不大，即在5次传递调整之后， 大多数属性的相似度趋于稳定，再进行调整的意义不大，因此实验结果中只对比了 N=0,1,5 时的结果，如图5-2至图5-4所示。

图 5 - 2 中noFD表示不使用SAWFD算法的相似度分布情况，SAWFD(N=1) 和SAWFD(N=5) 表示使用SAWFD算法后的相似度分布情况，N=1 和N=5 分别 表示取算法最大执行次数为1和5。图5-2(a) 对比了在 CleanDB 数据集上采用 Jaro距离计算属性相似度时，使用SAWFD算法前后记录对相似度的分布情况，从 图中可以看出使用SAWFD算法后，SAWFD(N=1) 和 SAWFD(N=5) 情况下的记 录对相似度分布更靠近0,而且随着算法最大执行次数的提高SAWFD(N=5) 情况

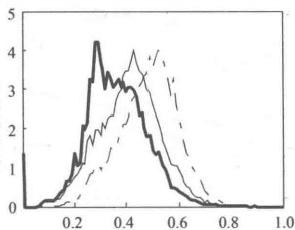
第 5 章 实体分辨中的相似度计算方法( 129

所占比例/%

所占比例/%

所占比例/%

所占比例/%



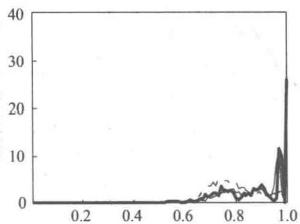
|  |  |
| --- | --- |
| 记录相似度 | 记录相似度 |
| (a)Jaro | (b)Atomic String |



图 5 - 2 SAWFD算法对 CleanDB 记录相似度影响



记录相似度



记录相似度

(a)Jaro (b)Atomic String

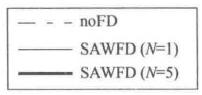
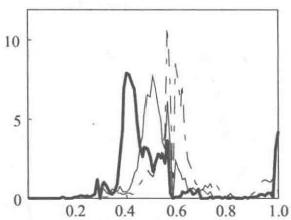


图 5 - 3 SAWFD算法对 DuplicateDB 记录相似度影响

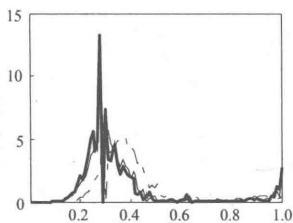
下得到的记录对相似度分布较SAWFD(N=1) 更靠近0。noFD 情况下所有不匹配

记录对的平均相似度为0.48,SAWFD(N=1) 和SAWFD(N=5) 情况下所有不匹配 记录对平均相似度分别为0.40和0.35。可以看出，对于不匹配记录对使用SAW- FD算法能够得到更小的相似度，且算法执行次数N=5 时效果更显著。图5-2 (b) 对比了采用Atomic String算法计算属性相似度的情况，从图中也可以得到与 图 5 - 2(a) 类似的结论，不过随着算法最大执行次数的增加调整效果没有明显变 化，这说明调整算法在达到最大执行次数前已经终止。

130)数据质量导论

所占比例/%

所占比例/%



|  |  |
| --- | --- |
| 记录相似度 | 记录相似度 |
| (a)Jaro | (b)Atomic String |

|  |
| --- |
| -noFD  —SAWFD(N=1)  SAWFD(N=5) |

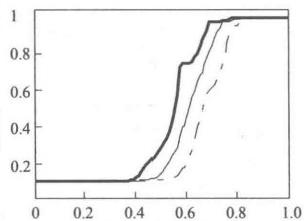
图5-4 SAWFD算法对 MixedDB 记录相似度影响

图5 - 3(a) 和 图 5 - 3(b) 对比了在 DuplicateDB 数据集上采用Jaro 距离和 Atomic String算法计算属性相似度时，使用SAWFD 算法前后记录对相似度的分布 情况。可以看出，与CleanDB 数据集上的结果相反，使用SAWFD 算法后，匹配记 录对的相似度分布更靠近1。这说明对于匹配记录对使用SAWFD 算法后能够得 到更大的相似度。

图 5 - 4(a) 对比了在MixedDB 上采用Jaro 距离计算属性相似度时，使用 SAWFD算法前后记录对相似度的分布情况。noFD情况下，记录对的相似度基本 上集中分布在两个区域，靠近0的为不匹配记录对，靠近1为匹配记录对，但也有 许多记录对落在两个区域之间，说明采用Jaro 距离不能很好描述这些记录对的相 似程度。使用SAWFD算法后，SAWFD(N=1) 和SAWFD(N=5) 情况下，匹配记录 和不匹配记录的相似度更加向两端靠拢，分离得更加明显，且落在两个区域之间的 记录对的数目也明显减少。noFD情况下匹配记录对和不匹配记录对相似度的平 均差别为0.37,SAWFD(N=1) 和SAWFD(N=5) 情况下的匹配记录对和不匹配记 录对相似度的平均差别分别为0.47和0.53。这说明使用SAWFD算法能够增加 匹配记录对和不匹配记录对的相似度差别，因此也就能够更好地区分匹配记录和 不匹配记录。图5-4(b) 也可以得到类似的结论。

[**5.4.4.3**](5.4.4.3) **准确性验证**

为了验证 SAWFD算法能否提高实体分辨结果的准确性，本实验基于开源实 体分辨框架 SERF(Benjelloun et al.,2007)对 SAWFD算法进行应用实现，并在 MixedDB数据集上进行验证。首先，在 SERF中分别用Jaro 距离和 Atomic String 算法计算属性相似度，对MixedDB 进行实体分辨；然后分别针对Jaro距离和 Atom-

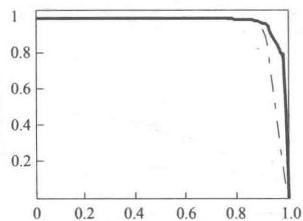
第5章实体分辩中的相似度计算方法(131

查全率/%

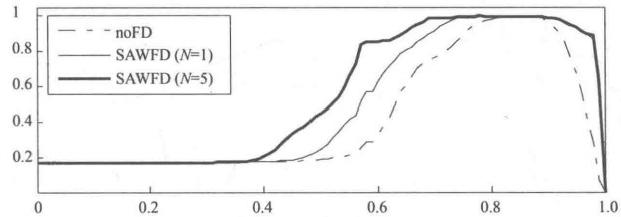
查准率

可值

ic String算法得到的相似度向量应用SAWFD 算法进行调整，再采用调整后的属性 相似度对 MixedDB 进行实体分辨；最后对比相同记录匹配阈值条件下 SAWFD 算 法对实体分辨准确性指标的影响(查全率、查准率和 F1 值)。实验结果如图5-5 和图5-6所示。



|  |  |
| --- | --- |
| 记录匹配阈值 | 记录匹配阈值 |
| (a) 查全率 | (b) 查准率 |



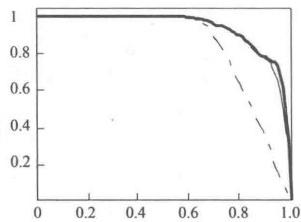
记录匹配阈值

(c)F1值

图5-5 SAWFD算法对Jaro 实体分辨准确性的影响

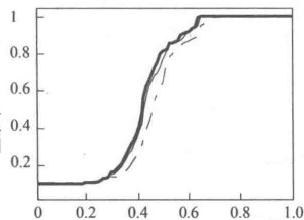
图5-5中的三个图依次对比了采用Jaro 距离计算属性相似度时，使用 SAW- FD 算法前后得到的查全率、查准率和 F1 值。可以看出，在相同的记录匹配阈值 条件下，无论是查全率、查准率还是 F1 值，使用SAWFD算法后都可以获得更高的 指标值，即在相同条件下，使用SAWFD算法可以获得更准确的实体分辨结果。而 且随着算法执行次数从N=1 增加到N=5, 实体分辨结果的查全率和查准率都得 到了相应的提高，尤其是查准率提高更为明显。

图5-6中的三个图依次对比了采用Atomic String算法计算属性相似度时，使 用SAWFD算法前后得到的查全率、查准率和F1 值。也可以看出，使用SAWFD算 法后，各项指标均得到了提高，但算法执行次数的增加对实体分辨结果影响不大， 这说明算法在一次执行后就基本上达到了终止条件。结合第<5.4.4.2>节的实验结 果，可以说明在使用Atomic String算法计算属性相似度时，调整算法在N=1 以后

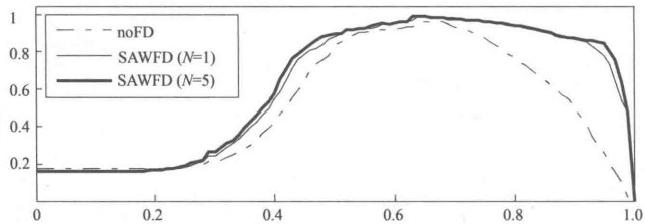
)数据质量导论

凡值 查全率

查准率



|  |  |
| --- | --- |
| 记录匹配阈值 | 记录匹配阈值 |
| (a) 查全率 | (b) 查准率 |



记录匹配阈值

(c)F¹值

图5-6 SAWFD算法对 Atomic String 实体分辨准确性的影响

就基本趋于稳定。

为了进一步验证 SAWFD算法具有更好的区分能力，在保证实体分辨结果正 确的前提下对记录匹配阈值的范围进行了统计对比，实验结果图5-5(a) 中可以 看出，在采用Jaro 距离计算属性相似度时，能够获得正确实体分辨结果的阈值范 围是[0.86,0.93],使用SAWFD算法后，可以获得正确实体分辨结果阈值范围达 到了[0.82,0.95](N=1) 和[0.78,0.96](N=5), 阈值范围扩大了约一倍。也就 是说，在相同匹配阈值条件下(如0.82),不使用SAWFD算法可能无法获得正确的 实体分辨结果，而使用SAWFD算法后则可以获得正确的实体分辨结果。从(b) 中 也可以得出类似的结论，虽然采用Atomic String算法情况下阈值范围增大效果不 如Jaro 距离明显，但也扩大了约50%左右。结果说明使用SAWFD算法对匹配阈 值的要求更加宽松，这从另一方面也验证本章提出的算法能更好地区分匹配记录 对和不匹配记录对。

[**5.4.4.4**](5.4.4.4) **算法效率验证**

为了验证 SAWFD 算法的效率和可扩展性，在 MixedDB 数据集上采取使用 SAWFD算法和不使用SAWFD算法两种情况计算所有记录对的相似度向量，然后

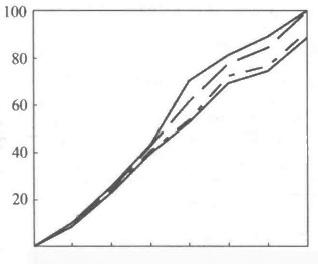
第5章实体分辨中的相似度计算方法(**133**

时 间 / s

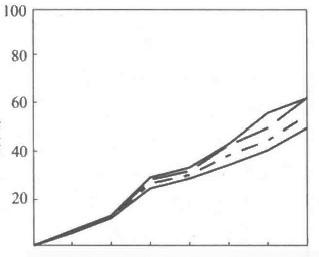
时间/s

对比它们消耗的时间。此外，为了验证最大执行次数N 对算法效率的影响，在相 同数据规模的情况下对比了不同的N 值情况下SAWFD 算法所消耗的时间。

MixedDB数据集中包含7171对记录对，随机取其中7000对，以1000 为增量 逐渐递增，验证在使用SAWFD算法(N=1,N=5 和N=10) 和不使用SAWFD算法 情况下计算记录对相似度向量消耗的时间。结果如图5-7所示。

0 1000200030004000500060007000

记录对数目/对



0 1000200030004000500060007000

记录对数目/对

(a)Jaro &MixedDB (b)Atomic String &MixedDB

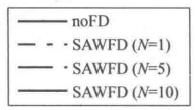


图5-7 SAWFD算法效率

从图5-7中可以看出，使用SAWFD 算法和不使用SAWFD 算法消耗的时间 基本上都随着记录对数目的增加成线性递增。图5-7中最下方的虚线是不使用 SAWFD 算法计算相似度向量的时间消耗曲线，通过与使用SAWFD 算法的情况对 比可以发现，因使用SAWFD算法进行相似度调整而增加的时间远小于原来计算 相似度向量所需的时间，这是由于SAWFD算法不涉及具体属性值，虽然需要两次 遍历相似度向量中的元素，但其计算复杂度仍远小于字符串相似度算法。图5-7 中上方的三条曲线对比了算法最大执行次数N=1,N=5 和N=10 时使用 SAWFD 算法的时间消耗，可以看出随着N 的增加，SAWFD 算法消耗的时间也会随之增 加，但增加的时间与N 的大小并没有线性递增的关系。这是因为在到达最大执行 次数之前，相似度向量的各个相似度值可能已经达到稳定状态，SAWFD 算法消耗 的时间随着实际执行次数的增加而增加，与最大执行次数之间没有必然联系。

**5.5** **本章小结**

因现有记录比较过程中大部分相似度函数都是针对西文字符提出的，不能很

*13)数据质量寻论*

好地反映中文环境下的字符串相似程度的问题，本章以编辑距离为基础提出了融 合多种编辑距离的字符串相似度计算方法。通过拼音距离、五笔距离和字符距离 等多种距离的融合减少了单独考虑拼音或五笔可能带来的偏差，有效提高了字符 串属性相似度计算的准确性，进而可以提高实体分辨的准确性。

在计算属性相似度时，各种用于字符串比较的相似度算法是正确判断两个属 性相似程度的基础，提高字符串相似度算法的准确性对于提高实体分辨的准确性 具有重要意义。但这些字符串相似度算法通常只用到属性值信息，准确性提升程 度有限，而且不同的相似度算法都有其适用范围，如果不能选择合适的相似度算 法，属性相似度的准确性也会受到影响。因此，本章又针对大多数属性相似度函数 只利用属性值信息，准确性提升程度有限，而且由于每种相似度函数都有一定的适 用范围，相似度函数选择不适当也会影响属性相似度的准确性的问题，提出了基于 函数依赖的属性相似度调整方法。基于函数依赖对属性相似度进行调整的方法扩 展了计算属性相似度所能用到的信息范围，能够在不确定属性相似度函数是否最 优的情况下提高属性相似度的准确性，为下一步的匹配决策提供更加准确的判断 依据。

**参考文献**

[1]俞荣华，田增平，周傲英.2002.一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，29

(1):118-121.

[2]曹犟，邬晓钧，夏云庆，等.2009.基于拼音索引的中文模糊匹配算法[J]. 清华大学学报：自然科学版，49

(S1):1328-1332.

[3]曹建军.2008.基于提升小波包和改进蚁群算法的自行火炮在线诊断研究[D]. 石家庄：解放军军械工程 学院 .

[4]Bachteler T,Reiher J,Schnell R.2013.Similarity Fitering with Mulibit Trees for Record Linkage[R].Germany

Nuremberg:German Record Linkage Center.

[5]Benjelloun O,Garcia -Molina H,Gong H,et al.2007.D-swoosh:A Family of Algorithms for Generic,Distribu- ted Entity Resolution [C]//Proceedings of the 27th International Conference on Distributed Computing Systems.Toronto,Ontario,Canada:IEEE:37-37.

[6]Bhattacharya I,Getoor L.2005.Relational Clustering for Multi-type Entity Resolution[C]//Proceedings of the

4th International Workshop on Multi-Relational Mining.New York,NY,USA:ACM:3-12.

[7]Bhattacharya I,Getoor L 2007.Collective Entity Resolution in Relational Data[J].ACM Transactions on Knowl-

edge Discovery from Data(TKDD),1(1):5.

[8]Chen Z,Kalashnikov D V,Mehrotra S.2005.Exploiting Relationships for Object Consolidation[C]//Proceedings of the 2nd International Workshop on Information Quality in Information Systems.Baltimore,Maryland,USA: ACM:47-58.

[9]Chen Z,Kalashnikov D V,Mehrotra S.2007.Adaptive Graphical Approach to Entity Resolution [C]//Proceedings

of the 7th ACM/IEEE-CS Joint Conference on Digital Libraries.Vancouver,BC,Canada:ACM:204-213.

第5章实体分辨中的相似度计算方法(135)

[10]Christen P.2012.Data Matching:Concepts and Techniques for Record Linkage,Entity Resolution,and Dupli- cate Detection[M].New York:Springer Science &Business Media.

[11]Cohen WW,Ravikumar P,Fienberg S E.2003.A Comparison of String Distance Metrics for Name -matching Tasks[C]//Proceedings of the IJCAI-2003 Workshop on Information Integration on the Web(IIWeb-03). Acapulco,Mexico[S.n.]:73-78.

[12]Fellegi I P,Sunter A B.1969.A Theory for Record Linkage[J].Journal of the American Statistical Association, 64(328):1183-1210.

[13]Gomaa W H,Fahmy A A.2013.A Survey of Text Similarity Approaches[J].International Journal of Computer Applications,68(13):13-18.

[14]Hong G,Lindsay B J,Asoke K N.2005.Feature Generation using Genetic Programming with Application to Fault Classification[J].IEEE Transactions on System,Man,and Cybernetics,35(1):89-99.

[15]Kalashnikov D V,Mehrotra S,Chen Z.2005.Exploiting Relationships for Domain -Independent Data Cleaning [C]//Proceedings of the 2005 SIAM International Conference on Data Mining Newport Beach,CA,USA:SDM: 262 -273.

[16]Koudas N,Sarawagi S,Srivastava D.2006.Record Linkage:Similarity Measures and Algorithms[C]//Proceedings of the 2006 ACM SIGMOD International Conference on Management of data.Illinois,USA:ACM:802-803.

[17]Lee M L,Ling T W,Low W L.2000.IntelliClean:A Knowledge -Based Intelligent Data Cleaner[J].Knowl- edge Discovery &Data Mining:290-294.

[18]Levenshtein V I.1966.Binary Codes Capable of Correcting Deletions,Insertions,and Reversals[J].Soviet

Physics,10(8):707-710.

[19]Nuray-Turan R,Kalashnikov D V,Mehrotra S.2013.Adaptive Connection Strength Models for Relationship - Based Entity Resolution[J].Joumal of Data and Information Quality(JDIQ),4(2):8.

[20]Paradies M,Malaika S,Simeon J,et al.2012.Entity Matching for Semistructured Data in the Cloud[C]//Pro-

ceedings of the 27th Annual ACM Symposium on Applied Computing[S.I.]:ACM:453-458.

[21]Snae C.2007.A Comparison and Analysis of Name Matching Algorithms[J].International Journal of Applied

Science,Engineering and Technology,4(1):252-257.

[22]Wagner R A,Fischer M J.1974.The String-to-String Correction Problem[J].Journal of the ACM(JACM), 21(1):168-173.

[23]Wang J,Li G,YuJX,et al.2011.Entity Matching:How Similar is Similar[J].Proceedings of the VLDB En- dowment,4(10):622-633.

[24]Weis M,Naumann F.2004.Detecting Duplicate Objects in XML Documents[C]//Proceedings of the 2004 In-

termational Workshop on Information Quality in Information Systems.Paris,France;ACM:10-19.

[25]Winkler W E,Thibaudeau Y.1991.An Application of the Fellegi-Sunter Model of Record Linkage to the 1990

US Decennial Census[J].US Bureau of the Census:1-22.

**第** **6** **章** **基于关系的实体分辨**

**6.1** **引言**

属性相似度的计算为匹配决策判断两条记录的匹配状态提供依据，匹配决策 模型能否准确判断记录对的匹配状态是决定实体分辨准确性的另一个重要因素。 传统的实体分辨中通常只涉及单一类型的客观实体，匹配决策模型通常只利用记 录对的属性相似度向量通过分类或者聚类的方法判断记录对的匹配状态。然而 随着互联网、电子商务和大数据技术的发展，实体分辨任务往往要对包含多种 类型实体的数据进行分辨，传统的匹配决策模型大多没有考虑实体之间的关 联关系，经常会产生大量匹配状态为可能匹配的记录对，需要大量的人工参与 进一步的判断。

基于关系的实体分辨是近几年来出现的一类实体分辨匹配决策模型，可以利 用实体间的关联关系对传统匹配决策模型不能准确判断的记录对做进一步处理。 此类方法通常将数据集看作实例化的实体关系图，利用图论方法计算两个实体描 述之间的连接强度(Connection Strength),并依据连接强度的大小判断它们是否描 述同一客观实体(Whang et al.,2012;Kalashnikov et al.,2008;Whang et al. 2013)。为了提高这类方法的通用性，利用训练数据来学习合适的路径权重是较 为常用的方法，但存在着对训练数据的数量和满足条件要求过于严格的问题 (Nuray -Turan et al.,2013;Li,2010)。为了降低基于关系的匹配决策模型对训练 数据的要求，并提高决策模型的准确性，本章针对基于关系的实体分辨，提出了一 种基于关系类型的自适应实体分辨方法。

基于关系的实体分辨(Relationship-based Data Cleaning,RelDC)针对的是利 用现有的基于特征相似度(Feature Based Similarity,FBS)方法不能有效识别的对 象，且它的计算复杂性较高，因此需要首先找出FBS方法不能有效识别的对象，再 利用 RelDC 对其进行分辨。FBS 方法中，两大类方法最为常见， 一类在计算记录 对的对应字段相似度的基础上计算记录相似度，根据记录相似度判断该记录 对是否相似；另一类仅计算记录对对应字段的相似度，再利用学习算法基于字 段相似度判断其是否相似。本章针对以上两大类 FBS 方法，研究提出两种记 录对划分技术。

第6章基于关系的实体分辨 (**137**

**6.2** **基于云模型的实体分辨记录对划分**

FBS 方法通过计算记录中各属性值的相似程度进行相似重复记录检测(Anan- thakrishna et al.,2002),由于属性提供的信息是有限的，因此尽管 FBS 方法能够较 好地完成检测，却往往不能给出高置信度的决策。RelDC 基本思想是用图对关系 数据库建模得到数据库的完整实体关系图，通过分析实体关系图，利用实体之间的 关系提高相似重复记录检测效果(Kalashnikov et al.,2003)。实体关系图中不仅包 含消歧对象的直接链接实体，还包含更多的间接链接实体，蕴涵着丰富的信息。相 比FBS 方法而言，RelDC 能够取得更高置信度的相似重复记录检测效果，但是实体 关系图的规模依赖于数据库的规模，运算复杂度非常高。因此一种可行的进行 相似重复记录检测的方法是当 FBS 方法置信度高时，采用FBS 方法进行相似 重复记录检测，在用FBS 方法不好判断时，采用 RelDC 。因此，需要对实体分 辨记录对进行划分，将记录对划分为易分辨的记录对和难分辨的记录对(曹建 军，等，2010b), 对易分辨的记录对使用FBS 方法进行分辨，对难分辨的记录 对使用 RelDC 进行分辨。

为研究记录对划分，采用曹建军等(2010a) 用到的数据集，首先人工对记录进 行比对，得出重复记录对和不重复记录对，再根据属性相似度计算方法计算属性的 相似度，并对各属性的相似度取均值得到记录的相似度。随后利用SPSS 软件分析 记录相似度分布情况，指出其类似于正态分布，但是不符合正态分布。

云模型研究随机性和模糊性的关联性，并统一用熵作为客观事物和主观认 知中不确定状态的度量，用超熵确定不确定状态的变化。超熵是正态云中用以 弱化正态分布条件的参数，表征云模型偏离正态分布的程度(李德毅，等， 2005)。目前云模型广泛应用于数据挖掘、图像分割等领域(邸凯昌，等，1999; 秦昆，等，2006)。由于图像切割与记录对划分有类似之处，本节将云模型运用 于记录对划分研究中。将不相似重复记录对和相似重复记录对映射为两朵半 云，再通过给定约束条件，求出两个满足目标的最优的记录相似度阈值，这两个 记录相似度阈值之间的相似度所对应的记录对即为难分辨的记录对，其余为易 分辨的记录对。

**6.2.1** **云模型简介**

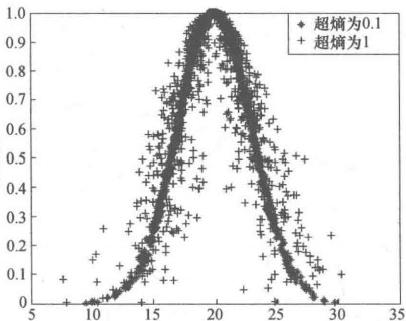
云是用语言值表示的某个定性概念与其定量表示之间的不确定性转换模型， 用以反映自然语言中概念的不确定性，它综合考虑随机性和模糊性，构成定性和定 量的映射，云模型是云的具体实现方法。

138)数据质量导论

确定度

设U是一个用精确数值表示的定量论域，C是 U上的定性概念，令定量值x ∈ U,且x 是定性概念C 的一次随机实现，若x 满足：x～N(Ex,En¹²), 其中 En'~ N(En,He²), 且x 对概念C 的确定度μ满足：

 (6-1) 则x 在论域U上的分布称为正态云。其中，Ex 为期望，En 为熵，He为超熵。为形 象地说明x 与确定度μ的关系，用图6-1表示期望为20,熵为3,超熵分别为0.1 和1时的1000 云滴。



x

图6-1 云模型示例

由图6-1可知，离期望值越近的点，确定度越大；超熵越大，数据点越分散，偏 离正态分布的程度越大。

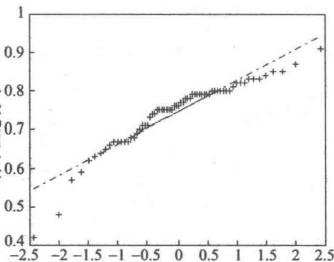
**6.2.2** **记录相似度的分布**

令相似重复为定性概念M, 不相似重复为定性概念U。 由预先给出的记录经 过人工比对，得出其中相似重复记录对α,不相似重复记录对β。α对应的记录相 似度为A,β对应的记录相似度为B,为直观说明记录相似度的分布，用直方图和 Q-Q 图分别表示匹配和不匹配记录对记录相似度的分布，结果如图6-2所示。

用K-S 检验定量说明相似度的分布，结果如表6-1所列。

表6 - 1 K-S 检验

|  |  |  |  |
| --- | --- | --- | --- |
| **相似度** | **统计量** | **样本个数** | **显著性水平** |
| A | 0.167 | 65 | 0.000 |
| B | 0.056 | 812 | 0.000 |

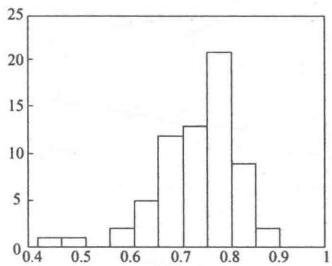
第6章基于关系的失体分班**(39**

出现次数

样本正态分布

出现次数

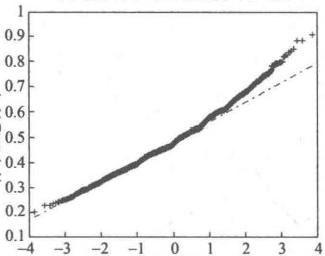
样本正态分布

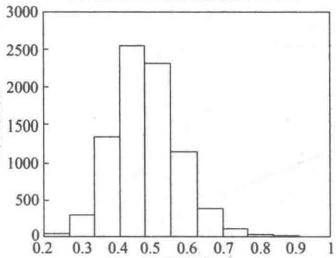


标准正态分布

相似度

**(a)相似记录对的相似度的直方图** (b)相似记录对的相似度的Q-Q图



相似度 标准正态分布

(c)不相似记录对的相似度的直方图 (d)不相似记录对的相似度的Q-Q图

图6-2 正态检验结果

从图6- 2的直方图和标准Q-Q 图可知，相似度向量接近于正态分布，但是 经过 K-S 检验发现，它并不属于正态分布。因此，本节假设记录相似度呈近似正 态分布。

**6.2.3** **记录相似度的云模型表示**

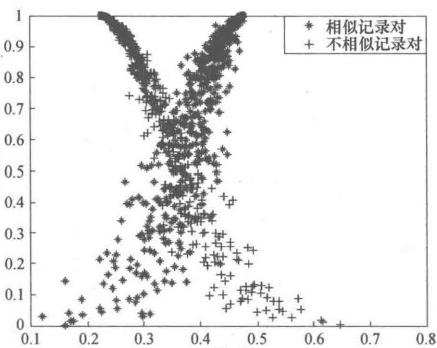
由于记录相似度的分布只是近似呈正态分布，可以用超熵来反映其偏离正态 分布的程度，用二维正态云模型描述总的分布情况，期望是所有云滴(相似度)在 二维坐标上的平均值的坐标，熵一方面反映相似度的随机性，即分别在水平和垂直 方向上相对于期望值的离散度，另一方面又反映了归属于相似或不相似的模糊 性——隶属度。超熵反映了离散程度，体现了隶属度的不确定性及偏离正态分布 的程度。

首先利用逆向云发生器计算记录相似度的数字特征，再利用正向云发生器模 拟生成不同数量的云滴(根据相似度向量的性质，仅取半云)(李德毅，等，

140)数据质量导论

确定度

2005)。为分析切换，将不相似重复记录和相似重复记录的云模型在一张图里 表示，得出图6-3。



相似度

图6-3 云模型

6.2.4 划分方法

由于目前用云模型进行图像分割的研究较多，而对边缘点的情况研究较少，因 此为寻找切换的边缘，设计约束条件如下：

(1)相似重复记录区间的下界要大于不相似重复记录区间的上界，即两个区 间不能重叠。

(2)记录相似度阈值所对应的确定度应该足够大，以保证划分出的易分辨记 录对能达到较高的准确性。

(3)能采用FBS方法清洗的尽量用FBS方法，因此拒绝域(即上边缘和下边 缘之间的区间)应该尽量小。

(4)由于已知某点是相似还是不相似，因此应使相似的点尽可能地落在相似 区间内，反之亦然。

令A对应的期望为 EA, 超熵为 HeA,某个下边缘为 a₁;B 对应的期望为 EB, 超熵为 HeB, 某个上边缘为b; 。因此，拒绝域可以表示为(a,b;), 归一化

后的拒绝域为：cr,=(a₁-b₁)/max(a₁-b;)。 因此，此时M 的置信区间为[a,

1],对应的置信度为Pm=1- φ((em -m₁)/enm) (由于仅是单边的云模型，因 此这样计算置信度),归一化后的概率为npm=Pm/max(pm),U 的置信区间为

[0,u,], 对应的置信度为Pg=φ((uy-eu)/enu) 。 归一化后的概率为npy=P /

max(p)。

第6章基于关系的实体分辨(141)

置信度

那么约束条件可表示为

a;>b; (6-2)

cm>0m 且 cu;>θ (6-3)





式中：0,和0,分别为匹配和不匹配记录对的记录相似度的确定度阈值。 为便于分析求解，将第3条和第4条约束条件合并为

max(np₈₁+nP₀-cri)

据此建立如式(6-7)所示的模型。

,4j *npa₁+npo,-cr*

s.t. *a;>b*

*cm;>0m&cu;>θ₄*

(6-4)

(6-5)

(6-6)

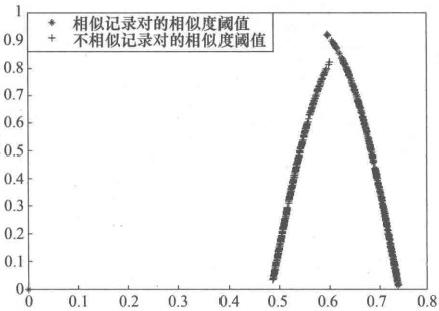
(6-7)

(6-7)

由于相似度向量和不相似度向量都不是正态分布，由图6-3可知，点是 离散的，且分布较散，同一个x 值可能有不同的u 值。因此，难以用正态分布 的性质求解。好在云滴为离散的，且数目较小，可以通过穷举搜索，找出最 优解。

为形象说明，将记录对相似度与置信度及拒绝域大小的关系用图6-4和 图6-5表示。令确定度阈值均为0.5,则

根据约束条件，求得最优解为：a。为0.6003,b。为0.5995。此时，A 对应的 确定度为0.6643,B 对应的确定度为0.5260。A 对应的置信度为0.9207,B 对应的置信度为0.7968。拒绝域大小为0.0007。



相似度阈值

图6-4 记录对相似度与置信度的关系

**142)**数据质量导论

拒绝域大小

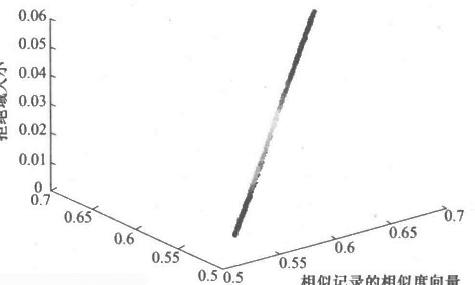
不相似记录的相似度向量

图6-5 记录对相似度与拒绝域大小的关系

**6.2.5** **结果分析**

在绝域中的相似度所对应的记录对为难分辨的记录对，其余为易分辨的记 录对。

相似重复记录的云滴分布较散，因为它对应的超熵大，即偏离正态分布的程度 较大。

直观上看，对不相似重复记录来说，其边缘取值越大，置信度越高。这是因为， 值越大，则包含更多的不相似重复记录的云滴，置信区间更大，因此置信度越高。 相似重复记录的情况同理。拒绝域有两个极小值， 一个出现在(0.6,0.6)附近， 一 个出现在0(此时对应的是不满足约束条件第一条或第二条的点)。

为进一步评价求得的边缘，将其用在原始的相似度向量中。由于相似记录相 对不相似记录特别少，查全率和查准率不适合本书的情况，因此利用分辨正确率 (正确分辨的记录对与总记录对之比)和分辨错误率(错误分辨的记录对与总记录 对之比)两个指标评价边缘的优劣。前述最优值对应的分辨正确率和分辨错误率 如表6-2所列。

表6 - 2 最优值对应的分辨正确率和分辨错误率

|  |  |  |
| --- | --- | --- |
| **记录对** | **分辨正确率** | **分辨错误率** |
| 相似重复记录对 | 0.9394 | 0.0606 |
| 不相似重复记录对 | 0.8879 | 0.1050 |

第6章基于关系的实体分辨 (143)

此时，分辨错误率较大，而分辨错误率较大主要是由于拒绝域较小，致使部分 不相似重复的点落在了相似重复的区间内。现仅考虑第四条，则：最优的下边缘为 0.6107,最优的上边缘为0.6053。此时拒绝域大小为0.0054。与前述最优解对比 可知，此时，分辨正确率提高，分辨错误率降低，是建立在整体边缘都提高的情况， 直观上看，这将引起相似重复记录对的分辨正确率降低，分辨错误率提高，事实上 相似重复记录对的分辨正确率和分辨错误率没有变化，是因为相似记录少(仅66 条),导致边缘提高一点对整体没有影响，如表6-3所列。

表 6 - 3 不考虑拒绝域时得到的分辨正确率和分辨错误率

|  |  |  |
| --- | --- | --- |
| **记录对** | **分辨正确率** | **分辨错误率** |
| 相似重复记录对 | 0.9394 | 0.0606 |
| 不相似重复记录对 | 0.9071 | 0.0814 |

考虑拒绝域进一步增大的情况，假设下边缘为0.5,上边缘为0.7,此时情况如 表6-4所列。

表 6 - 4 拒绝域较大时得到的分辨正确率和分辨错误率

|  |  |  |
| --- | --- | --- |
| **记录对** | **分辨正确率** | **分辨错误率** |
| 相似重复记录对 | 0.7273 | 0.0303 |
| 不相似重复记录对 | 0.5710 | 0.0162 |

可知，增大边缘大小有利于降低分辨错误率，但是分辨正确率也会降低。因此 本节的考虑是第三条和第四条约束条件的折中。

**6.3** **基于邻域粗糙集的实体分辨记录对划分**

6.2节首先利用云模型计算各记录相似度对应相似或不相似概念的置信度， 根据置信度求出相似记录对和不相似记录对的记录相似度阈值，并认为大于相似 记录对的记录相似度阈值，则为相似；小于不相似记录对的记录相似度阈值，则为 不相似；二者之间的则既可能为相似也可能为不相似，对应记录对为难分辨的记录 对。但是云模型假设了记录相似度呈近似正态分布，并且其仅通过记录相似度来 判断记录对是否相似，仅能发现线性可分或不可分的情况，不能发现非线性可分或 不可分的情况。

Zhou 等(2015)假设记录相似度呈正态分布，并利用三倍方差选择难分辨的记 录对，即记录相似度大于相似记录对的记录相似度的均值减去三倍方差，则认为相

144)数据质量导论

似，记录相似度小于不相似记录对的记录相似度的均值加上三倍方差，则认为不相 似。这种方法的好处是实现简单，但是缺点不仅和6.2节的方法一样，而且记录相 似度呈正态分布的假设还更为严格。

粗糙集中的边界域(Pawlak,1982) 表示根据已有知识，不能明确属于哪一类的 样本，它不需要假设样本的分布，且所求出的难以分类的样本与学习算法无关，效 率较高，因此本节研究粗糙集对记录对进行划分。

由于实体分辨中，字段相似度为[0,1]区间的连续值，因此采用Hu 等(2008) 提出的邻域粗糙集查找难分辨的记录对。

6.3.1 邻域粗糙集

粗糙集用一个四元组S=(U,A,V,f) 表示一个信息系统，其中U是一个称为 论域的非空有限对象集合，A 是一个非空有限属性集合，V=Uac₄V。,V。是属性a 的 值域，f:U×A→V 是一个从论域U, 属性集合A 到值域 V的信息函数。当A= CUD,C 是条件属性集合，D 是决策属性集合时，该信息系统称为决策信息系统 (Liang et al.,2014)。

对于离散值，每一个非空子集B 都决定了一个不可分辨关系，R={( x,y)∈ U×Ulf(x,a)=f(y,a),Va∈B}, 其 中U是论域，B 是属性集合，x,y 是论域上的对 象，不可分辨关系是粗糙集理论的基础；对于连续值，利用邻域关系将不可分辨 关系扩展为：Rg={(x,y)∈U×Ul△⁴(x,y)≤δ,Va∈B}, 其 中B 是属性集合，△ 是距离函数，δ是距离阈值。应用邻域关系的决策信息系统称为邻域决策信 息系统。

给定邻域决策信息系统S=(U,CUD,V,f), 令 X₁,X₂,…,Xw 为对应决策1到 N 的对象集合，δ(x₁)={y ∈Ul△°(x,y)≤ δ,Va∈B}为 x;在特征空间BCC 上的邻 域对象集合，D相对于属性集合B 的下近似集和上近似集的定义如下(Hu et al.,

2008):

N₆D=U\*NnX,NaD=U(₁NnX

其中

NX₁={x₁lδ(x)∩X₁≠②,x;εU},NnX₁={x;lδ(x)CX,x;∈U}

D 相对于属性集合B 的决策边界域定义为(Hu et al.,2008):

*BN(D)=N₁D-N₄D*

(6-8)

(6-9)

(6-10)

该定义对于噪声比较敏感，因此，Ziarko等(1993)提出了变精度粗糙集，它引 入包含度对邻域粗糙集进行泛化。给定论域中的两个集合A、B,包含度的定义

第6章基于关系的实体分辨(145

如下；

 (6-11)

式中：A≠0。

此时，X 的上下近似的定义为

N\*X={x;II(δ(x₁),X)≥k,x;∈U},N\*X={x;II(8(x₁),X)≥1-k,x;∈U}

(6-12)

式中：k 称为包含度阈值，取值范围为0.5≤k≤1。

6.3.2 基于邻域粗糙集的记录对划分

利用变精度邻域粗糙集求解记录对的边界域，并认为处于边界域的记录对为 边界记录对，对应难分辨的记录对；余下的记录对为正常记录对，对应易分辨的记 录对。其算法流程如下：

输入：匹配记录对M 和不匹配记录对U, 包含度阈值k, 距离阈值δ。

步骤1:计算匹配记录对M 和不匹配记录对U 的字段相似度向量X₁和X₂。

步骤2:根据式(6-9)分别计算X₁ 和 X, 的上下近似集NgX₁ 、NgX₁ 、NX, 和

NgX₂₀

步骤3:根据式(6-8)计算全体记录对的上下近似集N₄D 和N₁D。

步骤4:根据式(6-10)计算边界域，以及处于边界域的记录对BR, 余下为正 常域内的记录对NR。

输出：BR 和NR。

算法过程即是首先分别求出相似记录对和不相似记录对的上下近似集，再得 到全体记录对的上下近似集，最后根据定义，求出边界域。

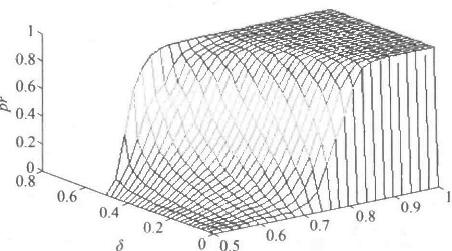
变精度和邻域关系的引入将导致找出的边界域的记录对为难分辨的记录对的 置信度下降。对于变精度粗糙集，包含度阈值k 的影响较大，k 越大，正常域里的 记录对属于易分辨的记录对的置信度越高，但是将使得边界记录对占比pr 过大， 其中pr= 边界样本数/总样本数。

另一方面，邻域关系对置信度影响也较大。当距离阈值δ=0时，仅完全相同 的记录对被判定为相似，因此所有记录对均为正常记录对，此时两个记录对同为相 似或不相似的置信度最高，pr=0。 距离阈值δ越小，置信度越大，但是也将导致 pr 过大。

为说明包含度阈值k 和距离阈值δ与边界记录对占比pr 的关系，以 Papadakis 等(2014)的 amazon\_gp 数据集为例。使δ在0～0.5之间以0.02的步长进行变

146)数据质量导论

化，k 在0.5~1之间以0.02的步长进行变化，计算相应的pr, 得出如图6-6所示 的关系。



k

图6-6 边界记录对占比pr 和包含度阈值k 及距离阈值δ的关系

从图6- 6可知，pr 与 k 呈递增关系，且随着δ增加，pr 相对k 增长更快。然 而，并不是所有的记录对都呈现严格递增关系，部分记录对存在波动。

综合来看，k 趋于1,δ趋于0时，利用粗糙集查找出的边界记录对确实 为难分辨记录对的置信度最高。如果记录对中难分辨记录对过多，可以选 择将δ设置为尽量小；如果为了使求出的边界记录对不要太多，可以将k 设 置较小。

**6.3.3** **实验分析**

本书采用Papadakis 等(2014)中用到的 dblp\_acm 、abt\_buy 和 amazon\_gp 数据 集。对数值型字段，利用sim(a,b)=abs(a-b)/max{(a,b)} 计算相似度，对字符 型字段，利用Jaccard 相似度(Xiao et al.,2011)计算相似度，将相似记录对的类标 设为1,不相似记录对的类标设为0,根据各数据中的pr 和k 及δ的关系，选择合适 的 k 和δ。

为比较记录对划分对实体分辨的影响，将记录对分为正常记录对、边界记录对 和原始记录对，比较各组记录对在利用相似度阈值分类和利用KNN分类时的准确 性，以说明记录对划分的有效性。利用相似度阈值分类，即认为相似度大于该阈值 则为相似，小于该阈值则为不相似。利用KNN分类，即利用KNN算法将记录对分 为相似和不相似两类。

实验环境为1台i7-4790 CPU,4GB内存的PC, 实验平台为MATLAB 7.0。

[**6.3.3.1**](6.3.3.1) **记录对划分对记录相似度阈值的影响**

为比较记录对划分对利用相似度阈值分类的影响，取各字段相似度的均值，得

第6章 基于关系的实体分辩 (47

到记录的相似度。

利用MATLAB 画出正常记录对、边界记录对和原始记录对中的相似记录对和 不相似记录对的记录相似度的散点图，以直观、定性地表示记录对划分对用相似度 阈值进行分类的影响。

测试记录对的参数及对应的pr 如表6-5所列。

表6- 5 测试记录对的参数及对应的pr

|  |  |  |  |
| --- | --- | --- | --- |
| **数据集** | k | δ | pr |
| dblp\_acm | 0.94 | 0.12 | 0.0814 |
| abt\_buy | 0.82 | 0.02 | 0.2345 |
| amazon\_gP | 0.88 | 0.04 | 0.0906 |

从图6-7至图6-9可知，尽管由于变精度阈值的存在，位于正常域的记录 对、相似记录对和不相似记录对的记录相似度存在部分交叉，但是相比原始记录 对，存在交叉的记录对已经大幅减少，更有利于用相似度阈值进行分类；而边界域 的记录对存在交叉的变多，更难用相似度阈值进行分类。

[**6.3.3.2**](6.3.3.2) **记录对划分对分类准确性的影响**

对正常记录对、边界记录对和原始记录对，分别利用KNN分类器运行10轮5 折交叉检验验证分类的准确性，取K=5。

对各记录对，求出分类正确率的均值和标准差，并采用置信度为0.05的双边 t 检验，将正常记录对、边界记录对与原始记录对进行对比，若结果明显好，则用 “●”标记；若结果明显差，则用“O” 标记，各数据最好结果加粗表示，最后一行表 示 win/tie/loss统计结果。准确性如表6-6所列。

表6 - 6 分类正确率比较

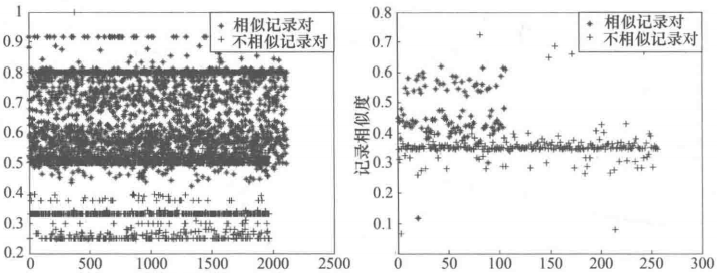
|  |  |  |  |
| --- | --- | --- | --- |
| **数据集** | **正常记录对** | **边界记录对** | **原始记录对** |
| dblp\_acm | **0.9995±0.0006** | 0.7571±0.19110 | 0.9978±0.0012 |
| abt\_buy | **0.9647±0.0129●** | 0.8087±0.0222O | 0.9030±0.0159 |
| amazon\_gp | **0.9891±0.0058●** | 0.7400±0.0611O | 0.9672±0.0070 |
| win/tie/loss | **2/1/0** | 0/0/3 |  |

从表6-6可知，正常记录对相比原始记录对，分类正确率得到了提高，而边界 记录对相比原始记录对，分类正确率显著降低，进一步说明了记录对划分的有 效性。

**148)**数据质量导论

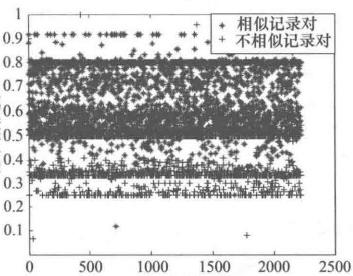
记录相似度

记录相似度



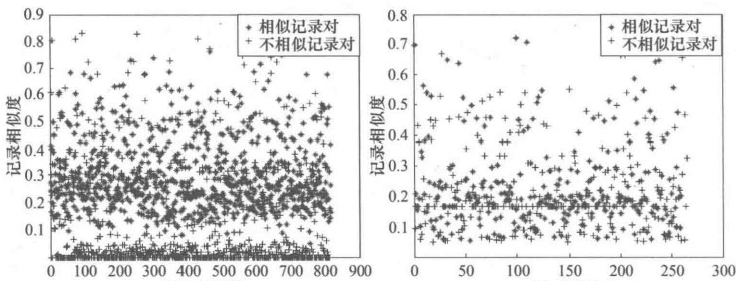
记录对编号 记录对编号

(a) 正常记录对 (b) 边界记录对

记录对编号

(c)原始记录对

图 6 - 7 dblp\_acm 的记录相似度比较



记录对编号 记录对编号

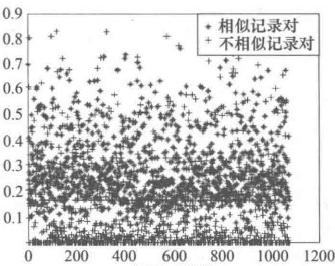
(a)正常记录对 (b)边界记录对

第6章基于关系的实体分辨 

记录相似度

记录相似度

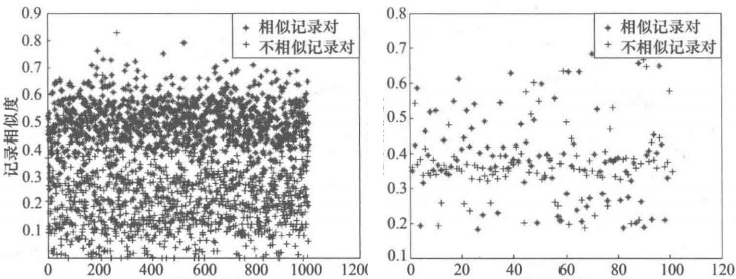
记录相似度



记录对编号

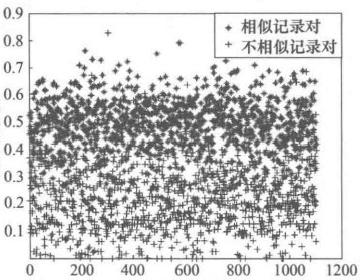
(c)原始记录对

图 6-8 abt\_buy 的记录相似度比较



记录对编号 记录对编号

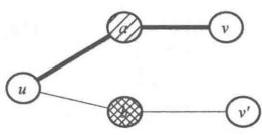
(a)正常记录对 (b)边界记录对



记录对编号

(c)原始记录对

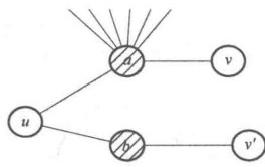
图 6 - 9 amazon\_gp 的记录相似度比较

150)数据质量导论

**6.4** **基于关系类型的自适应实体分辨**

在计算实体关系图中任意两个结点u,v 之间的连接强度时，通常认为两个结 点之间的路径越多，路径的长度越短，它们的连接强度越大。两个结点间的连接强 度的计算一般分为三个步骤：①寻找结点u,v 之间的所有L-short 路径R₁;② 计算 u,v之间的每条路径r的连接强度cs(r);③计算u,v之间的连接强度cs(u,v)。

影响单路径连接强度的因素主要有两个方面， 一是关系类型对于连接强度的 重要性，二是与同一个实体关联的同类型的关系数量。在实体关系图中，这两个因 素分别对应不同类型边的权重和同一结点上相同类型边的数目。为了说明这两个 因素对单路径连接强度的影响，图6-10给出一个实体关系图的部分示例。



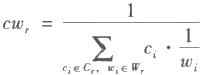
(a)关系类型 (b)关系数量

图6-10 影响单路径连接强度的两个因素

如图6-10(a) 所示，结点u 分别通过两条路径与结点v 和 v'相连，这两条路 径都只包含一个中间结点和两条边，但u-a-v 路径中的粗线边表示该路径中的 关系类型对于连接强度而言更为重要，因此，图6-10(a) 中 u,v 之间的连接强度 应该大于u,v'之间的连接强度。图6-10(b) 描述了所有的边都属于同一类型的 情况，图中u-a-v 路径的中间结点a 连接了更多的实体，而u-b-v '路径的中间 结点b 只连接了u 和v'两个结点，显然，在计算连接强度时，u,v 之间的连接强度应 该大于u,v'之间的连接强度。基于此，将一条路径的连接强度分为两个部分：路径 权重和路径概率。路径权重表示路径r 对于结点u,v 之间连接强度的重要程度， 而路径概率指的是由结点u 通过路径r 与结点v 连接的概率。

**6.4.1** **路径权重**

路径权重指的是一条路径r 对于其连接的节点u,v 之间的连接强度的重要程 度。Nuray-Turan 等(2013)提出了一种基于路径类型的权重设置方法(PTM 模 型),每条路径的类型由其包含的边对应的关系类型序列表示。对于一个具有n

第6章基于关系的实体分辨(

种关系类型的数据集，其路径类型有(n¹-1)/(n-1) 种之多(l 为结点间最长路径 的长度),难以采用人工方式对所有路径类型进行赋权，因此Nuray-Turan 等 提出了一种从训练数据中学习各类路径的权重的方法，然而实际应用中由于 无法保证训练数据中包含所有类型的路径，难以得到一个完备的路径类型权 重集合。本节提出一种关系类型模型(Realationship Type Model,RTM),认为 任何一条路径的权重由其包含的边对应的关系类型决定，而与其排列顺序无 关(实验中将进行验证),只要确定了所有关系类型的权重，便可以确定任何 一条路径的权重。

对于一个给定的数据集D, 及其实体关系图G, 数据集中的关系类型集合(即 实体关系图G 中的边类型集合)记为S={s₁,S₂,…,s₄}, 每种关系类型s;对应一个 权重w, 表示该类型的关系对于连接强度的重要程度，所有的关系类型权重组成 集合W={w₁,w₂,…,w 。} 。对于实体关系图中结点u,v 之间的任意一条路径r, 其 包含的关系类型记为S,={s₁,s₂,…,sm},m=1,2,…,n, 对应的类型权重集合记为 W,={w₁,w₂,…,w,}, SrCS,WrCW,路径r 上类型为s, 的边的数目记为c, 所有类 型边的数目组成集合C,={c₁,c₂,…,cm}, 则路径r 权重用下式计算：

(6-13)

式(6-13)中，类型为s;的关系类型权重的倒数1/w;称为关系类型s;的连阻系数， 路径r 的连阻系数按照类似于求串联电路电阻的方式计算，路径r 的权重为r 的连 阻系数的倒数。由式(6-13)可见，对于给定的数据集，只要能够确定所有关系类 型的权重W,就能够通过计算得到任何一条路径权重。通常情况下关系类型的权 重可以由领域专家给出，然而随着互联网技术的发展，数据集中包含的关系类型可 能涉及多个领域，很难由一个或几个专家确定它们的权重，而且对于不同的领域， 相同的关系类型可能有着不同的权重，本节将在6.4.4节提出一种基于有监督学 习的权重学习方法。

6.4.2 路径概率

路径r 的路径概率取决于路径中各个结点上连接的边的数目，Nuray-Turan 等(2013)和Li(2010) 采用Random Walk 算法计算结点u 通过路径r 到结点v 的路 径概率。如图6-11所示的路径，按照 Random Walk 算法路径u-a-v 的路径概 率为结点u 到结点a 的概率与结点a 到结点v 的概率的乘积。结点u 有2条边与 它相连，因此结点u 到 a 的概率p(u,a)=1/2, 同理p(a,v)=1/8, 则路径u-a-v

的路径概率Pa-a=p(u,a)·p(a,v)=1/16。

)数据质量导论

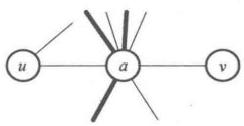


图6-11 关系类型对路径概率的影响示意图

这种方法根据当前结点的度计算两个相邻结点的路径概率，没有考虑关系类 型的区别。在图6-11中，结点a 的度为8,但这8条边分别隶属于两种不同的关 系类型，其中3条边与结点a 到结点v 的边不属同一类型。不同类型的边通常连 接不同类型的结点，即使连接相同类型的节点，也具有不同的物理含义，因此当计 算结点a 到结点v 的概率的时候不应该考虑与边a-v 不同类型的边，即P-a-=

p(u,a)·p(a,v)=1/2·1/5=1/10 。 基于此，提出一种区分关系类型的路径概率 计算方法。

对于实体关系图G 中的两个结点u 和v,r为连接u,v 的一条路径，A=<a₁,

a₂,…,a₄> 为路径r 上除u,p 结点之外的结点序列，S,={s₁,s₂,…,s} 为路径r 包 含的关系类型集合，记路径r 中以结点a,i=1,2,…,n 为起点以结点a 或 v 为 终点的边的关系类型为se,Sa∈S,, 与结点a; 连接的类型为s 的边的数目记为

Num(a₁,s₀₁),则结点u,v 间路径r 的路径概率按式(6-14)计算：



(6-14)

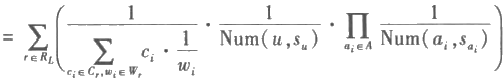
6.4.3 连接强度

得到路径r 的路径权重与路径概率后，可以由多种方法将两者进行融合，采用 式(6-15)计算路径r 的连接强度cs(r):

cs(r)=cw,·P, (6-15)

结点u,v 之间的连接强度为u,v 间所有路径的连接强度之和。





(6-16)

由式(6-16)可知，对于由不重要的关系组成的路径r, 其路径权重值相对较 小，因此减少了不重要的路径对于连接强度的干扰，同时，路径概率考虑了路径中 结点连接的同类型关系数目对连接强度的影响。

第6章基于关系的实体分辨 

6.4.4 自适应关系类型权重学习

6.4节的前三小节提出了基于关系类型的链接强度模型，但其中关系类型的 权重仍然需要确定。实际应用中并不总是有领域专家参与，而且对于不同的应用 领域，同一种关系类型可能具有不同的权重，因此，如何获得不同应用领域中关系 类型的权重仍然是需要解决的问题。本节提出一种有监督的自适应关系类型权重 学习模型。在实际应用中，要确保训练数据包含所有路径类型是十分困难的，但确 保训练数据中包含所有关系类型则相对容易，因此，这种方法通常都可以获得所有 关系类型的权重。

当给定一个训练数据时，其中包含的关系类型为S={s₁,s₂,…,s₄}, 权重学 习的目的是利用训练数据学习这些关系类型对应的权重 W={w₁ ,w₂ ,…,w,}。

训练数据中的任一实体描述d 对应的实体o, 是已知的，由CAP原则可知，d 到实 体o, 的连接强度应该明显大于d 到候选实体集O₄ 中其他候选实体的连接强 度，即

cs(d,o₁)>cs(d,oa),d∈D,oa∈O₄-{o₁ (6-17)

使用δ-band 方法可以描述为

cs(d,o₁).-cs(d,oa)>δ,d∈D,0m∈O₄-{o₁} (6-18)

然而，由于CAP原则是一个理想化的原则，实际应用中式(6-18)的约束过于 严格，并不一定总是成立的，因此，向式(6-18)中引入松弛因子，并使得该松弛因 子的和最小，则可以得到



s.t. cs(d,o₁)+5a-cs(d,oa)>δ,deD,oa∈O₄-1o₁|(6-19)

5≥0

1≥w;≥0

式(6-19)始终可以成立，其中ξ是引入的松弛因子，此处的目标是寻找合适 的关系类型权重集合W,使得松弛因子之和最小。由于式(6-13)计算路径连接 强度时需要对关系类型权重的倒数求和，因此，式(6-19)中的问题是非线性规划 问题，可以利用许多现有的方法进行求解。

值得注意的是，式(6-19)中的δ虽然需要领域专家进行设定，但该方 法对于δ的依赖并不大，δ只是一个常数，如果δ的值设定得比较大，则得到 的关系类型权重 w;普遍较大，如果δ的值设定得比较小，则得到的w;普遍 较小。

)数据质量导论

所占比例/%

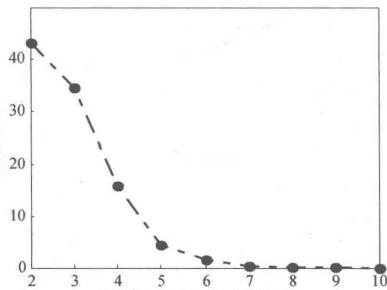
**6.4.5** **实验分析**

为了验证本章提出的方法，在真实数据集上对比了五种不同的实体分辨算法。

[**6.4.5.1**](6.4.5.1) **实验设置**

在本实验中，使用真实数据集 Stanford Movies Dataset(记为 MovData)作为实 验数据集，并依据不同实验方案的需要对该数据集进行调整和修改。MovData 数 据集包含三种不同的实体，分别是movies(11453 个实体)、studios(992个实体)、 people(22121 个实体),这些实体之间存在着五种不同的关系，分别是 actors、di- rectors、producers、producing studios 和 distributing studios。

MovData 数据集是一个干净的数据集，每一条记录都对应着一个确定的实体， 而实体分辨实验需要数据中包含不确定的相似记录，因此，需要向数据集中引入相 似记录。采用人工的方式向数据集中引入不确定性，这种方法已经得到诸多研究 人员的认可和广泛使用(Nuray-Turan et al.,2013;Li,2010)。在向数据集中引入 不确定性时，有2个参数可以进行控制调整， 一个是引入不确定记录的百分比，记 为f,O≤f≤100%; 另一个是引入的不确定记录对应的候选实体集中的实体数目， 记为c,c=1,2,3, … 。 参数f 通过人工设定为各种不同的值，参数c 可以通过人工 设定为固定值或者令其服从如图6-12所示的概率分布。



候选实体集中的记录数目/条

图6-12 候选实体集的概率分布

[**6.4.5.2**](6.4.5.2) **实验方法及评价测度**

为了检验基于关系类型的路径权重模型能否降低对训练数据的要求，对比基 于路径类型的路径权重模型(PTM) 与6.4.1节中的基于关系类型的路径权重模型 (RTM); 为了验证区分关系类型对 Random Walk 模型计算路径概率的准确性的影 响，对比 RandomWalk 模型(RW) 与6.4.2节中区分关系类型的 Random Walk 模型

第6章 基于关系的实体分辨 (**155)**

正确率/%

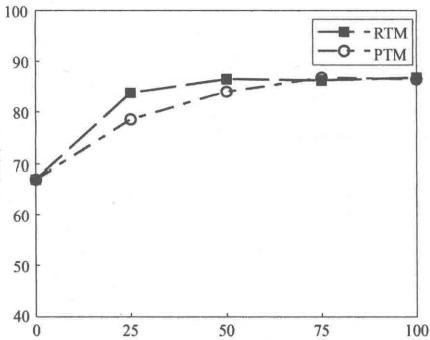
(RW\_T) 的准确性；为了验证融合基于关系类型的路径权重和路径概率的连接强 度对于实体分辨准确性的影响，对比6.4.3节中的连接强度模型(RTM+RW\_T)

与目前效果最好的基于关系的连接强度模型(PTM+RW)(Nuray -Turan et al., 2013)和基于上下文的实体分辨方法(简称 CB(Context Based)方法)的实体分辨准 确性(Bhattacharya et al.,2005)。

本实验主要关注实体分辨方法的准确性，用正确率作为评价实体分辨方法准 确性的量化指标，该指标定义为正确识别的记录占参与识别记录总数的百分比。 各个算法的效率不是关注的重点，本实验不详细对比各个方法的运行效率，不过， 在实验过程中发现 RTM 模型方法在训练过程中的耗时明显小于PTM模型，而在 实体分辨实验过程的耗时高于PTM模型。

[**6.4.5.3**](6.4.5.3) **路径权重模型验证**

如前文所述，PTM 模型必须学习全部路径类型的权重，对训练数据要求较高， 当实际数据中出现训练数据中没有的路径类型时，便有可能会影响该方法的准确 性；RTM模型只需要学习关系类型权重，对训练数据要求较低。为了验证训练数 据对PTM模型和RTM模型的影响，本实验通过向 MovData 引入不确定性构建两 个数据集，其中一个数据集作为训练数据，另一个数据集作为测试数据，这两个数 据集具有相同的不确定性参数f=25%,c=2。 通过选择不同比例的训练数据对 PTM模型和RTM模型进行训练，然后只利用路径权重进行实体分辨后对比两种 方法在相同测试数据上结果的正确率。实验结果如图6-13所示。



训练数据的比例/%

图6-13 训练数据对 PTM 模型与 RTM 模型的影响

从图6-13中可以看出，在没有训练数据的情况下(训练数据的比例为0),所 有的关系和路径都被看作属于同一类型，PTM 模型和 RTM 模型的正确率几乎相

**156)**数据质量导论

同，而且该情况下两种方法得到的结果的正确率是整个实验中最差的结果。随着 训练数据所占比例的逐渐提高，RTM 模型正确率提高的速度明显快于PTM 模型。 当训练数据比例达到30%左右时，RTM 方法便可以达到其最高正确率，而PTM模 型则需要训练数据大于75%才能达到。当训练数据的比例达到100%时，两个模 型都达到最高正确率，而且能达到的最高正确率相差无几。这说明在能达到最高 正确率的情况下，基于路径类型的PTM 模型和基于关系类型的 RTM 模型在正确 率上几乎没有差别，但就对训练数据的要求而言，RTM 模型更容易达到最高正 确率。

为了进一步分析导致这种现象的原因，表6-7中对比了在各个不同的训练数 据比例情况下PTM模型学习到的路径类型的数量和RTM 模型得到的关系类型的 数量。

表 6 - 7 PTM和 RTM方法中的路径类型和关系类型数目

|  |  |  |
| --- | --- | --- |
| **训练数据比例** | **PTM中路径类型数目** | **RTM中关系类型数目** |
| 0.00 | 0 | 0 |
| 0.25 | 339 | 5 |
| 0.50 | 417 | 5 |
| 0.75 | 431 | 5 |
| 1.00 | 431 | 5 |

如表6-7所列，当选用的训练数据所占比例是25%时，PTM 模型能够从中学 习得到339个路径类型的权重；当这个比例增加到50%时，可以得到417个路径 类型的权重；当训练数据的比例高于75%时，这个数目增加到431个，这个数字可 能是数据集中所有路径类型的数目。由于PTM 模型需要所有路径类型的权重才 能准确地计算两个节点间的连接强度，训练数据的比例至少要超过75%才能达到 其应有的准确度。在实际应用中，获得足够量的训练数据都是非常困难的，而确保 训练数据包含PTM 模型需要的所有路径类型是几乎不可能的。对于RTM 方法 而言，其对训练数据的要求是包含的所有关系类型，数量远远少于路径类型。 表6-7中显示，当训练数据所占的比例达到25%时，便能够得到数据集中全 部五种关系类型的权重。实际上，仅仅是获得五种关系类型的权重所需的训 练数据比例远小于25%,但要得到合适的关系类型权重，则需要30%左右的 训练数据。

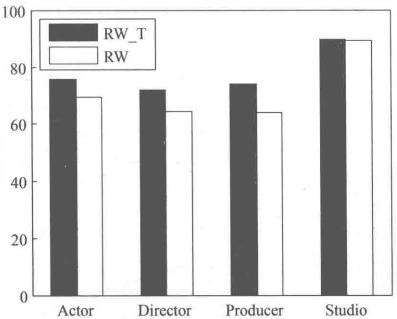
将表6-7与图6-13放在一起观察，可以发现表6-7中的结果与图6-13 中的曲线基本可以匹配，这说明路径类型的数量是影响 PTM模型正确率的重要因 素。RTM模型与 PTM模型相比，需要的关系类型数量远远小于路径类型数量，对 训练数据的数量和满足条件的约束更加宽松。

第6章 基于关系的实体分辨 

正确率/%

[**6.4.5.4**](6.4.5.4) **路径概率模型验证**

分别利用 Random Walk 模 型(RW) 和区分关系类型的 Random Walk 模型 (RW\_T) 计算路径概率，并只利用路径概率对 MovData 数据集中的不同实体(Ac- tor,Director,Producer 和 Studio) 进行基于关系的实体分辨，验证区分关系类型对于 路径概率正确率的影响。在向数据集中引入不确定性时设置c 服从图6- 12中的 概率分布，f=50% 。 实验结果如图6-14所示。



实体类型

图6-14 区分关系类型对于 Random Walk 模型正确率的影响

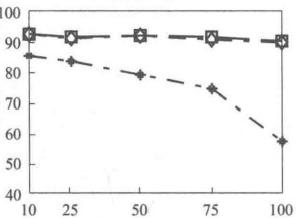
从图6-14可以看出，对于Actor、Director 和 Producer 三个类型的实体，区分 关系类型的 RW\_T模型得到的结果正确率明显高于不区分关系类型的 RW模型， 对于Studio 实体而言，两者区分并不明显。总体而言，可以认为利用区分关系类型 的 Random Walk 模型得到的路径概率可以更加准确地反映实体关系图中两个结 点的连接概率。

[**6.4.5.5**](6.4.5.5) **连接强度模型验证**

在本实验中，将提出的连接强度模型(RTM+RW) 与目前效果最好的基于关 系的连接强度模型(PTM+RW) 和基于上下文的实体分辨方法(CB) 进行对比。利 用不同的方法对MovData 数据集中不同实体(Actor,Director,Producer 和 Studio) 进 行识别，对比识别结果的正确率。

与<6.4.5.4>节类似，基于MovData 构建训练和测试数据集，设置c 服从图6 - 12中的概率分布，选择5个不同的相似记录比例，即f={10%,25%,50%,75%,

100%},对于每个f 的取值分别构建一个训练数据集和测试数据集。为了保证所有 的方法都能达到其最高准确度，根据<6.4.5.3>节中的分析，基于路径类型的 PTM+ RW 模型使用全部的训练数据集作为训练数据，基于关系类型的RTM+RW\_T 模 型使用30%的训练数据集作为训练数据，CB 方法无需进行学习，不需要训练数

**158)**数据质量导论

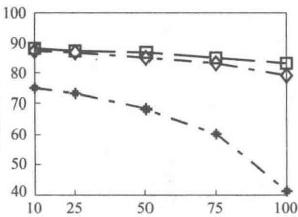
正确率/%

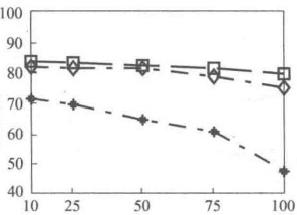
正确率/%

正确率/%

正确率/%

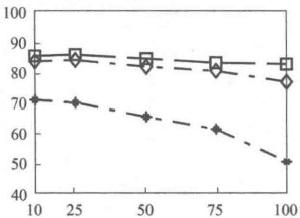
据。对于每一个f 值，所有的方法均在相同的测试数据集上进行实验，实验结果如 图6- 15所示。





相似记录所占比例/% 相似记录所占比例/%

(a)actor (b)director



相似记录所占比例/% 相似记录所占比例/%

(c)producer (d)studio

|  |
| --- |
| —日 -RTM+RW T  —— ◇-PTM+RW  一 \* -CB |

图6-15 三种实体分辨方法的正确率

图6-15对比了在不同的f 值下，不同的方法对 MovData 数据集中不同类型 实体进行识别的正确率。可以看出，图6-15(a) 、图6 - 15(b) 和图6- 15(c) 的结 果比较类似。在识别 actor,director 和 producer 这三种实体时，基于关系的方法 (RTM+RW\_T,PTM+RW) 的正确率均明显优于基于上下文的方法(CB) 。RTM+ RW\_T 方法在各个f 值下的识别正确率都是最高的，PTM\_RW 方法仅次于RTM+

RW\_T 方法。而值得注意的是，RTM+RW\_T 方法只使用了30%的训练数据进行 学习。这说明 RTM+RW\_T 方法与PTM+RW 方法相比在更少的训练数据的情况 下取得了更高的识别正确率。基于上下文的方法(CB) 在三种方法中正确率最差， 而且它的正确率随着f 值的增加而下降的速度明显快于另外两种方法。这主要是 由于RTM+RW\_T 和 PTM+RW 方法在不同相似记录比例的情况下都进行权重学 习，从而能够更好地适应不同的数据环境；而CB方法随着相似记录所占比例的增

第6章基于关系的实体分辨((**159**)

加，其利用的上下文信息的正确率随之降低，从而降低了 CB 方法的正确率。 图 6 - 1 5(d) 中显示了所有方法识别 studio 实体的结果，与前三个图区别较大，对 于 studio 实体而言，两种基于关系的方法得到的正确率几乎相同，且明显优于基于 上下文的 CB 方法。

**6.5** **本章小结**

本章为实现从 FBS 方法到 RelDC 的转换，考虑常见的两大类 FBS 方法，针对 利用记录相似度进行分辨的情况，利用云模型求解各记录相似度相对相似或不相 似概念的置信度，求解出两个相似度值，并根据这两个相似度值将记录对分为易分 辨和难分辨两种情况；针对利用学习算法进行分辨的情况，利用变精度邻域粗糙集 对实体分辨记录对进行划分，为对易分辨的记录对和难分辨的记录对分别应用不 同的分辨方法提供基础。

针对基于关系的实体分辨在计算路径权重过程中对训练数据的数量和必须满 足的条件要求过于严格的问题，提出一种基于关系类型的自适应实体分辨方法。

**参考文献**

[1]李德毅，杜鹚.2005.不确定性人工智能[M]. 北京：国防工业出版社.

[2]邸凯昌，李德毅，李德仁.1999.云理论及其在空间数据发掘和知识现中的应用[J]. 中国图象图形学 报，4(11):930-935.

[3]秦昆，李德毅，许凯.2006.基于云模型的图像分割方法研究[J]. 测绘信息与工程，31(5):3-5.

[4]曹建军，刁兴春.2010a 基于特征选择的相似重复记录分类检测及其蚁群算法实现[J]. 兵工学报，31 (9):1222-1226.

[5]曹建军，刁兴春，汪挺，等.2010b.领域无关记录对清洗研究综述[J]. 计算机科学，37(5):26-29.

[6]Ananthakrishna R,Chaudhuri S,Ganti V.2002.Eliminating Fuzry Duplicates in Data Warehouse[C]//Proceed-

ings of the VLDB Conference.Hong Kong.

[7]Bhattacharya I,Getoor L.2005.Relational Clustering for Multi-type Entity Resolution[J].Mrdm Workshop, 4(6):3-12.

[8]HuQH,Yu D R,Liu J F,et al.2008.Neighborhood Rough Set based Heterogeneous feature Subset Selection [J].Information Sciences,178(18):3577-3594.

[9]Kalashnikov D V,Chen Z,Mehrotra S,et al.2008.Web People Search via Connection Analysis[J].IEEE Trans- actions on Knowledge &Data Engineering,20(11):1550 -1565.

[10]Kalashnikov D V,Mehrotra S.2003.Exploiting Relationships for Data Cleaning[R].Newport Beach,Califor- nia,USA:SIAM:262-273.

[11]Li P.2010.Multiple Relationship Based Deduplication[C]//Proceedings of the 4th SIGMOD PhD Workshop on Innovative Database Research.New York,USA:ACM:25-30.

[12]LiangJY,Wang F,Dang C Y,et al.2014.A Group Incremental Approach to Feature Selection Applying Rough

**160)**数据质量导论

Set Technique[J].IEEE Transactions on Knowledge and Data Engineering,26(2):294-308.

[13]Nuray-Turan R,Kalashnikov D V,Mehrotra S.2013.Adaptive Connection Strength Models for Relationship-

Based Entity Resolution[J].Jourmal of Data and Information Quality(JDIQ),4(2):8.

[14]Papadakis G,Koutrika G,Palpanas T,et al.2014.Meta-Blocking:Taking Entiy Resolution to the Next Level [J].IEEE Transactions on Knowledge and Date Engineering,26(8):1946-1960.

[15]Pawlak Z.1982.Rough Sets[J].International Journal of Computer and Information Sciences,11(5):341 -356.

[16]Whang S E,Garcia-Molina H.2012.Joint Entity Resolution[C]//Proceedings of the 28th International Con- ference on Data Engineering(ICDE).Washington,USA:IEEE:294-305.

[17]Whang S E,Garcia-Molina H.2013.Joint Entity Resolution on Multiple Datasets[J].The International Jour- nal on Very Large Data Bases,22(6):773-795.

[18]Xiao C,Wang W,Lin X M,et al.2011.Eficient Similarity Joins for Near -Duplicate Detection[J].ACM

Transactions on Database Systems,36(3):15.

[19]Zhou X,Diao X C,CaoJJ.2014.A Data Cleaning Switch Technology Based on Cloud Model[C]//Internation-

al Conference on Information Quality.Xi'an,China:ACM:44-51.

[20]Zhou X,Diao X C,CaoJJ.2015.A High Accurate Multiple Classifier System for Entity Resolution Using Resa- mpling and Ensemble Selection[J].Mathematical Problems in Engineering,2015(2):1-6.

[21]Ziarko W.1993.Variable Precision Rough Sets Model[J].Journal of Computer and System Sciences,46(1): 39 -59.

**第** **7** **章** **不完整数据的分类与检测**

**7.1** **引言**

Wand等(1996)提出了真实世界系统到信息系统状态映射的四个问题：数 据是否完整，数据是否明确，数据是否有意义，数据是否正确。其中数据完整 性是数据质量的重要维度之一，并且在很多情况下，完整性会直接影响其他 维度。

知识的不完备性是知识不确定性的重要表现，而知识内容的不完整是知识不 完备性的具体体现之一(李德毅，等，2005);在数据挖掘、数据仓库和数据质量管 理中，缺失数据是影响数据质量的重要原因之一。

实际的数据经常含缺失值，当用来做报表、信息共享和决策支持时，缺 失值的存在可能导致严重问题(Li et al.,2009):首先，含缺失值的数据可能 提供不完整信息，例如，有关个人问题的调查问卷，其中的一些涉及隐私的 敏感问题可能得不到回答而造成对象描述不完整；其次，许多数据建模和数 据分析技术不能处理缺失值，从而不得不将仅缺失一个属性值的记录丢弃 (Michie et al.,1994;SAS Institute,1990),造成数据资源的浪费；再次，尽管 某些数据建模和分析工具能够处理缺失数据值，但在某些领域，缺失值是被 禁止的，例如，分类系统不允许分类属性值缺失(Breiman et al.,1984;Quinlan, 1993)。

从数据清洗的角度，缺失数据处理包括缺失数据的检测、分类和估计填充 等步骤，分别对应数据清洗的检测、分析和修正三类典型技术。在实施缺失数 据估计填充前，首先要完成缺失数据的检测和分类，以针对不同的缺失数据类 型选择合适的处理方法，当前这方面的研究尚不够深入。Chen 等(2003)将不 完整记录分为可用和不可用两种情况，但主要根据记录的字段缺失比来确定 记录是否可用，只适用于各属性重要程度相同的情况，并且没有考虑属性之间 的关系。

本章从不同角度提出两种不完整数据的分类与检测方法，通过对不完整数据 进行分类，提高对不完整数据处理的准确性和效率。

)数据质量导论

**7.2** **基于位运算的不完整数据分类与检测**

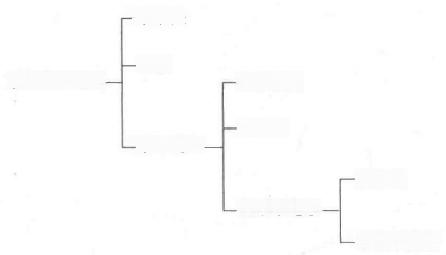
本节提出一种基于位运算的不完整数据分类与检测方法。

**7.2.1** **不完整数据及其分类**

在对不完整数据进行界定的基础上，给出了不完整记录的一种层次分类方法。

[**7.2.1.1**](7.2.1.1) **不完整数据的界定**

对一个关系数据集，数据集内容不完整的层次结构如图7-1所示。



—表缺失

-空表

-记录缺失

-空记录

—表不完整-

空字段

-记录不完整-

字段不完整

数据集不完整-

图7- 1 数据集内容不完整的层次结构

图7-1中，表缺失是指某表没有被创建；空表是指没有填记录的表；记录缺失 是指表中缺少记录；空记录是指记录只有主键等创建时生成的字段值，其余字段为 空字段；空字段是指字段值为空值的字段；字段不完整是指字段值内容不完整，如 中文姓名只填写一个中文字符等，此类情况不予讨论。

如图7-1所示，尽管数据集内容不完整包括多种情况，但是， 一般而言，对关 系数据集，字段是对关系数据集操作的基本单位，而记录由若干字段组成，是对基 本数据对象进行完整描述的基本单位，因此，关系数据集的不完整主要是指因字段 缺失造成的记录不完整。

将含缺失字段的记录称为不完整记录，不完整记录既包括空记录，又包括只有 一个字段缺失的记录。习惯上，不完整记录和空字段都称为缺失数据(Missing Da- ta), 空字段中未填写的属性值称为缺失值(Missing Value)。

[**7.2.1.2**](7.2.1.2) **记录的层次分类**

不完整记录检测首先应区分完整记录和不完整记录。将不完整记录进一步分 为不完整合格、不完整修正、不完整删除三类。记录分类简要描述见表7-1。

第7章不完整数据的分类与检测**163**

表 7 - 1 记录的分类及描述

|  |  |  |
| --- | --- | --- |
| **No.** | **名称** | **描述** |
| 1 | 完整记录 | 不含空字段的记录 |
| 2 | 不完整合格记录 | 含空字段但合格的记录 |
| 3 | 不完整修正记录 | 含空字段需要修正的记录 |
| 4 | 不完整删除记录 | 空记录，或失去修正价值的含空字段的记录 |

表7-1中，完整记录是指记录中没有空字段的记录。不完整合格记录，包括： 准完整记录，虽然记录中含空字段，但不影响对象描述的完整性，如未婚人员的配 偶信息相关字段为空；某些有价值字段不空，而空字段不影响当前应用，或不具有 修正价值，或当前无法修正的不完整记录。不完整修正记录，需对其中的某些空字 段进行估计填充的不完整记录。不完整删除记录，包括：空记录；无效记录，关键字 段为空无实际意义的不完整记录；不可修正空字段过多，无修正价值的记录。以上 三类不完整记录可以由专家针对具体数据集确定。

不完整记录检测是一个对记录进行层次分类的过程，其层次分类结构如 图7-2所示。

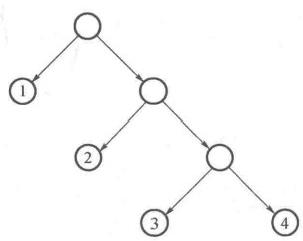


图7-2 记录的层次分类结构图

图7-2中的节点编号与表7-1 的分类编号相同。对图7-2中的不完整记 录做如下处理：对不完整合格记录的空字段做相应标识，如填充“NULL”(陈伟， 2004);对不完整修正记录，根据缺失字段的类型及选择合适的缺失值估计填充算 法进行修正(Li et al.,2009;SAS Institute,1990;Breiman et al.,1984;Quinlan,1993; Witten et al.,2005);对不完整删除记录，将该条记录删除(级联删除),注意此处的 删除是指逻辑删除，不是物理删除，数据清洗中所有对数据集的更改都必须有良好 的数据溯源和数据世系设计(Codd,1979), 以保证数据更改的可追溯性和数据的 可恢复性。

164)数据质量导论

7.2.2 记录的二进制表示

对不完整记录进行检测、分析和修正，需要对记录的不完整信息进行描述，本 节引入记录的二进制表示来描述记录的不完整信息。

假设某表有n个字段，m 条记录，记录r,i=1,2,…,m, 的字段记为ay,j=1, 2,…,n,则称二进制数B₁=b₁b₂…b, 为记录r;的二进制表示，b;的取值由式(7-1) 给出：

(7-1)

记录的二进制表示示例见表7-2。

表 7 - 2 不完整记录的二进制表示示例

|  |  |  |  |
| --- | --- | --- | --- |
| 记录 | | | **二进制表示** |
| 姓名 | 性别 | **年龄** |
| 王国庆 | 男 |  | 110 |
| 刘援朝 |  |  | 100 |

记录的二进制表示全面地描述了记录的完整性信息，b,=0 表示第i 条记录的 第j个字段为空，记录r;的缺失字段数由式(7-2)计算：



根据式(7-2)可以方便地求取记录r;的字段缺失比率(Chen,2003)。

(7-2)

7.2.3 不完整记录的位运算分类检测方法

不完整记录的位运算分类检测的实现，需要首先生成各类的标准记录二进制 表示集，然后通过将待检记录的二进制表示按图7-2的分类流程与各类的标准二 进制表示做位运算实现分类检测。

<7.2.3.1> 标准记录二进制表示集的生成

可以用两种方法生成各类的标准记录二进制表示集， 一种是直接生成法，根据 具体数据集直接生成，将2"个或所有可能出现的记录二进制表示分配到相应的 类，因实际数据集的不完整记录缺失字段具有一定分布规律，此方法在应用初期会 产生大量的冗余表示和冗余检测运算；另一种是样本生成法，根据实际的记录样本 生成标准记录二进制表示集，并根据记录二进制表示出现的次数确定该表示在相 应类中的检测优先级，此方法在应用初期可能出现不能识别二进制表示的情况，对

第7章不完整数据的分类与检测(165)

新出现的二进制表示需补充进相应的类标准二进制表示集，但此方法实现简单、效 率高。本节采用样本生成法生成各类的标准二进制表示集。

标准记录二进制样本集生成步骤如下：

(1)每一类的标准记录二进制表示集记为S,k=1,2,3,4,S₂ 中的二制表示记 为sμ,l=1,2 ,…,显然有

 (7-3) (2)对一个第k 类不完整记录样本的二进制表示B, 进行如下添加运算：

1.If S=φ

2. S={B};

3.Else

4. {For s

5. {If~xor(B₄,Sa)

6. {S=S₄+{B₄};

7. break;}

8. End if}

9. End for}

10.End if

第5行中xor 是异或运算符，第4～10行的功能是通过对 B₄ 和S 中的元素进 行异或运算，检测S,中是否已有Bα,若已有，B₄ 的出现次数加1,否则将B₄ 添加进 S, 并记其出现次数为1。

(3)对标准记录二进制表示集合S₄ 中的所有因缺失字段过多而要删除的记 录用以下检测规则代替：

n-E₁≤△ (7-4) 式中：E;为B₄ 的缺失字段数；△为正整数，其取值可视具体数据集给定，也可以取 此种样本记录的最小缺失字段数。

对S₄中某少数重要字段(如第x 和第y 字段)为空需删除的记录二进制表示 用下式代替：

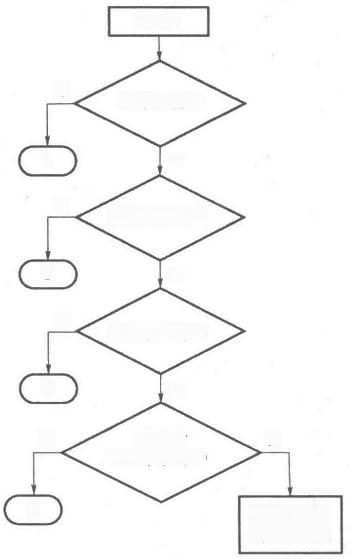
*B₁&B₂=0* (7-5) 式中：&为与运算符；B,为第x,y 位为1,其余位为0的n 位二进制数。对特定数据 集可有多个如式(7-5)的表达式。

(4)S₄ 中的标准二进制表示按样本中出现次数确定优先级，S₄中置式(7-4) 为最高优先级，置式(7-5)为次优先级。

<7.2.3.2> 不完整记录的位运算分类检测

对待检测记录的二进制表示B, 不完整记录分类检测流程图如图7-3所示。

**166)**数据质量导论



求取B;

Y

xor(B,su)=0

N

Y

3 xor(B₂,s₂)=0

N

Y

3 xor(B₁,su)=0

N

E,≥Aor

3 xor(B,su)=0

4

未检出二进 制表示处理

N

Y

2

3

1

图7-3 不完整记录检测流程图

如图7-3所示，对记录 r,进行检测，首先求取B,,然后逐类进行判定决策， 一旦 确定该记录类别，则算法停止；在每一决策节点，按照标准二进制表示的优先级逐一 进行异或运算， 一旦出现运算结果为零，则输出检测结果，算法停止；对未检出记录二 进制表示，由人工判定记录类别，并将其添加进相应的类标准二进制表示集。

记录的二进制表示，携带着全面的记录完整性信息，利于进一步修正处理；缺 失字段多但重要字段不为空的记录被检测为第2类，与第4类不会发生冲突，解决 了陈伟(2004)所提方法中存在的问题。

**7.2.4** **应用实例**

选择某信息系统的人员基本情况表为实验数据，该表有19个字段，首先选取 不完整不修正、不完整修正和不完整删除记录样本各100个按<7.2.3.1>的方法生 成标准二进制样本集，并用式(7-4)和式(7-5)对不完整删除标准二进制表示集 进行合并。然后用从表中选择连续的10000条记录按流程图7-3进行不完整记 录分类检测。所生成的标准二进制样本集、检测过程中处理的未检出二进制表示

第7章不完整数据的分类与检测(167)

以及检测结果列于表7-3。

表7 - 3 标准二进制样本集与检测结果表

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **类别** | **S,,k=1,2,3,4;A=7** | **处理的未检出** **二进制表示数** | **检测**  **结果** | **百分** **比** **/** **%** |
| 1 | S₁={1111111111111111111| |  | 2186 | 21.68 |
| 2 | S₂={1111111111111110111,1111111111111010111,1111111111111100111, 1111111111111111110,1111111111111000111,1111111110111111111, 1111111111111011111,1111111110111011111} | 20 | 6685 | 66.85 |
| 3 | S₃={1111111110111110110,1111111110111100110,1111111111111110110,  1111111111111010110,1111111111111000110,1111111111111100110, 1111111110101110111,1111111110111010111,1111111111111000110, 1111111111110010111,1111110111111010111,1111111111000010111, 1111111111100111111,1111111110111111111,1111111110111011110, 1111111111100110110,1111111110111110111,1111110111111111110} | 32 | 1032 | 10.32 |
| 4 | S₄={E;≥13,B;&1000000000000000000=0,1111111110100010111| | 2 | 97 | 0.97 |

由表7-3可知，不完整记录占到总记录的100%-21.68%=78.32%,说明在 该数据集中由空字段导致的不完整记录是普遍存在的，但85.55%的不完整记录 (占总记录的66.85%)只需做必要的标识，便可正常使用，并不需要修正，总记录 中10.32%的记录需要进一步根据各记录二进制表示记录的字段缺失情况调用具 体的缺失值估计与填充算法进行处理，而需删除的记录占记录数的0.97%。

在对10000条记录分类检测中，共出现54个未检测出二进制表示，将其分别 添加进标准二进制表示集后，选取另外10000条记录进行检测，没有出现未检出二 进制表示的情况。

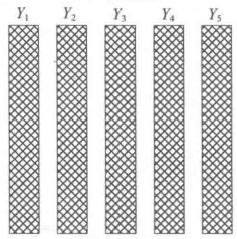
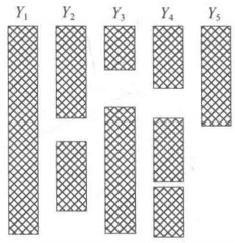
7.3 基于统计关系的不完整数据分类

本节从统计关系的角度讨论数据缺失模式和数据缺失机制的分类。

7.3.1 数据缺失模式分类

Rubin 提出使用数据缺失模式(Missing-Data Pattern)来描述数据集中哪些变 量是缺失的，哪些变量是未缺失可观测的，并且可以从一定程度上反映数据缺失机 制(见7.3.2节)。在数据清洗和数据分析中，对具有不同缺失模式的数据应该采 取不同的缺失处理方法(Rubin,1976;Little et al.,2002)。

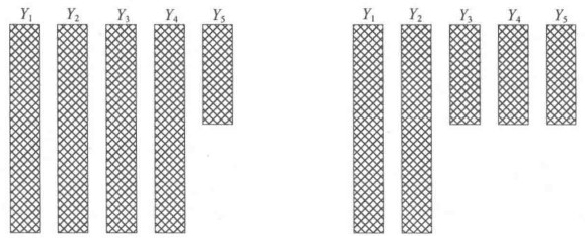
设Y=(y;) 为一个(n×k) 数据集，y;表示第i 个对象的第j 个属性。M=(m;)

**168)**数据质量导论

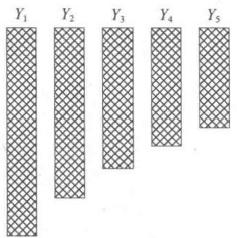
为缺失指示矩阵，当y;缺失时m,=1, 当 y,可观测时m;=0, 则 矩 阵M 就定义了关

于数据集 Y的缺失模式。

图7-4为常见的缺失模式，图中每一行代表一个观测对象，也就是关系数据 集中的一条记录，每一列代表一个属性，也就是关系数据集中的一个字段。阴影部 分表示数据的可观测部分，空白部分表示数据的缺失部分。按照不同缺失模式可 以将不完整数据分为以下几类。

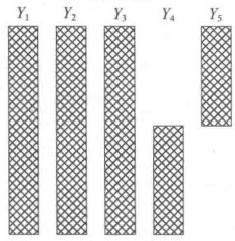


(a)单变量缺失 (b)组变量缺失



(d)一般情况

(c)单调缺失



(e)文件匹配

(d)隐变量

图7 - 4 缺失模式示例

1)单变量缺失

如 图 7 - 4(a) 所示的单变量缺失即缺失只在一个变量(即属性)上出现，而其

第7章 不完整数据的分类与检测(169

他变量是完整可观测的。这在农业试验中通常被称为缺区(Missing Plot)问题，此 时考虑的是因变量Y(如农作物产量)同其他自变量(如温度，湿度等)之间的关系 (示例图中K=5) 。 由于自变量是实验设置的故不会缺失，而因变量由于各种原因 可能会出现缺失，也就出现了单变量缺失。

2)组变量缺失

图7 - 4(b) 和(a) 中的区别在于(a) 中的一个因变量变成了一组因变量，也就 是缺失集中在一组变量上，除该组变量外的变量均是完整的。例如，在进行记名的 调查文件时， 一部分受调查者由于不愿意参加调查会拒绝回答问题，因为问卷的一 些属性(如姓名，单位等)是问卷设计好的，这部分信息不会缺失，所以只有需要受 调查者回答的属性会出现缺失。

3)单调缺失(Monotone)

在纵向研究(Longitudinal Study)中需要对一组对象按照时间重复进行调查， 如果一个对象在调查中途退出了调查(如受调查者去世),那么以后都不会再次回 到调查中。此时便会出现如图7-4(c) 中所示的单调缺失，单调缺失(或近似单调 缺失)在样本调查(Panel Survey)中十分常见，具有单调缺失模式的不完整数据通 常比具有一般缺失模式的数据更容易处理。

4)文件匹配(File Matching)

数据集中有两组变量不会同时被观测到，对每个对象来说也就是如果对象的 其中一组变量是可观测的，则另一组变量有很大概率会缺失，并且随着缺失比例的 增大这种趋势会更加明显。如图7-4(e) 所示正属于这种称为文件匹配的特殊缺 失模式。当这种缺失发生时，数据分析者需要更加谨慎，因为如果分析的过程涉及 到这两组变量则很有可能导致错误的结果。

5)隐变量

数据集中某些变量是无法观测的，也就是说该变量属于隐变量。将隐变量当 作完全缺失的变量，这样就可以使用缺失数据的理论和方法来分析隐变量。 图7 - 4(f)中 Y₃ 就是这样一个隐变量，这时可以使用如多元回归等方法来估计隐 变量 Y₃的真实取值。

在一些特殊的缺失模式中，数据的缺失模式往往隐含着对数据分析有用的真 实值，因此缺失模式的研究对不完整数据的分析具有重要意义(Litle et al., 2002)。

**7.3.2** **数据缺失机制分类**

数据缺失模式直观地体现了缺失数据中缺失变量的分布，但是这种直观的印 像难以定量地描述缺失变量同其他变量之间深层次的关系。为了更为准确地描述



170)

不完整数据中隐含的本质关系，需要引入数据缺失机制(Missing-Data Mecha- nisms)的概念。

数据缺失机制描述的是数据缺失概率同数据集中变量值的关系，从本质上说 明了缺失发生的原因和缺失发生的过程。缺失机制的概念由Rubin 在1976年首 次提出，并将不完整数据按照其缺失发生机制划分为三类：完全随机缺失、随机缺 失和非随机缺失。

设D 为完整的数据集，D, 表示第i 条记录的第j 个属性的值，D. 表示缺失变 量，D 表示可观测变量。M 为缺失指示矩阵(也就是缺失模式),M₂=1 当且仅当 D;缺失，否则M₃=0 。 则数据缺失概率与每个对象在各属性上取值的关系可以用 条件概率式(7-6)表示，如果D 为连续性变量则为概率密度式(7-7),其中φ表 示与缺失机制有关的未知参数。

*P(MID,φ)* (7-6)

*f(MID,φ)* (7-7)

1)完全随机缺失

P(MID,φ)=P(Mlφ) (7-8)

如式(7-8)所示，如果一个属性发生缺失的概率同数据集中的任何属性都无 关，其缺失完全是随机产生的，那么其缺失机制属于完全随机缺失。完全随机缺失 一般是由于系统原因或者随机噪声造成的，随着数据采集、存储手段的发展，系统 的可靠性越来越强，完全随机缺失也变得越来越少。

虽然完全随机缺失是一个相当强的假设，但有时候它也是合理的，特别是当数 据缺失是作为研究设计的一部分时。例如，当某个特殊变量度量起来太过昂贵时， 可以只针对全体样本中的随机子样本进行度量。这样也就意味着对剩下的数据而 言，该数据的缺失机制是完全随机缺失。

2)随机缺失

P(MID,φ)=P(MIDo,φ) (7-9)

如式(7-9)所示，如果一个属性发生缺失的概率同自身的真实取值无关，但 是同数据集中其他可观测的属性相关，那么其缺失机制属于随机缺失。需要注意 的是，虽然名称为随机缺失，但是这种缺失并不是真随机产生的。而之所以称为随 机缺失，是考虑到缺失发生的概率同数据缺失属性自身取值是无关的。

3)非随机缺失

*P(MID,φ)=P(M|Dm,φ)* (7-10)

P(MID,φ)=P(M|Dm,Da,φ) (7-11)

如果数据缺失发生的概率仅依赖于缺失属性本身的值(如式(7-10)所示), 或既依赖于发生缺失的属性的值也依赖于其他可观测属性的值(如式(7-11)所 示),那么其缺失机制属于非随机缺失。非随机缺失是最为普遍，同时也是最为复

第7章 不完整数据的分类与检测((171

杂的一种缺失机制。

举例说明，对于有“年龄”和“收入”两个属性的数据集，发生缺失的属性是“收 入”。如果“收入”属性发生缺失的概率与“年龄”和“收入”的属性值都无关，那么 该缺失属于完全随机缺失。如果对于不同“年龄”,“收入”发生缺失的概率不同， 而对应于同样“年龄”的不同“收入”,“收入”的缺失概率相同，即“收入”属性发生 缺失的概率与“收入”属性自身取值无关而与“年龄”属性取值相关，那么该缺失属 于随机缺失。如果对于同样“年龄”不同的“收入”,“收入”出现缺失的概率不同， 即“收入”属性发生缺失的概率与“收入”属性自身取值相关，那么该缺失属于非随 机缺失。

由于当数据缺失机制属于完全随机缺失和随机缺失时，基于似然的估计量是 无偏的，所以完全随机缺失和随机缺失被认为是可以忽略的。而当数据缺失机制 属于非随机缺失时，估计量一般是有偏的，缺失机制不可忽略。因此当缺失机制属 于非随机缺失时，通过完整实例法、填补法均难以得到无偏的聚类结果。Little 等 (2002)提出对非随机缺失的分析应该结合两个过程：数据产生的过程以及缺失产 生的过程，也就是说要对缺失发生的过程建模，并将缺失过程模型结合到分析模 型中。

**7.4** **本章小结**

本章提出两种不完整数据分类方法：第一种基于位运算的不完整记录层次分 类方式对记录进行层次分类，避免了类之间的冲突，层次分类和按标准二进制样本 优先级进行检测减少了运算冗余，方法以位运算为主，运算效率高，并且不破坏被 检测记录的二进制表示，方便数据清洗的数据溯源和数据世系设计；第二种基于统 计关系的不完整数据分类分别按照数据缺失模式和数据缺失机制对不完整数据分 类，相较于第一种分类方式更关注数据缺失的统计特征以及其体现出的深层次 信息。

**参考文献**

[1]李德毅，杜鹚.2005.不确定性人工智能[M]. 北京：国防工业出版社.

[2]陈伟.2004.数据清理关键技术及其软件平台的研究与应用[D]. 南京：南京航空航天大学.

[3]Breiman L,Friedman J H,Olshen R A,et al.1984.Classification and Regression Trees[M].Belmont;Wad- sworth International Group:203-215.

[4]Chen G,Astebro T.2003.How to Deal with Missing Categorical Data:Test of a Simple Bayesian Method[J].Or- ganiational Research Methods,6(3):309 -327.

[5]Codd E F.1979.Extending the Database Relational Model to Capture More Meaning[J].ACM Transactions on

172)数据质量导论

Database System,4(4):397-434.

[6]Institute S.1990.SAS Procedures Guide[M].Cary,North Carolina;SAS Institute.

[7]LiX B.2009.A Bayesian Approach for Estimating and Replacing Missing Categorical Data[J].ACM Joumal of

Data and Information Quality,1(1):1 -11.

[8]LittleR,Rubin DB.2002.Statistical Analysis with Missing Data[M].New Jersey:John Wiley &Sons:19 -20.

[9]MichieD,Spiegelhalter DJ,Taylor C C.1994.Machine Learning,Neural,and Statistical Classification[M].New York:Prentice Hall.

[10]Quinlan J R.1993.C4.5:Programs for Machine Leaming[M].San Mateo:Morgan Kaufmann of Elsevier. [11]Rubin D B.1976.Inference and Missing Data[J].Biometrika,63:581 -592

[12]Wand Y,Wang R.1996.Anchoring Data Quality Dimensions in Ontological Foundations [J].Communications of the ACM,36(11):86-95.

[13]Witten I H,Frank E.2005.Data Mining:Practical Machine Learning Tools and Techniques[M].San Francis-

co,Califomia:Morgan Kaufmann of Elsevier.

**第** **8** **章** **不完整数据的估计与填充**

**8.1** **引言**

常见的不完整数据处理方法主要可以分为：基于完整记录的方法，加权方法， 基于填补的方法和基于模型的方法。其中基于完整记录的方法和加权方法会丢弃 不完整的记录，导致可用信息大量丢失；基于模型的方法计算复杂度较高，且当模 型选择不当时会产生严重的错误。而基于填补的方法首先选择某种方法将数 据集填补“完整”再进行分析，充分利用了不完备数据集中的信息，而且不用 在后续的数据分析中考虑缺失数据的影响，便于实施，是处理不完备数据最为 常用的方法。

研究者提出的众多数据填补方法中最为简单的是全局常量(Global Constant) 填补和均值填补法，这两种方法分别使用全局常量和缺失属性的均值对缺失字段 进行填充，但在很多情况下这两种方法会得到和基于完整实例的方法同样有偏的 结果。回归分析也常被用在数据填补中，以不完整的属性对完整属性建立回归模 型，并依该模型估计缺失属性的值。热平台(Hot Deck)填补使用类似样本中点的 对应值代替缺失项的值，与热平台方法类似的冷平台(Cold Deck)不从当前数据集 而是从其他来源的数据集寻找填补值。以上填补都属于单一填补(Single Imputa- tion), 相对单一填补更为复杂的方法是多重填补(Multiple Imputation)。多重填补 使用多种单一填补方法对缺失数据进行填补，这样可以构造多个完整的数据集，最 终的分析结果(如参数的估计值)就是由多个完整数据集得到的估计值的一种综 合(如 Rubin 法则)。

为了提高不完整数据填充的准确性和效率，本章提出两种不同的数据填充方 法：基于统计关系学习的填充算法和基于机器学习方法的填充算法。此外，将数据 生成当作缺失的一种极端情况(即所有属性值都缺失),提出一种基于函数依赖一 致性的数据生成方法。

**8.2** **基于统计关系学习的缺失数据估计与填充**

在概述统计关系学习的基础上，分别基于马尔可夫模型和关系马尔可夫模型 提出了两种缺失数据估计与填充方法。

174)数据质量导论

**8.2.1** **统计关系学习概述**

本节概述统计关系学习的由来、研究内容、应用范围及实现方法。

[**8.2.1.1**](8.2.1.1) **统计关系学习的由来**

借助计算机程序从数据中学习，以得到隐含的信息，也就是传统上讲的数据挖 掘(Data Mining),一直是人工智能研究的一个重要领域。它节省了人力资源，提 高了知识获取的效率。目前，大部分学习算法总是对数据施加诸如独立、同分布等 限制，集中于扁平(Flat) 数据的处理，这些方法仅适于分析较简单的对象。而现实 世界的数据，除了一部分具有较好形式被称为结构化数据外，还有许多半结构(如 e-mail 消息和一些 HTML/XML网页等)或非结构化数据。为了处理这些复杂数 据，必须解决数据间多关系的处理问题。为此人们研究并提出了两类解决方法：关 系到命题的转换方法和直接关系方法。关系到命题的转换方法是将关系数据通过 某种方法压缩为扁平数据，再用传统的算法进行处理，这种方法不可避免地存在一 些局限，如数据的冗余、潜在信息的缺失、数据的多次复制，以及不恰当的属性来源 等；直接关系方法不对关系数据施加转换，而是对其直接进行处理，避免了转换方 法的缺陷。此外，由于现实世界数据存在不确定性，如数据中存在噪声或缺失值 等，这就要求学习算法必须具有处理不确定性的能力，基于统计的似然性理论为解 决不确定性问题提供了一种有效方法。

将关系表示、似然性理论和机器学习相结合能更好地解决现实世界复杂数据 的多关系性和不确定性，这一结合产生了人工智能领域的一个新子域——统计关 系学习(Statistical Relational Learning)。

20 世纪90年代末，国际上多个研究小组都开展了统计关系学习的研究。但 这些小组的侧重点不尽相同，因此对该研究子域的称谓也有所不同。侧重于关系 知识表示的研究者称之为“统计关系学习”,侧重于一阶谓词逻辑知识表示的研究 者称之为“似然逻辑学习”(Probabilistic Logic Learning),侧重于机器学习或数据挖 掘的研究者称之为“多关系数据挖掘”(Multi-relational Data Mining)或“关系学 习”(Relational Learning)。本章采用了国内普遍使用的“统计关系学习”这一名 称。统计关系学习集关系(逻辑)表示、似然性理论(不确定性理论)和机器学习 (数据挖掘)于一体，目的是获取关系数据中的似然模型或特定信息等，之后再用 获得的知识进行推理、预测和分类等。2000年以来，在IJCAI、ICML、AAAI等重要 国际学术会议上，都将统计关系学习列为一个独立专题，国外一些统计关系学习研 究项目已陆续展开，其研究结果已在生物信息学、系统生物学、Web 导航、社会网、 似然模型获取与利用、地理信息系统和自然语言理解等领域，取得了成功应用，统 计关系学习已成为人工智能领域的研究热点。

第8章不完整数据的估计与填充

[**8.2.1.2**](8.2.1.2) **统计关系学习的研究内容**

统计关系学习中的“统计”是指基于概率论的似然表示和推理机制。“学习” 是指基于数据来构建和调整统计关系模型。典型的模型学习算法，或者是已知模 型结构(包括图模型或逻辑子句)进行参数估计，或者是既学习模型结构又学习参 数。“关系”是指一阶谓词逻辑和关系表示，比之传统的基于命题的方法，使用“关 系”能更好地表示包括多个对象及对象间关系的复杂情况。统计关系学习的研究 领域如图8-1所示。

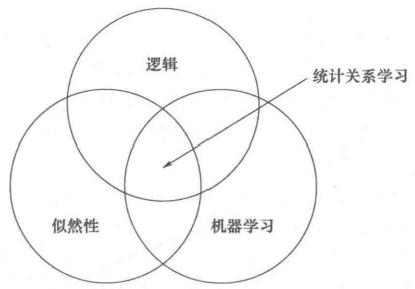


图8-1 统计关系学习的研究领域

图8-1中，三个圆分别表示逻辑(Logic)、似然性(Probability)和机器学习 (Machine Learning)三个领域，统计关系学习研究的是这三个领域的交集。三个领 域中两两之间都有交集，下面对两两交集部分说明如下：

1)似然性与机器学习的融合

基于统计的似然性与机器学习相结合，即是传统的基于命题的似然学习方法， 它提供了对带有不确定性信息数据的学习能力。其中基于梯度的方法、期望最大 化(EM) 算法系列以及马尔可夫链蒙特卡罗(Markov Chain Monte Carlo)方法等在 计算机视觉、语音识别、医疗诊断等领域已被广泛采用。该类方法主要是针对属 性-值表示的，因而无法处理现实的复杂关系数据。

2)逻辑与机器学习的融合

逻辑与机器学习相结合可自动构建关系模式，在计算生物学和化学等多关系 数据挖掘领域应用较广。逻辑有很强地描述对象间复杂关系的能力，使以归纳逻 辑程序设计(Inductive Logic Programming)为代表的技术成为挖掘关系模式的主要 手段，但该方法难于处理不确定性信息，限制了它的应用。

3)逻辑与似然性的融合

逻辑与似然性相结合的研究较早，其主要目标是从知识表示方面对逻辑 进行概率扩展，强调逻辑的似然性特征。该研究在专家系统中应用较广。然

)数据质量导论

而，由人工来构建模型既耗时又费力，用计算机从数据中学习是一种富有前景 的解决方案。

统计关系学习结合上述的交叉领域，克服了各自的不足，其主要目标是结合关 系表示与不确定性描述和推理机制，从现实世界错综复杂的关系数据中自动抽取 和构建具有相互依赖关系的事物模型并加以应用。此外，很多针对特定关系问题 的算法，如关系分类算法、关系强化学习算法等，结合了关系处理与学习机制，也是 人工智能领域研究的一个方面，将它们作为统计关系学习研究的一部分，可以更加 丰富这一研究领域。

[**8.2.1.3**](8.2.1.3) **统计关系学习的应用**

当前包括基因科学、计算机视觉、语音理解和识别、故障诊断和处理、信息检 索、软件调试等应用研究领域都存在着复杂的关系和不确定性，统计关系学习正是 解决这些领域问题的有效方法，具体如下。

1)聚集分类(Collective Classification)

传统的分类是给定某一对象的属性值，预测对象属于哪一类型。聚集分 类利用对象之间的联系来考虑所有相关对象是否属于同一类。例如，在论文 的引用文献方面，同类研究主题的论文相互引用的可能性很大，分为同类的可 能性也会很大。因此，同时考虑对象间的相关性进行分类，可提高效率和精 确度。

2)链接预测(Link Prediction)

链接预测是通过对象之间的属性和其他可能的关系确定两个对象间是否存在 某种联系。例如，通过某个研究生选课的情况可大体推断其研究方向等。

3)基于链接的聚类(Link-based Clustering)

聚类的目标是将具有相同或相近属性的对象分为一类，与分类不同的是预先 不知道类别的存在。在基于链接的聚类中，同一类的对象与其链接和属性相关。 与聚集分类相似，相互链接的对象被聚为一类的可能性更大。

4)对象识别(Object Identification)

基于对象的属性和关系来预测两个实体是否是同一对象，可以用来识别重复 对象，删除冗余。

5)复杂关系系统建模

在时空推理和社会网建模等领域中需要先刻画事物的相互联系和关系等特 征，而后才能展开其他更深入的研究，统计关系学习为该领域研究提供了强有力的 工具和手段。

[**8.2.1.4**](8.2.1.4) **统计关系学习实现方法**

统计关系学习是近十年内发展起来的新的数据挖掘方法，其实现方法由似 然关系模型和学习算法组成。似然关系模型是关系的似然表示形式，是通过

第8章 不完整数据的估计与填充 (177)

将不同的概率表示和推理机制，如贝叶斯(Bayesian) 网、随机文法、马尔可夫 网、(隐)马尔可夫模型等，与关系、 一阶谓词逻辑表示相结合得到的。根据统 计关系学习方法所用的概率表示和推理机制不同，统计关系学习实现方法主 要有以下几类。

1)基于贝叶斯网的统计关系学习方法

贝叶斯网是最重要且最有效使用概率进行表示和推理的模型。然而，传统贝 叶斯网是命题逻辑的概率扩展，该方法针对扁平数据，不能处理丰富的关系数据， 且表达能力有限。将传统贝叶斯网与一阶谓词逻辑相结合(一阶贝叶斯逻 辑)或将贝叶斯网与实体关系模型相结合(似然关系模型),就可以处理关系 数据，并具有更强的表达能力。研究者们从不同角度、采用不同的扩展方式提 出了众多基于贝叶斯网的统计关系学习方法。根据表示形式的不同，该类又 可以分成两小类： 一类是采用图型结构表示的方法，该类模型均采用图型结构 表示数据及数据间关系；另一类是采用逻辑子句结构表示的方法，该类方法将 逻辑与贝叶斯网相结合。

2)基于随机文法(Stochastic Grammars)的统计关系学习方法

基于随机文法的统计关系方法是将随机文法提升到一阶谓词逻辑的一类方 法，是随机文法的一个泛化。其主要包括随机逻辑程序(Stochastic Logic Program) 和统计建模程序设计(Programming in Statistical Modeling)。这类方法通过对逻辑 成分附加概率，以处理关系和不确定性。该类方法以逻辑程序来描述模型结构，因 而表示能力较强。随机逻辑程序通过在随机上下文无关文法(Stochastic Context Free Grammars)的每个子句附加一个概率值，对其进行直接升级；统计建模程序设 计是一个符号-统计建模语言(Symbolic-Statistical Modeling Language),它不仅对 逻辑程序进行了概率扩展，而且还能使用EM 学习算法从例子中学习，简单地说， 统计建模程序设计是在“事实”中附加概率分布的逻辑程序。

3)基于马尔可夫网的统计关系学习方法

由马尔可夫网扩展而来的方法，主要有两小类：2002年 Taskar 等人提出的关 系马尔可夫网和2004年 Richardson 等人提出的马尔可夫网逻辑网，它们允许循环 路径存在，与基于贝叶斯网的方法相比可以更加灵活地表示现实中的关系特性。 这些模型在本质上定义了一种马尔可夫网模板，因此，它们对于实例集合定义了一 致的概率分布。

4)基于(隐)马尔可夫模型的统计关系学习方法

2002年由 Anderson等人提出的关系马尔可夫模型，以及次年由Kersting 等人 提出的逻辑隐马尔可夫模型是主要基于(隐)马尔可夫模型的方法。这些模型在 马尔可夫模型的基础上进行了扩展，允许状态是不同类型，可用来解决动态的序列 化问题。

178)数据质量导论

5)其他统计关系学习方法

其他统计关系学习方法有基于决策树和关系向量的方法。其中基于决策树的 方法，利用一阶谓词逻辑框架来处理关系数据，它们的特征类限制于一阶谓词逻辑 的子集。

8.2.2 基于马尔可夫模型的缺失值估计方法

本节提出一种基于马尔可夫模型的缺失值估计方法。

<8.2.2.1> 基本马尔可夫过程原理

马尔可夫过程属于近代概率论的一个重要的新领域——随机过程理论，在现 实中常常遇到做随机运动的系统，系统不断地改变它的状态。马尔可夫模型是描 述做随机运动的系统的有力工具，它也是最简单的一类随机过程，在自然科学和社 会科学的各个领域都有广泛应用背景。

马尔可夫过程是下述的这样一种过程：在已知时刻t₀ 系统或过程所处状态的 条件下，在时刻t₀ 以后系统将到达的状态只与t₀ 时刻的状态相关，而与时刻t₀以前 系统所处的状态无关。这个特性称为无后效性或马尔可夫性。

以下给出马尔可夫过程及马尔可夫链的严格定义。

定 义 8 -1:马尔可夫过程。设有 一 随机过程 | ξ(t),teT},t₁ <t₂<t₃<…<

tm<tm+1∈T,若在t₁,t₂,t₃,…,m,m+ 对ξ(t) 观测得到相应的观测值x₁,x₂,x₃,…, xm,X+i 满足条件；

fln,….t(xm+ilx₁,x₂,…,xm)=fim,ll(xm+ilxm) (8-1)

式中：fm…(xm+1lx₁,x₂,…,xm) 为在ξ(t₁)=x,ξ(t₂)=x₂,…,ξ(1m)=xm 的

条件下ξ(tm+1)=xm+1的条件概率密度。

则称这类过程为具有马尔可夫性质的随机过程或马尔可夫过程。

定义8-2 :马尔可夫链。{ξ(n),n =0,1, 2,…} 是离散状态(状态空间为

I)、参数为非负整数的随机过程，且ξ(n) 满足条件：

P{ξ(n+1)=jls(0)=i₀,ξ(1)=i₁,…,ξ(n-1)=i-i,ξ(n)=i,}

(8-2)

=P{ξ(n+1)=jlξ(n)=i 。}

即在参数为n(n=0,1,2,…,n), 状态取ξ(0)=i₀,ξ(1)=i₁,…,ξ(n-1)=

i,ξ(n)=i 。 的条件下，ξ(n+1)=j 的条件概率与ξ(0),ξ(1),…,ξ(n-1) 无

关，而仅与ξ(n) 所取的值有关，这类随机过程称为马尔可夫链。

一阶马尔可夫链也可以描述为三元组(I,A,π), 其中I={q₁,q₂,…,q₁},leN

为状态集合；A 为概率转移矩阵，a;=P(ξ(n+1)=q;lξ(n)=q₁),i,j=1,2,…,l 表 示由状态q;转换为状态q,的概率(假设 a 与n 无关);π=(π₁,π₂,…,π₁)为初始

第8章 不完整数据的估计与填充(179)

概率向量，π₁=P (ξ( 0)=q₁)。

<8.2.2.2> 基于马尔可夫模型的缺失数据估计模型

一个关系中可能存在多个枚举型属性，但不是任两个属性的属性值之间都存 在依赖，不存在依赖的属性值之间的内分布是随机的，因此，在数据生成之前，首先 应根据以下原则对枚举型属性进行分组。

在分组中， 一个重要的信息是属性组中属性的优先级，关系数据中的属性优先 级确定应遵循以下原则：

1)时间先后

某些属性所描述的实体存在时间先后顺序。如出生年、入伍年；第一学历、最 高学历。

2)空间从属

有些属性所描述的实体存在空间从属关系。如城市、大学；省、市、学校、院系、 专业。

3)概念层次

某些属性间存在层次分类关系。如一级学科、二级学科；师、团、营、连等。

4)业务主次

在某些业务领域中，根据相关领域知识，存在某属性值受另一属性值约束的情 况。如部队中职务级别对军衔的限定，即军衔从属于职务级别，此时，职务级别优 先级应高于军衔；再如学历优先级应高于学位。

对一个枚举型属性(可以是属性的组成部分)分组，按照以上四个原则进行优 先级排序，排序一旦确定，该分组中的属性值按此顺序依次生成。

对一个有序枚举型属性集(元组)G= <a₁,a₂,…,a>, 各属性对应的域为

V=[V₁,V₂,…,V,], 其中V₁={v₁,v₂,…,vm,},i=1,2,…,n 为具体取值范围，各

属性域可以不相同。则有以下关系马尔可夫模型：

P(a₁=2u₁|a₁=Vi₁,Q₂=tzx₂,…,₁-1=Vu-1)k-₁)

=P(a₁=V₄₁|a₁-1=V-D)k-₁),i=2,3,…,n

(8-3)

式中：P(a₁=va₁ |a₁-₁=Va-1)k)为在属性a;-₁取值va-Dk 的条件下，属性 a;取值 vu的概率。式(8-3)表明，属性 a₁ 的取值只依赖于属性 a₁-₁ 的取值，与其他属性 的取值无关。

枚举型数据的缺失值估计公式为



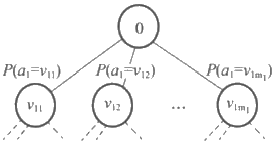


(8-4)

(8-5)

**180)**数据质量导论

在分组中数据全为缺失时，可以得到如图8-2所示的生成树。



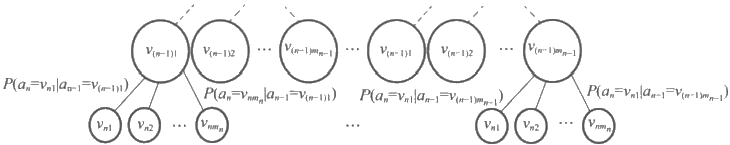


图8-2 基于马尔可夫模型的数据生成树

[**8.2.2.3**](8.2.2.3) **应用实例**

实验数据采用(Li, 等，2009)中的数据，如表8-1所列。

表 8 - 1 示例数据集

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **No** | **Income** | **Age** | **Gender** | **HomeOwner** |
| 1 | low | <30 | female | no |
| 2 | low | <30 | male | no |
| 3 | low | 30～55 | female | yes |
| 4 | low | 30～55 | female | no |
| 5 | low | >55 | female | no |
| 6 | high | <30 | male | yes |
| 7 | high | 30~55 | female | yes |
| 8 | high | 30~55 | male | yes |
| 9 | high | 30~55 | male | yes |
| 10 | high | 30～55 | male | no |
| 11 | high | >50 | male | yes |
| 12 |  | 30～55 | female | yes |
| 13 |  | 30～55 | female | yes |
| 14 |  | <30 | female |  |
| 15 |  |  | male | no |
| 16 |  |  | male | no |

第8章 不完整数据的估计与填充(181)

如表8-1所列，该数据集是包括有4个属性(Income,Age,Gender 和 Home- Owner(是否拥有家庭))的16条记录，其中共有8个空字段。

首先，表8-1中的数据集不可再分，按照<8.2.2.2>节的排序原则，对其排序： 〈Gender,Age,Income,HomeOwner)。

然后，选择前10条记录作为式(8-5)估计依赖概率的样本集：















最后，分别用MaxProp 和 PropProp(Li,2009) 两种方法填充缺失值，如表8-2 所列。

表 8 - 2 填充缺失值的结果

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| No. | Income | | Age | | HomeOwner | |
| MaxPost | PropProp | MaxPost | PropProp | MaxPost | **PropProp** |
| 12 | high | high | — | — | — | — |
| 13 | high | high | — | — | — | — |
| 14 | low | low | — | — | no | no |
| 15 | high | low | 30～55 | <30 | — | —— |
| 16 | high | high | 30~55 | 30~50 | —— | — |

显然表8-2中所列的结果，对PropProp 方法而言只是其中的一种情形。

8.2.3 基于关系马尔可夫模型的缺失值估计

马尔可夫模型被广泛用于时序过程建模，并在网络日志挖掘、生物计算、语音 识别、自然语言处理、机器人技术、故障诊断等领域得到了成功应用。2002年，华

)数据质量导论

盛顿大学的 Anderson 提出了关系马尔可夫模型(Relationship Markov Model, RMM), 并将其用于网站自适应导航，对网页间的链接情况进行了有效估计。关系 马尔可夫模型通过在马尔可夫模型的状态集上附加一个关系结构得来，能够描述 不同类型的状态，弥补了马尔可夫模型在数据预测时难以描述多种类型状态的不 足。即使某些状态在训练集中没有出现，也能够很好地估计状态间的转移概率，表 现出了更强的学习和推理能力。

[**8.2.3.1**](8.2.3.1) **关系马尔可夫模型相关定义**

关系马尔可夫模型用五元组< D,R,Q,A,π> 来表示。其中，D为定义域集， D∈D为其中的一个定义域。R 为关系集，D内的不同定义域作为关系R∈R的属性 集。Q为状态集，Q={q₁,q₂,…,qn},nεN, 关系R 中的属性取具体值时即为某一 状态。A 为状态转移概率矩阵，其中的元素为ag=P(S,=q;I S=q₁),i,j=1,

2,…,n, 表示t-1 时刻的状态q₁ 到 t 时刻状态 q, 的转移概率。π为初始概率向 量，π;=P (S₀=q₁ )表示初始状态为 q;时的概率。

在 RMM 中，利用关系和领域抽象概念定义一系列的状态，抽象概念将状态划 分为不同的类别。定义域D 根据抽象层次不同表示为树形结构，nodes(D) 代表域 D 中的所有中间节点d,leaves(d)为 d 的后代，表示D 中的叶子节点δ。假设R 表 示k 元关系，属性分别取自域 D,,…,D,d₁,…,d₄ 为相应域的节点。定义与抽象 集R(d₁,…,d₄) 一致的状态的概念：

{R(δ,,…,δ)∈Qlδ;eleaves(d₁),i=1,2,…,k}

对于状态q∈Q,令q=R(δ₁,…,δ),R 中属性的定义域分别为D₁,D₂…,D, 定

义状态q 的抽象集概念A(q):

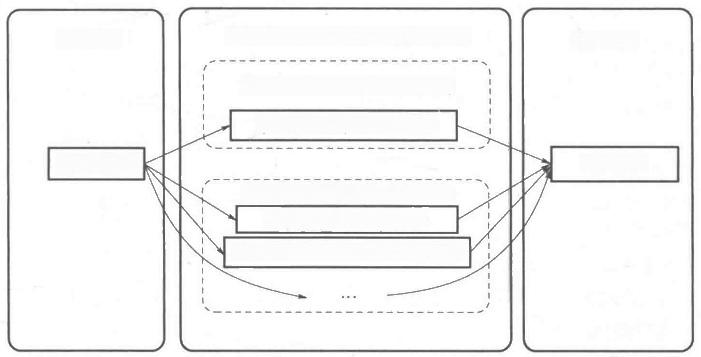
{R(d₁,…,d₄)CQld;enodes(D)Aδ;eleaves(d),i=1,2,…,k}

我们以购物网站的例子来解释关系马尔可夫模型中的相关定义。在图8-3 中，圆角矩形框(实线)为关系马尔可夫模型中定义的关系，如“网站入口”“产品页 面(所有产品，产品储备情况)”“结算网页”。虚线框代表了每一个抽象集，抽象集 中包含相应的状态，如抽象集“产品页面(所有产品，缺货状态)”。小的矩形框为 一个网页，表示了每个具体的状态，如状态“iphone页面(iphone4, 缺货)”。每个关 系、抽象集、状态都有相应的属性，各属性属于不同的定义域D,在这里有产品定义 域和产品储备情况定义域，定义域根据抽象层次不同表示为树形结构，如图8-4 和图8-5所示。状态之间由带箭头的线相连，表示状态间的转移。

[**8.2.3.2**](8.2.3.2) **关系马尔可夫模型状态转移概率**

关系数据中，把一个属性组的具体值定义为状态，相同类别的属性组定义为抽 象集。一对状态抽象集(α,β),α,β∈A(q), 定义α中的状态到β中的状态的转移 概率为

第8章 不完整数据的估计与填充 

产品页面(所有产品，产品储备情况)

网站入口

结算网页

产品页面(所有产品，缺货状态)

iphone页面(iphone4,缺货)

网站主页面

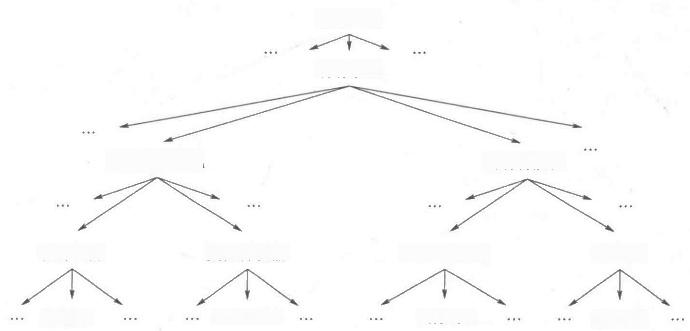
结算页面

产品页面(所有产品，有货状态)

htc页面(htc328D,有货)

黑白屏机器页面(nokia1100,有货)

图 8 - 3 关系马尔可夫模型状态抽象概念图



所有产品

所有手机

所有智能机

android 系统

htc328D

彩屏机器

联想E3

所有非智能机

黑白屏机器

ios系统

iphone4

nokial100

图 8 - 4 产品抽象层次

所有储备情况

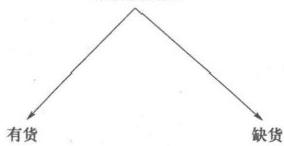


图 8 - 5 产品储备情况抽象层次

(8-6)

184)数据质量导论

式中：P(q;la) 表示抽象集α中状态为q; 的概率；a; 为马尔可夫模型中状态q; 到状 态q; 的转移概率，ag=P{X₄=q,lXm=q₁}, 即随机过程{X(t),t∈T} 在时刻m 处于 状态q;条件下，在时刻 n转移到状态 q;的转移概率。

假设目的状态为qu,源状态为 q, 得到相应的状态转移概率为

*a=a₄₁gP(qalβ)* (8-7) 式中：P(q₄Iβ) 表示抽象集β中状态为q₄ 的概率。如果α中的某个状态到β中某 一状态的转移情况在训练集中未出现，可假设P(q₄Iβ)=1/Iβl, 以实现状态转移 概率的计算。其中，Iβl 为抽象集β中的状态数。

<8.2.3.3> 动态属性选择方法

Lee.等(2000)提出了动态属性选择(Dynamic Attribute Selection,DAS)方法，该 方法选择最优的属性组，以估计某一属性中的缺失值。新的属性加入最优属性组 后，属性类别交互信息取最大值，通过这一原则，得到包含信息最大的最优属性组。 同时，Lee等(2000)采用决策树法，利用最优属性组包含的信息，估计某属性中的 缺失值。通过与其他方法的对比实验，验证了文中提出方法的优越性。

动态属性选择方法通过计算属性类别交互信息的最大值，确定包含信息量最 多的属性组合，并通过该组合估计缺失值。某属性包含的不同属性值为属性的类

别，类别不确定性由信息熵表示，其定义为

(8-8)

式中：C 为属性中的所有类别，cεC。

条件类别不确定性的信息熵定义为



即在部分属性已知的条件下类别的不确定性。其中，S 为所有已选定的属性 组合，f为新加入S 的属性。

不确定性的减少由交互信息表示，其定义为

GI(C;(f,S))=H(C)-H(Cl(f,S)) (8-10)

GI(C;(f,S)) 取最大值或不再增长时的S 为最优的属性组合，即用属性组S 来估计缺失的属性值。

<8.2.3.4> 基于关系马尔可夫模型的缺失值估计与填充方法

该方法综合了属性排序和动态属性选择方法，利用关系马尔可夫模型对缺失 值进行估计，并采用MaxPost 和 ProProp 两种方法对缺失值进行填充。

关系数据的属性间一般存在着某种关联性，首先，应根据这种关联关系对属性 进行排序，以利用前面的属性或属性组对后面的缺失值进行估计。通常，属性间存 在时序关系、隶属关系、概念层次关系和业务优先权关系。不同的数据集需按照相 应的关联关系进行属性排序，同一数据集中不同类别的属性组采取的排序方法也

第8章 不完整数据的估计与填充(185)

不尽相同。实际当中可能需要综合多种方法对属性进行排序。

其次，对排序后的属性利用动态属性选择方法，确定估计某属性中缺失值的最 优属性组合，并将属性组分为源属性组(源状态集)和目的属性组(目的状态集), 在两个属性组中按状态的类别划分不同的抽象集，并根据关系马尔可夫模型的状 态转移概率公式，估算状态间的转移概率，得到状态转移概率矩阵。

最后，根据转移概率矩阵，分别采用MaxPost 和 ProProp 两种方法对缺失值进 行填充。MaxPost 法将状态转移概率最大的目的状态作为填充值，此方法对非随 机缺失的数据效果较好。非随机缺失数据是指因敏感性、个人隐私等原因造成的 缺失数据。如个人收入、犯罪史等信息往往是故意令其缺失。ProProp 法根据状态 转移概率分布情况，进行缺失值填充，该方法根据状态间的转移概率随机选择一个 状态作为目的状态，转移概率大的状态被选为目的状态的可能性较大，转移概率小 的状态被选为目的状态的可能性相对较小。因此，该方法适合估计随机缺失的数 据。综上所述，缺失值估计与填充综合方法描述如表8-3所列。

表 8 - 3 缺失值估计与填充方法

|  |  |
| --- | --- |
| 步骤1 | 根据属性间关联关系，综合考虑，确定属性的估计顺序 |
| 步骤2 | 根据动态属性选择方法，计算属性的类别不确定性H(C)和条件不确定性H(CI(f,S)),得到 最大的交互信息GI(C;(f,S)),确定估计缺失值的最优属性组合，并将属性组分为源状态集 和目的状态集 |
| 步骤3 | 利用关系马尔可夫模型，确定源状态集和目的状态集中的抽象集，根据式(8-6)和式(8-7) 计算源状态到目的状态的转移概率，得到状态转移概率矩阵A |
| 步骤4 | 根据状态间的转移概率，分别采用MaxPost和ProProp两种方法对缺失值进行填充，验证方法 的有效性和优越性 |

[**8.2.3.5**](8.2.3.5) **实例说明**

为阐述该方法在缺失值估计中的应用，现举例说明。考虑一张学生信息数据 表，如表8-4所列。表8-4中包含20条记录，包括学生的平时成绩、性别、教师 经验、课程成绩4个属性。此例中将前15条记录作为训练数据，对16～20 条记录 中的缺失值进行估计。

表8 - 4 学生信息表

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **No.** | **平时成绩** | **性别** | **教师经验** | **课程成绩** |
| 1 | 好 | 女 | 差 | 良 |
| 2 | 好 | 男 | 好 | 优 |
| 3 | 中 | 女 | 好 | 中 |

186)数据质量导论

(续)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **No.** | **平时成绩** | **性别** | **教师经验** | **课程成绩** |
| 4 | 好 | 男 | 差 | 中 |
| 5 | 好 | 男 | 好 | 良 |
| 6 | 差 | 男 | 中 | 差 |
| 7 | 中 | 女 | 差 | 中 |
| 8 | 好 | 男 | 中 | 差 |
| 9 | 中 | 男 | 差 | 中 |
| 10 | 差 | 男 | 中 | 中 |
| 11 | 中 | 女 | 好 | 优 |
| 12 | 好 | 男 | 差 | 良 |
| 13 | 好 | 女 | 差 | 中 |
| 14 | 好 | 女 | 差 | 差 |
| 15 | 好 | 男 | 好 | 中 |
| 16 | 好 | 男 | 好 |  |
| 17 | 差 | 男 | 中 |  |
| 18 | 好 | 男 | 差 |  |
| 19 | 好 | 女 | 差 |  |
| 20 | 中 | 女 | 好 |  |

首先，针对学生信息数据表的具体情况，根据学生的平时成绩决定课程考试成 绩，教师经验影响学生的成绩，学生性别不同擅长的科目不同，确定属性的先后顺 序为<性别，教师经验，平时成绩，课程成绩>。即用前面的属性或属性组估计后 面属性中的缺失值。

其次，确定性别、教师经验、平时成绩三个属性的最优组合。根据式(8-8)计 算课程成绩属性的信息熵：



根据式(8-9),计算已知平时成绩条件下的信息熵：





=1.0288

如上述计算过程 H(Cl(f 师经验，S平时成绩))=0.7562,H(Cl( f在别，S教师经验，

第8章不完整数据的估计与填充 (187)

S 平时成绩))=0.7167。

由计算结果可知，当S 包含性别、教师经验、平时成绩三个属性时交互信息取 最大值，因此利用这三个属性的组合估计课程成绩属性中的缺失值。

将数据表中前三个属性作为源状态集，课程成绩属性作为目的状态集。源状 态集中每个状态作为一个抽象集，目的状态集中的“优”“良”作为一个抽象集， “中”“差”作为另一抽象集。根据式(8-6)和式(8-7)计算状态间的转移概率， 结果见表8-5。

表8 - 5 状态转移概率

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| No. | **平时成绩** | **性别** | **教师经验** | **课程成绩** | | | |
| **优** | **良** | 中 | 差 |
| 16 | 好 | 男 | 好 | 0.2667 | 0.4000 | 0.2333 | 0.1000 |
| 17 | 差 | 男 | 中 | 0 | 0 | 0.7000 | 0.3000 |
| 18 | 好 | 男 | 差 | 0.2000 | 0.3000 | 0.3500 | 0.1500 |
| 19 | 好 | 女 | 差 | 0.1333 | 0.2000 | 0.4667 | 0.2000 |
| 20 | 中 | 女 | 好 | 0.2000 | 03000 | 0.3500 | 0.1500 |

表8-5中的数值为第16～20条记录中，源状态到目的状态的转移概率。

根据状态转移概率，分别按照 MaxPost 和 ProProp 两种方法进行缺失值填充， 填充结果见表8-6。

表8 - 6 缺失值填充结果

|  |  |  |
| --- | --- | --- |
| No. | **课程成绩** | |
| **MaxPost** | **ProProp** |
| 16 | 良 | 良 |
| 17 | 中 | 中 |
| 18 | 中 | 中 |
| 19 | 中 | 中 |
| 20 | 中 | 差 |

由表8-6可知，两种方法效果相当。另外，因存在一定的概率分布，ProProp 方法得到的结果并不唯一，这里任取一种情况。

[**8.2.3.6**](8.2.3.6) **实验对比与分析**

为检验提出方法的普适性，使用公认数据集 UCI 中的 Adult 数据集和 Bank

188)数据质量导论

Marketing 数据集进行实验。在两个数据集上，分别进行关系马尔可夫模型、DAS RMM 和 Li(2009)的贝叶斯方法的对比实验，并采用MaxPost和 ProProp 两种方法 进行缺失值的填充。实验环境：Intel(R)Core(TM)2 Duo 2.4GHz CPU,2GB 内存， Windows XP 操作系统。

1)对比方法 ·

贝叶斯方法在数据挖掘和数据库管理方面应用于多个领域，如分类分析、信息 获取、数据集成等。贝叶斯方法的理论基础为贝叶斯定理。考虑一个数据集，具有 M 个属性，假设其中某个属性具有两种类别属性值c₁ 和c₂,其余的无类别属性为 X₁,X₂,…,Xw-1。 对于一个新的记录x=(x₁,x₂,…,xw-i), 如果P(c₁ |x)>P(c₂ l

x),则该属性值为c₁ 类别，否则为c₂ 类别。其中，后验概率P(clx) 可以由贝叶斯 定理得到，概率P(c₄) 和P(xlc) 可以从数据集计算得到。其中，P(c) 比较容易 估计，而数据集维数较多时P(xlc₄) 很难计算。为简化计算， 一般假设各属性之间 是条件独立的，基于此，P(xlc₄) 由下面的公式计算得到：



此种方法称为朴素贝叶斯方法。

(8-11)

Li(2009)给定一个具有N 条记录，M 个属性的数据集X,X₂,…,Xy, 令 L; 为 相应属性X, 中的类别数，N,为 X, 值已知的记录数，N₄ 为X, 值等于第k 个类别ca 的记录数，Nr 为 在X₁=ca,j≠i 的条件下，X, 等于第r 个类别的记录数。其缺失 值估计和替代算法如表8-7所列。

表8 - 7 缺失值填充算法

|  |
| --- |
| 1.计算每个属性的先验概率：P(X₁=cg)=Ng/N,i=1,2,…,L |
| 2.计算X;=ca时，X;的条件概率：  P(X;=crIX;=ca)=N₇ Ia/Na,j=1,2,…,M;j≠i;r=1,2,…,L |
| 3.给定一数据元组x,在属性X;中含有缺失值，令J为记录x所有属性的序号集，x,为x中相应的属 性值。根据式(8-4)、式(8-5)计算后验概率值：    其中，不必计算出P(xy)值，该值可在其后的计算中约掉 |
| 4.基于上步得到的概率，采取MaxPost或ProProp方法对x中的缺失值X,进行填充。 |

2)实验数据集

Adult 数据来自美国人口调查局数据库，描述了人员的年龄、性别、种族、教育 情况等信息。其中包含了48842条记录，9个枚举型属性(workclass,education,ma- rital-status,occupation,relationship,race,sex,native-country,class),6个数字型属 性(age,fnlwgt,education-num,capital-gain,capital -loss,hours -per-week)。关

第8章 不完整数据的估计与填充((189

系马尔可夫模型方法和贝叶斯方法主要用于枚举型缺失值的估计，实验中选取9 个枚举型属性中5个关联性强的属性进行实验，根据属性间的关联关系，确定其顺 序为<sex,race,education,occupation,workclass >。

Bank Marketing数据集为葡萄牙一家金融机构有关直接营销的数据，数据集 记录了该金融机构业务员与顾客沟通的信息，及客户的基本资料信息。其中包含 了45211条记录，10个枚举型属性(job,marital,education,default,housing,loan,

contact,month,poutcome,y),7 个数字型属性(age,balance,day,duration,campaign,

pdays,previous)。实验中，同样选择3个枚举型属性，根据属性间的决定和影响关 系，确定其顺序为<education,job,poutcome>。

综合方法分别对两个数据集进行统计学习，按照陈伟(2004)的方法，随机地 将其中25%的数据置空，并对缺失值进行估计。其中，Adult 数据集随机置空前、 后的部分数据分别如表8-8和表8-9所列。

表 8 - 8 随机置空前的部分数据

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| No. | sex | race | education | occupation | **workclass** |
| 1 | Male | Black | 11th | Machine-op-inspet | Private |
| 2 | Male | White | HS-grad | Farming-fishing | Private |
| 3 | Male | White | Assoc-acdm | Protective -serv | Local -gov |
| 4 | Male | Black | Some -college | Machine-op-inspct | Private |
| 5 | Male | White | 10th | Other -service | Private |
| : |  |  |  | : |  |

表8 - 9 随机置空后的部分数据

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| No. | sex | race | **education** | occupation | **workclass** |
| 1 | Male | Black |  | Machine-op-inspct | Private |
| 2 |  | White | HS-grad | Farming-fishing | Private |
| 3 |  | White | Assoc -acdm | Protective -serv | Local -gov |
| 4 |  | Black | Some-college | Machine-op-inspct | Private |
| 5 | Male |  | 10th | Other -service | Private |
|  |  |  | : | : | : |

由表8-8、表8-9可知，Adult 数据表的前5条记录分别随机地将序号1的 “education”、序号2的“sex”、序号3的“sex”、序 号 4 的“sex” 和序号5的“race” 属 性值置空。实验中，首先用属性sex 中的属性值对属性race 中的缺失值进行估计， 此时属性sex 为源状态集，属性 race 为目的状态集，在两个状态集中划分各自抽象 集，利用关系马尔可夫模型计算源状态集中状态到目的状态集中状态的转移概率，

190)数据质量导论

根据状态转移概率采用MaxPost 和 ProProp两种方法对缺失值进行填充；而后采用 动态属性选择方法确定估计属性 education 中缺失值的属性组合，即为属性 race 或属性组 <sex,race>, 此时属性 race 或属性组 <sex,race> 为源状态集，属性 education为目的状态集，在两个状态集中划分各自的抽象集，并利用关系马尔可 夫模型计算源状态集中状态到目的状态集中状态的转移概率，并根据状态转移 概率采用 MaxPost 和 ProProp 两种方法对缺失值进行填充。依此类推，利用动态 属性选择方法和关系马尔可夫模型方法按照排序的属性顺序估计下一属性中的 缺失值。

3)评价指标

对比实验中，采用Li(2009) 提到的单变量统计、关联规则指标和程序执行时 间三个指标对方法进行效果评估。

单变量统计指标常用于数据仓库相关领域，与类别数据相关的单变量统计指 标是指属性的分布率，因此引入均方根误差(Root Mean Square Error,RMSE)指标 进行，该指标度量了原始值和替代值单变量频率分布的接近程度。其定义如下：

(8-12)

式中：m₄ 为原始数据第i 个属性中第k 个类别值缺失的数量；m₄ 为填充后数据中 第i个属性中第k个类别值的数量；m 为缺失数据的总数；M 为含有缺失值的属性 个数；L₁ 为第i 个属性中属性的类别数。该指标越小说明估计的效果越好。

关联规则指标是关联规则挖掘中处理类别数据的重要评价指标，其体现了关

系数据中各属性间的关联性，其定义如下：

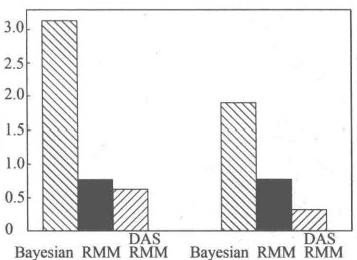


(8-13)

式中：T为一条记录中属性值组合；Q,为原始数据中第t 个项目集的数目；Q.为填 充后数据中第t 个项目集的数目。该指标度量了填充前后数据集中相同项目集出 现频数接近程度，该指标越小，填充效果越好。

4)实验结果与分析

Adult 数据集两个指标的实验结果如图8- 6中的(a) 、(b) 所示。从图 8-6中可以看出，关系马尔可夫模型的两个指标均优于贝叶斯法，因为关系马 尔可夫模型充分考虑了属性间的关联性，对缺失值的估计更准确。同时可以 看出，结合动态属性选择的关系马尔可夫模型方法估计效果更好，优于关系马 尔可夫模型方法。因为动态属性选择方法能够确定用于估计缺失值的最优属 性组合，该属性组合包含的信息量最大，能更准确地估计缺失值。同时，因为 该数据集中的缺失数据是随机缺失的，ProProp 填充方法较 MaxPost 填充方法 效果更好。

第8章 不完整数据的估计与填充(191)

RMSE

Eri lic

RMSE

Erilic



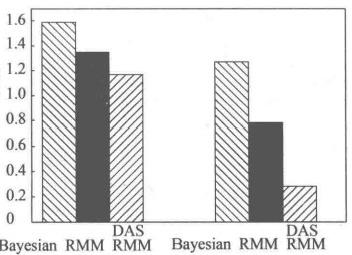
Bayesian RMM RMM Bayesian RMM RMM

ProProp

MaxPost

(a)Bayesian 、RMM 和

DAS RMM 的 RMSE 指标对比



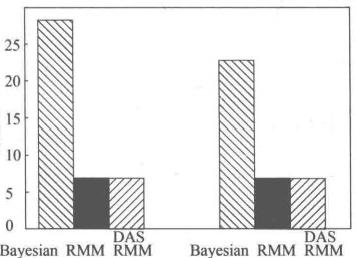
MaxPost ProProp

(b)Bayesian 、RMM 和

DAS RMM的 Erilic 指标对比

图8- 6 Adult 数据集上 Bayesian、RMM和DAS RMM的 RMSE、Erilic 指标对比

Bank Marketing数据集两个指标的实验结果如图8-7中的(a) 、(b) 所示，可 以看出关系马尔可夫模型方法的两个指标同样优于贝叶斯方法。DAS RMM 方法 的 RMSE 指标与 RMM 方法相同，但Erilic 指标优于关系马尔可夫模型法，同样说 明了DAS RMM方法的优越性。因为 Bank Marketing数据集里属性的类别较少，并 且不同类别值分布不均，某一类别的值较多，使得同一方法两种填充方式的差别不 大，但总体看来ProProp填充方式效果更好一些。



MaxPost ProProp MaxPost ProProp

(a)Bayesian 、RMM 和 (b)Bayesian 、RMM 和

DAS RMM 的 RMSE 指标对比 DAS RMM 的 Erilic 指标对比

图8-7 Bank Marketing 数据集上 Bayesian、RMM 和DAS RMM的 RMSE、Erilic 指标对比

针对程序执行的结果，以关系马尔可夫模型的运行时间为准，其他方法运行时 间除以该时间值，得到各方法的相对运行时间如表8-10所列。从中可以看出关 系马尔可夫模型和 DAS RMM 两种方法的运行时间较贝叶斯方法短，其主要原因 在于贝叶斯法需计算所有已知属性的条件概率，计算量更大。同时，由于DAS

192)数据质量导论

RMM法用多属性来预测缺失值，其运行时间较关系马尔可夫模型方法长。

表8 - 10 不同方法的运行时间

|  |  |  |
| --- | --- | --- |
| **数据集** | **方法** | **相对执行时间** |
| Adult | Bayesian | 2.62 |
| RMM | 1.00 |
| DAS RMM | 1.82 |
| Bank Marketing | Bayesian | 1.51 |
| RMM | 1.00 |
| DAS RMM | 1.22 |

8.3 基于机器学习的缺失数据估计与填充

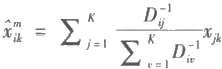
本节基于k- 近邻算法和局部敏感哈希技术，提出一种缺失数据估计与填充 方法。

8.3.1 基于k- 近邻的填补算法

k- 近邻填补算法是机器学习中常用的算法，通常被用在分类等问题中， 其主要思想是：如果一个样本在特征空间中的k 个最相似(即在特征空间中最 接近)的样本中的大多数样本属于一个类别，那么该样本也很有可能属于这个 类别。

Aittokallio(2010) 中将k- 近邻算法运用到数据填补中，设一个待填补项x;= (x;,x;“),其中x;为其中的完整部分，x”为缺失的部分，D,表示项i和项j之间的距

离，则x;”的填补值为



(8-14)

式(8-14)中取k 个距离x;最近的记录，并以其距离的倒数(即相似度)作为 权重，通过计算这k 条记录的加权均值作为估计值来进行填补。当待填补的属性 为枚举型时，可以选择k 条记录中出现最多的值作为填补值。

k- 近邻算法具有精度高，对异常值不敏感的特点，但是由于需要计算所有记 录间的相似度，具有较高的计算复杂度，填补效率较低(Anagnostopoulos,2014),在 数据量不断增长、数据实时性要求越来越高的背景下，已经难以满足实际应用的 需求。

第8章 不完整数据的估计与填充(193)

8.3.2 局部敏感哈希技术

局部敏感哈希(Locality Sensitive Hash,LSH)理论用于在海量高维数据中寻找 近似相似项。其基本思想是：将原始特征空间中相邻的两个数据点通过相同的投 影或者变换后，这两个点具有较大概率仍然相邻，而在原始特征空间中不相邻的点 具有较大概率仍然不相邻(Rajaraman,2012)。

局部敏感哈希的一般做法是对目标项进行多次哈希处理，这样相似项会比不 相似项更有可能哈希到同一个桶中。设d(x,y) 表示x,y 之间的距离，则将这个过 程中使用的多个哈希函数f 组成的一个函数族F 称为(d₁,d₂,p₁,p₂) 敏感的函数 族，如果其中的每一个函数f 都满足下列条件：

(1)如果d(x,y)≤d₁, 那么f(x)=f(y) 的概率至少为p₁;

(2)如果d(x,y)≥d₂, 那么f(x)=f(y) 的概率最大为P₂。

不同的距离度量下，需要选择不同的局部敏感哈希函数族，如 Jaccard 距离下 的最小哈希函数族，欧式距离下的基于p- 稳定分布的哈希函数族。

8.3.3 LSH\_KNN 数据填补算法

由于在使用k- 近邻算法进行填补时需要计算待填补记录与其他所有完整记 录之间的距离，当数据量较大时计算时间难以承受。在现实应用中的相似记录搜 索中，往往不需要真的获取真正的“最近邻”,而是满足用户指标的近似相似记录 即可。局部敏感哈希可以用来在大规模数据中寻找近似的相似记录，考虑使用局 部敏感哈希来加速k- 近邻算法。

定理8-1:对于所有距离不大于d₁ 的两条记录，通过(d₁,d₂,p₁,p₂) 敏感的哈 希函数族F 映射之后，被认为不相似的概率不大于(1-p₁)'”'。

证明：由8.3.2节可知，对(d₁,d₂,P₁,P₂) 敏感的哈希函数族来说，当两条记录 的距离不大于d₁时，发生哈希“碰撞”的概率至少为p₁, 则不发生碰撞的概率至多 为1 -p₁ 。由于F 中的哈希函数互相独立，经过函数组F 中每个哈希函数映射后 都不发生碰撞的概率至多为(1-p₁)'"。

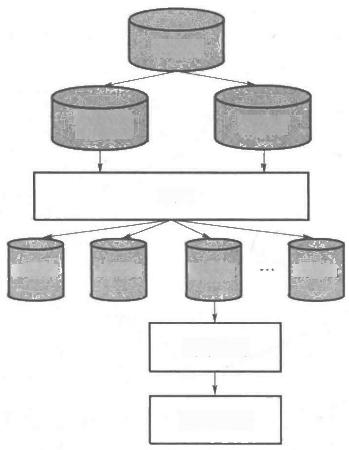
定理8-2:对于所有距离不小于d₂ 的两条记录，通过(d₁,d ₂,P₁,P₂ )敏感的哈 希函数族F 映射之后，被认为相似的概率不大于1-(1-p₂)''。

证明：由8.3.2节可知，对(d₁,d₂,P₁,p₂) 敏感的哈希函数族来说，当两条记录 的距离不小于d₂时，发生哈希“碰撞”的概率至多为p₂, 不发生碰撞的概率至少为 1-p₂ 。 由于F 中的哈希函数互相独立，经过函数组F 中每个哈希函数映射后都不 发生碰撞的概率至少为(1-p₂)', 则被认为相似的概率为1-(1-p₂)'。

194)数据质量导论

定理8-1和定理8-2分别考虑的是相似记录搜索中的“假阴性”和“假阳 性”两种错误，通过选择合适的哈希函数族以及函数族的“与”和“或”构造，可以尽 量降低两类错误发生的概率，获得满足用户相似记录搜索需求的局部敏感哈希函 数族，并用来进行k- 近邻填补。

如图8-8所示，为了提高k- 近邻数据填补算法的效率，在填补之前首先为 所有完整的数据建立哈希索引，也就是利用局部敏感哈希将其分配到各个哈希桶 中。然后将每条待填补的记录，同样哈希到桶中，然后从桶中寻找k 条最相似的数 据，并利用式(8-14)对目标项进行填补。



数据集

完整集

LSH

Bucket Bucket Buckel Bucket

近邻查找

KNN 填补

缺失集

图8-8 LSH\_KNN 数据填补算法示意图

由于不同的距离度量下寻找近似相似候选项的局部敏感哈希函数族是不同 的，下面讨论在两种常见的距离度量下的利用局部敏感哈希寻找近似相似候选项 的方法：Jaccard 距离下的最小哈希方法和欧式距离下的基于p- 稳定分布的局部 敏感哈希方法，混合型距离度量下的局部敏感哈希策略。

<8.3.3.1> Jaccard 距离与最小哈希

当数据集中的属性都为枚举型且每个属性的值域都不相交时，每条记录都可以 表示为一个集合。这时可以使用Jaccard 距离来度量两条记录x,和x;之间的距离：

 (8-15)

集合的Jaccard 距离与集合最小哈希(Min-hashing) 值之间存在十分密切的 关联：两个集合经过随机排列转换后得到的两个最小哈希值相等的概率等于这两

第8章 不完整数据的估计与填充(195)

个集合的 Jaccard 相似度(王斌译，2012)。

表8-11是四个集合S₁,S₂,S₃,S₄ 的矩阵表示，每一列表示一个集合，当一列的 某一行取值为1时表示该行元素属于这一行表示的集合，如S₁为{A,D} 。 为了计算 这四个集合的一个最小哈希值，随机选择n 个哈希函数h₁,h₂,…,h 。, 令 sig(i) 为签 名矩阵中第i 个哈希函数在目标项上的值，首先对所有i 初始化sig(i) 为 ，然后 对目标项的每一行r做如下处理：

(1)计算h₁(r),h₂(r),…,h,(r);

(2)如果该行取值不为0,更新sig(i)=min{sig(i),h₁(r)}。

表8- 11 集合的矩阵表示

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **元素** | S₁ | S₂ | S₃ | S₄ |
| A | 1 | 0 | 0 | 1 |
| B | 0 | 0 | 1 | 0 |
| C | 0 | 1 | 0 | 0 |
| D | 1 | 0 | 1 | 1 |
| E | 0 | 0 | 1 | 1 |

在得到目标项的最小哈希签名向量之后，将其划分为b 段，每段的长度为r。 每段最小哈希签名对应一个独立的桶数组，对每段签名构造一个哈希函数，将该段 签名组成的向量映射到一个桶数组中。只要两个项最小哈希之后，在b 段签名中 任一段对应的哈希函数值相等，那么这两个项就是相似候选对。在b 段签名产生 的所有候选对中，选择其中k 个距离目标项最近的项之后再使用k- 近邻算法进 行填补即可。

<8.3.3.2> 欧式距离与 E2LS H

当数据集中的属性为数值型时，数据集中的记录可以表示为欧式空间中的数 据点，使用欧式距离来度量两个记录之间的距离，其中x₁ 表示第i 条记录的第v 个

属性值：



(8-16)

Datar 提出了一种欧式距离下基于p- 稳定分布的局部敏感哈希方法(Datar, 2004;Andoni,2014), 并在 Datar的基础上实现了其开源版本——E2LSH(Exact Eu- clidean Locality-sensitive Hashing)。基于p- 稳定分布的局部敏感哈希是一种随 机化的局部敏感哈希方式，其原理是利用基于p- 稳定分布的局部敏感函数对数 据进行降维，使原本就接近的两个点在降维之后仍然接近。

其具体算法是将数据点投影到随机的直线上，随机直线的方向矢量的分量互

196)数据质量导论

相独立且服从p- 稳定分布。由于p- 稳定分布的特性，两个点在欧式空间下距离 与其投影到一个随机向量之后的距离是同分布的，因此两个点的投影距离越小则 其原始距离也就越小。将直线按等长划分，且对每条线段进行编号，则投影点所在 的线段编号即为哈希函数取值。

由于数据的不完整性，在欧式距离度量下进行结合局部敏感哈希的k- 近邻 数据填补时有两个问题需要解决：

(1)p- 稳定 LSH 不能对存在缺失的数据进行哈希。这是因为当数据缺 失时，记录表示的不再是一个点，而是退化为1条线(当有1个属性缺失时)甚 至1个面(当有2个属性缺失时)。根据缺失的情况不同可以采用以下两种处 理策略：

忽略缺失属性值：当数据缺失集中在数据集中的某几个属性上，可在计算记录 (包括完整的记录和存在缺失的记录)哈希值时只计算记录的完整属性值，而令缺 失属性值为0。

随机或者均值填补：当缺失属性分布较为分散，第一种策略不再适用时，可以 用随机值或属性均值将目标项补充完整后，再计算其哈希值。

(2)在获得候选的相似项之后，需要计算候选项与待填补项的欧式距离。而 在数据有缺失的情况下，由于记录的部分属性取值未知，难以计算两个记录之间的 欧式距离。

Hathaway 等(2001)提出了一种针对不完整数据的局部距离度量方法，设x 是 不完整的记录，y 为完整记录，则x 和y 之间的局部欧式距离为

(8-17)

(8-18)

利用式(8-17)和式(8-18)中的局部欧式距离可以在数据不完整的情况下 度量数据间的距离。

<8.3.3.3> 混合距离度量下的局部敏感哈希技术

实际中的数据集中枚举型属性和数值型属性可能同时存在，大部分局部敏感 哈希技术都不能同时对这两种类型的属性进行局部敏感哈希。

因此，需要将数据集中的属性分为两个部分，然后对两个部分的属性分别使用 对应的局部敏感哈希方法建立哈希索引。具体做法如下：

设两部分属性对应的局部敏感哈希函数族分别为F={f,f₂,…,f.} 和 G= {gi,g₂,…,g m},对任意两个项x={x⁶,x^} 和 y={y°,y\*} (其 中 x“表示其中的数 值型部分，x^表示其中的枚举型部分),如果存在函数he(FUG), 有 h(x)=h(y),

第8章不完整数据的估计与填无(197

则x 和y 是候选的相似记录。

确定候选的相似项之后，为了获取近似的近邻项集，需要计算候选对的真实距 离。设d° 和d\* 分别为数值型距离度量和枚举型距离度量，则项x 和y 的混合距

离为



(8-19)

式(8-19)分母中max(d⁶) 和 max(d\*) 是为了对距离度量进行归一化。如果 其中一类属性占属性集的比例很小，也可以忽略这类属性，只对属性集中的多数类 属性进行局部敏感哈希。

<8.3.3.4> LSH\_KNN 算法复杂度分析

与经典的k- 近邻算法相比，LSH\_KNN 所耗费的时间分为两个部分：建立局 部敏感哈希索引以及搜索k- 近邻。设数据集中的完整记录有c 条，存在缺失的记 录有m 条，局部敏感哈希函数组容量为f, 每个局部敏感哈希函数对应的桶数组为 h 个，每个桶数组的容量为n, 由于在n 个数据中寻找最小k 个项的复杂度为n×

logk,因此搜索k- 近邻并进行填补耗费的时间为m×f×n×(1+logk) 。 在平均情

况下有n≈c/h,因此总耗费的时间约为c×f×d+m×f×n×(1+logk), 而经典的

k- 近邻算法所消耗的时间为c×(1+logk)×m。

因此，当需要填补的记录数m>f×h/[(h-f)×(1+logk)] 时，结合局部敏感 哈希的k- 近邻算法的填补时间将会低于经典的k- 近邻算法。

8.3.4 实 验 验 证

为了说明结合局部敏感哈希的k- 近邻填补算法(LSH\_KNN) 的准确性和效 率，以UCI的 Abalone 、Gas sensor array under dynamic gas mixtures(包括 ethylene\_ methane 和 ethylene\_CO 两个数据集)和 HIGGS 三个数据集为例，对数值型缺失数 据进行填补，其中使用到的 E2LSH 算法是基于开源程序E2LSH0.1 进行的修改 (Andoni,2014)。

<8.3.4.1> 实验环境

实验环境：Inter Xeon E52609 V2处理器(主频2.5GHz), 内存128GB, 操作系 统为 Centos 6.4(64 位)。

实验数据集一：Abalone数据集有9个属性，其中1个枚举型属性，8个数值型 属性。由于枚举型属性所占比例较小，所以在局部敏感哈希时忽略枚举型属性而 只是用数值型属性。数据集有4177条记录，将按照2:1划分为训练集和测试集， 将测试集中的 Rings(年龄)属性置空后再进行填补。

实验数据集二：Gas sensor array under dynamic gas mixtures 数据集有19个属

198)数据质量导论

性，全部为实数型属性，共有4178504×2条记录，数据文件大小为1.19GB。 该数 据集按照时间序列记录了16个气体传感器获取到的空气中的乙烯\_甲烷混合物 (对应数据文件 ethylene\_methane)以及乙烯\_一氧化碳混合物(对应数据文件 eth- ylene\_CO)的浓度。两个文件的结构相同，前3个属性分别为时间(单位为s), 甲 烷(或一氧化碳)浓度(单位为10~⁶),乙烯浓度(单位为10~⁶),后16个属性对应 16个传感器记录的数据。原数据集不存在缺失，通过随机不放回抽样将ethylene\_ methane 和 ethylene\_CO 两个数据集按照4:1划分为训练集和测试集，将测试集的 最后一个属性置空然后进行填补。

实验数据集三：HIGGS 数据集有28 个属性，全部为实数型属性，共有 11000000 条记录，数据文件大小为7.48GB。该数据集通过蒙特卡罗方法模拟生 成，前21个属性为粒子加速器测量的动力学特征，后7个属性为前21个属性的函 数。原数据集是不存在缺失的，选择后500000条记录为测试集，随机将测试集中 每条记录后7个属性中的一个属性置空再进行填补。

[**8.3.4.2**](8.3.4.2) **实验方法**

实验中分别使用结合局部敏感哈希的k- 近邻填补算法(LSH\_KNN)、经典 k- 近邻(KNN)、多元线性回归(REGRESS)3 种算法对4个数据集(Abalone,Ethyl- ene\_CO,Ethylene\_Methane,HIGGS)进行填补。通过计算填补的均方根误差和填补 时间，比较3种算法的填补准确性和填补效率。

为了避免近邻个数k 可能带来的偶然性，实验中设置参数k 为3、5、7、9,多次 进行实验。实验中使用到的多元线性回归算法直接调用Matlab 回归分析工具包 regress,使用训练集对中其他未缺失属性对缺失属性建立回归模型后，利用建立的 模型对缺失值进行预测。

实验中对算法的评价主要分为准确性和效率，用均方根误差度量准确性，对同 一数据集填补的完整过程耗费的时间来评价效率。对结合局部敏感哈希的k- 近 邻算法来说耗费的时间包括对完整数据集进行局部敏感哈希，以及搜索待填补数 据的近似k- 近邻并进行填补两部分时间。原始的k- 近邻填补算法耗费时间为 搜索待填补项k- 近邻，然后进行填补的时间；而多元线性回归算法耗费时间为模 型建立和利用模型估计两个部分。

[**8.3.4.3**](8.3.4.3) **实验结果**

表8-12至表8-15分别是在四个数据集上进行缺失填补实验的填补 均方根误差和填补用时。其中KNN 代表经典的k- 近邻算法，LSH\_KNN 代 表结合局部敏感哈希的k- 近邻算法，其后的数字代表近邻个数，如 KNN\_3 表示设置近邻个数k 为3的经典k- 近邻算法，REGERESS 代表多元线性回 归方法。

第8章不完整数据的估计与填充 

表8 - 12 Abalone 数据集实验结果

|  |  |  |
| --- | --- | --- |
| **算法** | **准确性(RMSE)** | **填补时间/s** |
| KNN\_3 | 0.0132 | 821.5900 |
| KNN\_5 | 0.0235 | 712.1500 |
| KNN\_7 | 0.0283 | 660.5000 |
| KNN\_9 | 0.0307 | 699.0800 |
| LSH\_KNN\_3 | 0.0132 | 129.7300 |
| LSH\_KNN\_5 | 0.0235 | 92.0700 |
| LSH\_KNN\_7 | 0.0283 | 108.8600 |
| LSH\_KNN\_9 | 0.0307 | 104.5600 |
| REGERESS | 2.0703 | 0.0200 |

表8- 13 Ethylene\_CO 数据集实验结果

|  |  |  |
| --- | --- | --- |
| **算法** | **准确性(RMSE)** | **填补时间/s** |
| KNN\_3 | 6.4594 | 208374.3800 |
| KNN\_5 | 6.3218 | 208340.2900 |
| KNN\_7 | 6.3435 | 209321.6000 |
| KNN\_9 | 6.4109 | 209267.7800 |
| LSH\_KNN\_3 | 6.6249 | 3071.7900 |
| LSH\_KNN\_5 | 6.3756 | 3093.9400 |
| LSH\_KNN\_7 | 6.3598 | 3095.8900 |
| LSH\_KNN\_9 | 6.4023 | 3083.5500 |
| REGERESS | 203.3312 | 6.8600 |

表8-14 Ethylene\_Methane 数据集实验结果

|  |  |  |
| --- | --- | --- |
| 算法 | **准确性(RMSE)** | 填补时间/s |
| KNN\_3 | 6.4578 | 183684.6900 |
| KNN\_5 | 6.3211 | 184630.3700 |
| KNN\_7 | 6.3425 | 174005.7800 |
| KNN\_9 | 6.4073 | 185947.0400 |
| LSH\_KNN\_3 | 6.6260 | 7587.4300 |
| LSH\_KNN\_5 | 6.3767 | 7657.8900 |
| LSH\_KNN\_7 | 6.3608 | 7682.6200 |
| LSH\_KNN\_9 | 6.4034 | 7615.6900 |
| REGERESS | 48.3738 | 6.4600 |

**200)**数据质量导论

表8- 15 HIGGS数据集实验结果

|  |  |  |
| --- | --- | --- |
| **算法** | **准确性(RMSE)** | **填补时间/s** |
| KNN\_3 | 0.1159 | 408723.6600 |
| KNN\_5 | 0.1148 | 408798.3100 |
| KNN\_7 | 0.1150 | 409015.2900 |
| KNN\_9 | 0.1160 | 409265.3000 |
| LSH\_KNN\_3 | 0.1168 | 11203.1100 |
| LSH\_KNN\_5 | 0.1151 | 11951.2800 |
| LSH\_KNN\_7 | 0.1158 | 11829.6400 |
| LSH\_KNN\_9 | 0.1153 | 14241.3900 |
| REGERESS | 0.1268 | 40.3200 |

表8-12至表8-15中的实验结果表明：

在多个数据集上LSH\_KNN 算法误差水平基本保持与KNN算法相同，而RE- GERESS算法在四个数据集上的误差水平都高于KNN 和 LSH\_KNN。可能的原因 一是数据集其他属性与缺失属性之间的关系不一定是线性关系；二是基于回归的 填补方法对于异常值敏感。

在所有数据集上 REGERESS填补用时最短，KNN 用时最长，特别是在三个较 大的数据集上填补时间都超过了1天。LSH\_KNN 填补用时介于两者之间，并且最 长不超过4小时。

综合填补准确性和填补效率两个指标可以发现，多元线性回归填补虽然 用时很短，但是准确性较差难以满足应用需求，而结合局部敏感哈希的k- 近 邻填补算法的准确性和传统的k- 近邻填补算法相差无几，并且保持较高的 填补效率。

**8.4** **函数依赖一致性数据生成**

本节将数据生成作为所有属性值均缺失的极端情况，提出了一种基于函数依 赖的数据生成方法。

**8.4.1** **函数依赖一致性**

函数依赖描述的是数据库中两条或两条以上记录的属性之间的多对一(函 数)映射关系。对于关系模式R,∑ ={F₁,F₂,…,F m},m≥0 为定义在属性集

atr(R)={A₁,A₂,…,A,},n≥1 上的函数依赖集合，F,e∑ 为任一函数依赖，F;形

第8章不完整数据的估计与填充(201)

式为F;=X;→Y, 其中X;,Y;Cattr(R) 分别是 F;的左部属性集和右部属性集，分别 记为X;=LHS(F)={x₁,x₂,…,xixn} 和 Y,=RHS(F)={y₁,y₂,…,yim},X; 和 Y; 之间可以存在交集。X,( 或Y) 中的任一属性xr,k=1,2,…,IX;I ( 或yi,l=1,

2,…,IY;1) 记为X(k) ( 或Y,(l)) 。1={t₁,t₂,…,t,}, p≥1 为关系模式R 的一个 实例，t,q=1,2,…,p 为实例1的一个元组，元组t 中属性 A;对应的属性值记为 t₀(A;) 。给出函数依赖一致性的形式化定义如下。

**定义8-3(函数依赖一致性):**对于任一关系模式R, 及定义在R 上的函数依 赖集合 ∑,F;为≥中的任一函数依赖，I 为 R 的一个实例， ,t₂ 为1中的任意两个元 组，称关系模式R 的实例I 满足函数依赖一致性(记为I|=2), 当且仅当以下条 件成立：

(1)如果t₁(x)=t₂(x₄),x₄eLHS(F),k=1,2,…,ILHS(F;)I 成立，则有 t₁(y₁)=t₂(y₁),yi₁∈RHS(F),l=1,2,…,IRHS(F;)l 成立。

(2)如果t₁(y₁)=t₂(y₁), 允许t₁(x₄)≠t₂(x₄)。

定义8-3表明，在满足函数依赖一致性的实例中，对于任一LHS(F;) 属性值， 只有一个RHS(F) 属性值与之对应，但对于一个 RHS(F;) 属性值可以存在多个 LHS(F;)属性值与之对应。

基于定义8-3,函数依赖一致性数据生成问题可以描述如下：对于关系模式 R,及其R 上定义的函数依赖集合 ，为R 的各个属性生成指定数量的属性值，并 保证由生成的属性值构成的实例II=2。

目前关于函数依赖一致性数据生成的研究相对较少。Alexander(2004) 针 对确定的关系模式，采用构建临时表(Temporary Table)和指定有效对(Valid Pair)的方法，预先构造满足函数依赖约束的属性值字典和映射关系，通过SQL 语句向数据库插入满足函数依赖约束的属性值对。这种方法具有较高的生成 效率，但一旦关系模式发生改变必须重新构建属性值字典和映射关系，不具通 用性。Fan 等(2008,2014a) 及 Fagin等(2014)提出了多种函数依赖一致性修 复方法，利用“先生成后修复”的方法可以获得满足函数依赖约束的一致性数 据，但由于函数依赖一致性修复问题是 NP-hard 问 题(Fan et al.,2008, 2014b), 此类方法的实际效率往往难以满足应用需求。针对现有函数依赖一 致性数据生成方法的不足，本章提出一种通用、高效的函数依赖一致性数据生 成方法。

**8.4.2** **单函数依赖一致性数据生成算法**

对于关系模式R, 它包含的属性集用 attr(R)={A₁,A₂,…,A。},n≥1表示，任 一属性A,eattr(R),i=1,2,…,n 是一个二元组<N,,C;>, 其 中N;为该属性的属

202)数据质量导论

性名称，C;为该属性必须满足的约束集合，如类型、长度、格式、取值范围和概率分 布等。这里并不关心单属性的具体生成算法，而是将单属性生成算法视为一个抽 象方法 GNA(), 对于任意A,eattr(R),GNA(A₁) 总能返回一个满足约束集 C;的随 机值作为A,的属性值。

基于此，本节中不考虑单个属性的实际语义和具体生成实现，对任意类型的单 属性生成算法均以相同的抽象符号表示，任何一种单属性生成算法均可应用于本 章提出的方法。对于不同关系模式和函数依赖集合，数据生成方法无需进行修改， 只是单属性生成算法GNA() 的具体实现不同，因此本节提出的方法通用于各种不 同关系模式和函数依赖集合。

在函数依赖一致性数据生成的各种情况中，最简单的情况是函数依赖集合为 空，即II=0, 只需使用各种单属性生成算法为各个属性生成指定数量的属性值 即可。当|ZI>0 时，必须对单属性生成算法加以约束，保证每个函数依赖左部属 性集与右部属性集之间的多对一映射关系。当|I=1 时 ，F 为乙中的唯一的函数 依赖，为了使生成的数据满足函数依赖一致性，对于F 中包含的属性必须逐个元 组生成数据，且必须先生成LHS(F) 属性值，再通过LHS(F) 属性值确定 RHS(F) 属性值，从而保证左部、右部属性值之间的多对一映射关系，具体单函数依赖一致 性数据生成算法(Tuple Generator with Single FD,TGSFD)如表8-16所列。

表8 - 16 TupleGeneratorwithSingleFD 算法

|  |
| --- |
| **Name:TupleGeneratorwithSingleFD(R**,F) |
| **Input:**关系模式R,函数依赖F  **Output:**R的一个元组 |
| 1.1=φ  2.For eachA;r(RHS(F))and t(A;)==**NULL**  3.t(A₁)=GNA(A₁)  4.For each t,el  5.Ift₄ (LHS(F))==t(LHS(F)) **then**  6. e=**TRUE;**  7. t(RHS(F))=t(RHS(F));  8.Else  9. For each A;e RHS(F)and e≠**TRUE**  10. (A;)=GNA(A;);  11.End if  12.I.add(t);  13.**Return** t; |

第8章 不完整数据的估计与填充(203)

如表8 - 16所列，TGSFD 算法首先生成元组t 的 LHS(F) 属性值(第2～3 行),然后遍历已生成的实例I, 如果存在与t 具有相同LHS(F) 属性值的元组t, 则 将t的 RHS(F) 属性值设为与t 中相同(第4~7行),如果不存在，则使用单属性 生成算法为t 的每个RHS(F) 属性生成属性值(第9～10行)。由于TGSFD算法 需要遍历所有已生成的元组，因此生成n 个元组的时间复杂度为0(n²)。

8.4.3 基于有向无环图的多函数依赖 一致性数据生成

当|El≥2 时，最直观的生成方法是对于每一个函数依赖应用TGSFD算法生 ₁ :A→B,F₂ :C→

A},如果按照F₁ 先于F₂ 的顺序应用TGSFD算法生成数据，在对F₁ 的处理过程中 生成满足F₁ 约束的A 和B 的属性值，然后对F₂ 进行处理。由于A 的属性值已经生 成，为了保证生成的数据满足F₂ 的约束，算法会依据生成的C 的属性值重新生成 A的属性值，显然这样便无法再保证最终生成的数据满足F₁ 的约束。有两种方法 可以解决上述问题： 一种是采用函数依赖递归的方法，当A 的属性值发生改变后， 重新对 F₁应用TGSFD算法生成B 的属性值；另一种是采用函数依赖排序的方法， 依据函数依赖的传递关系对三中的函数依赖进行排序，使得F₂ 先于F₁ 被处理。对 于本例中的两个函数依赖，函数依赖递归方法需要运行三次 TGSFD 算法，但实际

应用中的函数依赖的数量和关系要复杂得多，在最差的情况下需要运

次 TGSFD算 法(n=|ZI), 如此的复杂度在大规模数据生成中是难以接受的。对 于函数依赖排序的方法，由于在实际应用中的函数依赖可能存在冗余属性和 冗余依赖，因此可能无法直接依据函数依赖的传递关系对所有的函数依赖进 行排序。

TGSFD算法需要依据函数依赖的左部属性值确定其右部属性值，多函数依赖 一致性数据生成的问题本质上是属性生成顺序的问题。任一函数依赖的左部属性 要先于其右部属性生成数据，否则就可能生成违反函数依赖约束的数据，或者需要 额外的工作进行纠正。对于函数依赖一致性数据生成而言，单个函数依赖描述了 其左部、右部属性之间的先后顺序，当存在多个函数依赖时，需要根据各个函数依 赖中属性的先后顺序计算所有属性的全序关系。为此，将函数依赖集合描述为有 向无环图，并利用图论中的方法对所有属性进行排序。

<8.4.3.1> 函数依赖集合的有向无环图模型

由于函数依赖集合 中可能包含冗余属性和依赖，并不直接对 进行建模， 而是首先计算与 等价的极小覆盖，使每个函数依赖的右部只包含一个属性， 并去除三中的冗余属性和依赖，然后用有向图对 的极小覆盖进行建模。在

204)数据质量导论

本章中，除非特殊说明，之均指的是三的极小覆盖。对于函数依赖一致性数据 生成而言， 一般情况下，要生成的元组数目远远大于关系模式中属性的数目以 及其上存在的函数依赖的数目，函数依赖集合的极小覆盖求解时间与数据生 成时间相比可以忽略不计，在统计数据生成时间时可以不考虑极小覆盖求解 的时间。

函数依赖集合2(IZI≥1) 包含的属性集合记为 V,则

 (8-20)

定义8-4(函数依赖关系):对于任一函数依赖 F,称属性xεLHS(F) 与 y ε RHS(F) 之间存在函数依赖关系，即y 依赖于x,记为<x,y>。

函数依赖关系表示同一个函数依赖中，右部属性依赖于左部属性，在数据生成 过程中，左部属性先于右部属性生成。显然，对于数据生成而言，函数依赖关系是 偏序关系。

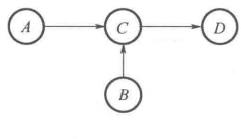
函数依赖集合三中函数依赖关系的集合记为E, 则

E={<x,y>lx∈LHS(F),y∈RHS(F),Fe2} (8-21)

对于任一函数依赖集合2,其有向图模型记为G=(V,E), 其中结点集V是有 限非空集合，元素为 包含的属性集合，边集E 是三中函数依赖关系的集合，任一 条边e ∈E 的起点为某一函数依赖的左部属性，终点为同一函数依赖的右部属性， V和 E 中的元素存在对应关系。

例如，对于函数依赖集合2={ F₁:(A,B)→C,F₂:C→D,F₃:H→ J}, 其包含的

属性集合V={A,B,C,D,H,J}, 关系集合E={<A,C>,<B,C>,<C,D>,<H, J>}, 其有向图模型如图8-9所示。



H  J

图8-9 函数依赖集合的有向图模型

命题8 - 1:任一不存在循环依赖的函数依赖集合三的有向图模型是有向无 环图。

证明：如上文所述，这里函数依赖集合三指的是三的极小覆盖。通过反证法 证明。假设函数依赖集合三中不存在循环依赖，且其有向图模型中存在一个环。 由于 ∑是极小覆盖，其中任一函数依赖只有一个右部属性，对于该环上任一结点

第8章不完整数据的估计与填充(205)

A,依据 Armstrong公理肯定可以推出如下形式的函数依赖：{QUA}→A, 其中Q 为 以该环上结点为终点的边的起点组成的属性集合。{QUA}→A 说明属性A 依赖 于其自身，即函数依赖集合 中存在循环依赖，与假设矛盾，因此命题8-1成立。

由于关系数据库中很少存在循环依赖，下文中出现的函数依赖集合中均默认 不存在循环依赖，因此，函数依赖集合对应的有向图模型均为有向无环图。对于任 一有向无环图模型G=(V,E),V={0₁,v₂,…,v₀}, 采用nxn 邻接矩阵N=[a;] 进 行表示：

 (8-22) 任一结点号的入度 表示该结点函数依赖的结点数目。

<8.4.3.2> 属性排序

函数依赖关系定义了同一个函数依赖中左部、右部属性间的偏序关系，为了 保证所有函数依赖的左部属性均先于其右部属性生成，必须依据此函数依赖 关系得到函数依赖集合乏中所有属性的一个全序。基于上文建立的有向无环 图模型，将此问题转化为有向无环图中的拓扑排序问题进行求解，具体算法如 表8- 17所列。

表8- 17 TopoSort 算法

|  |
| --- |
| Name:TopoSort(V,N) |
| Input:2的属性集合V,有向无环图的邻接矩阵N  Output:拓扑序topo\_order |
| 1.topo\_order =中  2.While V≠φ  3.For each t,eV  4. If d~(v₁)==(  5. topo\_order.append(v;);  6 V=V-t  7. Fori=1 to n  8 a=0;  9 **Break;**  10 **End** **for**  11. **End** **if**  12. End for  13.End while  14.Return top\_order; |

**206)**数据质量导论

如表8-17所列，算法选择任意一个入度为0的结点且将其输出(第4～5 行),然后从图中删除该结点和以它为起点的边(第6～8行),重复这两个步骤直 到所有的结点都已输出(第2行)。

TopoSort 算法以函数依赖集合三的有向无环图模型为输入，每次选择一个入 度为0的结点依次输出，如果某一结点被输出，说明该结点不依赖任何其他属性， 或者它依赖的左部属性在此之前已经被输出，在最终得到的属性序列中，工中的任 一函数依赖的右部属性均排在其依赖的左部属性之后。值的注意的是 TopoSort 算 法得到的结果并不唯一，例如图8-9中的函数依赖集合的属性排序结果可能是 <A,B,C,D,H,J>, 也可能是<A,H,J,B,C,D> 、<A,B,C,H,J,D> 等，但任何一 个排序结果均能保证任一函数依赖的左部属性先于其右部属性(如本例中所有排 序结果中属性A,B 均排在属性C 之前，属性C 排在属性D 之前),因此任何一个排 序结果均能符合属性值生成顺序的要求。

得到三中所有属性的全序关系后，便可按照全序关系中的先后顺序使用单属性 生成算法为各个属性生成数据，只有当一个属性是某一函数依赖的右部属性时才需调 用TGSFD算法，由于极小覆盖中的每个函数依赖只包含一个右部属性，因此只需运行 n=I2I 次TGSFD算法。然而由于TCSFD算法本身具有O(n²) 的时间复杂度，生成效 率依然不高，为了进一步提高生成的效率，依据有向无环图模型，将三中的所有属性划 分为多个互不相关的最小独立属性子集(见8.4.1节),使用流水线技术提高生成效率。

**8.4.4 属性集划分和数据生成流水线**

在三中包含的属性集中，并不是任意两个属性之间都存在依赖关系。例如图 8-9中的函数依赖集合2={F₁:(A,B)→C,F₂:C→D,F₃:H→J}, 属性A 和B 之 间不存在依赖关系，属性集{A,B,C,D} 中的任一属性与属性集{H,J} 中的任一属 性之间都不存在依赖关系。对于A 和 B, 由于A,B 都与C 存在依赖关系，在调用 TGSFD 算法生成C 的属性值时，无法在没有A或B 的属性值的情况下生成满足F₁ 约束的数据；而对于{A,B,C,D} 和{H,J},{H,J} 的属性值并不影响 TGSFD算法 为{A,B,C,D} 中的属性生成满足F₁ 和F₂约束的数据，反之亦然。因此，可以将V 划分为{A,B,C,D} 和{H,J} 两个属性子集，分别生成数据。

[**8.4.4.1**](8.4.4.1) **最小独立属性子集划分**

**定义8-5(独立属性子集):给**定函数依赖集合，V 为三中包含的属性集合， E为三中存在的函数依赖关系集合，SCV, 对于任意属性xεS 和yɛV-S, 如果有 <x,y>≠E 且 <y,x>éE 成立，则称S 为 V的一个独立属性子集。

**定义8-6(最小独立属性子集):**给定函数依赖集合2,V为 中包含的属性 集合，E 为三中存在的函数依赖关系集合，有S 为 V的一个独立属性子集，如果不

第8章不完整数据的估计与填充 

存在 V的独立属性子集S'CS, 则称S 为 V的一个最小独立属性子集。

显然，两个独立属性子集可能存在交集，而两个最小独立属性子集之间没有交 集，且所有最小独立属性子集构成属性集V的一个划分。

基于乏的有向无环图模型，将最小独立属性子集划分问题转化为有向无环图 中的弱分图划分问题进行求解，具体算法如表8-18所列。

表8- 18 MinIndpPart 算法

|  |
| --- |
| **Name:MinIndpPart(**G) |
| **Input:2**的有向无环图模型G  **Output:**包含的属性集V的一个划分 |
| 1.UDG=Undirected(G);  2.Partitions =φ;  3.VistNodes =φ;  4**.While(**V-VistNodes≠φ)  5. node=PickNodeFrom(V-VistNodes);  6.S-DFSTraverse(node,G);  7. Partitions.add(S);  8. VistNodes =VistNodes US;  **9.End** **while**  **10.Return** Partitions; |

如表8-18所列，算法首先将有向无环图模型中的有向边转换为无向边(第1 行),以任一结点为起点进行深度优先遍历，直到 V中所有结点都被遍历过一遍， 每次遍历经过的结点就是V的一个最小独立属性子集，(第4～7行)。为了避免 结点被重复遍历，使用VistNodes 记录遍历过结点，在选择起点时只从没有遍历过 的结点中进行选择(第4行，第8行)。

[**8.4.4.2**](8.4.4.2) **数据生成流水线**

**定理8-3:** 给定函数依赖集合 ∑,V为∑ 中包含的属性集合，E 为乏中存在的 函数依赖关系集合，P={S₁,S₂,…,S,} 为 V 的一个划分，其中S,i=1,2,…,n 为 V 的最小独立属性子集，2(S;) 为每个最小独立属性子集上的函数依赖集合，如果 Il=2(S) 为每个最小独立属性子集对应的一致性实例，则由各个 I;连接得到的 实例I 为属性集 V 的一个一致性实例，即Il=2。

证明(反证法证明):由于P 是 V 的划分，显然I 为属性集V 的一个实例，只需 证 明Il=2 。 假设每个最小独立属性子集对应的实例I1=2(S;), 由各个I;连接 得到的属性集V的实例Il≠2 。 由定义8 - 3知I 中必然存在两个元组t₁ 和t₂ 违反 之中的某一函数依赖的约束，不妨设该函数依赖为F;。 由定义8 - 5知F,中包含的 属性必然属于某个最小独立属性子集，不妨假设该最小独立属性子集为S, 则有

208)数据质量导论

FjεZ(S),I 在S 上的投影为S₂对应的实例I, 且Il≠Fj 。 又由于F,∈2(S), 所 以有Il≠2(S), 与假设每个最小独立属性子集对应的实例I,I=E(S;) 矛盾。

由定理8-3可知，将 ∑ 中包含的属性集V 划分为多个最小独立属性子集，可 以依据<8.4.3.2>节得到的全序关系分别为每个最小独立属性子集生成函数依赖一 致性实例，然后将各个实例连接得到满足函数依赖集合三约束的一致性数据。可 见，生成一条元组的任务可以划分为多个相互独立的模块，每个模块对应一个最小 独立属性子集，各个模块可以并行地为各个最小属性子集生成属性值，从而提高整 条元组的生成效率。

在实际数据生成过程中，可以采用主从结构或流水线结构等并行方法。值得 注意的是，由于每个最小独立属性子集彼此都没有共同属性，定理8-3 中它们的 连接I 实际是各个最小独立属性子集实例I;的广义笛卡儿积。若采用主从结构的 并行方法，需要额外增加一个汇集模块，保存各个最小独立属性子集的实例， 并对各个实例进行连接，当生成的数据量较大时需要花费大量的时间并占用 大量的存储空间。而这里的目的是获得一定数量的满足指定函数依赖约束的 实例，而不是所有可能的实例。因此，采用流水线结构的并行方法，每个最小 独立属性子集对应的生成模块每生成一个元组，便将其与上一个模块传来的 元组进行连接，并将连接后的元组交给下一个生成模块，最后一个生成模块完 成后，得到的就是完整的元组，可以将其直接写入数据库。这样，各个最小独 立属性子集实例之间的连接在生成单条元组时进行，无需对所有的元组进行 连接操作。最终便可以得到整个属性集的一个记录数为 N 的实例，该实例是 各个I;连接的一个子集，根据函数依赖的特性可知，该实例必然满足指定函数 依赖集的约束(Rosen,2011)。

虽然在流水线方法中，同一条元组中的不同独立属性子集无法同时生成属性 值，但对于不同元组而言，各个最小独立属性子集可以同时生成属性值，当生成大 量元组时，由此增加的时间可以忽略不计。图8-10对比了对函数依赖集合 = {F₁ :(A,B)→C,F₂ :C→D,F₃ :H→J| 和 V={A,B,C,D,H,J} 使用流水线技术和不 使用流水线技术生成数据的时间消耗。

如图8-10所示，不使用流水线技术生成{A,B,C,D,H,J} 的一个元组需要的 时间为t=(t′+t”),生成n 个元组的时间为n·t=n·(t′+t"), 而通过属性集划 分，使用流水线技术需要的时间减少为n·t'+t"” 。 在实际应用中，生成元组的数量 远大于独立属性子集的数目，因此，独立属性子集的数目越多，每个独立属性子集 中属性的数目越少，使用流水线技术所能节省的时间就越多，数据生成的效率就越 高。在理想情况下(每个独立属性子集只包含两个属性),使用流水线技术大约可

以提高 倍的效率。

第8章 不完整数据的估计与填充(209)

{A,B,C,D,E,F} {A,B,C,D} *{H,J]*

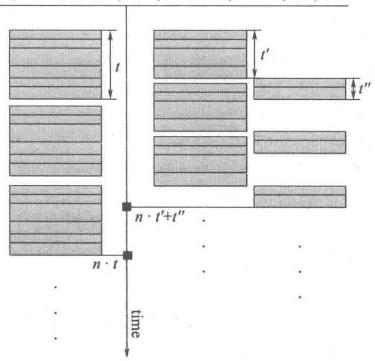


图 8 - 1 0 流水线技术对数据生成时间的影响示意图

**8.5** **本章小结**

填充法是最为常用的不完整数据处理方法，为了提高填充的准确性和效率，本 章分别提出了基于统计关系学习的缺失数据填充算法和基于机器学习的缺失数据 填充算法，通过理论论证和实验分析，证明使用这两种方法对存在缺失的不完整数 据进行填充时具有较高的准确性和效率。将数据生成当作缺失的一种极端情况 ( 即 所 有 属 性 值 都 缺 失 ) , 提 出 一 种 基 于 函 数 依 赖 的 数 据 生 成 方 法 。

**参考文献**

[1]陈伟.2004.数据清理关键技术及其软件平台的研究与应用[D]. 南京：南京航空航天大学.

[2]Rajaraman A,Ullman J.2012. 互联网大规模数据挖掘与分布式处理[M]. 王斌，译.北京：人民邮电出版 社，50-67.

[3]Aittokallio T.2010.Dealing with Missing Values in Large -Scale Studies;Microarray Data Imputation and Beyond [J].Briefings in Bioinformatics,11(2):253-264.

[4]Alexander K.2004.Generate Test Data using SQL[EB/OL].<http://www.IBM.com/developerworks/data/li>- brary/techarticle/dm -0405kuznetsov/.

[5]Anagnostopoulos C,Triantafillou P.2014.Scaling Out Big Data Missing Value Imputations:Pythia vs.Godzilla [C]//Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Min- ing.New York:ACM:651-660.

[6]Andoni A,Indyk P.2014.E2LSH0.1[EB/OL].<http://web.mit.edu/andoni/www/LSH>.

210)数据质量导论

[7]Datar M,Immorlica N,Indyk P,et al.2004.Locality-Sensitive Hashing Scheme Based on P-Stable Distributions

[C]//Proceedings of The 20th Annual Symposium on Computational Geometry.New York:ACM:253 -262.

[8]Fagin R,Kimelfeld B,Reiss F,et al.2014.Cleaning Inconsistencies in Information Extraction via Prioritized Re- pairs[C]//Proceedings of the 33th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems.Snowbird,USA:ACM:164-175.

[9]Fan W,Greets F,Jia X B.2008.Conditional Functional Dependencies for Capturing Data Inconsistencies[J]. ACM Transactions on Database Systems,33(2):1-48.

[10]Fan W,Geerts F,Tang N,et al.2014a.Conflict Resolution with Data Currency and Consistency[J].Joumal of

Data &Information Quality,5(1-2):1-37.

[11]Fan W,Ma S,Tang N,et al.2014b.Interaction between Record Matching and Data Repairing[J].Joumal of Data &Information Quality,4(4):469-480.

[12]Hathaway R J,Bezdek J C.2001.Fuzzy C-means Clustering of Incomplete Data [J].IEEE Transactions on System,Man,and Cybernetics,31(5):735-744.

[13]Lee K C,ParkJS,Kim Y S,et al.2000.Missing Value Estimation Based on Dynamic Attribute Selection[C]// Proceedings of the PAKDD.Berlin,Heidelberg:Springer -Verlag:134-137.

[14]LiX B.2009.A Bayesian Approach for Estimating and Replacing Missing Categorical Data[J].ACM Journal of Data and Information Quality,1(1):1-11.

[15]Rosen K.2011.Discrete Mathematics and Its Applications(7th Edition)[M].New York,USA:MeGraw-Hill Science.

**第9章** **条件函数依赖挖掘及其优化方法**

**9.1** **引言**

对数据约束的表达是设计和管理数据库系统的基本任务。数据结构设 计、数据输入控制、数据修复以及查询等均与数据约束密切相关。数据依赖是 数据约束的一种形式化表达，用于描述属性间取值相互依赖关系。20世纪提 出的针对数据库模式设计的数据依赖有：函数依赖、包含依赖、多值依赖、连接 依赖等。虽然一般的关系数据库又要满足第一、二、三范式，但实际上基于这 些数据依赖，关系数据的模式设计规范，目前已有第六范式(6NF) 的定义。基 于数据依赖的范式理论为数据模式设计提供了理论基础，能够有效地保证数 据完整性， 一致性等。

近年来研究人员开始将数据约束用于数据清洗，希望发现违反数据约束的问 题数据。传统的函数依赖同样可以用于检测不一致数据，但这种数据依赖只能约 束到模式层，对实例层不一致数据的检测能力则稍不足(胡艳丽，等，2009)。针对 传统函数依赖在数据质量控制中的弱点，樊文飞教授在函数依赖的基础上进行了 扩展，提出了条件函数依赖理论(Fan,2008) 。 条件函数依赖通过绑定数据实例层 具体的属性值，定义数据子集上成立的约束，较传统函数而言，条件函数依赖对数 据约束的表达要精细得多，更能适应数据质量控制的需求。从一定意义上说，传统 函数依赖对关系数据的约束是全局而抽象的，而条件函数依赖则是局部而具体的。 因此，条件函数依赖在数据质量控制上更有优势。

本章从传统函数依赖引出了扩展的条件函数依赖，并对函数依赖和条件函数 依赖进行了详细的比较，并讨论了函数依赖控掘的主要思路，介绍了变量条件函数 依赖和常量条件函数依赖挖掘算法，在此基础上提出了一种新的基于开项集剪枝 的常量条件函数依赖挖掘算法。

**9.2** **条件函数依赖挖掘及其常用算法**

本节首行讨论条件函数依赖的概念及其挖掘问题，然后详细介绍函数依赖挖 掘，认识常量 CFD和一般 CFD算法。

212)数据质量导论

t: t: tj:

t₄:

ts:

t6:

t7:

tg:

**9.2.1** **条件函数依赖及其挖掘问题**

函数依赖、条件函数依赖(常量、变量)的概念见<3.4.1.2>节。

前面已经比较了函数依赖、条件函数依赖、关联规则在数据一致性上约束强度 的差异，下面通过具体的数据实例进一步对这些概念进行直观说明(Fan et al., 2011)。下面的关系模式cust, 它指定了一个客户的客户电话(国家代码(CC), 地 区代码(AC), 电话号码(PN), 名称(NM), 和地址(街道(STR), 城市(CT), 邮政编 码(ZIP))) 。 一 个 cust 的实例r₀ 显示在图9- 1中。

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| CC | AC | PN | NM | STR | CT | ZIP |
| 01 | 908 | 1111111 | Mike | Tree Ave. | MH | 07974 |
| 01 | 908 | 1111111 | Rick | Tree Ave | MH | 07974 |
| 01 | 212 | 2222222 | Joe | 5th Ave. | NYC | 01202 |
| 01 | 908 | 2222222 | Jim | Elm Str | MH | 07974 |
| 44 | 131 | 3333333 | Ben | High Str. | EDI | EH41DT |
| 44 | 131 | 4444444 | lan | High Str. | EDI | EH41DT |
| 44 | 908 | 4444444 | lan | Port PI | MH | WIB 1JH |
| 01 | 131 | 2222222 | Sean | 3rd Str | UN | 01202 |

图9-1 一个cust关系实例

在r₀上成立的传统依赖函数(FD) 包括：

fi:[CC,AC]→CT

f₂:[CC,AC,PN]→STR

这里，函数依赖f; 要求具有相同国家和地区代码的两个客户，也要具有相同的 城市；函数依赖f₂ 与之类似。

与此相反，在r₀ 上成立的条件依赖函数(CFD) 不止包括了f; 和f₂, 还包括了以 下(以及更多)关系：

40:([CC,ZIP]→STR,(44, | ))

φ₁:([CC,AC]→CT,(01,908 ||MH))

92:([CC,AC]→CT,(44,131Ⅱ EDI))

φ₃:([CC,AC]→CT,(01,212Ⅱ NYC))

在条件依赖函数q 中，(44,\_ Ⅱ\_)是取值模式，它将一个元组内的属性(CC, ZIP,STR) 及其语义相关的常量进行绑定。它指出在英国的客户，ZIP 确定了唯一 的STR。这是一个仅仅在模式值“CC=44” 的子集上成立的函数依赖，而不是在

第 9 章 条件函数依赖挖掘及其优化方法(

整个关系r₀ 上都成立。条件依赖函数φ₁保证了任何来自美国(国家代码为01)且 地区代码为908的客户，该客户所在的城市必须为MH, 这是由它的取值模式(01, 908 ||MH)所强制的。条件依赖函数φ₂和φ₃与之类似。上述这些情形都无法用 函数依赖来表示。

更具体来说， 一个 CFD是这样一种形式(X→A,tp), 其 中X→A 是一个FD, t,是一个X 和A的属性构成的取值模式，这个取值模式包含了常量以及未命名的 变量“\_ ”,这些变量可以与任何值相匹配。要发现一个 CFD,不仅需要发现传统的 函数依赖X→A, 而且发现它的取值模式t,也是非常有必要的。拥有相同的函数 依赖X→A, 很可能会有由不同取值模式定义的多个不同CFD,例如φ₁~φ₃ 。因 此，在r₀ 上成立的 CFD正则覆盖集通常远远大于它对应的 FD。事实上，假设已经 给出一个固定的函数依赖X→A,仅发现与 FD相关的合理模式就已经是一个 NP 完全问题。

注意，每个φ₁~φ₃ 中的取值模式在它的左边和右边都只包含了常量。如定义 9-3所述，这种 CFD 被称为常量条件函数依赖(CCFD) 。CCFD是实例级的 FD (Lim et al.,1996),它在实体识别(一个对数据质量和数据集成都非常关键的问 题)中非常有用。

此外，从定义以及举例也可以看出，FD 和 CCFD 实际上是 CFD 的2个特例 (FD 的模式全部为变量，CFD 的模式全部为常量)。

实际上，当前在数据质量研究中 CCFD 受关注度最高，主要是因为(Fei, 2008):

(1)CCFD 是最底层模式的表达，不仅可以检查数据的一致性，同时可以检查 一般条件函数依赖的一致性。

(2)利用CCFD 可以产生一般条件函数依赖CFD(Fan et al.,2008)。 (3)数据清洗中，使用CCFD 检查数据一致性简单可行，应用广泛。

[**9.2.1.1**](9.2.1.1) **FD** **与** **CFD** **的比较**

程录庆(2011)对传统条件函数依赖FD 和条件函数依赖 CFD从多个方面进 行了详细比较：

1)表达能力的比较

与传统函数依赖FD相比，条件函数依赖 CFD 对数据约束的表达能力更强。 FD 只能表达在整个数据库范围内成立的数据约束，在一个数据库实例中，只要存 在一条记录不满足约束，则 FD不能成立，而 CFD 可以表达某个局部数据集合内成 立的数据约束，比 FD 的表达更精细。在 CFD 中，将形式如“李新给函授生只上英 语课”这样的一类右部属性取确定值的约束纳入到条件函数依赖的表达框架内， 增强约束表达能力。但是，CFD 表达数据约束能力增强的同时，也带来一个问题，

214)数据质量导论

即CFD规则集本身的一致性需要预先检查，删除相互矛盾的 CFD,而 PD规则集在 任何情况下都是一致的(Fan et al.,2008),CFD 本身的不一致表示与数据约束之 间存在相互矛盾，而 FD在任何情况下都是一致的。

2)形式上的比较

CFD扩大了函数依赖的表达范围，但表达的形式比 FD复杂，给实际应用 增加了复杂度。FD 的表达形式简洁、 一致，采用统一的简单符号“→ ”表示两 组属性集合间的函数依赖关系，而 CFD 为表示限制数据集合的条件，采用了 型表(Pattern Tables),并且增加了“匹配”(≈)的判断，提高了程序设计的复 杂度和计算的工作量。另外，依据 Armstrong 公理，可在线性时间内推导 FD 集合的蕴含，而CFD的蕴含推导的时间复杂度是NP 完全的，即使在具体案例 下，其时间复杂度也是多项式时间 O(n²) 的，蕴含推导的困难使得 CFD 的实 际应用复杂。

3)对数据约束规律的揭示

传统 FD的定义揭示了一种普遍存在的数据约束性质，即属性集合X 的取值 函数确定了属性集合Y 的取值(X→Y), 随后定义的多值依赖、连接依赖、广义依赖 等也都揭示了各自所指的数据约束规律。这些数据约束对应着现实信息世界的固 有的语义规则，具有普遍性，数据依赖理论也由此而产生。比较之下，CFD仅是对 传统 FD 在表达形式上做了改进，增加了局部数据集合下的函数依赖表达方式，但 本质上，CFD 就是条件表达式与传统 FD的结合，并没有揭示函数依赖以外的新的 数据约束特性，所以，CFD的提出并不能得到新的数据设计范式。CFD采用型表 表达数据约束是一种新的数据约束表达方法，这种表达是对传统 FD的一个形式 上扩展，但其对于数据依赖的理论价值不大。

4)应用范围的比较

在数据库的局部数据集合下成立的函数依赖依然会导致数据冗余，违反 此类数据依赖的实例存在数据不一致，这种数据不一致在数据集成的环境下 是广泛存在的。由于传统函数依赖没能表达此类数据约束，所以在传统的函 数依赖理论框架内，不能通过模式分解的办法解决这一难题，而条件函数依赖 由于不能提供新的规范化模式，也无法在数据的整体结构层面解决这一难题。 但是，CFD表达了这一数据依赖，使得在数据修复中解决此类数据不一致问题 成为可能。

[**9.2.1.2**](9.2.1.2) **CFD挖掘问题**

第3章已经介绍了条件函数依赖挖掘的相关定义，并指出挖掘条件函数依赖， 实际上是挖掘数据实例上的 CFD正则覆盖集。很显然，返回在r 成立的所有 CFD 的集合不是一个好办法。因为这个集合包括了不重要的和冗余的 CFD, 而且过于

条件函数依赖挖掘及其优化方法 (215

庞大。因此，希望找到一个正则覆盖，即一个只包括最小CFD 的非冗余集合，从这 个正则覆盖出发，r 上所有的CFD可以通过关联分析推导得出。此外，现实生活中 的数据往往是脏数据，包含很多误差和噪声。为了将那些只与误差和噪声有关的 CFD 排除在外，我们考虑频繁CFD, 它的取值模式有r 上的支持度必须高于阈值 (Fan et al.,2011)。

CFD在数据清洗领域中潜在的应用显著增强了对 CFD 发现进行深入研究的 必要性。但是研究 CFD 挖掘必须要考虑以下几个问题：

(1)如前所述，常量 CFD对实体识别尤其重要，因此值得作为一个单独的问 题进行研究。人们希望拥有有效的方法来单独发现常量CFD,而不需要为发现所 有 CFD付出额外代价。事实上，常量 CFD的发现往往比普通 CFD的发现要快好 几个数量级。

(2)对于高维样本关系来说，levelwise 算法由于其固有的指数复杂性，它的性 能可能不是很好，所以必须有更有效的办法来解决高维数据集面临的问题。

(3)许多关联规则挖掘的技术已经被开发出来，而且利用这些技术来进行 CFD的挖掘是非常自然的。这些技术不仅可以被用于常量 CFD 发现，而且还大大 加快了普通CFD的发现。

从以上考虑出发，Fan 等人(2011)提供了三种 CFD 挖掘算法，分别是用于常 量 CFD挖掘的 CFDMiner 算法，以及用于一般 CFD 挖掘的 CTANE和 FastCFD算 法。目前这几种算法已经成为CFD 挖掘的经典算法，很多研究都是基于这几种经 典方法展开的。另外，由于函数依赖挖掘为条件函数依赖挖掘研究提供了基础，因 此也有必要对其进行简单介绍。

**9.2.2** **函数依赖挖掘**

条件函数依赖挖掘的思想基本源于传统函数依赖、关联规则的挖掘，例如挖掘 常量条件函数依赖的算法 CFD Miner 算法主要基于两类特殊的频繁项集，挖掘变 量条件函数依赖的算法 CTANE 则是由传统函数依赖挖掘算法 TANE 扩展而来。 因此，这里首先介绍函数依赖挖掘的基本思想和一般步骤。

给定一个R 上的实例r,挖掘函数依赖的核心任务是找出r中的所有的函数依 赖。这个过程包括两个子问题(侯林娇，2011):

**子问题1:**对给定的属性集X与 Y,建立一个判断 FD X→Y 是否成立的模型；

**子问题2:**使用已经建立的模型找出r 中所有的函数依赖。因为R 上的任意 属性的组合X 都可能成为FD X→Y的前驱，使得X→Y在r 上成立。

Attr(R) 的所有子集，即X 的所有可能值形成了函数依赖挖掘的搜索空间，这 个搜索空间的规模随着属性数目增加而呈指数趋势增长，因此，子问题2的关键是

216)数据质量导论

利用修剪规则修剪搜索空间。例如，图9-2展示了4个属性A、B、C、D的搜索空 间，它展示了4个属性的所有可能的非空组合，这4个属性的组合可产生2⁴=16 种可能的子集，其中有2⁴-2=14个属性子集成为函数依赖左侧的候选集，因为全 集{ABCD}和空集不可能出现在任何函数依赖的左侧。对于n 个属性，其网格图 有n2°-'条边。因为所有的搜索空间都是从第一层开始的，而不是从空集开始的， 所以网格图中共有 n2\*-1-n 条边。



*ABCD*

图9-2 Attr(R)={A,B,C,D }的层次网格图

函数依赖集可分为三部分(Wen et al., 2009):

(1)由子问题1建立的模型找出的函数依赖集；

(2)利用Armstrong公理从已发现的函数依赖中推导的函数依赖集；

(3)未被发现的函数依赖集。

因为存在着大量的第2种函数依赖，这种函数依赖是冗余的，而通过子问题1 检测元组的方法发现第3种函数依赖集非常费时，因此挖掘函数依赖的主要任务 是找到更多有效的推理规则修剪冗余的函数依赖集。

现有的函数依赖挖掘算法可以分为三类(侯林娇，2011;张方舟，等，2015):

(1)先产生候选集，再检测候选集是否成立(The Candidate -Generate -and - Test);

(2)最小覆盖法(The Minimal-Cover);

(3)形式概念分析法(The Formal Concept Analysis)。

第一类方法，首先采用算法逐层搜索属性空间，在每一层产生函数依赖的候选 集，再检测候选集是否为函数依赖。这种挖掘方法可以通过修剪规则(Pruning Rule) 减少搜索空间。经典的函数依赖挖掘算法 TANE 和 FUN 都是采用第1种挖 掘方法，算法TANE(Huhtala et al.,1999)和算法 FUN(Novelli et al.,2001)挖掘函 数依赖的方法为：从小规模属性开始检测是否满足函数依赖，然后逐层地增加属性 的规模(如图9-2所示，从第1层开始逐层检测),并在每层中尽可能修剪搜索空

第9章 条件函数依赖挖掘及其优化方法 (217

间，这两种方法都是首先根据元组的属性值对元组进行划分，这样，即使在元组数 非常大的情况下，TANE 和 FUN 也可以快速地找出函数依赖集。这两种算法都是 通过计算第k 层的候选闭包找到第k 层的函数依赖集，根据第k 层的函数依赖集 产生第k+1 层的候选集，不同的是两种算法使用了不同的修剪规则来修剪搜索 空间。

第二类方法称为最小覆盖法，该类方法是发现函数依赖的最小覆盖集的挖掘 方法。FDEP(Hammon et al.,2006)由3个算法组成：自底向上的算法、双向算法和 自顶而下的算法。相关研究表明，自底向上的算法最有效。自底向上的方法步骤 为：第一步，得到所有使数据集不满足函数依赖条件的覆盖，在这些函数依赖不成 立的覆盖中，可能包含有其他一般的依赖关系；第二步，重复第一步操作，对于候选 集来说，每重复一次，满足函数依赖条件的候选集比例增加；第三步，得出函数依赖 成立的覆盖集。算法 FastFDs(Catharine et al.,2001)和算法 Dep-Miner(Anupama et al.,2013)是通过判断每两个元组是否满足函数依赖条件的方法来发现函数依 赖集，如 Agree-Sets, 首先，从关系表中去掉划分中的单个元组组成的等价类后， 得到新的关系表；然后，对划分中的每个等价类计算函数依赖集，最后得到最大的 函数依赖集。因此，可以根据函数依赖集产生得到最小的函数依赖覆盖集。算法 FastFDs 和算法 Dep-Miner 的不同在于两者在检测函数依赖时，用的搜索方法不 同：FastFDs的搜索方法是深度优先搜索，而 Dep-Miner 的搜索方法为逐层搜索。 Dep-Miner 在计算 Agree-Sets 时消耗的时间很多，随着元组数的增加，时间复杂 度呈指数平方的速度增加；而在算法TANE( 或FUN)中，检测元组是否满足函数依 赖涉及到大量的操作。

第三类是形式概念分析方法，该方法采用形式概念分析的观点，它最早 是由德国学者 Wille 提出的。它的核心思想是，考虑关系数据库理论与形式 概念分析的关系，使用预定义的形式概念分析闭包操作提取数据库中的函 数依赖。

**9.2.3** **CTANE** **算法**

接下来将介绍 CTANE算法， 一个发现最小的、k- 频繁(变量和常量)CFD的 层方法(Levelwise)算法。它是FD挖掘的经典算法的TANE算法(Huhtala et al., 1999)的扩展。

[**9.2.3.1**](9.2.3.1) **基本原理**

CTANE 算法的基本原理如表9-1(Fan et al.,2011;周健昌，等，2012)。

**218)**数据质量导论

表9 - 1 CTANE 算法

|  |
| --- |
| **CTANE算法** |
| **输** **入** **：**关系R上的数据实例1,大于1的自然数k  **输** **出** **：**实例I上k-频繁的条件函数依赖正则覆盖集 |
| 1.L;=|(A,-)}IAeattr(R)|U{A,a}lsupp((A,a),I)≥k,Aeattr(R)|;  2.CandRHS(φ):=L;e=1;  3.While L≠φde  4.For all(X,t,)eL do  5. CanndRHS(X,tp):=∩gexCandRHS(X\}B},[X\{B}]);  6. End For  7. **For** all(X,to)eL,A∈X,(A,C₄)ECandRHS(X,t。)dc  8. φ=([X\{A|]→A,(tp[X\{A|]ⅡC₄));  9. If l=φ then  10. 输出φ:  11. 根据下式更新CandRHS(X,tp)中的u,  u[A]=C₄和up[X\{A}]≤tp[X\|A}]  12. End If  13. **End** **For**  14. 从L中移除所有使CandRHS(X,tp)=φ的(X,t。);  15. 从L中成成L{-1,同时令l:=1+1;  **16.End** **While** |

[**9.2.3.2**](9.2.3.2) **算法改进**

研究者针对CTANE算法进行分析，并针对其缺陷提出了相应的措施，以降低 其在实际场景中应用的难度(周健昌，等，2012)。

CTANE算法的主要缺陷表现在以下几点：

(1)需要对数据库进行多次扫描；

(2)判断(X,s) 、(Y,t,) 是否有相同前缀，需要消耗较多时间；

(3)上层节点查找率低；

(4)字段编码操作效率低；

(5)多余节点占用大量空间。

针对以上问题采取了相应的措施。

(1)对每个节点(X,sp) 保留相应等价类信息。为了避免对数据库进行多次扫 描，可为每个节点(X,s,) 保留相应等价类信息，通过这个集合快速判断 CFD是否 成立，通过上层的相关节点来生成下层的节点，避免数据库的扫描。

(2)生成节点(X,s,) 时提供并保留一个前缀码。在判定(X,sp) 、(Y,t) 是否

第9章 条件函数依赖挖掘及其优化方法(219

有相同前缀期间，需要消耗较多的时间，可以在生成(X,s,) 的同时为这个节点生 成一个前缀码，并使在同一层的节点中相同前缀对应相同前缀码，而不同的前缀对 应不同的前缀码。

(3)提高上层节点的查找效率。在CTANE算法中，查找LATTICE 格数上层节 点是每一步都必须做的，对于每一个n 层节点，都要查找n 个 n-1 层节点。如此看 来，查找效率显得尤为重要。为此，可以采取如下措施来提高查找效率：建立一个哈 希表用来存放已经查找过的节点，表结构必须映射到n-1 层各节点；需查找某个 n-1 层的节点时先到哈希表中查，若找到则直接返回此节点，否则顺序查找n-1 层 的所有节点，直到找到或者为空再把结果插入哈希表，并返回这个节点或空标记。

(4)由于数据库中大部分信息是以字符串的形式存储的，对大量的字符串信 息进行操作会影响到数据挖掘操作的效率，同时也会占用大量的空间。如果在同 一值域，可将相同字符串映射成相同的数值，不同字符串则对应不同的数值。如果 在不同值域，映射关系则彼此独立。这样可以在数据挖掘的过程中直接对数值进 行操作，不仅可以提高效率也可以节省存储空间。

(5)删除多余节点信息。在前n-1 层节点中，只用到第一层以及n-1 层的 节点，由于LATTICE格树中每个节点(X,s) 都包含大量信息，然而其中的划分信 息是没用的，所以可以删除第二层到第n-1 层所有节点的全部信息，删除第一层 和第n-1 层的划分信息，来降低空间复杂度。

**9.2.4** **CFDMiner 算法**

本小节主要介绍 CFDMiner,一种常量条件函数依赖(CCFD)挖掘的算法。给定一 个R上的实例r 和一个支持度阈值k,CFDMiner 发现一个形式为(X→A,(t,lla)) 的 k- 频繁最小CCFD的正则覆盖集(Fan et al.,2011)。

该算法是基于左极简CFD与开项集和闭项集之间的连接。首先我们来介绍 开项集和闭项集的相关概念。

定义9-1(项集):形式如(X,t;[X]) 的称为项集(Itemset), 简写作(X,t)。

其中属性子集XCAtr(R),t, 为X 上的某个具体的取值模式，即对于A∈XUY,t; [A] 为Dom(A)中某个特定的常量模式，或者记作t[X]CDom(X) 。 实例I 上匹 配(X,t;[X]) 的所有元组构成的集合，记为 supp((X,t),I), 即为该项集的支持。 集合内元组的数目为项集的支持度，记作supp((X,t),I)l。

**定义9-2(开项集、闭项集):**对于某个项集(X,t;[X]), 若存在项集(Y,

s;[Y]),YCX,s;[Y]=t;[Y], 其 中t;[X] 和s,[Y] 分别是X 和 Y上具体的取值模

式，则称(X,t,[X]) 比(Y,s[Y]) 更加具体(Specific), 或者(Y,s[Y]) 比 (X,

220)数据质量导论

t;[X]) 更加一般化(General), 记作(X,t;[X])<(Y,s;[Y]) 或者(Y,s;[Y])>(X,

t;[X]) 。 显然，supp((Y,s;[Y]),I)2supp((X,t;[x]),I)。

(1)若6(Y,s₇)>(X,tr), 使得supp((Y,s),I)=supp((X,t),I), 则(X,t)

为开项集(Free Itemset);开项集也称为生成器(Generator)。

(2)若6(Z,uī)<(X,t), 使得supp((Z,uī),I)=supp((X,t),I), 则(X,t) 为闭项集(Closed Itemset)。

(3)若存在一个开项集(Y,s) 和闭项集(Z,ui),(Y,sr)>(X,t)>(Z,u;)

且supp((Y,s),I)=supp((X,t;),I)=supp((Z,uī),I), 则(Y,s₁) 和(Z,u;) 分别

称为(X,t) 的开项集和闭项集。(X,t) 的闭项集可以记作col(X,tr)=(Z,ui)。

表9-2是应用最为广泛的挖掘CCFD 的 CFDMiner算法的具体步骤(Fan, 2009)。

表9 - 2 常量条件函数依赖挖掘算法 CFDMiner

|  |
| --- |
| CFDMiner算法 |
| 输入：关系R上的数据实例1,大于1的自然数k;  输出：实例1上k-频繁的常量条件函数依赖正则覆盖集。 |
| 1.利用GeGrowth算法挖掘出实例I上所有k-频繁的开项集(X,t)、闭项集(Y,sg)及其配对关系 2.令L为所有的开项集(X,)的集合；  3.For all k-频繁的闭项集(Y,sg),do  4.For all(X,tr)eL do  5. 令(X,t)为CCFD的LHS,利用其对应的(Y,)求得CCFD初始化的RHS,即：RHS(X,t)=  (Y\X,s[YLX])  6. Endfor  7.For all(X,t)eL do  8. For all(X',t[X'])eL,且X'CX do  9. RHS(X,t)=RHS(X,t₇)\RHS(X',r;[X'])  10. End for  11. If RHS(X,t)≠Q  12. 输出(X→A,t[X]la),其中(A,a)eRHS(X,g);  13. End if  14. End for  15.End for |

GcGrowth 算法(Li et al.,2005)输出的开项集、闭项集以及配对关系作为CFD- Miner算法的输入(见表9-2第1行)。

根据表9-2第4~6行，首先按照配对关系，生成初始化的CCFD, 假设算法的

第9章条件函数依赖挖掘及其优化方法 

输入中有一初始的CCFD为(a,b)→(d,e), 初始的 CCFD为非平凡但可能冗余的

CCFD。

根据表9-2第8～10行，将初始 CCFD中的冗余元素删去，则需要找出其 LHS 所有的非空真子集(如a→d,b→d), 并将其RHS 中的元素减去，获得最终的 CCFD( 如(a,b)→e) 即为 CFDMiner 的输出。

9.3 基于开项集剪枝的常量条件函数依赖挖掘算法

CFDMiner 能够从开项集和相应的闭项集中高效地计算数据中的CCFD。挖掘 开项集和闭项集的算法很多(Li et al.,2005;Agrawal,1994;Goethals et al.,2003; Pasquier et al.,2000;Wang et al.,2003;Zaki,2004;Calders et al.,2007;Li et al., 2006;Li et al.,2007)般采用能够同时挖掘开项集和闭项集的算法如 GcGrowth。

过去对CCFD 挖掘算法和CFDMiner 优化的研究并不多， 一方面是由于 CFDMiner 算法本身已经有很高的效率，通过其他方式(如通过关联规则挖掘(Diallo et al., 2012))挖掘CCFD的算法很难超越 CFDMiner;另一方面，对于CFDMiner 算法优化 通常都集中在对候选开项集的研究(Li et al.,2013)。但是 GcGrowth 挖掘全部开 项集和闭项集已经非常高效，因此通过只产生有效的候选开项集的方式提高 CCFD 挖掘效率的效果非常有限。

CFDMiner 的输入为数据的全部开项集和闭项集，本节首先证明了搜索全部开 项集对于产生有效的 CCFD存在大量的无效和冗余搜索，通过合理剪枝可以避免 进行这些搜索，并且产生与原算法一致的CCFD 正则覆盖集。实验证明：提出的剪 枝策略能缩小搜索空间，提高算法效率，使算法性能得到显著改善。

9.3.1 剪枝与优化策略

对于CFDMiner,本节给出两个剪枝策略。

<9.3.1.1> 剪枝策略1

定义9-3:若3开项集(X,tp),其相应的闭项集 clo(X,t)=( X,t ),则称(X, t) 为平凡项集。

引理9-1:平凡项集不产生有意义的CCFD。

证明：根据 CFDMiner,对于平凡项集产生的φ,有LHS(φ)=(X,t),RHS(φ)= (Y\X,s;[Y\X])=(X\X,s;[X\X])=Q 。 因此φ为无意义的 CCFD。

显然，由于平凡项集产生的φ的RH S(φ)=② ,因此对于LHS(φ)的任意超集 supper(LHS(φ)),RHS(φ) 不可能包含RHS(supper(LHS(φ))) 中的任意(冗余)

222)数据质量导论

元素。因此，计算 CCFD不需要考虑平凡项集。

策略1:删除 CFDMiner 输入中所有的平凡项集。

<9.3.1.2> 剪枝策略2

在下文的叙述中，(X,;[X]) 将简记作(X,t,[X])。

引理9-2:对于一个开项集(X,t,[X])与其对应的闭项集clo(X,t,[X])= (Y,s,[Y]), 若存在该开项集的超集X'2X,(X',t’[X]), 则对于其对应的闭项集 clo(X',t;[X'])=(Y,s;[Y']), 有s,[Y\x']2s,[Y\x]。

证明：假设s;[Y'\X']Üs,[Y\X],分2种情况讨论：①3A∈Y\X且A₄Y'\X', s,[A]és;[Y'];②3A∈Y\X且AeY\X',s,[A]≠és;[A]。

对于假设1,因为X'2X, 且(Y',s'[Y']) 是(X',t,[X']) 的闭项集，所以 supp (Y',s;[Y'])=supp(X',t;[X'])supp(X,t,[X])=supp(Y,s,[Y]), 即supp(Y', s;[Y'])Csupp(Y,s,[Y]), 即对于Vtesupp(X',t;[X']), 都有tesupp(Y,sp [Y]) 。因为3A∈Y\X且A≠Y\X', 即s,[A]és’[Y'], 对于(Y',s’[Y]) 的超集模 式(Y'UA,(s'[Y'],s,[A])),Vtesupp(Y',s’[Y']), 都有tesupp(Y'UA,

(s;[Y'],s,[A])), 所以 supp(Y',s;[Y'])Csupp(Y'UA,(s;[Y'],s,[A])) 。 因为 supp(Y',s’[Y])2supp(Y'UA,(s’[Y'],s,[A])), 所以 supp(Y',s'[Y'])=supp (Y'UA,(s',[Y'],s,[A]))。所以至少存在一个超集模式(Y'UA,(s',[Y'], s₉ [A])),其支持度与(X',t;[X'])相等，这与“(Y,s;[Y'])是(X',t;[X'])的闭项 集”矛盾，所以假设1不成立；

对于假设2,因为X'2X, 且(Y',s,[Y'])是(X',t;[X'])的闭项集，所以 supp (Y',s;[Y'])=supp(X',t’[X'])Csupp(X,t,[X])=supp(Y,s,[Y]), 对于Vt∈ supp(Y”,s;[Y']),都有tesupp(Y,s,[Y]) 。 但是s,[A]≠s’[A], 所以有t,[A]≠ t;[A], 不等号左右两边是同一条记录(不同常量模式，A为不同模式的属性交 集),所以假设2不成立。

综上，引理9-2得证。

定义9-4:对于某开项集(X,t,[X]),lXI=n,若其子集(Y,t,[Y]),lYl = n-p(0<p<n), 则称(Y,t,[Y]) 为(X,t,[X]) 的 -p 阶子集。

求某个开项集的-p 阶全部子集记作sub-”(X,t,[X])。若 -p>-q(p<q ), 则称为sub-”(X,t,[X]) 是比sub~°(X,t,[X])高阶的子集。对(X,t,[X]) 的 -p 阶 子集按指定顺序排序，第j个子集记作sub-”(X,t,[X]);,sub~P(X,t,[X])中一共 有C” 个子集，即je{1,2,…,Ca-}。

如开项集(a,b,c,d) 的全部-2阶子集为sub-²(a,b,c,d)={(a,b),(a,

c),(a,d),(b,c),(b,d),(c,d) },-2 阶子集合的第2个和第3个子集分别

为sub-²(a,b,c,d)₂=(a,c) 和sub-²(a,b,c,d)₃=(a,d)。

第9章 条件函数依赖挖掘及其优化方法(223)

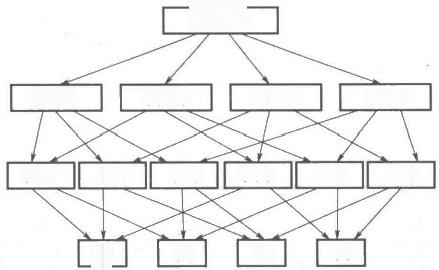
定义9-5:某开项集(X,t,[X]),IXI=n,其所有不包括空集的真子集按从高 阶到低阶的顺序形成的集合树，称为子集全局树，简称子集树。子集树相邻层按隶 属关系连线。

图9-3为开项集(a,b,c,d) 的子集树。

显然，对于(X,t,[X]) 树中的sub~\*层中每个子集的-1阶子集的并集为







(a.b.cd)

(a,b,c) *(a,b,d) (a,c,d)* (b,c;d)

(a,b) (a,c) (b,c) *(a,d) (a,d)* (c,d)

(a) (b) (c) (d)

图9- 3 开项集(a,b,c,d)的子集树

引理9-3:开项集的-1阶子集模式必然是开项集。

证明：假设(X,t,[X])是开项集，存在一个(X,t,[X])esub~¹(X,t,[X

])不

是开项集。

匹配(X,t,[X]) 的元组(即项集的支持)记为T₁=supp((X,t,[x]),T),T₁C T,因为(X,t,[X]) 是开项集，所以有：①T\T₁的记录中不包含(X,t,[X]);②T\T₁ 的记录中有包含(X₁,t,[X₁]) 的记录(否则(X,t,[X]) 不是开项集)。

不妨设T\T₁中匹配(X₁,t,[X₁]) 且不匹配((X\X₁),t,[X\X₁]) 的记录为T₂, 所以T中匹配(X₁,t,[X₁]) 的记录supp((X₁,t,[X₁]),T)=T₁UT₂, 且T\(T₁U T₂)的记录不匹配(X₁,t,[X₁])。

因为(X₁,t,[X₁]) 不是开项集，至少存在一个该项集的子集(X₂,t,[X₂]) 满足 supp((X₂,t,[X₂ ]),T)=supp((X₁,t,[X₁ ]),T),所以T\(T₁UT₂) 不匹配(Xz, t,[X₂]),T 中匹配(X₂,t,[X₂]) 的记录也为T₁UT₂。分2种情况讨论：

(1)假设T\(T₁UT₂) 中有匹配((X\X₁),t,[X\X₁]) 的记录集T₃,则T中匹配 (X\X₁,t,[X\X₁ ])的记录为T₁UT, 包含(X₂,t,[X₂ ])的记录为T₁UT₂, 匹配((X\ X₁,X₂),t,[X\X₁,X₂]) 的记录supp(((X\X₁,X₂),t,[X\x,X₂]),T)=supp((X,

t,[X]),T),而((X\X₁,X₂),t,[X\X₁,X₂ ])c(X,t,[X]),所以(X,t,[X])不是开 项集，与题设矛盾。

224)数据质量导论

(2)假设T\(T₁UT₂) 中没有匹配((X\X₁),t,[X\X₁]) 的记录集，则T中匹配 (X\X₁,t,[X\X₁ ])的记录为T, 包含(X₂,t,[X₂ ])的记录为 T₁UT₂, 匹配((X\X₁, X₂),t,[X\X₁,X₂]) 的记录supp(((X\X₁,X₂),t,[X\X₁,X₂]),T)=T₁, 即

supp(((X\X₁,X₂),t,[X\X₁,X₂]),T)=supp((X,t,[X]),T),((X\X₁,X₂), t,[X\X₁,X₂])C(X,t,[X]), 所以(X,t,[X]) 不是开项集，与题设矛盾。

因此，假设均不成立，即若(X,t,[X]) 是开项集，则任意的(X₁,t,[X₁ ]) ∈ sub~¹(X,t,[X]) 都是开项集。

推论9- 1:开项集的所有子集都是开项集。

证明：开项集(X,,[X]) 的 -(k+1) 阶子集与-k阶子集的关系为sub-“\*1) ,sub-¹(sub-\*(X,t,[X]),), 即s  (sub~¹(X,t,[X]),),sub~³(X,t,[X])=u;,sub-'(sub-²(X,t,[X]),),…, 即开 项集的-(k+1) 阶子集可以通过对-k 阶子集求得。根据引理9-3,开项集的- 1 阶子集模式必然是开项集。因此，开项集的所有子集必然是开项集。

推论9-2:对于开项集(X,t,[X]),其-1阶子开项集为sub~¹(X,t,[X]),-1 阶子集所对应的闭项集记作 clo(sub¹(X,t,[X]))=U;21'clo(sub~¹

(X,t,[X]),), 则有



证明：根据引理9-2,对于一个开项集(X,t,[X]) 与其对应的闭项集clo(X, t,[X]),若存在该开项集的超集(X',t,[X']),X′≥X,则其对应的闭项集clo(X', t,[X])2col(X,t,[X]) 。 因为sub-¹(X,t,[X])2sub-\*(X,t,[X])(k=1,2,…,

n-1),clo(sub~¹(X,t,[X]))2clo(sub-\*(X,t,[X])), 所以有 clo(sub~¹(X, t,[X]))2Ux{clo(sub-\*(X,t,[X]))。

策略2:CFDMiner 算法中(表9-2第8行)只需要搜索该开项集的- 1阶子集 即可。

根据引理9-3,开项集的子集必然是开项集。因此，GcGrowth 输出的开项集 能够覆盖全部开项集的子集。

根据推论9-2,开项集的全部-1阶子集对应的闭项集，包含了所有非空真子 集对应的闭项集的元素。因此只需搜索此开项集的-1阶子集，即能获得全部子 集对应的闭项集模式。

<9.3.1.3> 优化算法 prCFDMiner

根据上述的剪枝策略，对CFDMiner进行优化，表9-3是剪枝后的 prCFDMin- er 算法。

第9章 条件函数依赖挖掘及其优化方法(225)

表 9 - 3 基于剪枝策略的优化算法 prCFDMiner

|  |
| --- |
| **prCFDMiner算法** |
| 输入：关系R上的数据实例1,大于1的自然数k;  输出：实例I上k-频繁的常量条件函数依赖正则覆盖集。 |
| 1.利用GcGrowth算法挖掘出实例I上所有k-频繁的开项集(X,t)、闭项集(Y,sg)及其配对关系 根据配对关系建立开闭项集的哈希映射C2F:<(X,t),(Y,sp)〉;令L为所有的开项集(X,t)的 集 合 ；  2.For all k-频繁的闭项集(Y,sg),do  3.For all(X,r;)CL d  4. 令(X,t)为CCFD的LHS,利用其对应的(Y,sg)求得CCFD初始化的RHS,即：RHS(X,t)= (Y\X,s;[Y\x]);  5.IfRHS(X,t₁)≠②  6.删除C2F中相应的<(X,tg),(Y,sg))  7. End if  8.End for  9.For all(X,)CL do  10.For sub-¹(X,tg[X]);CL,je{1,2,…,C|xi)do  11. RHS(X,)=RHS(X,tp)/RHS(sub-¹(X,r[X]););  12. End for  13.IfRHS(X,rr)≠〇  14. 输出(X→A,tg[X]lla),其中(A,a)eRHS(X,ts)  15. End if  16. End for  17.End for |

由表9-3的第5～7行和10～12行可见，prCFDMiner 主要是根据策略1和2 分别对 CFDMiner 进行了优化。

9.3.2 优化前后复杂度对比

由对 CFDMiner的分析可见，算法分为三步：

步骤1:根据配对关系，计算初始化的CCFD, 其复杂度为0(n),n 为输入开项 集数目，见表9-2第4~6行。

步骤2:对于每一个开项集，搜索其在列表中的子集，见表9-2第8行，有两 种搜索策略：

( 1 ) 将n 个开项集以列表形式存储，对于每一个开项集，在整个列表中搜索非空 真子集，则每一个开项集的搜索空间为(n-1), 所有开项集的搜索空间为n(n-1);

( 2 ) 将n 个开项集以哈希映射(HashMap) 的形式存储，对于每一个开项集(假

226)数据质量导论

设项集的平均长度为l), 首先计算其所有的非空真子集，假设计算和搜索一个真 子集均需要1个时间单位，则所有开项集的计算和搜索时间为2n(2¹-2)。

当数据集的属性值固定，支持度变化时，挖掘出的开项集的平均长度通常是固 定的(一般小于属性个数的50%)。但是随着支持度的降低，前者的搜索策略的时 间复杂度为O(n²), 而后者为0(n) 。 因此，通常采用后者的搜索策略。

**步骤3:**对于每一个开项集(假设平均长度为l),假设在整个列表中找到的真 子集数目为(2¹-2),计算最终的CCFD 的时间为(2¹-2),所有n 个开项集的计算 时间为n(2¹-2);

由以上分析可知，通过合理的选择存储方式和搜索策略，CFDMiner 算法对于 输入开项集的复杂度为0(n)。

本章给出的优化和剪枝策略主要是针对CFDMiner算法的步骤1和步骤2。

在步骤1 中剪去了无意义的输入(即不会产生有效 CCFD 的开项集),见 表9-3第5~7行。假设全部开项集的数目为n, 其中平凡项集的数目为n₁(n₁< n), 则在步骤1中节省了n₁ 个时间单位。

在步骤2中，CFDMiner算法需要搜索全部的真子集，在所有开项集的平均长 度为l的情况下，计算和搜索全部真子集需2n(2¹-2) 个时间单位。经过剪枝后的 prCFDMiner 算法，只需搜索每个开项集的-1阶真子集，所以计算和搜索全部-1 阶真子集需2nl个时间单位，两者时间比为(2'¹-2)/l。可见，若开项集的平均长度 为5和10,则理论效率分别约提高6倍和100倍。

**9.3.3** **实验验证与结果分析**

[**9.3.3.1**](9.3.3.1) **实验数据集及配置**

为了验证优化后的算法 prCFDMiner 是否与CFDMiner 得到一致的结果以及是 否在效率上有所提升，选取了3个UCI公共数据集(<http://archive.ics.uci.edu/>

ml/) 进行测试。由于算法的输入是由GcGrowth算法根据预设的支持度输出的开 项集和闭项集的配对，因此表9-4列出了数据集的相关属性，以及不同支持度下 输出的开项集和闭项集的配对数目。

表9 - 4 数据集属性及不同支持度下的开闭项集数目

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **数据集** | **属性**  **个数** | **记录**  **条数** | **开项集和闭项集配对数** | | | | | | | |
| **支持度/** **%=50** | **30** | **10** | 5 | 1 | 0.5 | 0.3 | 0.1 |
| Adult | 12 | 32561 | 18 | 72 | 725 | 2360 | 24260 | 57869 | 101990 | 296250 |
| Mushroom | 23 | 8124 | 51 | 558 | 7631 | 21160 | 103517 | 164526 | 235013 | 360166 |
| Chess | 7 | 28076 | 1 | 5 | 47 | 122 | 1210 | 2997 | 4955 | 18908 |

第9章 条件函数依赖挖掘及其优化方法 (**227**)

对于 Mushroom 和 Chess 数据集，在预处理时只需根据GcGrowth 算法要求对 数据进行数字化即可；对于Adult 数据集，首先删去了3个连续型属性，对剩下的 属性进行数字化。实验代码使用Java 编写，在JVM 上运行。硬件平台为Intel Xe- on处理器(频率为2.5GHz), 内存为64GB。

[**9.3.3.2**](9.3.3.2) **实验结果**

经验证，prCFDMiner与 CFDMiner 得到结果是一致的，即输出的 CCFD 相同， 表9-5列出了不同支持度下输出的 CCFD数目。

表9 - 5 不同支持度下输出的 CCFD 数目

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 数据集 | **输出CCFD的数目** | | | | | | | |
| **支持度(%)=50** | 30 | 10 | 5 | 1 | 0.5 | 0.3 | 0.1 |
| Adult | 0 | 2 | 7 | 27 | 288 | 692 | 1324 | 7700 |
| Mushroom | 27 | 95 | 2774 | 7510 | 26029 | 37040 | 46498 | 60886 |
| Chess | 0 | 0 | 0 | 2 | 7 | 15 | 36 | 327 |

为了比较策略1和2对算法优化的影响，分别对原算法 CFDMiner、经策略1 优化后的算法CFDMinerS、经策略1和2结合优化后的算法 prCFDMiner 的搜索空 间和搜索时间等指标进行测试，实验结果如图9-4所示。

可以看出：

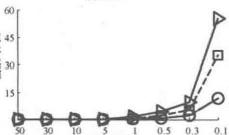
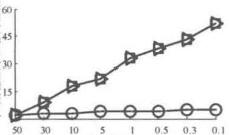
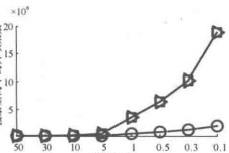
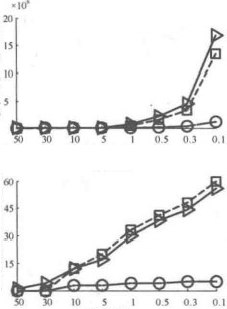
(1)随着支持度的降低，输入的开闭项集对的增多，算法的搜索空间和搜索时 间逐渐增大，表9-5也同时说明了挖掘出的 CCFD 逐渐增多。

(2)同一支持度下，prCFDMiner算法搜索开项集的数目和搜索时间总是要明 显少于 CFDMiner;CFDMinerS 算法在 Adult 和 Chess 数据集上的性能有明显的提 升，而在 Mushroom数据集上性能几乎没有改善，这是因为Mushroom 没有产生“平 凡项集”。

( 3 ) 图 9 - 4(d)~(f) 的平均存储空间实际上反映的是项集的平均长度，即对 于每一个开项集平均所需计算的子集的数目。例如在图9-4(d) 中，0.5%的支持 度下，CFDMiner算法的平均存储空间多于30个开项集，而prCFDMiner算法为4~ 5个。这与理论分析基本一致：CFDMiner 对于每个项集需计算2'-2个项集，而 prCFDMiner算法只需要计算l 个项集。

(4)根据(3)可知，prCFDMiner算法与 CFDMiner 的搜索时间比约为l(2¹- 2),这在图9-4(g)~(i) 中也得到了相应的验证。

需要说明的是，Diallo等(2012)和Li 等(2013)是直接在数据集上挖掘CCFD, 而 CFDMiner则是在已挖掘出的开闭项集上计算 CCFD(prCFDMiner 在此基础上剪 枝),因此 prCFDMiner 与他们工作的可比性不强。CFDMiner 和 prCFDMiner 的优 点在于借助了已有丰富的挖掘开闭项集的算法成果，使得CCFD 的挖掘更加简便。 在 Li 等(2013)的研究中， FACD算法仍然是通过从数据中直接限制产生冗余的候

228)数据质量导论

搜索空间中的开项集数

平均存储空间中的开项集数

搜索时间/s

搜索空间中的开项集数

平均存储空间中的开项集数

搜索空间中的开项集数

平均存储空间中的开项集数

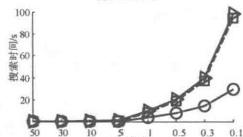
支持度/%

支持度/% (a)Adut

(b)Mushroom

支持度/% (e)Mushroom

支 持 度 / % (d)Adult



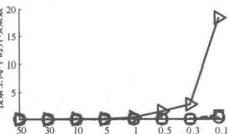
支持度/%

支持度/% (g)Adult

(h)Mushroom

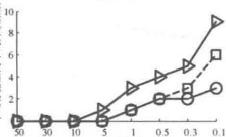
-CFDMine - -CFDMinxS

x10



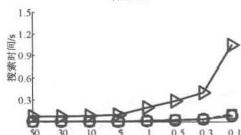
支持度/%

(c)Chess



支持度/%

(D Chess



支接度/%

(0Chess

一prCFDMiner

图9-4 不同优化策略在搜索空间、平均存储空间和搜索时间上的性能对比

选项集的方式计算CCFD, 但 FACD会随着数据集属性的增多，算法性能下降。而 CFDMiner算法是在开闭项集基础上搜索 CCFD,其计算能力受属性个数影响不大 (从上述实验选取的不同数目属性的数据集的实验中可得到验证)。

**9.4** **本章小结**

本章首先介绍了条件依赖函数及其挖掘的基本概念，指出较传统函数依赖而 言，条件依赖函数在对数据约束的表达方面要精细得多，可以广泛用于数据的输入 控制、数据的维护及数据的自动清洗等方面。然后，介绍了函数依赖挖掘算法的研 究现状，并重点介绍了两种条件函数依赖挖掘算法 CTANE 和 CFDMiner,其中后者 专门用于常量条件函数依赖挖掘。最后，提出了一种基于开项集剪枝的常量条件 函数依赖挖掘算法 prCFDMiner,该算法对常量条件函数依赖的常用算法 CFDMiner 进行了剪枝和优化，通过比较搜索策略的复杂度，证明了可以在0(n) 的复杂度下 完成CFDMiner的运算，这也是实施优化策略的基础；证明了剪枝优化策略能够得 到与原算法一致的结果，并通过实验得到了相应的验证；优化后的算法在搜索空 间、平均存储空间和搜索时间上的性能都得到了明显的提升；实验证明了不同策略 在不同的数据集下对算法优化的影响程度，策略2具有普适性即总是会提升算法

第9章 条件函数依赖挖掘及其优化方法 (**229)**

效率，而策略1对数据集的特点有一定要求即数据集需要存在“平凡项集”。

**参考文献**

[1]张方舟，高晓松.2015.基于条件函数依赖的挖掘算法研究[J]. 计算机技术与发展，25(5):56-59.

[2]周健昌，刘波.2012.一种条件函数依赖挖掘算法的分析和实现[J]. 计算机与数字工程，40(9):8-11. [3]胡艳丽，张维明.2009.条件依赖理论及其应用展望[J]. 计算机科学，36(12):1115-1118.

[4]侯林娇.2011.基于函数依赖的成批处理模式挖掘方法研究[D]. 湘潭：湖南科技大学. [5]程录庆.2011.数据约束表达研究[J]. 洛阳：洛阳师范学院学报，30(11):61-64.

[6]Agrawal R.1994.Fast Algorithms for Mining Association Rules[C]//Proceedings of 20th International Confer- ence on Very Large Data Bases.Santiago,Chile:ACM:487-499.

[7]Anupama A C,vijay K V,2013.Mining Functional Dependency in Relational Databases using FUN and Dep- miner:A Comparative study[J].International Joumal of Computer Applications,78(15):34 -36.

[8]CaldersT,Goethals B.2007.Non-derivable Itemset Mining[J].Data Mining &Knowledge Discovery,14(1): 171 -206.

[9]Catharine M W,Chris G,Edward L R.2001.Fast FDS:A Heuristic-driven,Depth -firsst Algorithm for Mining Functional Dependencies From Relation Instance[C]//Proceedings of the 3th International Conference on Data Warehousing and Knaoledge Discovering Mnnich,Germang:101 -110.

[10]Diallo T,Novelli N,Petit J M.2012.Discovering(Frequent)Constant Conditional Functional Dependencies

[J].Intemational Jounal of Data Mining,Modelling and Management,4(3):205 -223.

[11]Fan W F,Jia X B,Anastasios Kementsietsidis.2008.Conditional Functional Dependencies for Capturing Data Inconsistencies[J].ACM Transactions on Database Systems,33(2):1-48.

[12]Fan W,Geerts F,Lakshmanan L V S,et al.2009.Discovering Conditional Functional Dependencies[C]//Proceed- ings of the 25th International Conference on Data Engineering(ICDE).Shanghai,China;IEEE:1231-1234.

[13]Fan W,Geerts F,Li J,et al.2011.Discovering Conditional Functional Dependencies[J].IEEE Transactions on Knowledge &Data Engineering,23(5):683-698.

[14]Fei C,Miller R J.2008.Discovering Data Quality Rules[J].Proceedings of the VLDB Endowment,1(1):1166- 1177.

[15]Goethals B,Zaki M J.2003.Frequent Itemset Mining Implementations[C]//Proceedings of the ICDM 2003 Workshop on Frequent Itemset Mining Implementations.Melboume,Florida,USA:IEEE,1 -13.

[16]Hammon M,Herbst J,Leiner NK.2006.In tercctive workflow Mining-Requirements,Concepts and Implemen-

tation[J]Data and krowledge Engineering,56(1):41 -63.

[17]Huhtala Y,Karkkainen J,Porkka P,et al.1999.TANE;An EFicient Algorithm for Discovering Functional and

Approximate Dependencies[J].The Computer Jourmal,42(2):100-111.

[18]Huhtala Y,Karkkainen J,Porkka P,et al.1998.Eficient Discovery of Functional and Approximate[C].

[S.l]:International Interence on Data Mining.Orlando,Florida,USA:IEEE,392-401.

[19]Li H,Li J,Wong L,et al.2005.Relative Risk and Odds Ratio:A Data Mining Perspective[C]//Proceedings of 24th ACM SIGMOD -SIGACT -SIGART Symposium on Principles of Database Systems,PODS 2005.Baltimore,Massachusetts,USA:ACM:368-377.

[20]Li J,Li H,Wong L.2006.Minimum Description Length Principle;Generators are Preferable to Closed Patterms [C]//Proceedings of 21st AAAI Conference on Artificial Inteligence and 18th Innovative Applications of Ari- ficial Intelligence Conference.Las Vegas,Nebraska,USA;AAAI Press;409-414.

[21]LiJ,Liu G,Wong L.2007.Mining Statistically Important Equivalence Classes and Dela-discriminative Emer-

**230**)数据质量导论

ging Pattems[C]//Proceedings of 13th SIGKDD Intemational Conference on Knowledge Discovery and Data Mining.San Jose,Califomia,USA:ACM,430-439.

[22]LiJ,Liu J,Toivonen H,et al.2013.Effective Pruning for the Discovery of Conditional Functional Dependencies [J].The Computer Journal,56(3):378-392.

[23]Lim E,Srivastava J,et al.1996.Entity Identification in Databaselntegration[J].Information Scientices an Inter- notional Joumal,89(1):1-38.

[24]Novelli N,Ciccheti K.2001.FUN:An Efficient Algorithm for Mining Functional and Embedded Dependencies [C]//Proceedings of the 8th International Conference on Database Theory,London,UK:189-203.

[25]Pasquier N,Pasquier N,Bastide Y.2000.Discovering Frequent Closed Itemsets for Association Rules[J].Lec- ture Notes in Computer Science,1540;398-416.

[26]Wang J,Han J,Pei J.2003.CLOSET+:Searching for the Best Strategies for Mining Frequent Closedltemsets [C]//Proceedings of 9th International Conference on Knowledge Discovery and Data Mining Washington,Dis- trict of Columbia,USA:ACM,236-245.

[26]Wen L,Wang J.2009.A Novel Approach for Process Mining Based on Event Types[J].Joumal of Intelligent Information Systems,32(2):163-190.

[27]Zaki M J.2004.Mining Non-redundant Association Rules[J].Data Mining &Knowledge Discovery,9(3):

223-248.

**第10章** **基于规则的不一致** **数据检测与修复方法**

**10.1** **引言**

数据质量规则，即数据约束，是客观世界的数据所应遵循的限制条件(程录 庆，2011;郭志懋，2002;袁满，2013)。通常，数据质量规则的来源有三种途径： 一是 由业内专家根据先验知识指定质量规则，主要适用于数据属性较少、关系简单或专 家经验可信度较高的场合；二是从数据中自动挖掘出规则集，主要适用于数据关系 复杂且专家不可得的场合；三是结合专家领域知识，给定先验挖掘准则或部分依赖规 则，从数据中挖掘出规则集后，再由专家对规则集进行检查和筛选。基于规则的数据质 量检查，就是利用规则约束，对不同类型的数据问题进行专项检查(Zhang et al.,2011; 石少敏，2011)。对于给定的某一数据集，用户根据检查对象数据特点以及检查目的设 置合适的规则，然后进行数据质量检查，将不满足数据质量规则的问题数据检出。

数据质量规则的分类方法和具体种类有很多。由于不满足一致性是常见的数 据质量问题，因此数据一致性是数据质量约束规则中的一类重要约束。所谓数据 一致性(Data Consistency)是指数据对实体描述的有效性(Validity) 和完整性(In- tegrity), 主要用于检查数据中的不一致和冲突(Madnick et al.,2009)。在关系数 据表中，单表关系中的单条记录、多条记录和跨表关系的多条记录都可能存在不一 致的数据。常用的15类数据质量约束规则定义见第11章的表11-1,从表11-1 中可以看出与数据一致性相关的规则占相当的比例。

数据质量规则也可以从完整性、 一致性、准确性、时效性等四个维度进行分类 见第11章的表11-2。从表11-2中也可看出，违反数据一致性的问题确实是一 类重要的数据质量问题。

因此数据一致性是数据质量约束规则中的一类重要约束，不满足一致性是常 见的数据质量问题，本章拟对基于规则的不一致数据检测与修复方法进行研究。 首先研究了基于Fellegi-Holt 方法(Fellegi et al.,1976)的不一致数据检测方法， 基于已知规则自动推演出完备的最小规则集，并基于该规则集进行不一致数据检 测；然后，提出了一种基于Evidence-Rules 模型的不一致数据修复方法，基于上述 检测结果，充分利用原有记录的信息，有效提高修复过程的正确性和高效性。

)数据质量导论

**10.2** **基于Fellegi-Holt 方法的不一致数据检测**

对于实际的信息系统而言，它们往往都是面向某个特定的领域，系统中的 数据通常都应是合乎业务逻辑的。对于这样的具体应用，如何采用自动的方 法来解决数据中不符合业务逻辑的错误是一个具有实际应用价值的问题。基 于规则的不一致数据检查过程就是对数据库中的数据逐条检测记录是否满足 给定的规则约束。当然，这样的检查建立在具有一组合法的规则集的基础上。 然而，当所给出的规则集本身就存在错误或者规则之间相互矛盾，则会影响到 检查结果的准确性。

因此，在进行不一致数据检测之前，必须首先对检测的依据——规则集本身进 行考察，以保证其正确性。很显然，人工处理该问题往往面临着效率低且正确性难 以保证的问题。因此，有必要引入规则的自动逻辑检查。所谓规则的自动逻辑检 查，就是对规则集自身进行正确性校正的处理机制(Fan,2008)。 这主要针对初始 规则集中存在的规则冗余、规则冲突、规则不完备等情况。通常是采用形式化的方 法对规则进行统一的表述，然后通过符号计算的方式进行相关的逻辑推理与校验， 最终得到一个完备的规则集。Fellegi-Holt 方法就是一种专门针对逻辑规则推演 提出的模型(Fellegi et al.,1976)。

**10.2.1** **Fellegi-Holt方法**

1976年，Fellegi 等人提出一个严格的形式化数学模型 Fellegi-Holt 方法(Fel- legi et al.,1976),对于具体的应用，该模型根据领域知识定义显式约束规则，然后 根据数学方法求出规则闭集。对于每条记录，自动判断是否违反规则约束。该方 法的正确性已经得到证明，并且广泛用于政府机构的相关统计审核软件系统中，如 统计软件 SPEER 和 DISCRETE。近些年来，有关对该方法的形式化分析以及算法 改进和效率提升等方面的研究层出不穷(William,2002)。

[**10.2.1.1**](10.2.1.1) **Fellegi-Holt** **方法简介**

通过一个简单的例子说明 Fellegi-Holt 方法以及它所解决的主要问题。 **示例1:**考虑一张某公司职员信息统计表中的数据，具体模式如下：

员工信息(姓名，年龄，优秀员工，工龄)表示某个员工的姓名、年龄、工龄以及 是否被评为优秀员工的信息。由于数据采集过程的某些原因，部分记录的属性取 值存在缺失。具体数据实例如表10-1所列。

第10章基于规则的不一致数据检测与修复方法 

表10- 1 某公司员工信息数据实例

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **No.** | **姓名** | **年龄** | **优秀员工** | **工龄** |
| 1 | Tim | 20 |  | 0 |
| 2 | Tom |  | 1 | 1 |
| 3 | Jim | 20 |  | 6 |
| 4 | Lily | 18 | 1 |  |

该公司对员工的招收以及评优有一些要求和条件，从而进一步约束了信息表 中部分属性值的取值范围：

规则1:优秀员工=1 →工龄>5

规则2:工龄>1 → 年龄>21

即如果某一员工被评为优秀员工，那么其工龄至少为5年；如果某一员工工龄 超过了1年，那么其年龄至少21岁。

通常情况下，进行数据质量检查时，含有缺失值的记录往往被视为问题数据被 舍弃，那么表10-1中的四条记录会被舍弃。然而实际上记录1并没有违反任何 一条约束规则。相反，可以通过属性值“工龄”=0以及规则1进一步推断出该员 工不是优秀员工，进而对缺失字段“优秀员工”进行填充，填充值为0。

按照常用的函数依赖方法，直接利用上述规则对表10-1中的记录进行检查， 不难发现记录2违反了规则1(员工Tom是优秀员工，但是工龄不满5年),记录3 违反了规则2(员工Jim 工龄6年，但是年龄只有20岁)。

实际上，记录4也是有问题的，然而由于记录中工龄字段值的缺失，上述两条 规则并不能有效检测出该条记录中存在的问题。

仔细分析不难发现，结合规则1和规则2,可以推断出其隐含的第3条规则。

规则3:优秀员工=1 →年龄>21

即如果某一员工是优秀员工，那么其年龄至少21岁。

显然，记录4违反了规则3(员工Lily 是优秀员工，但是年龄只有18岁)。

简言之，Fellegi-Holt 方法就是在已知规则的基础上，进一步找出规则所蕴含 的其他规则，以更加准确地对问题数据进行检测的方法。

[**10.2.1.2**](10.2.1.2) **Fellegi-Holt 方法的符号表示及计算**

Fellegi-Holt 方法适用于以记录形式存在的数据，且相关属性值的取值范围 为有限集，同时要求各字段存在清晰的逻辑规则。该方法主要解决以下两个方面 的问题： 一是如何根据已知规则挖掘出所有潜在的规则，二是对给定的问题记录， 找出最小错误字段集合。

对包含N 个字段的记录T(T,T₂,…,T), T, 产生新规则的方法 FHG 如下：

对于起始规则集E 以及任一字段

)数据质量导论



式10 - 1中，A,表示第j 属性的值域(T;∈A;);A;表示不满足规则e的第j属性的值 域；另外，记 edit(e) 为不满足某条规则e 的记录取值空间。

重复上述操作，直到不再有新的edit产生，此过程结束后产生的集合称为最大 产生集(Maximal Generated Edits),记作 MGE(E)。

在上述例子中，规则1对应的 edit 可以表示成

e₁={A 优秀员工：1}×{ A工龄： 0,1,2,3,4,5}×{A年龄} (10-2)

即任何一个优秀员工，并且工龄小于5年的记录都不满足规则1,无论年龄字 段取值为多少。

规则2对应的 edit 可以表示成

e₂={Aπ 龄：2,3,…}×{A年龄：0,1,…,21}×{A优秀员r} (10-3)

即任何工龄大于1年，且年龄小于等于21 的记录都不满足规则2,无论其是 不是优秀员工。

那么通过计算新产生了如下 edit:

FHG(A 工龄，{e₁,e₂})

={A 年龄}∩{A年龄：0,1,…,21}×{A工龄：2,3,…}U{A 工的：0,1,2,3,4,5} ×{A优秀员r }∩{A 优秀员工：1}

={A年龄：0,1,…,21}×{A工龄}×{A优秀员工：1} (10-4)

其对应的规则3:优秀员工=1 →年龄>21。

对于给定的某一规则集，利用FHG算式产生新规则的过程是一个不断重复计 算的过程。可终止性主要指对于某次计算，由已知规则集和某一字段，算法必将停 止并得到一个新的规则集。

FHG 的计算过程中，每次选取一个属性A,将与之相关的规则集进行计算，产 生新的n 条规则。因此，最多进行INI次计算可进行一轮遍历。同样地，因为所有 计算均为集合运算，集合的运算复杂度在常数范围内，因此单次计算不会陷入死循 环状态中。然而，当规则基数较大时，此计算过程工作量相对较大，有时为了考虑 效率问题(Yeh,2010), 当某一轮计算新产生的规则数小于某一界限时，可认为剩 余规则量对于数据检测影响不大，算法达到某一收敛状态，停止计算。

总而言之，当规则基数较大或者规则形式较为复杂时，每一轮的计算需要很大 的时间与空间资源，从而降低了算法在实际运行中的效率。减少迭代次数可以降 低工作量，但同时也会导致部分规则无法被计算得出。因此，很多实际操作需要针 对具体需求在算法精度与效率之间进行权衡。

<10.2.1.3> Fellegi-Holt 方法适用范围研究

从前面的论述可知，Fellegi-Holt 方法根据已知规则计算得到其隐含的其他

第10章 基于规则的不一致数据检测与修复方法(235

规则，从而避免了因规则不完善导致部分问题数据无法检测的问题。在实际操作 中，因为受到算法所应用数据类型特点的限制，该算法主要针对属性之间具有较强 逻辑关系的结构化数据。对于这一类型的数据，算法在丰富规则集以及提高检测 准确性方面具有较明显的效果。然而该方法主要适用于属性值取值范围为有限集 的数据，即具有离散特点的枚举类型数据。

在实际的数据集中，属性值取值范围为连续值的情况相当普遍，例如，物品的 长度、寿命等。这些属性通常也与数据集中的某些其他属性(如等级划分)存在明 显的逻辑关系，可以进一步用来进行问题数据检测的规则依据。因此考虑对 Fel- legi-Holt 方法应用于连续属性值域的情况进行分析讨论。

这里主要分析两种情况： 一是规则涉及到的属性取值为某一连续区间；二是规 则涉及到的属性取值为某一连续区间内的若干个特殊值。

1)连续区间情况

假设现有如下两条规则：

规则1:x<a→y>b。

规则2:z<c→d<x<f。

其中A₄=(0,+);A,=(0,+~);A,=(- ,+);d≥a。a,b,c,d,f 为

常数。

那么,规则1对应的 edit 可以表示成

e₁={A₄:(0,a)}×{A,:(0,b)}×{A,} (10-5)

规则2对应的 edit 可以表示成

e₂={A:(-o,c)}×{A:(0,d)U[f,+o)}×{A,}

那么通过计算新产生如下edit:

FHG(A₄,{e₁,e₂})

={A,}∩{A,:[0,b]}×{A:(0,a)}U{A:[0,d]u[f,+w)} ×{A,}n{A₂:(-o,c)}

={A,:[0,b]}×{A,}U{A,:(-o,c)}

即得到规则3:z<c→y>b。

2)连续区间内特殊值情况

假设现有如下两条规则：

规则1:xe{a,b,c,d}→ye{df}。

规则2:ze{f,g,a}→xe{b,d}。

其中A₄=(0,+ );A,=(0,+ );A₂=(-,+~);a,b,c,d,f,g 那么,规则1对应的edit 可以表示成

e₁={A₄:{a,b,c,d}}×{A,:(0,+簾) - {d,f}}×{A,} 规则2对应的 edit 可以表示成

(10-6)

(10-7)

为常数。

(10-8)

236)数据质量导论

e₂={A:|f,g,a}}×{A:(0,+o)-{b,d}}×{A,} (10-9)

那么通过计算新产生如下 edit:

FHG(A,,{e₁,e₂})

={A,}n{A,:(0,+m)-{d,f}}×{A:{a,b,c,d}}U{A:(0,+x)

*-{b,d}{×{A,}n{A:{f,g,a}}*

={A,:(0,+簾) - {df}}×{A,}U{A:{f,g,a}} (10-10)

即得到规则3:ze{f,g,a}→ye{dfì。

通过观察可知 Fellegi-Holt 方法在对规则进行推演计算中，实质是对规则所 描述的属性取值范围进行的集合运算。而对于集合运算中的“交集”与“并集”运 算，其具体运算操作过程与结果和集合的元素类型没有关系。因此，对于数据质量 规则所涉及的属性取值为某一连续区间或者区间内的某些特殊值，集合的运算方 法不受其数据类型的影响。所以，将 Fellegi-Holt 方法的适用范围在属性值取值 范围上进行推广是切实可行的。

10.2.2 检测流程及策略

数据检测是指采取某些质量检查算法，发现数据集中不满足预设规则的问题 数据，其中问题数据是指不完整数据、错误数据和不一致的数据。本节主要研究对 不一致数据的检测。为了获得更加准确的问题数据检测效果，对于设定好的规则 集用 Fellegi-Holt 方法进行规则的一致性检查。下文所讨论的不一致数据检测的 依据，是基于Fellegi-Holt 方法处理完以后所生成的完备的规则集。

<10.2.2.1> 不一致数据检测流程

问题数据检测的目的在于发现数据中的问题记录，即不一致数据，检测过程可 以看作是一个二分类的过程：该过程的输入是一条记录，输出是一个结果标记，即该 条记录“是”或“不是”不一致数据。整个过程中的核心部分是二分类的具体操作。 在数据质量检测过程中，分类器可对应于一个数据质量规则集。分类器的处理部分将 输入的记录与规则集进行比较验证，并根据检测的具体需求选择终止条件和输出格式。

<10.2.2.2> 两种检测策略和算法

数据质量规则作为分类器中的核心部分，是判断记录是否为问题数据的参照 标准，可以有效地发现数据中存在的相关问题。现实情况下的问题数据检测工作 往往有不同的要求：有些检测工作旨在发现问题数据，即对于一条记录得出其 “是”或者“不是”问题数据的结论；有些则需要进一步地对问题数据进行修复等操 作，因此需要提供其不满足哪些约束规则的具体细节。因此，需要针对不同的检测 目的选取最合适的策略，以此获得更高的运行效率。

本节设计了两种不同的检测策略，分别针对问题数据的发现以及问题数据的修复。

第10章 基于规则的不一致数据检测与修复方法 (**237**)

对于数据库中的某张数据表，用给定的规则集进行数据质量检侧，检测的目的 在于发现并找出其中存在问题的数据记录。算法主要分为两个部分：首先，利用 Fellegi-Holt 方法生成给定规则的最大产生集MGE;然后利用生成的规则最大产 生集MGE检测并发现待检数据集中的问题数据。在检测过程中，为了避免某条规 则重复用于检测，对规则进行标记。

面向问题数据发现的检测算法如表10-2所列。

表10 - 2 检测算法1(面向问题数据发现)

|  |
| --- |
| Name:检测算法1 |
| Input:一条记录R,一个给定规则集 ∑  Output:问题记录R |
| 1.R'=null;  2.calculate MGE(2);  3.While MGE(E)is not null  4.For each pe MGE(2)do  5. If R does not match q  6. R'=R;  7.Else MGE(2)=MGE(2)-φ  8.Return R'; |

算法1适用于只需要找出数据表中问题记录的方案，为了进一步提高问题数 据的检测效率，该算法对于每条记录，与规则集中的每条规则依次进行比对。当发 现其不满足某条规则时终止比对，将该条记录标记成问题数据并返回，并取出下一 条记录进行比对。重复这个过程，直至所有数据记录检测完毕。

如表10-3所列为面向问题数据修复的检测算法。

表10- 3 检测算法2(面向问题数据修复)

|  |
| --- |
| Name:检测算法2 |
| Input:一条记录R,一个给定规则集  Output:问题记录R',违反的规则集 |
| 1.R'=null;X'=null;  2.calculate MGE(2);  3.For each φe MGE(2)do  4.If R does not match p  5. E=xU;  6. MGE(2)=MGE(Z)-φ;  7. R'=R;  8.Return R'and "; |

238)数据质量导论

算法2适用于以问题数据修复为目标的数据检测， 一方面需要找出数据表中 的问题记录，另一方面还要对哪些属性的取值违背了相应的规则进行标记。因此， 需要进行逐条记录与规则集中所有规则进行比对，这也会进一步增加算法复杂性 和运行时间。在后续的讨论中，我们还将就如何获得问题记录修复的最优方案提 供简单思路。

**10.2.3** **实验及分析**

为了检验上述方法的可行性与有效性，分别采用生成数据与实际数据进行了 实验。具体实验内容如下。

[**10.2.3.1**](10.2.3.1) **实验数据集**

数据集1:为了更加贴合实验需求，以某设备维修保养信息数据为依据，参考 其中具有明显逻辑关系的字段进行数据生成。其中主要字段含义及取值如 表10-4所列。

表10-4 设备数据表字段信息

|  |  |  |  |
| --- | --- | --- | --- |
| **序号** | **字段名称** | **值域** | **含义** |
| 1 | 设备编号 | 000～999 |  |
| 2 | 使用状态 | {0,1,2} | {未启用，使用中，已停用 |
| 3 | 使用时间/天 | 0~99 |  |
| 4 | 维修次数/次 | 0~10 |  |

结合实际情况以及实验需求，在满足各属性值域范围内利用现有的数据生成 软件产生了共850条记录，并控制使得共有100条记录含属性值缺失(其中含有2 个及以上缺失属性值的记录共25条)。

数据集2:实际数据采用来源于UCI 网站的有关疾病统计方面数据(Breast Cancer Wisconsin(Original)Data Set)(Mangasarian et al.,1990)。该数据集含有 11个属性(序号来自系统内部)共699条记录，数据集本身含有部分属性值缺失 的情况且各属性取值均为整型，便于逻辑关系设定。具体字段信息如表10-5 所列。

表10- 5 Breast Cancer 数据表字段信息

|  |  |  |  |
| --- | --- | --- | --- |
| **序号** | **字段名称** | **值域** | **含义** |
| 1 | Sample code number | 按序递增 |  |
| 2 | Clump Thickness | 1~10 |  |

第10章基于规则的不一致数据检测与修复方法(239)

(续)

|  |  |  |  |
| --- | --- | --- | --- |
| **序号** | **字段名称** | **值域** | **含义** |
| 3 | Uniformity of Cell Size | 1~10 |  |
| 4 | Uniformity of Cell Shape | 1~10 |  |
| 5 | Marginal Adhesion | 1~10 |  |
| 6 | Single Epithelial Cell Size | 1~10 |  |
| 7 | Bare Nuclei | 1~10 |  |
| 8 | Bland Chromatin | 1~10 |  |
| 9 | Normal Nucleoli | 1~10 |  |
| 10 | Mitoses | 1~10 |  |
| 11 | Class | {2,4} | |benign,malignant |

<10.2.3.2> 实验过程

对于数据集1,我们用来检验采用MGE 算法进行错误数据检查时的改进效 果。在进行对比时，主要采用了以下3种检查策略。

(1)含空即错：只要记录属性值中包含空值即检测为问题数据。

(2)既定规则：使用预先定义好的规则对字段取值进行检测，不满足任一规则 即检测为问题数据(规则涉及属性值缺失时，不进行检测)。

(3)MGE: 对既定规则采用Fellegi-Holt 方法计算出规则的最大产生集，然后 对字段取值进行检测，不满足任一规则即检测为问题数据(规则涉及属性值缺失 时，不进行检测)。

结合属性之间的具体关系，设定如下4条规则：

规则1:使用状态=0 →使用时间=0。

规则2:使用状态>0 → 使用时间>0。

规则3:使用时间≤3 →维修次数=0。

规则4:维修次数>0 →使用状态>0。

采用Fellegi-Holt 方法进行计算，得到的 MGE 最大产生集中新增了如下两条 规则：

规则5:使用状态=0 → 维修次数=0。

规则6:维修次数>0 →使用时间>0。

对于检测结果，用查准率Pr 和查全率 Re 两个指标进行方法评价，属性值之间 不满足逻辑关系的记录视为问题数据，此处，指标定义如下。

(10-11)

(10-12)

240)数据质量导论

式中：X 表示检测到的问题数据数；Y表示检测到的正确数据数；Z 表示未检测到 的问题数据数。

对于数据集2,我们用来检验问题数据检测算法分别在进行问题数据定位以 及违反规则确定时的效率，以此说明针对不同检测需要选择合适算法策略的必要 性。鉴于实验需要，对数据集2中的数据进行了简单扩充，对该数据集进行重复生 成使得其大小为原来10倍。

通过分析表中的字段关系以及统计数据的特点，根据病情严重程度与症状特 点，归纳出可能存在的相关规则，并以此为依据对数据进行检查。其中：

**规则7:**Class =2→Bare Nuclei≤5。

**规则8:C**lass=4→Bare Nuclei≥4。

**规则9:**Uniformity of Cell Size=1→Uniformity of Cell Shape =1 。 **规则10:**Uniformity of Cell Size>7→Uniformity of Cell Shape>7。 **规则11:**Class =2→Uniformity of Cell Size <2。

采用 Fellegi-Holt 方法进行计算，得到的 MGE最大产生集中新增了如下一条 规则：

**规则12:**Class =2→Uniformity of Cell Shape =1。

分别利用10.2.3节中的两种问题数据检测策略，以上述6条规则为依据对 Breast Cancer Wisconsin(Original)Data Set 进行检测。在得出检测结果的基础上， 进一步比较两种策略在工作量以及时间耗费上的差异。

[**10.2.3.3**](10.2.3.3) **结果分析**

数据集1的实验结果如表10-6所列。其中，将违背 MGE 中任何一条规则或 者超过两个属性值缺失的记录视为问题数据。由于方法各自的缺陷，存在漏检和 错检的现象。

表10- 6 数据集1实验结果记录

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **策略** | **总记录数** | **问题数** | **检出错误** | **实际错误** | **查全率** | **查准率** |
| 含空即错 | 850 | 250 | 100 | 75 | 0.30 | 0.75 |
| 既定规则 | 850 | 250 | 165 | 165 | 0.66 | 1 |
| MGE | 850 | 250 | 200 | 200 | 0.80 | 1 |

通过观察图10-1不难发现，盲目地排除含缺失属性值的数据其查全率和查 准率都非常低下，极大地造成了原始数据信息的丢失。

利用Fellegi-Holt 方法可以根据已知规则推演出所有的隐含规则，利用规则 最大产生集能检测出所有不满足逻辑规则的问题数据，从查全率指标上来看，其检 测结果优于仅仅利用既定规则进行的问题数据检测。但是对于缺失值过多的问题 数据无法检出，因此查全率方面有所不足。如果将MGE方法结合空值检测，则能

第10章基于规则的不一致数据检测与修复方法 

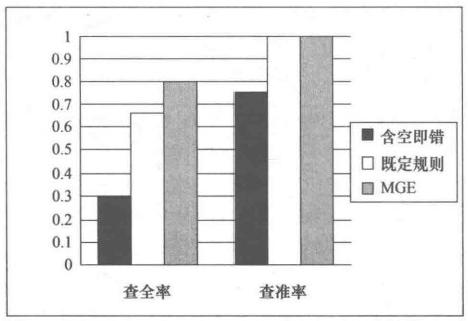


图10-1 不同检测算法的效果比较

够进一步提高检测结果的查全率。

数据集2分别利用两种策略下的问题数据检测算法对数据进行检测，并在算 法中加入代码对检测过程中记录与规则比对次数以及算法耗时进行统计，具体结 果如表10-7和10-8所示。

表10 - 7 数据集1实验结果记录(策略1)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **规则数** | **总记录数** | **问题记录数** | **比较次数** | **用时/ms** |
| 1 | 6990 | 530 | 6990 | 47 |
| 2 | 6990 | 810 | 7520 | 47 |
| 3 | 6990 | 1210 | 7800 | 48 |
| 4 | 6990 | 1220 | 8200 | 47 |
| 6 | 6990 | 1790 | 8540 | 48 |

表10- 8 数据集1实验结果记录(策略2)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **规则数** | **总记录数** | **问题记录数** | **比较次数** | **用时/ms** |
| 1 | 6990 | 530 | 6990 | 62 |
| 2 | 6990 | 810 | 13980 | 75 |
| 3 | 6990 | 1210 | 20970 | 78 |
| 4 | 6990 | 1220 | 27960 | 94 |
| 6 | 6990 | 1790 | 41940 | 110 |

通过观察图10-2和图10-3可知：给定同一个规则集，两种策略都能够有效

242)数据质量导论

地检测出数据集中存在的问题记录。随着规则数的不断增加，两种算法的比较次 数和用时都呈上升趋势。而相同规则数情况下的比较次数和用时，算法2都明显 高于算法1。因此算法1更适合于仅仅需要问题数据定位的情况。

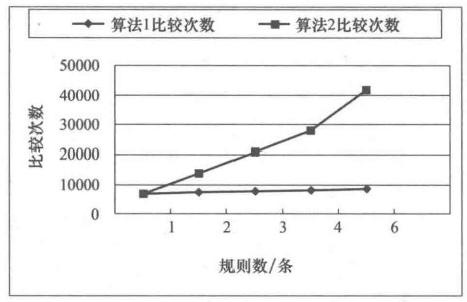


图10-2 不同检测算法的效果比较

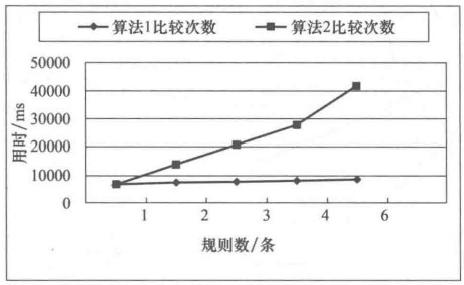


图10-3 不同检测算法的用时比较

**10.3** **基于** **Evidence-Rules 模型的不一致数据修复**

数据修复是指产生一条与原记录对应的记录，使其满足所有的规则约束并且 尽可能与原记录的相关信息保持一致(Fan et al.,2011;Fan et al.,2012)。 为了达 到上述目标，通过修改问题记录的某些字段值来使得其满足所有的规则约束。问 题数据修复作为数据质量提高的关键操作之一，用来作为修复依据的规则通 常是数据之间的函数依赖以及专家知识等(Beskales et al.,2010;武森，等， 2012)。而根据基本的函数依赖来完成数据修复工作在实际操作过程中虽然

第10章基于规则的不一致数据检测与修复方法(243

简单易行，但是由于单一的函数依赖通常缺乏足够多的知识，使得数据修复的 结果不可避免地产生错误(Cong et al.,2007)。为了能够尽可能地利用记录 当中现有的信息，选择较好的数据修复方案，本节设计了一种新的基于规则的 问题数据修复模型。

10.3.1 确定问题记录中待修改属性集

为了更加清楚地了解问题数据修复的具体操作过程，下面我们通过符号形式 化的方法进行简单描述(Fellegi et al.,1976)。

根据10.2.2节中 Fellegi-Holt 方法介绍可知，edit 表示不满足某条规则的记 录取值空间。当记录R属于e, 表示记录R 不满足e 所表示的规则。相反，当 R不 属于e,表示记录R 满足e所表示的规则。

不妨设e={A₁:A₄}×{A:A}×A, 记录R 不满足规则e,(R, 表示记录R 第 i 属性的取值)那么必然有R;属于A₂, 并且R,属于A₂。通过修复使得记录R 满足规 则e,只需要改变A,或者A,的取值，使得A,不属于A,或者A,不属于A,即可。

下面举例说明：

e₁={A 优秀员工： 1}×{A工龄： 0,1,2,3,4,5}×A年龄。

e₂={Arm:1,2,3,…}×{A 年的：0,1,…,21}×A秀员工。

e₃={A 年龄： 0,1,…,21}×A工龄×{A优秀员工：1}。

记录(Tom,/,1,1) 属于e₁, 因为Tom 是优秀员工，工龄却只有1年。为了修复 记录，只要修改规则e₁ 中涉及的两个字段“优秀员工”与“工龄”其中之一，使得其 不属于e₁。

如表10-9所列记录(Helen,18,1,1) 中，该记录违反了3条规则，即同时属于 ej,e₂,e₃, 其中与之相关的属性如表10-10所列。

表10-9 数据实例

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **序号** | **姓名** | **年龄** | **优秀员工** | **工龄** |
| **5** | **Helen** | **18** | **1** | **1** |

表10-10 规则与属性对应关系

|  |  |
| --- | --- |
| **规则** | **涉及属性** |
| e₁ | 优秀员工、工龄 |
| e2 | 工龄、年龄 |
| e₃ | 年龄、优秀员工 |

为了修复记录5,需要修改e₁,e₂和e₃ 涉及的相关属性值，使得记录5不属于任

244)数据质量导论

何一条edit 对应的记录取值空间。通过观察可知，通过选择3个字段中的任意2 个即可使得记录覆盖上述3条规则，进一步修改这两个属性的取值可以得到合乎 规则的正确记录。

因此，为了使数据修复的代价最小(Bohannon et al.,2005;Chomicki,2005),需 要通过修改最少的属性值得到合乎规则的记录。

根据算法2的检测结果，已知一条问题记录 R'以及该条记录违背的约束规则 集2”,要想使得记录符合某条规则，只需要修改这条规则所对应的edit 中的任意 字段值，使得其不属于这条 edit 的取值空间。问题进一步转化为找出”中所有规 则的最小字段覆盖集。

对于最大产生集 MGE 中的每一条规则e,记A(e)={A,,A;} 表示与规则e 相 关的属性值集合。对于问题记录R '以及该条记录违背的约束规则集∑ ”,可以计算 得到该条记录中每个属性值A;违反的规则数count(A),以及2'中与属性值A,相

关联的规则集合2(A;) 。确定待修改属性集的算法如表10-11所列。

表10-11 确定待修改属性集的算法

|  |
| --- |
| **Name:修改属性集算法** |
| Input:问题记录R',违反的规则集"  Output:需要修改的最小属性集F(R') |
| 1.F(R')=null;  2. For each A;eA do  3. count(A;);  4. while 2"is not null  5. select the maximum count(A;);  6. X=X-E(A₁);  7. A=A-A  8. F(R')=F(R')+A;:  9. End while  10. End for  11.Return F(R'); |

简言之，已知一条问题记录 R'以及该条记录违背的约束规则集2”,上述算法 给出了对该问题记录进行修复时需要改动的最小属性集。

**10.3.2** **基于函数依赖规则的属性值修复**

在已知某条问题记录需要修改的属性后，函数依赖通常是最常用的问题 数据修复依据。函数依赖表示数据记录中的某些属性之间的取值存在一定的

第10章基于规则的不一致数据检测与修复方法(245

联系，通常成对出现，因此当某种组合的出现不符合函数依赖的规则，如“国 家”和“首都”这一属性对取值为“中国”和“东京”,很显然是一对问题取值， 不满足一致性要求，需要对其中的某个取值进行修改。然而究竟是根据“中 国”将“东京”改成“北京”,还是根据“东京”将“中国”改成“日本”,则需要进 一步研究决定。

在通常的修复工作中，我们更多地采用统一的修复方法。例如，根据函数依赖 属性对取值中的前者来对后者进行修复，那么此时问题数据修复的过程，可以表示 为：{(A:p(A),B:p(B));(D:p(D)→D:tp'(D))}。

其中，A和 B 是函数依赖规则中涉及的属性集，tp(A) 和 tp(B) 对应某一组 取值。

D是一个问题数据属性集，p(D) 是问题取值，D=B 且 p(D)=p(B),tp'(D) 是正确取值。

即当属性集A取值为p(A) 时，根据函数依赖关系，属性集B 的取值应当为

*tp'(D)。*

这里的D 表示单个属性。因为即使D 表示一个属性的集合，我们也可以将其 拆分形成多条规则，同时不影响规则集的基本特性，如下所示：

{(A:p(A),B:tp(B));(d₁:tp(d₁)→d₁:p'(d₁))};

*{(A:tp(A),B:p(B));(d₄:tp(dn)→d₄:p'(dn))};*

其中，D={d₁,d₂,…,d,}。

下面，我们通过一个例子来了解一下函数依赖关系如何用来进行问题数据检 测与自动修复。考虑一张人员参会的相关信息记录表，其模式为：参会信息(姓 名、国家、首都、城市、会议)。分别表示了参会人员的姓名、会议举行的国家、该国 家的首都、会议举行的城市以及会议的名称。具体数据实例如表10-12所列。

表10-12 人员参会信息记录举例

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **No** | **姓名** | **国家** | **首都** | **城市** | **会议** |
| 6 | George | 中国 | 北京 | 北京 | SIGMOD |
| 7 | lan | 中国 | 上海(北京) | 上海 | ICDE |
| 8 | Peter | 中国(日本) | 东京 | 东京 | ICDE |
| 9 | Mike | 中国 | 东京(北京) | 北京 | SIGMOD |
| 10 | Jim | 加拿大 | 多伦多(渥太华) | 多伦多 | VLDB |

表10-12所列数据实例中，国家和首都之间的依赖关系如表10-13所列。

**246)**数据质量导论

表10-13 国家—首都函数依赖关系举例

|  |  |  |
| --- | --- | --- |
| **函数依赖规则** | **国家** | **首都** |
| 91 | 中国 | 北京 |
| 92 | 加拿大 | 渥太华 |
| 93 | 日本 | 东京 |

不难看出7、8、9、10这几条记录都存在问题数据。针对上述情况，模型根据规 则提出以下修正：

上述函数依赖关系中，φ,表示如果一条记录的“国家”字段取值为“中国”,那 么“首都”字段取值应当为“北京”;反之亦然。同理，φ₂表示如果一条记录的“国 家”字段取值为“加拿大”,那么“首都”字段取值应当为“渥太华”。φ₃表示如果一 条记录的“国家”字段取值为“日本”,那么“首都”字段取值应当为“东京”。

根据上述规则，我们将记录7、9、10中的“首都”字段值按照上述表格中的对 应关系进行修改即可得到正确的记录。

然而对于记录8,我们发现，通过将“首都”字段的取值“东京”改成“北京”得 到记录(Peter, 中国，北京，东京，ICDE), 该条记录仍然存在问题。因为“城市”和 “会议”字段的取值“东京”和“ICDE”与“国家”的取值“中国”存在矛盾。因此，要 想得到正确的记录，还需要对“城市”和“会议”字段的取值进行修改。

事实上，我们只需要将该条记录的“国家”字段取值改为“日本”即可得到一条 正确的记录。根据“代价最小化”原则，显然后者所做的修改更加合理，保持了更 多的与原记录一致的信息。

在上述例子中，当记录中“国家”与“首都”字段取值与参照表中函数依赖信息 发生矛盾时，为了进一步确定对其中的哪个字段进行修改，我们需要借助其他字段 的取值作为辅助决策的重要信息，使得修改方案中在保证记录正确的前提下修改 代价最小，即对最少的字段进行修改。

通过这个例子，我们可以了解到根据函数依赖规则进行数据修复在某些情况 下仍然存在问题，将对修复结果的质量产生影响。因此，针对上述修复模型的缺 点，我们设计了基于Evidence-Rules 的问题数据修正模型，用于问题数据的检测 与修复。

**10.3.3** **Evidence-Rules** **模型与问题数据修复**

本节中，首先给出 Evidence-Rules 模型的相关定义(见定义10 -1),再讨论 如何运用其进行问题数据修正的相关问题。

定义 10-1:考虑一个模式R, 其所有属性集合为attr(R),AεR 表示A 是

第10章 基于规则的不一致数据检测与修复方法(247

attr(R) 中的一个属性，对每一个属性AεR, 其值域定义为 dom(A) 。那么定义在

模式R 上的 Evidence-Rules 模型可以表示为：{(A:p(A),B:φ(B));(C:p (C));(D:p(D)→D:tp'(D))} 。 其中：

A、B、C、D分别表示属性集合，p(X) 表示属性集X 的一组取值；p(A) 和tp(B) 称为“冲突模式”,是值域内的一组相互矛盾的取值；p(C) 称为“证据模式”,是值 域内的某一组正确取值，作为问题数据修复方案选择的依据；D是一个问题数据属 性集，p(D) 是问题取值，(D=A 且 p(D)=tp(A)) 或者(D=B 且 p(D)=

tp(B)),φ'(D) 是正确取值；p(D)→φ'(D) 是问题取值的修改方法，即“解决方 案”。

即，“冲突模式”中的两个属性组的取值存在冲突，需要进行问题数据修复，修 改方案是将属性集D 的取值p(D) 改为p'(D), 而选择这种修改方案的依据因为 属性集 C 的取值为p(C)。

例如，可以将上述例子中的问题数据通过以下规则进行自动修复：

4₄ :{(国家：中国，首都：东京);(城市：北京);(首都：东京→首都：北京)}; qs:|(国家：中国，首都：东京);(城市：东京);(国家：中国→ 国家：日本)}。

即，根据规则φ₄将记录4中的“首都”取值“东京”改成“北京”,根据规则φs将 记录8中的国家取值“中国”改成“日本”,从而得到两条正确的记录。

传统的基于函数依赖的数据质量规则更多的是用来对问题数据进行检测，至 于如何对检测到的问题数据进行修复需要人为确定具体使用哪一条规则作为依 据。而 Evidence-Rules模型参考了比普通规则更多的记录信息，可以实现对问题 数据的自动修复。

利用Evidence-Rules 模型对问题数据进行修复，为了确保修复过程的正确性 和高效性，需要对修复过程进行记录，防止重复工作，因此采取对属性进行标记的 方法来进行控制。

我们将记录t 通过规则φ作用于属性集A 得到新的记录 t'这一过程表示如 下：t→(A,φ)t'。

[**10.3.3.1**](10.3.3.1) **模型分析**

1)可终止性

考虑一条记录的修复过程：t₀→(A₁,φ₁),t;…→(A;,φ₁),t …, 每经过一次修复 操作，被修改的属性个数至少增加1,当被修改的属性个数增加到最大值(即模式 中属性个数)时，修复过程必然终止。

当然被修改的属性个数不应该达到IRI, 否则原记录的所有信息都被修改了， 这样的数据修复是毫无意义的。

2)一致性

一致性问题是指对于给定的规则集，规则之间不存在相互冲突的情况，即所有

)数据质量导论

规则作用于同一条记录时得到同样的修复结果。规则与规则之间的一致性主要包 括以下3种情况。

**情况1:**如果{(A:p(A),B:p(B));(C:p(C));(B:φ(B)→B:p'

(B))}, 并且{(A:p(A),B:p(B));(C:p(C));(B:p(B)→B:p”(B))}, 应

当有 tp'(B)=tp”(B)。

**情况2:**如果有规则{(A:印(A),B:p(B));(C:p(C));(B:p(B)→B:p'(B))}, 不能存在规则{(A:p(A),C:p(C));(B:p(B));(C:p(C)→C:p'(C))}。

**情况3:**如果有规则{(A:p(A),B:p(B));(C:p(C));(B:p(B)→

B:p'(B))}, 不能存在规则{(A:p(A),B:φ'(B));(C:p(C));(B:p'(B)→

*B:tp(B))}。*

对给定的规则集进行一致性分析是很有必要的，因为当规则集本身存在矛盾 时，其作用于记录得到的修复结果必然不能够保证正确性。 一致性分析经常被用 于条件函数依赖等方法的检测中。

3)最小性

最小性是针对规则集中的规则冗余问题提出的。对于一个给定的 Evidence - Rules集合 以及某条规则φ,当ZU{φ} 是一致的，并且对于任意一条记录t,t₀→ (2)t₁ 和t₀→(2U{φ\)t₂ 的修复结果是一样的，我们说规则φ可以由集合三推出。 此时，规则φ可以被认为是冗余的。在进行数据修复时，冗余规则不会对修复结 果的正确性造成影响，但是将规则集中的冗余规则预先进行筛选并删除可以有效 提高修复工作的效率。

4)确定性

确定性问题是指对于一条给定问题记录t, 任何修复过程必定以同样的修复结 果结束。根据可终止性和一致性的定义可以很容易得出：如果一个 Evidence -Rules 集合是一致的，将其作用于记录t, 其过程必将终止，并且所得结果是唯一确定的。

[**10.3.3.2**](10.3.3.2) **问题数据修复的空间表示法**

问题数据修复是一个不断对记录中的属性值进行修改的过程(Fan et al., 2012)。不妨假设任意一条规则的作用仅对一个属性的取值进行修改，即{(A:p(A), B:p(B));(C:p(C));(D:p(D)→D:p'(D))} 模型中的D 表示单个属性。因 为即使 D 表示一个属性的集合，也可以将其拆分形成多条规则：

{(A:p(A),B:p(B));(C:tp(C));(d₁:p(d₁)→d₁:tp'(d₁))};

{(A:p(A),B:tp(B));(C:tp(C));(d₄:p(d₄)→d₄:tp'(d₄))};

D={d₁,d₂,…,d₄}。

同时不影响规则集的基本特性。

我们考虑问题数据修复的基本过程，每次对记录中的一个属性值进行修改。

第10章 基于规则的不一致数据检测与修复方法 (**249)**

根据算法的可终止性原则，每个属性值最多被修改一次，因此记录中的每个属性值 的取值最多存在两种情况：原始值与修正后的正确值。当然属性值也可能不会被 修改，此时这两种情况下的取值相等。

对于每一条记录R, 将其在整个修复过程中的值域空间记为dom(R), 则其大 小为2R| 。我们用一个二进制的“01”数字串表示修复过程中的记录状态，假设记 录初始态为全“0”数字串，每一次对属性值进行修改对应于将这个串中的某一位 “0”改成“1”。

例如： 一条含有4个属性的问题记录t, 分别利用两条 Evidence -Rules 对第2 个和第3个属性值进行修改后得到了正确记录。则其修复过程用域空间内“01” 串可以表示为表10-14。

表10-14 数据修复的空间表示法举例

|  |  |  |
| --- | --- | --- |
| **No.** | **使用规则** | **状态** |
| 0 |  | 0000 |
| 1 | 91 | 0100 |
| 2 | 92 | 0110 |

根据上述表示法可知，对于同一记录由某一条规则作用前后的两个二进制表 示串，有且仅有一位存在“0”和“1”的取值差异。因此，如果将记录在修复过程中 的某一状态看作是一个点，整个修复过程应该是一条单向无回路的路径。模型分 析中一致性的三种特殊情况可以表述如下。

**情** **况** **1:**如果同时存在两条修复规则，对应着分别将属性组(A,B,C) 的取值 (tp(A),tp(B),p(C)) 修改成(p(A),tp'(B),p(C)) 和(tp(A),tp”(B),p(C)),

应当有tp'(B)=p”(B), 即这两条修复规则应当等价。对应于域空间内的修复路 径，从一个状态到达下一个状态只有一条路径。

情况2:不存在模式相同的两条规则，其中一对属性值互为彼此的问题值和参 考值。对应于域空间内的修复路径，不存在两个属性值有差异的最终态。

**情况3:**不存在模式相同的两条规则，其中一对属性值互为彼此的问题值和修 复值。对应于域空间内的修复路径，不存在任何形式的回路。

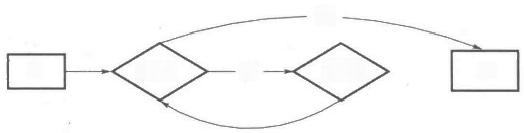
[**10.3.3.3**](10.3.3.3) **模型一致性检查**

为了确保问题数据修复的正确性，首先需要保证所给 Evidence-Rules 规则集 是一致的(Ebaid et al.,2013),即规则集本身不存在相互矛盾。因此需要采取一些 方法对规则之间的一致性进行校验，当发现不一致的规则时，需要采取相应的处理 方法消除其中的矛盾。

1)总体流程

对于一个给定的规则集，其一致性确认的主要过程如图10-4所示。

**250)**数据质量导论

是

步骤2

步骤1

否

∑

2”

图10-4 规则一致性检查流程一

步骤1:该步骤对于给定的规则集三进行一致性检查，如果证明其自身是一致 的则可以退出，返回规则集本身，否则交由步骤2处理。

步骤2:该步骤通过采取相应的自动算法或专家知识对步骤1中检测出的不 一致规则进行处理，处理完毕以后返回步骤1重新进行一致性检查。

为了确保最终得到的规则集能够用于问题数据检测和修复，需要确保一致性 检测过程的完备性，即能够发现其中所有的规则冲突。其次，对于检测到的有冲突 的规则，在对其进行处理时，应当选取合适的算法确保不会引入新的冲突。

2)检查算法

在进行一致性检查算法设计之前，先明确以下概念：根据一致性定义可知，对 于一个Evidence-Rules 规则集2,若三是一致的，其充分必要条件是对于 中的 任意两条规则φ;和φ,,它们也是一致的。

因此，根据上述思路，将对规则集三的一致性检查转化成对规则集中任意两条 规则φ之间的一致性检查。具体检查方案为：设n 是规则集 中规则个数，当n=1 时，显然规则集是一致的；当n=2 时，规则集 ∑是一致的等价于检查规则φ₁和φ₂ 是一致的；当n≥3 时，需要对规则集 ∑ 中的任意两条规则φ;和φ,(i≠j) 进行一致 性检查。

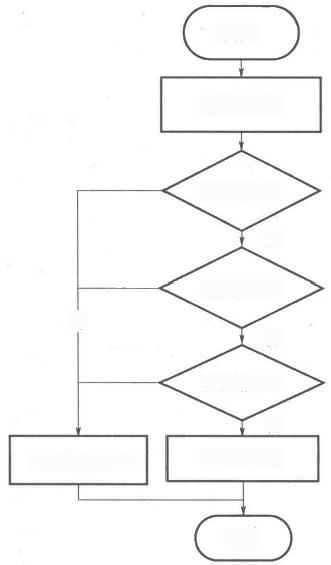
因此，对于规则集三的一致性检查主要针对其中的规则对<10.3.3.1>节中一致 性分析部分的三种特殊情况进行检查。主要检查过程如图10-5所示。

具体算法如表10-15所列。

表10- 15 规则一致性检验算法

|  |
| --- |
| **Name:一致性检验算法** |
| Input:一个给定规则集  Output:是否一致 |
| 1.Intf=1  2.If(caselCheck()and case2Check(2)and case3Check(2))  3. **Return** true  4**.Else** **Return** false;  5.**End** **if** |

第10章基于规则的不一致数据检测与修复方法(251

开始

输入规则集

是

规则集不一致

存在情况一

否

存在情况二

否

存在情况三

否

规则集一致

结束

图10-5 规则一致性检查流程二

进行规则校验之前，需要先统一规则的表示方式。这里采用矩阵的形式对规 则集进行存储，对于一个含有n 条规则的规则集，构造一个n×9 的矩阵，矩阵的每 一行对应A,(a,an₂,…,ag) 一条规则{(A:p(A),B: 如(B));(C:p(C));(D:

tp(D)→D:tp'(D))}, 每一列对应规则中一个元素的信息，具体情况如表10-16 所列。

表10- 16 规则的矩阵表示

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ml | an2 | an3 | m4 | am5 | am6 | ann | m8 | 9 |
| A | tp(A) | B | tp(B) | C | tp(C) | D | tp(D) | tp'(D) |

对于输入的规则集，分别利用表10- 17～表10- 19中的 caseCheck() 算法对其进行三种特殊情况的检查，同时对于检测到的相互冲突的规则，利 用outToText() 方法将其写入到预先定义的文档文件中，方便后续对其进行 处理。

**252)** 数据质量导论

表10- 17 caselCheck() 算法

|  |
| --- |
| **Name:caselCheck()** |
| **Input:**一个给定规则集 ∑  **Output:**是否一致 |
| 1.Int f=1;  2.For(i=0;i<n;i++)  3. **For**(j=i;j<n;j++)  4. **If**[aa,a₂,…,ag]==[a₁,az,…,ag]andag!=ap  5. f=0;outToText(A;,A;)  6. **End** **for**  7. **End** **for**  8.Iff==0 return false;  **9.Else** **Return** true |

表10- 18 case2Check() 算法

|  |
| --- |
| Name:case2Check() |
| **Input:**一个给定规则集  **Output:**是否一致信息 |
| 1.int f=1;  2.For(i=0;i<n;i++)  3.**For**(j=i;j<n;j++)  4. If[an,az,ag,da,ag,Q₆]==[a₁,ap,ag,ag,aβ,a₄]  5. f=0;outToText(A₁,A,);  6.If f==0 return false;  7.**Else** **Return** true |

表10- 19 case3Check() 算法

|  |
| --- |
| Name:case3Check() |
| **Input:**一个给定规则集  **Output:**是否一致信息 |
| 1.intf=1;  2.For(i=0;i<n;i++)  3.For(j=i;j<n;j++  4.If[aa,a₂,…,an,ag,g]==[a₁,az,…,aη,ap,ag]  5. f=0;outToText(A;,A₁)  6.Iff==0 **Return** false;  7.Else Return true; |

第10章 基于规则的不一致数据检测与修复方法(

[**10.3.3.4**](10.3.3.4) **冲突规则处理**

对于检查中发现的相互冲突的规则，需要采取一定的措施进行处理。由于规 则的种类多样，涉及到具体的领域知识，对于检测到的相互冲突的规则，通常需要 人为参与来确定如何对冲突规则进行取舍或修正来保持一致。

[**10.3.3.5**](10.3.3.5) **基于Evidence-Rules 的数据修复算法**

在完成对规则集的一致性校验之后，可将得到的 Evidence-Rules 规则集用于 问题数据的修复。为简化修复过程，避免规则在过程中的重复使用，需要对修复过 程中的相关变量进行标记。即要求每一条记录，每一条修复规则至多允许被使用 一次。这里采取以下策略：当一条规则φ被用于问题数据的修复，则将规则φ中 所涉及的属性进行标记，不允许做任何其他改动。因此，每进行一次属性值的修改 操作，需要完成以下两个步骤：①确定下一次修复可使用的规则集；②确定下一次 修复可修改的属性集。具体修复算法设计如表10-20所列。

表10-20 问题数据修复算法

|  |
| --- |
| **Name:问题数据修复算法** |
| Input:一个给定正确规则集乏， 一条问题记录R  Output:正确记录R' |
| 1.A:=φ;T=2;R'=R;updated:=true  2.While updated do  3.updated:=false;  4.For eachφ e ∑do  5.If R'matches g and Bφe A  6.R'[Bφ]:=tp+[Bφ];  7.A:=AUCpUBφ;T=T\φ;updated:=true;  8.Return R'; |

对于一条问题记录，依次对问题属性值进行修改操作，当找不到可以用来进一 步修复的规则或者没有可以继续被修改的属性值时，算法终止。认为此时得到的 记录即修复后的正确记录。

**10.3.4** **实验及分析**

为了检验上述方法的可行性与有效性，使用实例数据进行了实验。具体实验 内容如下。

[**10.3.4.1**](10.3.4.1) **数据来源**

实验采用的数据按以下模式生成：序号、国家、首都、城市、会议名称、主办单 位、会议级别。通过统计近年国内外召开的有关信息和数据质量方面的会议，选择

**254**)数据质量导论

具有函数依赖关系的字段并选择代表性的取值进行噪声引入，以此产生问题数据。

[**10.3.4.2**](10.3.4.2) **Evidence-Rules** **规则生成**

根据字段之间的函数依赖关系：“国家”字段的取值与“首都”字段取值互 相依赖；“会议名称”字段取值与“主办单位”字段的取值互相依赖。当两个互 相依赖的属性取值不满足对应关系时，通过“城市”字段的取值确定对“国家” 或“首都”字段的取值进行修改；通过“城市”字段的取值确定对“会议名称” 或“主办单位”字段的取值进行修改。由此确定基础的 Evidence -Rules 规 则集：

{(国家：p(国家),首都：p(首都));(城市：p(城市));(国家：tp(国家) → 国家：p' (国家))};

{(国家：p(国家),首都：p(首都));(城市：p(城市));(首都：p(首都) → 首都：p' (首都))};

{(会议名称：p(会议名称),主办单位：p(主办单位));(城市：p (城市)); (主办单位：p(主办单位)→主办单位：p' (主办单位))};

{(会议名称：p(会议名称),主办单位：p(主办单位));(城市：p (城市)); (会议名称：p(会议名称)→会议名称：p' (会议名称))};

对于上述每一条 Evidence-Rules 规则，通过改变规则实例中 Contradictory patterns tp(B)的取值不断得到新的 Evidence-Rules 规则。

例如对于记录(中国，北京，北京，SIGMOD 26TH,清华大学， 一级),由此产生 的规则可以为：

{(国家：中国，首都：上海);(城市：北京);(首都：上海→首都：北京)};

{(国家：美国，首都：北京);(城市：北京);(国家：美国→ 国家：中国)};

{(会议名称：SIGMOD26TH,主办单位：复旦大学);(城市：北京);(主办单 位：复旦大学→主办单位：清华大学)};

{(会议名称：VLDB26TH,主办单位：清华大学);(城市：北京);(会议名称： VLDB 26TH→会议名称；SIGMOD 26TH)};

根据统计数据中可能存在的其他错误，我们还可以定义其他字段的相关规则 来对问题数据进行修复操作。

[**10.3.4.3**](10.3.4.3) **结果分析**

通过对上述过程产生的Evidence-Rules 规则进行一致性检查，在得到较为完 备的规则集后，生成问题数据进行修复工作。对实验过程中算法发现的问题数据 以及修复后的数据进行记录，计算模型对于问题数据检测的查全率和查准率，以及 问题数据修复的查全率。实验结果如图10-6～图10-9所示。

当然，对于同一数据集的检测，其查准率和查全率的效果很大程度上取决于用 户设定的Evidence -Rules 规则集的规范性与完备性。

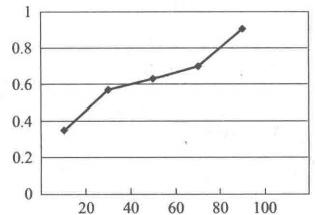
第10章 基于规则的不一致数据检测与修复方法(**255**

查准率 查全率

查全率

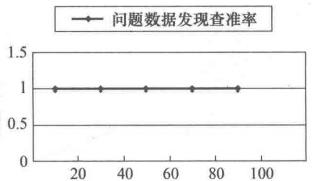
查准率

|  |
| --- |
| **→一问题数据发现查全率** |



规则数/条

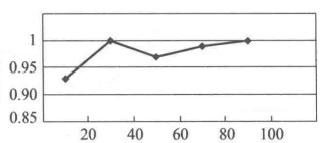
图10 - 6 Evidence-Rules 模型问题数据发现查全率



规则数/条

图10- 7 Evidence-Rules 模型问题数据发现查准率

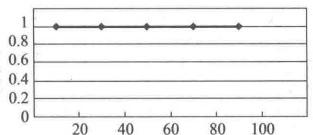
|  |
| --- |
| → 一问题数据修复查全率 |



规则数/条

图10 - 8 Evidence-Rules 模型问题数据修复查全率

|  |
| --- |
| **→—问题数据修复查准率** |



规则数/条

图10-9 Evidence-Rules 模型问题数据修复查准率

**256)**数据质量导论

查全率

为了进一步了解 Evidence-Rules 模型在对问题数据进行修复时比一般函数 依赖方法的改进效果，同时利用函数依赖关系规则对问题数据进行修复。其中，函 数依赖关系规则是“国家”和“首都”字段取值以及“会议名称”和“主办单位”字段 取值之间所对应的关系。分别利用关系表中的规则与 Evidence-Rules 规则集对 问题数据进行修复，在利用关系表中的规则进行问题数据修复时分别采用了下面 两种方法：①根据“国家”字段取值修改“首都”字段取值，根据“会议名称”字段取 值修改“主办单位”字段取值；②根据“首都”字段取值修改“国家”字段取值，根据 “主办单位”字段取值修改“会议名称”字段取值。其中实验结果记录如图10-10 所示。

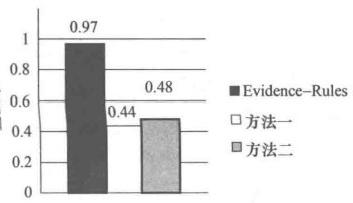


图10-10 不同问题数据修复方法效果比较

函数依赖关系可以很容易地找出不一致的属性对取值，较为全面地发现问题 数据。但是对于发生冲突的属性对，并不能决定修改哪一个属性的取值，方法一和 方法二所采取的统一的处理方法并不适合所有问题数据的情况，因此两种方法的 修复查全率都十分低；Evidence-Rules 模型通过参考其他字段的取值来选择需要 修改的冲突字段值，这样的修改更加合理，修复效果也更好。然而 Evidence - Rules 规则集产生时需要人为参与使得规则集尽可能地完备，否则对于问题数据的 部分特例不能很好地检测出，会产生漏检现象。

**10.4** **本章小结**

本章从规则自身的正确性角度考虑，对用于数据质量规则逻辑检查的 Fellegi - Holt 方法进行重点分析，在研究其基本理论的基础上，讨论方法的适用范围，对属 性值取值范围为连续区间情况以及连续区间内特殊值情况如何计算进行了分析， 并从集合运算的原理上分析其可行性；同时针对不同的问题数据检测需求设计了 面向问题数据发现与问题数据修复的两种检测策略。就如何利用数据质量规则进 行问题数据的修复进行了研究与实现，指出数据之间的函数依赖规则作为最常用 的问题数据检测与修复依据，其在问题数据的自动修复方面存在不足，因此设计了

第10章基于规则的不一致数据检测与修复方法 (2

基于 Evidence -Rules 的问题数据修复模型。该模型借鉴更多检测出的问题数据 中的信息，能够实现问题数据的自动修复，相比普通的函数依赖规则能够得到更加 准确合理的修复结果。

**参考文献**

[1]石少敏.2011.基于规则和数据学习的清洗模型研究[J]. 陕西教育学院学报，27(3);89-93.

[2]武森，冯小东，单志广.2012.基于不完备数据聚类的缺失数据填补方法[J]. 计算机学报，35(8):1726-173 8. [3]房强.2008.面向半结构化数据的数据质量控制系统的研究与实现[M]. 沈阳：东北大学.

[4]袁满，张雪.2013.一种基于规则的数据质量评价模型[J]. 计算机技术与发展，23(3):81-84. [5]郭志懋，周傲英.2002.数据质量和数据清洗研究综述[J]. 软件学报，13(11):2076-2082.

[6]程录庆.2011.数据约束对数据质量的影响研究[J]. 长江大学学报(自然科学版),8(5):100-102.

[7]Agnes B,Rajeev G,Markus H.2003.A Logical Formalization of the Fellegi-Holt Method of Data Cleaning [M]//Springer Berlin Heidelberg:Advances in Intelligent Data Analysis V:554-565.

[8]Beskales G,Ilyas I F,Golab L.2010.Sampling the Repairs of Functional Dependency Violations under Hard Constraints[J].Proceedings of the VLDB Endowment,3(1):197-207.

[9]Bohannon P,FanW,Flaster M,et al.2005.A Cost -based Model and Effective Heuristic for Repairing Constraints

by Value Modification[C]//SIGMOD Conference.Balimore,Maryland,USA:SIGMOD Conference:143-154. [10]Chomicki J,Marcinkowski J.2005.Minimal -change Integrity Maintenance Using Tuple Deletions [J].

Information&Computation,197(1-2):90-121.

[11]Cong G,Fan W,Geerts F,et al.2007.Improving Data Quality:Consistency and Accuracy[C]//Proceedings of the 33th International Conference on Very Large Data Bases,University of Vienna,Austria:VLDB:315-326.

[12]Ebaid A,Elmagarmid A K,llyas I F,et al.2013.NADEEF:A Generalized Data Cleaning System[J].Proceed- ings of the VLDB Endowment,6(12):1218-1221.

[13]Fan W.2008.Dependencies Revisited for Improving Data Qualiy[C]//Proceeding of the 27th ACM Sigmod -

Sigact -Sigart Symposium on Principles of Database Systems.Vancouver,BC,Canada;PODS,2008:159-170.

[14]Fan W,LiJ,MaS,et al.2011.Interaction Between Record Matching and Data Repairing[C]//SIGMOD.Ath- ens,Greece:SIGMOD Conference;469-480.

[15]Fan W,Li J,Ma S,et al.2012.Towards Certain Fixes with Editing Rules and Master Data[J].VLDB J,21 (2):213-238.

[16]FellegiI,Holt D.1976.A Systematic Approach to Automatic Edit and Imputation[J].Americ Statistics of Asso-

ciation:17 -35.

[17]Madnick S E,Wang R Y.2009.Overview and Framework for Data and Information Quality Research[J].ACM

Joumal of Data and Information Quality(JDIQ),1(1):1-22.

[18]Mangasarian O L,Wolberg W H.1990.Cancer Diagnosis Via Linear Programming[J].SIAM News,23(5):1-18.

[19]Peter Z Y,Colin A P.2010.An Efficient and Robust Approach for Discovering Data Quality Rules[C]//Pro-

ceeding of the 22th International Conference on Tools with Artificial Intelligence.Aras,France:ICTAI(1); 248-255.

[20]Stuart E M,Richard Y W.2009.Overview and Framework for Data and Information Quality Research[J].Jour-

nal of Data &Information Quality,1(1):1 -22.

**258)**数据质量导论

[21]William E W,Chen B C.2002.Extending the Fellegi-Holt Model of Statistical Data Editing[C]//SRD re- search report Statistics RRS2002/02,Washington D.C.USA:Researach Report Series.

[22]Zhang Z B,Zhou Y H,Liu Y Z.2011.Research of Error Data Detection Algorithm Based on Rules[C]//Com- munication Software and Networks(ICCSN),2011 IEEE 3th International Conference.Xi'an,China:IEEE:

159 -163.

**第11章** **数据质量工具**

**11.1** **引言**

当前，随着信息技术及网络应用的不断深入，信息资源的不断积累，信息和数 据质量问题日益突显，正在引起越来越广泛的关注。虽然大多数企业对数据质量 已经给予了足够的重视，但仍缺乏数据质量管理的具体手段。根据 TechTarget 商 务智能发起的2012年中国数据管理优先度调查显示，仍有18.8%的企业表示根 本没有任何的数据质量或者数据治理项目；在已经开展数据质量管理实践的企业 中，有42.2%的企业是采用手动编码方式进行数据质量管理；只有28.7%的企业 使用厂商提供的专业数据质量工具进行数据质量管理(TechTarget,2012) 。 本章 重点分析当前主要的数据质量工具厂商及其代表性数据质量工具软件，并提供了 两种数据质量工具具体实现方式，为各类组织机构的数据质量管理实施和同类产 品的研发提供参考。

**11.2** **数据质量工具发展概况**

本节基于Gartner 分析报告，对当前主流数据质量管理工具进行比较分析。

**11.2.1** **Gartner** **分析报告**

全球权威分析机构 Gartner公司每年都会根据前瞻性(Completeness of Vision) 和执行力(Ability to Execute)两方面15项衡量指标对厂商提供的数据质量工具进 行评估，将其评估结果分为四个级别：领导者(Leaders) 、挑战者(Challengers)、有 远见者(Visionaries) 和特定领域者(Niche Players)。

Gartner 认为，位列“领导者”象限的厂商在数据质量功能的各个方面，包括剖 析、解析、标准化、匹配、验证和扩充等，均展示出全方位的优势。它们对市场趋势 表现出清晰的理解与洞察力，能识别多域的数据质量问题和提供企业级数据质量 解决方案。此外，作为领导者的企业，在其本国市场上占有较强的地位和较高的市 场份额，同时在国际市场上也具备较强的竞争实力。

位列“挑战者”象限的厂商虽然可以提供强大的产品功能，但在功能广度方面与

260)数据质量导论

执行力

“领导者”相比要弱一些。例如，它们可能缺乏提供完整数据质量解决方案的能力。 “挑战者”企业已经建立了较强的地位、可信度和发展能力，但有可能仅局限在特定 的领域(如仅对客户名称和地址数据的清理),在领导和创新理念上没有突出表现。

位列“有远见者”象限的厂商对当前和未来市场趋势及方向，如对非传统部署 方式(像 SaaS 的重要性、对大数据集的支持、对业务问题解决专家的利用以及数据 质量服务的交付方式等),都具有深刻的理解。“有远见者”企业能够紧贴发展趋 势，但是缺少市场地位、品牌知名度、客户群以及大厂商的资源。

位列“特定领域者”象限的厂商具有有限的产品功能，缺乏快速开展功能领域 提升，如数据剖析、跨国支持的实力。另外，它们仅关注某一个特定的分块市场(如 中型企业)、有限的地理范围或者单独的一个领域(如客户数据),而没有定位到更宽 的应用。“特定领域者”企业可能具有一定的产品功能广度，但在市场地位方面还处 于发展的早期阶段，缺少客户群，并且资源有限，它们可能在其关注地理范围或数据 领域内具有功能齐全、性能强大的产品，并且能为客户提供实用的解决方案。

根据最新的 Gartner 2015 年度数据质量工具魔力象限(Magic Quadrant)报告， 除了有连续多年位于数据质量领导者之列的SAS、Informatica、Trillium Software、 IBM 和 SAP以外，Oracle 和 Information Builders 也跻身于领导者象限，Pitney Bowes Software 和 Experian则处在挑战者象限，特定领域者象限则包括 Uniserv、RedPoint、 Innovative Systems 、DataMentors 和 BackOffice Associates,而 Ataccama、Neopost、 Talend 及 MIOsoft 则被评定为有远见者。2015年度 Gartner 数据质量工具魔力象 限如图11- 1所示(Saul et al.,2015)。

|  |  |
| --- | --- |
| 挑战者  itney Bo  Expenan | 领导者  38  Tclmt  Builde |
| nnovative Systen  Red Pon  Data Mentors(  Back Offi Ass  特定领域者 | Talend  MIOsoft  有远见者 |

前瞻性

图11-1 2015年度 Gartner 数据质量工具魔力象限(截止2015年11月)

第11章 数据质量工具 

**11.2.2** **数据质量管理工具分析**

从数据质量工具核心功能需求来看，当前数据质量工具软件功能已到了相对 成熟阶段。但随着大数据和云服务等技术的发展，各厂商除了提供传统的本地部 署等方式外，还提供集群产品、SaaS 交付模式和云服务等形式的场外部署解决方 案。另外数据质量工具的应用场景也得到了拓宽，除了传统的事务处理、主数据管 理、商务智能和分析等应用场景以外，在信息/数据治理、大数据平台方面也在逐步 深入，能够使面向数据质量角色的人员，如数据管理员，完成数据质量改进工作。 随着数据质量软件需求的增大，各厂商推出的工具及产品众多，每款产品都具有各 自的特色和侧重点，下面将分析几个典型厂商的数据质量工具。

[**11.2.2.1**](11.2.2.1) **主要数据质量工具**

1)IBM 公司产品

IBM公司凭借其在 IT界的多年经验， 一直受到关注，其数据质量工具也一直 被公认为业界标准，可以和其他的IBM 产品集成于一体，应用于多数据域场景。 IBM 数据质量工具包括 InfoSphere Information Analyzer、InfoSphere QualityStage 和 InfoSphere DataStage 等 ，Information Analyzer 用于发现、归档并分析数据，Quali- tyStage 可进行数据标准化、合并和纠正，DataStage 用于转换和重构数据以适用于 新的用途。具体功能如下(IBM,2008):

(1)Information Analyzer:提供基于字段分析、表分析、交叉表分析以及基线分 析的数据质量评估、数据分析规则定义和数据质量监控功能。能够提供30余种数 据质量分析报表。

(2)QualityStage: 用于实现批量的或实时的数据标准化和清理，根据一定的规 则，将数据按照统一的格式进行标准化，然后对数据进行匹配，将不满足规则的或 者重复的数据进行去除。QualityStage 提供了一组用于数据再造任务的集成模块： 数据核查(Investigate)、数据标准化(Standardize)、数据匹配(Match) 和确定哪些数 据继续存在(Survive) 。另外 QualityStage 能够提供直观的用户界面简化数据质量 规则的设计。

(3)DataStage: 用于以批处理、实时或 Web service 方式实现从外部数据源集 成数据，支持数据校验规则合并、复杂的数据变换、元数据分析和维护等功能。

2)Oracle 公司产品

Oracle 公司进入数据质量软件市场相对较晚，但其通过一系列的关键收购策 略，扩充了其在数据质量方面的产品线。目前所属的产品有Oracle Enterprise Data Quality 和 Oracle Enterprise Data Quality for Product Data,可以提供数据分析、数据 清洗以及跨领域数据(如客户数据、产品数据)匹配等功能。

**262)**数据质量导论

Oracle Enterprise Data Quality 提供剖析和审计、解析和标准化以及匹配和合并 等功能(Oracle,2011), 其功能特点有：

(1)理解数据，建立数据质量规则，发现和量化数据库、数据表和平面文件中 存在的数据问题。

(2)系统性审计检测关键质量指标、缺失数据、重复记录和不一致数据。

(3)利用参考数据转化和标准化数据，如对姓名、地址、日期、电话号码等 数据。

(4)能够实现从无格式文本(Free-form Text)中抽取结构化信息。

(5)提供单个、部分和整体级别的匹配。

Oracle Enterprise Data Quality for Product Data 是针对特定问题提出的专门解 决方案，具有产品数据解析和标准化、产品数据匹配和合并等功能(Oracle,2010),

具体特点有：

(1)具有自动学习的语义识别能力，能够快速识别产品分类并且实施修正 规则。

(2)能够处理多个产品分类。

(3)提供对产品条目类别、属性进行抽取和标准化。

(4)提供准确、相似和有关联的匹配，依据已定义的去留原则合并记录。

3)SAS 公司产品

美国SAS 软件研究所(SAS Institute Inc.)自1976年成立起就开始进行统计分 析系统(Statistical Analysis System,SAS)的维护、开发、销售和培训，数据分析能力 强大。在数据质量工具方面SAS 公司凭借 DataFlux(SAS 子公司)产品已连续十年 位于领导者象限。

SAS 数据质量解决方案能够为数据管理人员和技术开发人员提供一套完整的 数据管理和数据质量工具，建立完整的数据质量生命周期管理。SAS 数据质量工 具具有多域数据处理能力，提供数据剖析、监控和流程策划等功能，以及SaaS 和云 部署方式，并且支持包括 Hadoop、Impala 和其他大数据系统在内的数据源(SAS, 2011)。SAS 数据质量解决方案由一系列SAS 产品和 DataFlux 产品组成，其中服务 器端组件包括 SAS Data Quality Server 和 Quality Knowledge Base,客户端组件有 dfPower Studio 、dfPower Match 、dfPower Profle 、dfPower Customize,其主要功能特点有：

(1)具有强大的、简单易用的图形用户界面：业务用户和数据管理员能够在友 好的 Windows环境下分析数据、定义业务规则、建立数据标准及数据匹配/集成规 范。其简单易用界面能够形象化地展现低劣数据所产生的影响，用户通过可视化 界面能够方便地定义可重复使用的数据质量改进流程。

(2)数据剖析功能：提供完整的分析企业数据的环境，确定数据中存在的细微 差别(Nuances) 、差 异(Discrepancies) 和误差(Inaccuracies), 评价企业的数据质量；

第11章数据质量工具(263

利用简单易用的接口和交互式报表机制能容易地确定低劣数据质量范围。

(3)数据匹配/标准化和清洗：对多个数据源中的数据，通过模糊逻辑聚合一 个域或多个域的相似值，建立唯一键值，实现数据合并，消除重复数据。

(4)客户化定制数据解析、标准化和匹配算法：支持个性化的数据解析、匹配 和标准化算法，能够创建或扩展用于解析姓名、地址、Email 地址、产品编码以及其 他业务数据的规则；在数据匹配算法中可以定制匹配规则权重；在SAS 产品(服务 器端)和 DataFlux产品(客户端)之间提供一个通用的质量知识库(Quality Knowl- edge Base),共享信息。

(5)国际支持：SAS 数据质量解决方案能够支持多种语言，包括在名字、地址 以及其他业务数据上的差别。能以多种语言提供正确的标准化数据，支持 Unicode 和双字节字符集数据。

4)Informatica 公司产品

Informatica 公司数据质量解决方案旨在帮助企业建立合适的数据质量计划， 为业务部门与IT 部门的相关人员分配明确的角色和责任，并配备合适的技术和工 具，以应对数据质量控制的挑战。从初始的数据探查到持续监测以及持续进行的 数据优化，业务部门与 IT 部门的数据使用者、业务分析师、数据管理员、IT 开发人 员和管理员，能够协同使用 Informatica 数据质量解决方案，并在整个企业的所有数 据领域和应用程序中嵌入数据质量控制。Informatica 数据质量解决方案中几个特 定用途的组件有：Informatica Data Explorer/Profiler、Informatica Data Quality 和 Infor- matica Identity Resolution(Informatica,2010)。

(1)Informatica Data Explorer/Profiler:通过基于角色的工具促进业务部门与 IT部门之间的协作，分析和发现多种数据来源的任何类型数据的内容、结构和缺 陷，监控数据质量问题。

(2)Informatica Data Quality:提供数据解析、清洗、匹配、报告、监控等功能，结 合可视化记分卡和仪表盘等可视化界面，支持在整个企业范围内实施和管理数据 质量计划。

(3)Informatica Identity Resolution:能使各机构从60 多个国家/地区以及各企 业和第三方应用程序中搜寻和匹配一致数据。

Informatica 公司数据质量解决方案提供基于角色的工具(Informatica Analyst、 Informatica Developer 和 Informatica Administrator),使得业务分析师和数据管理员、 IT 开发人员、IT管理员能同时针对业务部门和 IT 部门就相同数据提供不同视图。 三个基于角色的工具都是 Informatica Data Explorer/Profiler 和 Informatica Data Quality 的通用工具。其具体适用范围如下：

(1)Informatica Analyst:适用于业务分析师和数据管理员。通过用语义术语 表述数据，使分析师和数据管理员能够探查数据、创建和分析质量记分卡、管理异

264)数据质量导论

常记录、开发和使用规则，以及与 IT 部门展开协作。

(2)Informatica Developer:适用于IT 开发人员。基于 Eclipse 的开发环境允许 开发人员发现、访问、分析、探查和清洗处于任何位置的数据。开发人员可以对逻 辑数据对象建模、将数据质量规则与复杂转换逻辑合并，并在逻辑制定后，进行中 游探查以验证和调试逻辑。

(3)Informatica Administrator:适用于 IT 管理员。该工具为 IT 管理员带来集 中配置和管理的能力。管理员可以监测和管理安全性、用户访问、数据服务、网格 和高可用性配置。

5)Talend 公司产品

Talend公司是一家针对数据集成市场提供数据 ETL(Data Extraction,Transfor- mation and Loading,数据抽取、转换和加载)开源软件的供应商。其提供的开源软 件以套件的形式向外提供，数据质量工具是其中的一部分，包括Talend Open Studio for Data Quality 和Talend Enterprise Data Quality,提供的功能能够满足通用要求，支 持数据解析、标准化、匹配和数据剖析(Talend,2011) 。 其特点如下：

(1)数据解析，提供对当前数据质量的评估以及一段时间数据质量的测量，通 过数据质量门户展现数据质量处理过程中需要的关键信息。

(2)数据标准化和扩展，利用内部或外部参考数据、规范的表达式设置数值标 准和关于数据模型及大小的标准，然后通过集成的分解技术对数据进行结构解析， 达到数据质量的改进和提高。

(3)数据匹配，业务用户能够在 Talend 使用用户环境下配置匹配规则。

6)Data Cleaner

Data Cleaner 是一个开源的数据质量分析工具，用于管理和监测数据质量 (Human,2010) 。 其功能由两部分组成：

(1)数据剖析(Data Profling):针对源数据中数据概貌进行统计，包括标准度 量、数值分析、字符串分析、模式字符串匹配、值分布分析等功能。

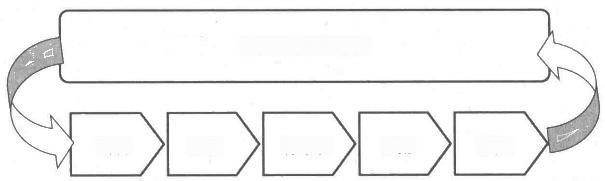
(2)数据验证(Data Validating):根据用户对数据理解定义出的数据规则进行 数据验证，以找出不满足数据规则的异常数据。其提供的数据验证包括：字段的非 空检验、字段值域检验、基于正则表达式的字段检验、基于脚本的复杂规则(如关 联字段)的检验。

[**11.2.2.2**](11.2.2.2) **数据质量工具核心功能及其基本流程**

通过以上典型数据质量工具的功能分析，可以看出全面的数据质量解决方案 应该把数据质量改进作为日常业务的一部分，涵盖数据质量剖析和监控以及一些 基本的数据清洗操作如解析、标准化、匹配以及数据扩充等，将这些核心功能集成 在一个数据管理平台中。在具体功能实现上通常还会提供一些特殊功能，如支持 广泛的姓名、地址格式，支持针对特定应用的知识库，支持以仪表盘方式显示数据

第11章 数据质量工具(265

质量指标等功能。另外还要综合考虑数据剖析及监控结果的可视化展现、简化规 则建立过程以及数据质量问题修正和跟踪等因素，更多地面向业务用户使用。数 据质量管理基本流程如图11-2所示。

审计、监控、报告

剖析 解析 标准化 匹配 扩充

图11-2 数据质量管理基本流程

**11.3** **基于表达式树的数据质量工具设计**

**11.3.1** **数据质量规则的分类与表达**

数据质量检查的目的是为了检查数据是否满足相关数据质量指标(准确性、 完整性、 一致性、有效性),对数据质量指标进行评价需要有具体的参照准则。这 些参照准则就是所谓的数据质量规则。

基于规则的数据质量检查，就是针对不同类型的数据问题进行专项检查。基 于规则的数据质量检查的一般思路是：针对数据集的自身数据特征以及用户需要 了解的有关待检查数据集的指标特性，从设定的数据质量规则中选取适当的规则 (例如，字段值是否为空、格式是否正确、字段间是否满足函数关系等)来进行检 查。对于同一数据集，检查结果的优劣，很大程度上取决于用于检查的数据质量规 则集是否足够全面与完善。自然，在进行有关数据质量检查的相关工作之前，需要 建立科学完备的规则集。因此，研究与讨论如何对种类繁多的数据质量规则进行 整理分类十分必要。

[**11.3.1.1**](11.3.1.1) **数据质量规则的定义**

数据质量规则，即数据约束，是客观世界的数据所应遵循的限制条件(程录 庆，2011;郭志懋，等，2002;袁满，等，2013)。不同层次上来讲，这种规则由表及里 又可以分成两个层次：“形”和“义”。所谓的“形”规则，是指数据在格式和语法上 应当满足的最基本的规范，例如数据库软件中对每个字段进行的字段类型和大小 的定义。所谓的“义”规则，对于商业系统中的数据来讲，就是领域知识和业务规 则。具体的业务规则限定了数据的属性值应符合其所反映现实的上下文。例如，

**266**)数据质量导论

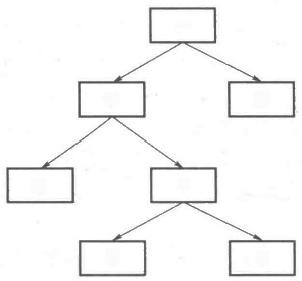
一个商店的客户关系数据库可能存在这样的规则：①一个新的客户在第一次购买 时享有85%的折扣；而一个VIP 客户在任何时候购买任何产品都享有75%的折 扣；②一个地址为中国的客户，“省”、“市”、“区”和“乡镇”字段确定“邮政编码”。 反映到数据库上来，就是对数据取值的约束。

现实中对此类数据约束的规则可能较为复杂，有些约束是在整个数据库范围 内成立的，有普遍性；有些是在某个局部数据集内成立；有些是需要满足特定条件 才成立；还可能存在规则之间的相互矛盾情况(如条件函数依赖自身的不一致)。 数据建模的一个重要任务是将数据库所描述的对象所提供的关于数据的上下文语 义表达成规范而系统的数据约束(梁吉胜，等，2012)。目前，这种数据约束的表达 形式主要是数据依赖，随着研究的深入，将会出现更多形式的数据约束表达。

[**11.3.1.2**](11.3.1.2) **质量规则表达式树**

表达式树是遗传编程领域中的一个概念，遗传编程(Genetic Programming)又 称基因编程，是一种从生物进化过程得到灵感的自动化生成和选择计算机程序来 完成用户定义的任务的技术。1989年，美国斯坦福大学的Koza教授发展了遗传 编程的概念，其基本思想是：采用树形结构表示计算机程序，运用遗传算法的思想， 通过自动生成计算机程序来解决问题。从理论上讲，人类用遗传编程只需要告诉 计算机“需要完成什么”,而不用告诉它“如何去完成”,最终实现真正意义上的人 工智能，即自动化的发明机器。

遗传编程的个体表示法可根据其基本结构大致分为三种：树形的、线性的和基 于图形的结构。其中最常见的是树形结构。这种结构系统中的每个个体都是一个 以树形结构来表示的程序(或表达式),个体具有分层结构。树中的结点可分为两 类，其中位于内部的结点称为“函数”(通常是数学运算符),而位于端点的叶结点 称为“终止符”(通常为函数的运算对象)。 一个一般函数式a+b\*c-d 的表达式 树描述结果如图11-3所示。



+ d

a \*

b c

图11-3 表达式树示例

第11章 数据质量工具 (2

[**11.3.1.3**](11.3.1.3) **规则的分类**

由于数据可能出现的质量问题多种多样，数据质量的评价需要尽可能从多个 方面发现数据中存在的问题。因此，数据质量规则也种类繁多，袁满等(2013)对 常用的15类数据质量约束规则定义如表11-1所列。

表11 - 1 数据质量约束规则描述

|  |  |  |
| --- | --- | --- |
| **No** | **名称** | **描述** |
| 1 | 非空约束规则 | 描述为该数据项上的数据不允许出现空值 |
| 2 | 连续性约束规则 | 描述为该数据项上的数据必须满足一定的连续取值 |
| 3 | 完整性约束规则 | 描述为该数据项上的数据必须满足完整性要求 |
| 4 | 历史性约束规则 | 描述为该数据项上的数据必须满足历史性要求 |
| 5 | 代码约束规则 | 描述为该数据项上的数据必须满足特定的代码规范 |
| 6 | 词法约束规则 | 描述为该数据项上的数据必须满足特定的词法要求 |
| 7 | 值域约束规则 | 描述为该数据项上的数据必须满足特定的取值规则 |
| 8 | 逻辑依赖约束规则 | 描述为该数据项上的数据必须与其他数据项上的数据满足某种 逻辑关系(如大于、小于等) |
| 9 | 及时性约束规则 | 描述为该数据项上的数据必须满足及时性要求 |
| 10 | 冗余性约束规则 | 描述为该数据项上的数据必须满足冗余性要求 |
| 11 | 等值函数依赖约束规则 | 描述为该数据项上的数据必须通过同一表中其他数据项上的数 据计算得来 |
| 12 | 等值一致性依赖约束规则 | 描述为该数据项上的数据必须通过不同表中其他数据项上的数 据计算得来 |
| 13 | 存在一致性依赖约束规则 | 描述为该数据项上的数据必须出现在其他表中的某一数据项中 |
| 14 | 逻辑一致性依赖约束规则 | 描述为该数据项上的数据与其他表中的数据项满足某种逻辑关 系(如大于、小于等) |
| 15 | 关联性约束规则 | 描述为该数据项上的数据必须满足关联性要求 |

既然数据质量规则是进行数据质量检查时的参考依据，其在某一个角度上反 映了数据是否满足某种指标特征，那么一种很自然的想法是根据数据质量的评价 指标对规则进行分类。房强(2008)和石少敏(2011)对相关的约束规则在完整性、 一致性、准确性、时效性等四个维度上进行了分类，如表11-2所列。

**268)**数据质量导论

表11- 2 数据质量约束规则分类

|  |  |
| --- | --- |
| **维度** | **具体规则** |
| 完整性约束 | 1.主键约束  2.联合主键约束  3.唯一约束  4.非空约束  5.候选码约束 |
| 一致性约束 | 6.外键约束  7.代码约束  8.等值依赖约束  9.逻辑依赖约束  10.等值一致性约束  11.逻辑一致性约束  12.存在一致性约束 |
| 准确性约束 | 13.数据类型约束  14.数据格式约束  15.数据精度约束  16.值域约束  17.固定值约束 |
| 时效性约束 | 18.时间段约束 |

上述分类方法以数据质量的评价指标为标准，这样的分类在理解上十分自然， 可是在操作上，因为每一指标下的规则在其适用对象和使用方法上相对而言比较 杂乱，并没有很好地反映出规则本身的特征。因此从实现的角度，考虑另一种分类 方法，该方法以规则的适用对象和规则结构特点为基础。

通常，数据质量规则检查的是存储在数据库中的结构化数据。由于字段是能 够检查的最小单元，因此有以下几类数据质量检查对象：①单个字段；②单个字段 与某个常数值；③两个字段；④多个字段；⑤一整条记录。具体分类结果如表11-3 所列。

表11- 3 数据质量规则结构分类

|  |  |
| --- | --- |
| **规则对象** | **具体规则类型** |
| **单个字段** | **1.主键约束**  **2.唯一约束**  **3.非空约束**  **4.数据类型约束**  **5.数据格式约束**  **6.数据精度约束** |

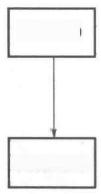
第11章 数据质量工具 

(续)

|  |  |
| --- | --- |
| **规则对象** | **具体规则类型** |
| 单个字段与某个常数值 | 7.值域约束  8.固定值约束 |
| 两个字段 | 9.简单逻辑依赖约束  10.等值依赖约束 |
| 多个字段 | 11.函数依赖约束 |
| 一整条记录 | 12.相似重复记录(曹建军，等，2010) |

[**11.3.1.4**](11.3.1.4) **规则的表达**

根据上述数据质量规则的分类方法，结合表达式树的结构

特点，提出采用一种称为质量规则表达式树的基本结构对数据 Method0

质量规则进行形式化描述(Koza et al.,1999)。根据数据质量

检查对象的五个基本分类，提出下列4 种树形结构，如

图11-4～图11-7所示。

1)单叶子树 Field.value

这种树具有以下结构特点：单叶子树由一个父节点和一个

叶子节点组成，主要适用于对单字段进行检查的规则。其中叶

图11-4 单叶

子树结构图

子节点用来表示待检查的字段，父节点表示用来对字段进行检

查的方法或函数名。

例如：

规则1:检查数据表中“姓名”字段取值是否为空值。

此时，Method() 方法表示内容为IsNull() 函数，Field.value 表示内容为待检查 的字段名称“姓名”。

检查执行时，程序根据规则内容，获取当前连接数据库表中记录的姓名字段 值，并调用IsNull() 函数对字段取值是否为空进行检查，当取值为空时，则将该条 记录作为问题数据记录到检查结果中。

2)叶子树

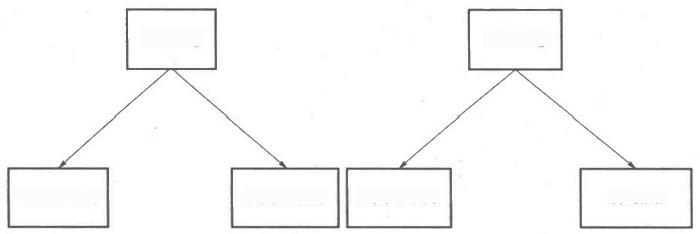
这种树具有以下结构特点：双叶子树由一个父节点和两个叶子节点组成，主要 适用于检查规则中涉及两个字段属性值或者一个字段值与一个常量之间的比较。 其中两个叶子节点分别用来表示字段名或常量，父节点表示用来对字段进行检查 的方法或函数名。

例如：

规则2:检查数据表中“年龄”字段取值是否大于“25”。

此时，Method() 方法表示内容为IsMoreThan() 函 数 ，Field1.value 表示内容为

270)数据质量导论



Method)

Field1.value constant

Field2.value

Field1.value

Method)

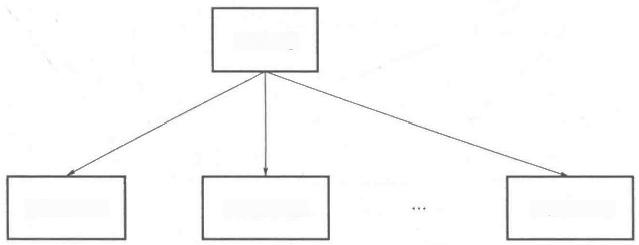
图11 - 5 双叶子树结构图

待检查的字段名称“年龄”,Field2.value 表示内容为待检查的常数值“25”。

检查执行时，程序根据规则内容，获取当前连接数据库表中记录的年龄字段 值，并调用IsMoreThan()函数对字段取值是否大于“25”进行检查，当年龄字段取 值小于“25”时，则将该条记录作为问题数据记录到检查结果中。

3)多叶子树

这种树具有以下结构特点：多叶子树由一个父节点和多个叶子节点组成，主要 适用于对多个字段进行检查的规则。其中叶子节点用来存储待检查的多个字段， 父节点存储用来对字段进行检查的方法或函数名。

Method0

Field1.value

Field2.value

Fieldn.value

图11 - 6 多叶子树结构图

例如：

规则3:检查数据表中“合格率”字段取值是否等于“合格人数”字段取值除以 “总人数”字段取值。

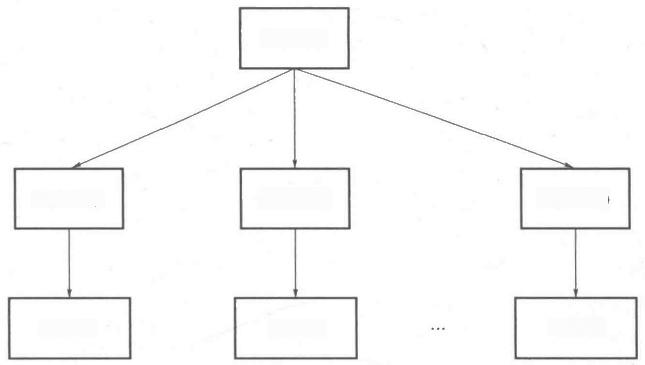
此时，Method() 方法表示内容为 FunctionDependence()函数，该函数包含五个 参数，对应于函数计算涉及到的字段以及数学符号。Field1.value 、Field2.value、 Field3.value 表示内容分别为待检查的字段名称“合格率”“合格人数”“总人数”。

检查执行时，程序根据规则内容，获取当前连接数据库表中记录的“合格率” “合格人数”“总人数”字段值，并调用FunctionDependence() 函 数 验 证 各 字 段 是 否

第11章数据质量工具(27)

满足对应的函数关系，并将不满足关系的记录作为问题数据记录到检查结果中。 4)多分支叶子树

这种树具有以下结构特点：整个树结构分为三层，其中叶子节点存储待检查的 各个字段名称，中间层存储信息为各个字段的相似度检查方法名称，根节点存储信 息为记录相似度的计算方法名称。



Method)

Methodn)

Fldn.val

Method2()

Method10

Fld2.val

Fld1.val

图11-7 多分支叶子树结构图

例如：

规则4:检查学生信息数据表中的相似重复记录。

此时，Field1.value,Field2.value,…,Fieldn.value 表示内容分别为待检查的字 段名称“姓名”“年龄”“性别”等，Method1(),Method2(),…,Methodn() 分别是根 据待检查字段的类型(数值型、字符串型、枚举型……)所选取的字段相似度计算 函数，Method()是计算整条记录相似度的方法名(平均值法、加权平均值法、欧氏 距离等)。

检查执行时，程序根据规则内容，获取当前连接数据库表中两条记录对应各个 字段的属性值，根据对应的字段类型调用各个字段的相似度计算方法进行计算，最 后综合计算得到整条记录的相似度。系统可以由用户自己设置判断两条记录是否 相似的相似度阈值区间，根据计算得到的结果返回这两条记录是否相似的结论。

**11.3.2** **数据质量规则的存储与识别**

根据数据质量规则的结构化特点对其进行分类描述，为了更好地将其应用于

数据质量的检查，需要考虑规则的存储与识别问题。考虑到不同类别数据质量规

272)数据质量导论

则的结构特点，分别设计了相应的XML文档对其进行存储并采用XQuery 语言对 其进行识别。

[**11.3.2.1**](11.3.2.1) **XML 与** **XQuery 简介**

XML的全称是 eXtensible Markup Language,中文为可扩展的标记语言(孟小 峰，2009)。自全球信息网协会W3C 于1998年2月通过XML1.0 规范的审核后， XML已经成为继 HTML、JAVA 之后在 Internet 上最受广泛关注的技术。

XML是一种标记语言，所谓标记语言(Markup Language)是由一些码或控制标 记所组成的，这些码或控制标记本身若单独存在是无意义的，它们必须结合文件信 息后才能形成一份有用的电子文件。目前所使用的标记语言可分为一般通用的标 记语言和特殊用途的标记语言两类。XML由称为“实体”的存储单元组成，每个实 体包含文本或者二进制数据，但不能同时存在。文本数据由字符组成，二进制数据 用于存储图片和小程序等类内容。

考虑到XML语言良好的结构规范性以及规则的结构化特点，利用XML文件 对规则所适用对象、规则类型、规则检查方法等进行结构化的存储，方便机器识别。

XQuery 是一种可以查询结构化或半结构化 XML数据的语言(廖湖声，2013)。 现有的大部分数据库引擎中都提供XML 数据类型支持，因此可以将文档存储在数 据库中，然后使用XQuery进行查询。XQuery 基于现有的XPath 查询语言，并支持 更好的迭代、更好的排序结果以及构造必需的XML的功能。XQuery 在 XQuery 数 据模型上运行。此模型是 XML 文档以及可能为类型化也可能为非类型化的 XQuery 结果的抽象概念。类型信息基于 W3C XML架构语言所提供的类型。

XQuery 语言是在 XML 数据中搜索特定结构信息的功能非常强大的一种查询 语言。它是由国际化的标准组织W3C 提出的。它的主要查询语句是 FLWR(For-

Let-Where-Return) 语句。FLWR描述了典型XQuery 查询语言的结构(廖湖声， 2013)。其中描述了哪些节点在满足哪些条件时生成什么格式的 XML 文档。Let 语句可以将数据绑定到变量，以供后续步骤使用。例如，下面例子中的XQuery语 句可以查询出版商为 Addison-Wesley 出版社的所有书籍，如图11-8所示。

For $book in document("<http://www.bn.com/bib.xml>")/book

Let Stitle :=Sbook/title

Where Sbook/publisher='Addison-Wesley'

Return <bookInfo>

{$title }

</bookInfo>

图11-8 XQuery 举例

第11章 数据质量工具 

从上例中可以看出XQuery 查询语言很复杂，而且查询的时候需要知道 XML 文件的结构信息，如 publisher、title 等标签，不适合于普通用户学习和使用。而本 章所讨论的XML文档存储内容为规则表达式树的相关信息，其标签内容以及嵌套 层次相对简单固定，因此能够对规则有效识别。

本节的设计思路是，首先对有关数据质量规则进行统一的结构设计，利用 XML实现对规则的结构化存储。然后，利用XQuery查询语言，对用户设定的规则 进行识别提取，同时在进行算法编写过程中定义良好的接口对规则进行算法实现， 以对数据进行检查。

[**11.3.2.2**](11.3.2.2) **规则的存储与识别**

将数据质量规则按照其适用对象和规则结构特点进行分类，采用数据质量规 则树的形式对其进行表达，按照<11.3.1.4>中提出的四种树形结构，分别设计了以 下四种XML文档模板，用来存储对应结构的数据质量规则。

1)单叶子树结构模板

单叶子树结构由一对父子节点构成，分别表示规则对应的函数名以及规则的 作用对象，因此其XML文件中，除基本的数据库表信息之外，规则部分标签含有两 层嵌套，元素<function> 和 <terminal> 分别表示了规则处理的函数名以及检查对 象。对应的XML文件的存储格式如图11-9所示。

|  |
| --- |
| <datasource id=“studentdb”type="rdb"name="学生信息库">  <darule id=“rule001”rule\_type="NullCheck"rule\_name="空值检查">  <ref\_object object\_name="STUDENT"type="rtable">  <function type="function"name="IsNull">  <terminal type=“varible"datatype="String"name="姓名”>  </function>  </dqrule>  </datasource> |

图11-9 单叶子树结构模板

上述模板中存储的规则含义为：检查“学生信息库”表中的“姓名”字段取值是 否为空。

2)双叶子树结构模板

双叶子树结构由一个父节点和一对子节点构成，分别表示规则对应的函数名 以及规则作用的两个对象，因此其XML文件中，除基本的数据库表信息之外，规则 部分标签含有两层嵌套，元素<function> 和 <terminal> 分别表示了规则处理的函 数名以及检查对象。与单叶子树模板不同的是，最里层含有两个<terminal> 标

274)数据质量导论

签。对应的XML文件的存储格式如图11-10所示。

<datasource id="studentdb"type="rdb”name=“学生信息库">

<dqrule id="rule002"rule\_type="LogicalCheck"rule\_name="逻辑检查">

<ref\_object object\_name="STUDENT"type="rtable">

<function type="function"name="BiggerThan">

<terminal type="varible"datatype="Int"name=“年龄”>

<terminal type="constant"datatype-"Int"name="15">

</function>

</dqrule>

</datasource>

图11 - 10 双叶子树结构模板

上述模板中存储的规则含义为：检查“学生信息库”表中的“年龄”字段取值是 否大于“15”。

3)多叶子树结构模板

多叶子树结构由一个父节点和多个子节点构成，分别表示规则对应的函数名 以及规则的作用对象，因此其XML文件中，除基本的数据库表信息之外，规则部分 标签含有两层嵌套，元素<function> 和<terminal> 分别表示了规则处理的函数名 以及检查对象，且最里层含有多个 <terminal> 标签。对应的XML文件的存储格 式如图11-11所示。

|  |
| --- |
| <datasource id="studentdb”type=“rdb”name=“学生信息库">  <dqrule id="rule002"rule\_type="FunDependencyCheck"rule\_name=“函数依赖检查”>  <ref\_object object\_name="STUDENT"type="rtable">  <function type="function"name="FunctionDependency">  <terminal type="varible"datatype="Float"name-“合格率"人  <terminal type="varible"datatype="Int"name="合格人数">  <terminal type="varible"datatype="Int"name=“总人数”>  </function>  </dqrule>  </datasource> |

图11 - 11 多叶子树结构模板

上述模板中存储的规则含义为：检查“学生信息库”表中的“合格率”字段取值 是否等于“合格人数”与“总人数”字段取值的比值。

第11章数据质量工具(275)

4)多分支叶子树结构模板

多分支叶子树有三层结构，其中根节点的每个子树都是一个单叶子树。根节 点表示相似重复记录检测中记录相似度的计算方法，每一个子树表示一个待处理 字段名称以及相似度的计算方法。因此其XML文件中，除基本的数据库表信息之 外，规则部分标签含有三层嵌套，最外层的元素<function> 存储记录相似度的计 算方法，内层含有多个并列的<function> 和 <terminal> 的结构，分别表示了各个 字段名称以及对应的相似度计算方法。对应的XML 文件的存储格式如图11-12 所示。

//记录相似度方法为平均值法

<datasource id=“studentdb"type="rdb”name=“学生信息库”>

<dqrule id="rule003"rule\_type-"XSDCheck"rule\_name-“相似度检查”>

<ref\_object object\_name="STUDENT""type="rtable">

<function type="Redfunction"name="AverageXSD">

<function type="fldfunction"name="StringfldXSD">

<terminal type="varible"datatype="String"name="姓名.value">

</function>

<function type="fldfunction"name="NumberfldXSD">

<terminal type="varible"datatype="Int"name=“年龄.value">

</function>

<function type="fldfunction"name-"EnufldXSD">

<terminal type="varible"datatype="Enumeration"name="性别.value"人 </function>

</function>

</dqrule>

</datasource>

图11-12 多分支叶子树结构模板一

XSD=AverageXSD(StringfldXSD(A.value,A'.value),NumberfldXSD(A.value, A'.value),EnufldXSD(C.value,C'.value))。

上述模板中存储的规则含义为：检查“学生信息库”表中的相似重复记录，其 中选取的三个字段为“姓名”“年龄”和“性别”;对应的字段相似度检查方法分别 为“字符串型”“数字型”“枚举型”;记录相似度的计算方法为平均值法。对应的 XML文件的存储格式如图11-13所示。

XSD =weightAverageXSD(StringfldXSD(A.value,A'.value),NumberfldXSD (A.value,A'.value),EnufldXSD(C.value,C'.value),0.3,0.3,0.4)。

上述模板中存储的规则含义为：检查“学生信息库”表中的相似重复记录，其 中选取的三个字段为“姓名”“年龄”“性别”;对应的字段相似度检查方法分别为

**276)**数据质量导论

//记录相似度方法为加权平均值法

<datasource id="studentdb”type-“rdb”name=“学生信息库">

<dqrule id="rule003"rule\_type="XSDCheck"rule\_name="相似度检查">

<ref\_object object\_name-"STUDENT"type="rtable">

<function type="Redfunction"name-"weightAverageXSD">

<function type="fldfunction"name="StringfldXSD"weight="0.3">

<terminal type="varible"datatype="String"name="姓名.value"/>

</function>

<function type="fldfunction"name="NumberfldXSD"weight="0.3">

<terminal type="varible"datatype="Int"name="年龄.value"D

</function>

<function type="fldfunction"name="EnufldXSD"weight="0.4">

<terminal type="varible"datatype="Enumeration"name="枚举.value"人

</function>

</function>

<dqrule>

</datasource>

图11-13 多分支叶子树结构模板二

“字符串型”“数字型”“枚举型”;记录相似度的计算方法为加权平均值法(三个字 段的权重分别为：0.3、0.3、0.4)。

文档对象模型(Document Object Model,DOM)是一种与浏览器、平台、语言无 关的接口，其给开发者提供一个标准的方法，让开发者可以访问相关的数据、脚本 和表现层对象。

XML文档是基于树形层次结构的，DOM 提供了一个 API,允许开发人员添加、 编辑、移动、删除树中的任意位置的节点，利用 Element 接口自带的 selectSingleNo- de() 方法可以很方便地查询出XML文档中相关节点的信息。

**11.4** **基于流程的数据质量工具设计**

数据 ETL在数据仓库、数据迁移等领域扮演着至关重要的角色，如果没有数 据 ETL 工具，数据管理人员就需要编写大量特定于数据存储的脚本进行数据处 理。设计实现一种通用的数据 ETL工具，提供可配置的数据处理功能，可以避免 重复编写数据处理脚本，显著提高数据管理员的工作效率。设计数据 ETL工具的 关键在于充分考虑异构性，提供灵活方案，针对不同的数据源和数据处理环节，为

第11章 数据质量工具(

用户提供统一的视图。

数据ETL工具应具备以下特点：

(1)灵活：能够表达复杂的数据处理逻辑。

(2)高效：能够应对大数据量的数据ETL。

(3)可扩展：能够根据需要增加新的数据处理功能。

(4)可管理：能够对数据处理流程进行静态和动态管理。

下面介绍一种数据 ETL模型，包括数据模型、作业模型、执行方案模型3个部 分。数据模型屏蔽不同数据源的差异，提供统一的数据表示；作业模型定义数据 ETL过程表示，以及数据 ETL过程的执行；执行方案模型以批处理方式组织和调 度作业的执行。

**11.4.1** **数据模型**

数据ETL过程需要从异构数据源抽取、加载数据，因此，需要统一的数据组织 形式，以屏蔽不同数据源的差异。数据模型描述了表示和组织数据的方式，如图 11-14所示。

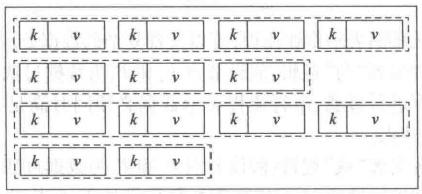


图11-14 数据模型

图11-14中的数据模型，包括元素、元组和视图三个部分：

(1)数据元素：键值对形式的最小数据单元；

(2)数据元组：多个元素组成的序列；

(3)数据视图：多个元组组成的序列。

数据视图用于封装数据。尽管数据视图与二维关系表存在一定的相似 性，但数据视图没有约束，数据视图中的数据元组可以是异构的，不同的数据 元组可以包含不同数目的数据元素，数据元素的键以及数据元素的值的类型 也可以不同。数据处理过程根据键获取数据元素值，并可以根据定义解释数 据视图所封装的数据：如果不存在相应的键值对，就认为该键对应的值为空； 如果数据元素值不是所期望的类型，就进行类型转换，或者抛出异常。

**278)**数据质量导论

**11.4.2** **作业模型**

作业是数据 ETL过程的具体表现形式，作业模型提供数据 ETL过程的定义方 法，并负责数据 ETL过程定义到线程与内存对象的转换。作业模型包括作业表示 模型和作业转换模型两个部分。

[**11.4.2.1**](11.4.2.1) **作业表示模型**

作业表示模型通过一个有向无环图将作业定义为一个流程，每个数据的抽取、 转换、加载操作都作为流程中的一个步骤。作业表示模型包括图、节点和有向边三 个部分：

(1)节点：表示具体的数据抽取、转换和加载步骤。

(2)有向边：表示数据在节点之间的流向。

(3)图：表示由节点和有向边所组成的作业流程。

其中，有向边又分为NORMAL 边、TRUE 边和 FALSE 边 ，NORMAL 边表示数 据无差别地从前置节点流向后续节点，TRUE 边表示满足前置节点所定义约束的 数据流向后续节点，FALSE 边表示不满足前置节点所定义约束的数据流向后续 节点。

通过以有向无环图表示作业流程，可以实现复杂的数据处理逻辑：

(1)串行流程实现“与”逻辑：假设节点 F₁和 F₂ 为数据过滤节点，可以根据数 据过滤条件对数据进行过滤，串行流程表示数据必须同时满足F₁ 和 F₂ 的过滤条 件，才能进入下一节点。

(2)并行流程实现“或”逻辑：假设节点F₁ 和 F₂ 为数据过滤节点，可以根据数 据过滤条件对数据进行过滤，并行流程表示数据满足F₁或者 F₂的过滤条件，即可 进入下一节点。

(3)流程分支实现“条件”逻辑：假设节点F 为数据过滤节点，可以根据数据 过滤条件对数据进行过滤，流程分支表示当数据满足F 的过滤条件时进入TRUE 边所连接的节点，当数据不满足F 的过滤条件时进入FALSE边所连接的节点。

[**11.4.2.2**](11.4.2.2) **作业转换模型**

作业转换模型负责将作业定义转化为线程以及内存对象，包括任务线程、交换 区和作业三个部分。

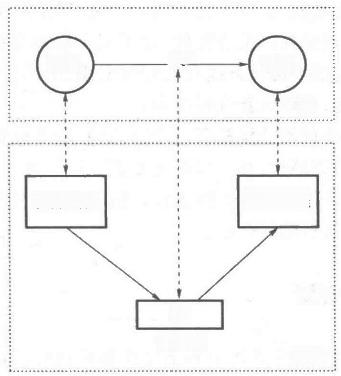
(1)任务线程：作为独立线程执行具体数据抽取、转换和加载操作。

(2)交换区：两个任务共享的内存区域，采用生产者—消费者模式， 一个任务 向交换区中写入数据，另一个任务从交换区读取数据。

(3)作业：包括多个任务和多个交换区，负责根据作业定义实例化任务和交换 区，并负责管理任务和交换区资源。

第11章数据质量工具(279

作业表示模型与作业转换模型之间的映射如图11- 15所示。



节点

图

任务线程

O

交换区

作业

任务线程

节点

边\_

图11-15 作业转换模型

作业表示模型中的节点被映射到任务线程，作业表示模型中的有向边被映射 到交换区，从而实现将作业定义转换为线程和内存对象。任务线程拥有 NORMAL 交换区列表、TRUE交换区列表以及FALSE 交换区列表，分别对应于作业表示模型 中的 NORMAL边、TRUE 边以及FALSE边。

从为作业分配资源，到作业资源被回收，作业生命周期包括初始化、启动、运 行、停止四个阶段。

1)作业初始化

根据作业定义，利用反射机制实例化任务对象，并建立交换区。

2)作业启动

从线程池中获取空闲线程，每个任务作为独立线程运行。

3)作业运行

数据抽取任务线程开始从数据源分页读取数据，并将数据写入输出交换区，而 其他任务线程则等待输入交换区中的数据，当从输入交换区读取数据后即进行处 理，最后将处理结果写入输出交换区。重复上述过程，直至数据抽取任务线程结 束，后续任务线程依次结束。

4)作业停止

作业的停止分为自动停止和强制停止两种情形。

(1)自动停止：首先，当任务线程执行结束或者因运行异常结束，就会将所有 与该任务线程相关的交换区置为无效状态；当某个任务线程从输入交换区读取数 据时，首先检查输入交换区是否有效，如果某个输入交换区无效，则表明与该输入

**280** )数据质量导论

交换区所对应的上游任务线程已经结束，如果某个任务线程已经执行完毕，且其所 有输入线程均已处于无效状态，则该任务线程结束；当任务线程将数据写入输出交 换区时，同样先检查输出交换区是否有效，如果某个输出交换区处于无效状态，则 表明该输出交换区对应的下游任务线程已经异常退出，当任务线程所有的输出线 程均已处于无效状态时，则该任务线程结束。

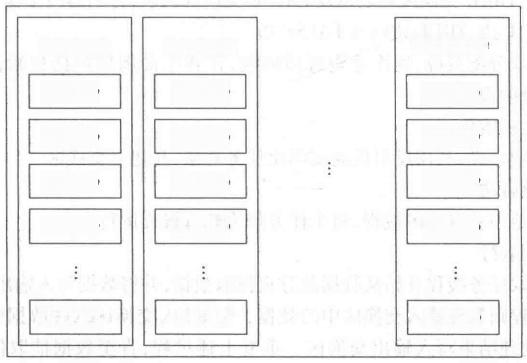
(2)强制停止：每个任务线程都有一个标志位，用于标识是否停止，任务线程 进入每次循环之前会检查该标志位，以确定是否立即停止。当需要强制终止作业 执行时，只需要设置每个任务线程的标志位，当任务线程检测到状态发生改变时， 就会退出循环，该任务线程随即停止。

**11.4.3 执行方案模型**

执行方案模型以批处理方式组织和调度作业的执行，包括执行方案表示和执 行方案调度两部分。

[**11.4.3.1**](11.4.3.1) **执行方案表示**

执行方案是由多个作业组成的作业序列，是作业执行引擎接收并执行的基本 单元。如图11-16所示为执行方案组织作业的示意图。



Solution\_m

Job\_n,p

Job\_n,p

Job\_n,P

Job\_n,β

Job\_n,p

Solution\_2

Job\_n,p

Job\_n,P

Job\_n,p

Job\_n,p

Solution\_1

Job\_n,p

Job\_n,p

Job\_n,p

Job\_n,p

Job\_n,p

Job\_n,p

图11-16 执行方案组织

每个执行方案中包括一组作业，且作业具有优先级，优先级高的作业先执行， 优先级低的作业后执行，且只有当优先级高的作业执行完毕后，优先级低的作业才 开始执行，而优先级相同的作业可以并发执行。执行引擎在接收到执行方案定义 时，先按照优先级从高到底对执行方案中的作业进行排序，并将执行方案定义添加

第11章数据质量工具(**281**

到列表中。 一旦作业执行完毕，就从执行方案定义中标记该作业为执行完毕； 一旦 执行方案中的所有作业都执行完毕，就从执行方案列表中移除该执行方案定义。

[**11.4.3.2**](11.4.3.2) **执行方案调度**

对于采用多线程技术的软件，其启动的线程不是越多越好，最优线程数取决于 系统所处环境，以及系统自身的特点(Brian,2012) 。 因此，系统设置了最大线程数 上限。每轮调度从第一个执行方案开始，依次选择作业，被选择的作业是否能提交 执行取决于两个约束条件：

(1)如果当前执行方案中有比被选择作业优先级高的作业正在执行，则该被 选择作业不能提交执行。

(2)为被选择作业预分配资源，如果被选择作业的执行将使系统线程数超过 最大线程数，则为被选择作业预分配资源失败，被选择作业不能提交执行。

为了提高调度效率，如果被选择作业不满足约束(1),则选择下一个执行方 案；如果被选择作业不满足约束(2),则选择下一个作业。执行方案调度算法如表 11-4 所列。

表11- 4 执行方案调度算法

|  |
| --- |
| **Name:执行方案调度算法** |
| Input;执行方案列表  Output:待执行作业 |
| 1.从执行方案序列中选择第一个执行方案，作为当前执行方案  2.如果当前执行方案为NULL,则转到步骤14  3.当前作业置为NULL  4.从当前执行方案中选择第一个作业，作为候选作业  5.如果候选作业为NULL,则从执行方案序列中删除当前执行方案，转到步骤1  6.如果候选作业执行完毕，则从当前执行方案中删除候选作业，转到步骤3  7.如果候选作业尚未执行，转到步骤10  8.将候选作业作为当前作业，从当前执行方案中选择当前作业的下一个作业作为候选作业 9.如果候选作业为NULL,转到步骤12  10.如果当前作业不为NULL,且候选作业比当前作业优先级低，转到步骤12  11.如果为候选作业预分配资源成功，转到步骤13,否则，转到步骤8  12.从执行方案序列中选择当前执行方案的下一个执行方案，作为当前执行方案，转到步骤3 13.将候选作业提交执行，转到步骤1  14.休眠t时间后转到步骤1 |

**11.5** **本章小结**

随着数据、应用的深入，数据质量和数据治理工作不断加强，对数据质量工具

282)数据质量导论

的需求也日益紧迫。本章分析了目前处于行业领先地位的数据质量工具及其功能 特点，归纳出当前数据质量解决方案应具备的核心功能，并提供了两种典型的数据 质量工具实现方式，基于表达式树的方式实现比较简单，基于流程的方式实现比较 复杂，但在通用性、可扩展性等方面具有明显的优势。

**参考文献**

[1]石少敏.2011.基于规则和数据学习的清洗模型研究[J]. 陕西教育学院学报，27(3):89-93. [2]房强.2008.面向半结构化数据的数据质量控制系统的研究与实现[M]. 沈阳：东北大学.

[3]孟小峰.2009.XML数据管理概念与技术[M]. 北京：清华大学出版社.

[4]袁满，张雪.2013.一种基于规则的数据质量评价模型[J]. 计算机技术与发展，23(3):81-84. [5]郭志懋，周傲英.2002.数据质量和数据清洗研究综述[J]. 软件学报，13(11):2076-2082.

[6]曹建军，刁兴春，王芳潇，等.2010.基于蚁群特征选择的相似重复记录分类检测[J]. 兵工学报，

31(9):1222-1227.

[7]梁吉胜，李天阳，王慧霞，等.2012.基于约束的数据质量评估算法研究[J]. 科学技术与工程，12(3):

551-554.

[8]程录庆.2011.数据约束对数据质量的影响研究[J]. 长江大学学报(自然科学版),8(5):100-102. [9]廖湖声.2013.XQuery 语言原理和实现技术[M]. 北京：科学出版社.

[10]TechTarget 中国.2012.2012年数据管理优先度与挑战调查报告[R]. 北京，中国：TechTarget. [11]Brian G,Tim P.2012.Java并发编程实战[M]. 童云兰，译.北京：机械工业出版社.

[12]Human Inference.2010.DataCleaner Reference[R].Netherlands:Human Inference.

[13]IBM.2008.IBM Information Server 简介[R]. 北京：IBM 中国 .

[14]Informatica.2010.Informatica 数据质量控制方法白皮书[R]. 北京：Informatica 中国 .

[15]KozaJR,Bennett H B,Andre D,et al.1999.Genetic Programming Ill:Darwini an Invention and Problem Solving

[J].IEEE Trans.Evolutionary Computation,3(3):251 -253.

[16]Oracle.2010.Oracle Data sheet:Oracle Product Data Quality[R].USA:Oracle.

[17]Oracle.2011.Oracle Data sheet:Oracle Enterprise Data Quality Product Family[R].USA:Oracle. [18]SAS.2011.Data Quality Solution[R].USA:SAS.

[19]Saul J,Ted F.2015.Magic Quadrant for Data Quality Tools [R].USA:Gartner.

[20]Talend.2011.Talend Technical Note:DQ Solution Checklist[R].France:Talend.

**第12章** **大数据与大数据质量问题**

**12.1** **引言**

众所周知，互联网改变了企业经营、政府行动以及人们生活的方式，但是一种 新的技术趋势却有着同样巨大的变革能力，那就是“大数据(Big Data)”。在社会 计算、工业计算、商业计算、科学计算等诸多领域，数据资源的价值逐步为人所知， 许多决策依据转向数据。大数据正逐渐成为现代社会基础设施的组成部分，就像 公路、铁路、港口、水电和通信网络一样不可或缺。人们甚至认为：数据，这个快速 增长的集合，已经上升到了与石油一样的同等重要的战略地位。

“大数据”源于以下事实：如今到处传播的信息比以往任何时候都多出了许 多，而且这一趋势正在应用于非同寻常的新用途。尽管互联网使数据的收集和共 享变得十分方便，但大数据与互联网截然不同。大数据的意义并不仅仅是通信，其 本质是从大量数据中学习到较少数据中无法获取的东西。“大数据”这个术语一 般用来描述对海量数据进行分析，从而发现规律、收集有价值的见解和预言复杂问 题答案的技能与科学。

麦肯锡公司2011年报告推测，如果把大数据用于美国的医疗保健， 一年可产 生潜在价值3000亿美元，用于欧洲的公共管理可获得年度潜在价值2500亿欧元； 服务提供商利用个人位置数据可获得潜在的消费者年度盈余6000亿美元；利用大 数据分析，零售商可增加运营利润60%,制造业设备装配成本会减少50%。

本章阐述大数据带来的革命性变革和大数据时代的核心任务，重点归纳梳理 大数据质量面临的挑战，为进一步的大数据研究与工程应用提供指导。在分析数 据治理动机的基础上，介绍了数据治理的一般性流程；针对国内信息环境特点，设 计了基于相对不变过程的数据治理系统框架，借助相对稳定的数据生命周期和项 目周期实现了项目数据的单独治理，以及单位或业务领域数据的统一治理，为进一 步的数据治理研究和系统实现提供基础。

**12.2** **大数据时代的特征**

本节对大数据与大数据质量问题进行了讨论，阐明了大数据的定义及其内含，

**284)**数据质量导论

并分析了大数据具有的规模大、类型多、速度快和价值高的特征；提出数据持续 积累、内容充分公开和大众广泛参与是进入大数据时代的三个必要条件；讨论 了大数据质量问题，重点讨论企业全方位数据质量管理框架。所得出的观点 和结论为下一步工作提供有益参考。总结大数据的含义与特征，进而分析了 进入大数据时代的必要条件，并讨论大数据质量问题，旨在解析大数据时代的 转变及其核心任务。

**12.2.1** **大数据的含义**

大数据指传统软件工具难以收集、存储、保留、管理、分析、共享和可视化的大容 量数据(涂子沛，2012)。但是具体多大的数据才能称为“大”,并没有普遍适用的定 义，例如，涂子沛(2012)认为大数据的数据量级应该是“太字节(1TB=2B)”, 而周 晓方等(2012)认为，“超大规模”表示的是“吉字节(1GB=2°B)” 级别的数据，“海 量”表示的是“太字节”级别的数据，而“大数据”则是“拍字节(1PB=2°B)” 级别 的数据。

尽管数据容量的激增，给数据的收集、保存、维护、共享等任务带来了现实挑 战，出现了诸多有待研究解决的新问题，然而，“大数据”之“大”,并不仅仅在于“容 量之大”。麦肯锡全球研究所认为，我们并不需要给“什么是大”定出一个具体的 “尺寸”,因为随着技术的进步，这个尺寸本身还在不断地增大，此外，对于不同领 域，“大”的定义也不同。“大数据”更大的意义在于：人们可以“分析和利用”的数 据在大量增加，通过数据交换、整合和分析，可以发现新的知识，创造新的价值，带 来“大知识”“大科技”“大利润”和“大发展”。

大数据的以上特征给数据处理、数据存储、数据安全、数据表示、数据展现、数 据质量等问题带来了前所未有的挑战。

**12.2.2** **大数据的特征**

大数据具有4V特征，即规模大(Volume)、类型多(Variety)、速度快(Velocity) 和价值高(Value)(Laney et al.,2001)。

1)规模大

通过各种设备产生的数据，其数据规模极为庞大，拍数量级将是常态。国际数 据公司(IDC) 的研究表明，截止2011年，全球数据总量约为1.8ZB(1ZB=2B), 其中 75%来自于个人(主要是图片、视频和音乐),远远超过人类有史以来所有印刷材料 的数据总量(200PB),1.8ZB 数据假如用9GB的 DVD光盘和1TB2.5 寸的硬盘分别 存储，然后将它们叠加起来，DVD 的高度超过2.6×10⁵km, 约是地球到月球距离的

第12章大数据与大数据质量问题(285)

2/3,而硬盘的高度超过1.7×10\*km,接近地球周长的一半(阳振坤，等，2012)。 2)类型多

数据种类繁多，在编码方式、数据格式、应用特征等多个方面存在差异性，多信 息源并发形成大量的异构数据。结构化数据可以在关系数据库中找到，多年来一 直主导着 IT 应用；半结构化数据包括电子邮件、文字处理文件以及大量发布在网. 络上的新闻等，以内容为基础，如谷歌和百度主要以半结构化数据为业务支撑；而 非结构化数据广泛存在于社交网络、物联网、电子商务之中。伴随着社交网络、移 动计算和传感器等新技术不断推广应用，目前获取的数据85%以上是半结构化和 非结构化数据。

3)速度快

数据量增长速度越来越快，要求的数据处理速度越来越快、响应时间越来越 短，对系统的延时和数据服务质量要求越来越高。企业数据正在以55%的速度逐 年增长；如今，只需两天就能产生出自人类文明诞生以来到2003年所产生数据的 总量。谷歌公司通过大规模集群和 MapReduce 软件，每个月处理的数据量超过 400PB;百度每天大约要处理几十拍字节数据；Facebook 注册用户超过10亿，每月 上传的照片超过10亿张，每天生成300TB 以上的日志数据；淘宝网会员超过3.7 亿，在线商品超过8.8亿件，每天发生数千万笔交易，产生约20TB 数据；雅虎的总 存储容量超过100PB。传感网和物联网的蓬勃发展是大数据的又一推动力，各 个城市的视频监控每时每刻都在采集巨量的流媒体数据。工业设备的监控也 是大数据的重要来源。例如，劳斯莱斯公司对全世界数以万计的飞机引擎进 行实时监控，每年传送拍字节量级的数据。大数据也导致信息系统的高可扩 展性成为最本质需求，并发执行(同时执行的线程)的规模从现在的千万量级 提高到10亿级以上。

4)价值高

麦肯锡的报告指出(James et al.,2011):只要实施正确的政策和激励，大数据 将成为竞争的关键性基础资源，并成为生产率提高、科技创新以及为消费者创造价 值的新支柱。仅2009年，谷歌公司通过大数据业务对美国经济的贡献就为540亿 美元；充分利用大数据可帮助全球个人定位服务提供商增收1000 亿美元，欧洲公 共部门管理者每年增收2500亿美元，美国医疗保健行业每年增收3000 亿美元，并 可帮助美国零售业获得60%以上的净利润增长率。

**12.2.3** **进入大数据时代的必要条件**

在大数据时代，社会资源的配置将更加精细、优化，社会运行的总成本将会降 低，同时，新的数据开发工作将创造新的就业机会，可谓既开源又节流，全社会受

286)数据质量导论

益。然而，尽管当前以美国为代表的发达国家的大数据战略风起云涌，但具体到国 内，还应冷静审视实际信息环境和所处的信息化阶段。在关注国际大数据战略的 同时，从国内实际出发逐步构建属于我们的大数据时代。作者认为，进入大数据时 代，必须具备以下三个必要条件：

1)数据持续积累

如前所述，数据持续不断地有序快速积累是大数据战略的最直接推动力。 一 是要常态化搜集数据，最为典型的例子是美国、日本等国家在我南海和东海进行常 态化、高密度电子情报监视与侦察，此类行动投入巨大，但利在长远，应纳入国家信 息基础设施发展战略规划；二是要有目标的积累数据，避免产生“现有的数据没 用，要用的数据没有”的尴尬局面；三是要更细粒度和更高密度地获取数据，随着 相关技术的发展，数据获取手段的精度不断提高，获取数据的时间和空间密度也在 不断增加，加快了数据量增长速度。

2)内容充分公开

只有数据向所有合法用户完整公开，数据才能被充分利用，进而创造出新价 值。 一是对原始数据的公开，对原始数据的任何修改都可能直接破坏数据的使用 价值；二是公开完整数据，数据的完整性直接影响数据分析结果，尽管公开完整数 据以及利用数据整合技术会带来潜在安全风险，但公开完整数据会带来更大收益； 三是对知识产权保护，在数据公开过程中，数据生产者的权益应得到保护，以形成 良性循环。最具代表性的是由美国首任首席信息官Kundra 主持的 Data.Gov 项 目，该项目的目标是由政府主导、向全社会开放政府拥有的公共数据，该项目代表着 数据在社会的自由流动、知识向大众的自由流动，这为更多的大众创新、社会创新提 供了一个平台。截止2011年12月，Data.Gov项目共开放了原始数据3721项、地 理数据38642项，同年，Kundra 也因该项目获得了美国EMC公司的首届“数据英 雄奖”。

3)大众广泛参与

各类企业掌握的数据量巨大，仅依靠内部员工已难以完全挖掘数据价值，利用 大众智慧是实现从大数据向“大利润”转变的必然途径。 一是参与的广度，Da- ta.Gov上线的第一天，便接受了210万次的点击量，第二天收获了250万次的点击 量，前两个月创下了2000 多万次点击访问总量；二是参与的深度，针对 Data.Gov 上开放的数据，阳光基金会(Sunlight Foundation)举办了程序员公共数据开发大 赛，出现了一些相当经典的数据开发工具，再如2012年的诺基亚移动数据挖掘竞 赛，共吸引了108支队伍参赛(马帅，等，2012)。

客观上讲，国内当前至少在多数领域还不具备以上条件，因此，建议坚持将主 要资源与当前实际结合来创造以上条件。

第12章 大数据与大数据质量问题 (287

**12.2.4** **大数据时代的革命性转变**

有了大数据的帮助，不会再将世界看作是一连串或是自然或是社会现象的事 件，我们会意识到本质上世界是由信息构成的。正如互联网通过给计算添加通信 功能而改变世界那样，因为大数据给人类社会创造了前所未有的可量化的维度，大 数据也必将引发世界巨变。从信息分析的视角，大数据带来了的三个革命性转变， 从根本上改变了人类理解和认知社会的方法基础。

[**12.2.4.1**](12.2.4.1) **全体数据而不是随机样本**

在大数据时代，可以分析更多的数据，甚至可以处理所有相关数据，而不再依 赖随机采样。19世纪以来，当面临大量数据时，都依赖于采样分析，然而，采样分 析是信息缺乏时代和信息流通受限制的模拟数据时代的产物。长期以来，采样分 析被看成理所当然的限制，但随着高性能数据技术的盛行，说明采样分析其实是一 种人为限制。与局限于一定数据量范围相比，使用一切数据会带来更高的准确性， 并且可以发现样本无法揭示的信息。例如，根据数据异常可以识别信用卡欺诈，做 到这一点需要掌握所有数据，这种情况下，异常值的信息价值最高。

尽管某些问题并不需要全体数据，如某通信公司，10%的客户数据就可以满足 客户关系管理(Customer Relationship Management,CRM)团队的分析需求，但是，销 售部门需要分析不同区域的销售情况，需要抽取10%区域的交易数据。销售部门 和CRM团队抽样对象完全不同。如果为了不同问题抽取不同的样本数据，最终会 需要100%的数据。不要因目光短浅而把某分析主题暂时不需要的数据丢弃。抽 样并不是反对收集并存储所有能获得的数据。大数据环境不是为了样本数据存在 的，相反，样本数据来自于大数据环境。因此，所有数据都应该是可用的，即使某些 数据至今都没有使用过(Bill,2013)。

[**12.2.4.2**](12.2.4.2) **从精确分析转向趋势分析**

一个小商店晚上打烊时要数清每分钱，而国民生产总值不需要，也不可能精确 到“分”,随着规模的增大，对精确度的要求大为减弱。

大数据时代的很多时候，追求精确度已经变得不可行，甚至不受欢迎。大数据 纷繁多样，优劣掺杂，分布在全球多个服务器上。拥有了大数据后，不再需要对细 节精益求精，只要掌握大体发展方向即可。当然，并不是完全放弃精度，只是不再 沉迷于此。忽略微观层面的精确度，会让我们在宏观层面拥有更好的洞察力。

沉迷于精确性是信息缺乏时代和模拟时代的产物。大数据时代，掌握的数据 越来越全，不必再担心某个数据点对整个分析的不利影响，更有效的做法是，接受 这些纷繁的数据并从中受益，而不是以高昂的代价消除所有不确定性。

**288)**数据质量导论

[**12.2.4.3**](12.2.4.3) **相关关系取代因果关系**

这是颠覆传统的变革，长期以来，特别是在中国传统文化中，常用“知其然而 不知其所以然”(没弄清因果关系)讽刺对问题认识的肤浅。

在大数据时代，无须再紧盯事物之间的因果关系(知其所以然),而应该寻找 事物之间的相关关系(知其然),以提供非常新颖且有价值的观点。相关关系也许 不能准确地告知我们某件事情为何会发生，但它会提醒我们这件事情正在发生。 有了大数据，不再需要在还没有收集数据之前，就把分析建立在早已设立的少量假 设的基础之上，尽管因果推理蕴含其中。

大数据揭示“是什么”(知其然)而不是“为什么”(知其所以然),而往往知道 “是什么”就够了。如果数百万条电子医疗记录显示橙汁和阿司匹林的特定组合 可以治疗癌症，那么找出具体的药理机制就远没有此治疗方法本身重要。

相关关系还是预测的关键。相关关系的核心是量化两个数据值之间的数理关 系。相关关系强是指当一个数据值增加时，另一个数据值很有可能随之增加或减 少；相反，相关关系弱就意味着当一个数据值增加时，另一个数据值没有随之变化 的规律。通过找出一个现象的良好关联物，相关关系可以帮助捕捉现在和预测 未来。

大数据时代开启了一场寻宝游戏，而对数据的看法以及对因果关系转向相关 关系时释放价值的认识，正是这场游戏的主宰。

**12.2.5** **大数据时代的核心任务**

数据在规划、获取、存储与共享、维护、应用、报废的全生命周期过程中，只有在 应用时才能得到数据的价值，信息的使用方式最终决定数据和信息价值(Dabette, 2010)。

[**12.2.5.1**](12.2.5.1) **推进大数据应用**

在大数据时代，数据存储和分析数据的方法将取代电脑硬件成为价值源泉。 然而，拥有大量数据本身并不会增加任何价值，拥有任何一个数据集，无论大小，其 自身都不会带来任何价值。只有将数据投入具体环境中并付诸使用，特别是看似 无限的再利用，而不单纯只是为了“数据”而“数据”,大数据才能产生价值，甚至是 潜在价值，变成强大的武器(Bill,2013)。

大数据应用的核心是预测。预测通常作为人工智能的一部分，或者更确切地 作为一种机器学习。但是，大数据不是教机器像人一样思考，而是把数学算法运用 到海量数据上来预测事件发生的可能性。海量数据是预测系统取得成功的基础， 并随着数据越来越多，预测系统会不断自我完善。可预见的将来，许多现在单纯依 靠人类判断力的领域都会被计算机系统改变甚至取代，相关技术可以用于疾病诊

第12章大数据与大数据质量问题(289)

断并推荐治疗措施，甚至识别潜在犯罪分子。

[**12.2. 5.2**](12.2.5.2) **发展大数据技术手段**

大数据是指大小超出了典型数据库软件工具收集、存储、管理和分析能力的数 据集(Merv A.,2012)。换言之，大数据超出了常用硬件环境和软件工具在可接受 的时间内为其用户收集、管理和处理数据的能力(Mckinsey Global Institute,2011)。 如前所述，对于大数据，不必纠结于数据量的大，也不必停留在对其内在价值的认 识上，而是要重点关注使用这些数据时需要的新的、强大的技术手段。

1)Hadoop 和 MapReduce

Hadoop 是以开源形式发布的一种对大规模数据进行分布式处理的技术，特别 是处理大数据中的非结构化数据时，Hadoop 在性能和成本方面都具有优势，并易 于通过横向扩展进行扩容。MapReduce 是一种分布式处理方法，Hadoop 是将 Ma- pReduce 通过开源方式进行实现的框架。

Hadoop 由三个组件构成：用于分布式存储大容量文件的分布式文件系统(Ha- doop Distributed File System,HDFS),用于对大规模数据进行高效分布式处理的 Hadoop MapReduce 框架，以及超大型数据表 HBase。从数据处理的角度，Hadoop MapReduce 最为重要，其工作在多台通用型计算机组成的集群上，对大规模数据进 行分布式处理。

在 Hadoop 中，将应用程序划分为集群中任意节点都可执行的成百上千个工作 负载，并分配给多个节点来执行。然后，通过对各节点瞬间返回的信息进行重组， 得出最终的结果。Hadoop 改变了因软硬件成本、处理时间限制不能处理大量非结 构化数据的状况。Hadoop 集群的规模可以很容易地扩展到 PB 级甚至EP 级，往常 只能依赖抽样数据来进行分析，现在则可以将分析对象扩展到全部数据的范围。 因处理能力质的飞跃，多种方法的重复分析以及多种方式的查询测试也变得切实 可行，从而获得前所未有的价值。

2)NoSQL 数据库

传统的关系型数据库管理系统(RDBMS), 是通过SQL 这种标准语言操作数 据库的，相比较而言，NoSQL 数据库不使用SQL 语言。NoSQL 数据库不是对现有 RDBMS的否定，而是对 RDBMS的有力补充，二者的区别列于表12-1(城田真琴， 2013)。

表12- 1 RDBMS与 NoSQL数据库的区别

|  |  |  |
| --- | --- | --- |
| **列项** | **RDBMS** | **NoSQL数据库** |
| 数据类型 | 结构化数据 | 以非结构化数据为主 |
| 数据库结构 | 需要事先定义，固定 | 无需事先定义，灵活可变 |

)数据质量导论

(续)

|  |  |  |
| --- | --- | --- |
| **列项** | **RDBMS** | **NoSQL数据库** |
| 数据一致性 | 通过ACID特性保持严密的一致性 | 存在临时的不保持严密一致性的状态(结果匹 配性) |
| 扩展性 | 基本是向上扩展。由于需要保持数据 的一致性，因此性能下降明显 | 通过横向扩展可以在不降低性能的前提下应 对大量访问，实现线性扩展 |
| 服务器 | 以在一台服务器上工作为前提 | 以分布、协作式工作为前提 |
| 容错性 | 为了提高容错性需要很高的成本 | 有很多无单一故障点的解决方案，成本低 |
| 查询语言 | SQL | 支持多种非SQL |
| 数据量 | 较小规模 | 较大规模 |

如表12-1所列，在 NoSQL数据库中，数据是通过键及其对应的值的组合，或 者是键值对和追加键(Colunm Family)来描述的，因此结构非常简单，无需一开始 就固定数据库结构，且方便随时修改。NoSQL数据库不遵循原子性、 一致性、独立性 和持久性(Atomicity,Consistency,Isolation and Durability,ACID)这种严格原则，而是 采用结果一致性(Eventual Consistency)原则，允许存在临时的、非严密一致性的状 态。从架构而言，RDBMS在数据量增加的时候，基本上是用更大的服务器向上扩展 (Scale Up),很难横向扩展(Scale Out),并且因需要严密保证数据一致性，对性能的 影响十分明显，如果进行非正则化处理而提升性能，则会降低数据库的维护性和操作 性；NoSQL数据库则很容易横向扩展，对性能影响很小，并且设计针对通用型硬件构 成的集群，在成本方面具有优势。RDBMS 可通过复制在多台服务器上保留数据副 本，从而提高容错性，但数据不匹配及要增加副本时，维护上的负荷与成本都会提高。

作为现有Oracle 数据库产品的补充，RDBMS的霸主Oracle 推出了 Oracle No- SQL Database 11g,也说明现有 RDBMS 并不是大数据基础的最佳选择。然而，整体 上来说，NoSQL数据库市场的产品还不够成熟，许多 NoSQL 是一些互联网企业以 内部使用为目的自行开发的，如 Amazon 的 Dynamo,Facebook 的 Cassanda 等，与商 用产品相比还不够成熟，并且与 RDBMS相比，成熟的数据工程师也很稀缺；许多 NoSQL数据库是开源项目，如 Dynamo的分支项目Project Voldemort 和 Cassandra 等，很难得到与商用产品一样的支持。

**12.3** **大数据质量面临的挑战**

**12.3.1** **数据安全问题**

大数据的隐私问题远远超出了常规的身份确认风险的范畴。我们时刻都暴露

第12章大数据与大数据质量问题(291

在“第三只眼”之下：亚马逊监视着我们的购物习惯，谷歌监视着我们的网页浏览 习惯，而微博似乎什么都知道，不仅窃听到了我们心中所想，还有我们的社交关 系网。

大数据的价值不再单纯来源于它的基本用途，而更多源于它的二次利用。这 就颠覆了以个人为中心的思想：数据收集者必须告知个人，他们收集了哪些数据、 做何用途，也必需在收集工作之前征得个人同意。大数据时代，很多数据在收集的 时候并无意用作其他用途，而最终却产生了很多创新性用途。所以，企业无法告知 个人尚未想到的用途，而个人也无法同意这种尚未知的用途。

如果信息在数据库中客观存在，有意识地避免某些信息就是“此地无银三百 两”。如谷歌街景，其中包含了道路和房屋的图像，民众认为这些图片会给黑帮窃 贼提供帮助，遭到强烈抗议；谷歌因此将一些目标模糊化，但这种模糊化的效果却 适得其反，使黑帮窃贼更容易发现行动目标。另一种隐私保护技术是匿名化，让所 有能揭示个人情况的信息都不出现在数据集里，如名字、生日、住址、信用卡号或者 社会保险号等，以确保数据分析利用时，不威胁个人隐私。但是大数据促进了数据 内容的交叉检验，使匿名化技术失效，如研究发现，通过对六部一般电影进行排序， 就有84%的可能性识别出奈飞公司的顾客，如果知道进行排序的日期的话，就有 99%的可能性从50万人的数据库中识别出此顾客。

大数据时代，不管是告知与许可、模糊化还是匿名化，这三大隐私保护策略都 失效了。如今很多用户的隐私已经受到了威胁，当大数据变得更为普遍的时候，情 况更加不堪设想。

大数据时代，隐私安全不再局限于泄露，还增加了被预知隐私的可能性，如生 病、拖欠还款和犯罪的预测算法会让我们无法购买保险、无法贷款，甚至在实施犯 罪前就被预告逮捕。

**12.3.2** **大数据的偏见和盲区**

对于数据集，不管是什么样的规模，仍然是人类设计的产物，而大数据技术工 具(如Hadoop)并不能使人们摆脱曲解、隔阂和错误。当大数据试图反映人类所生 活的社会化世界时，认清这些因素尤为重要。偏见和盲区存在于大数据中，就像它 们存在于个人的感觉和经验中一样，有时从大数据得出的结论并不比人为的意见 更客观。

例如，社交媒体是大数据分析的一个普遍的信息源，那里无疑有许多信息可以 挖掘。我们被告知，推特网的数据显示人们在离家越远的时候越快乐，而且在周四 晚上最为沮丧。但是存在许多理由对这些数据的含义提出质疑。首先，我们从皮 尤研究中心获悉，美国上网的成年人中只有16%使用推特网，因而与整体人口相

292)数据质量导论

比，他们绝对不是一个具有代表性的样本，他们中年轻人和城市人的比例偏多。此 外，我们知道许多推特账号是被称作“机器人”程序的自动程序、虚假账号或是“半 机器人”系统(得到机器人程序辅助的人为控制账号)。最近的估计显示，可能存 在多达2000万个虚假账号。因此在研究如何评估推特网用户情绪之前，应弄清这 些情绪究竟是来自真人，还是来自自动化算法系统。

这种情况下的大数据应用一开始就被数据误导了。

**12.3.3** **非结构化数据的质量控制**

通常非结构化数据本身是无法分析的。几乎没有哪种分析过程能够直接对非 结构化数据进行分析，也无法直接从非结构化的数据中得出结论。然而，非结构化 的数据可以通过某些特定方式进行结构化处理，并得到可以直接进行分析的结构 结果。对原始的非结构化数据进行解析和结构化处理的过程通常称为信息抽取。

文本分析的例子很好地说明了该过程：获取非结构化数据，然后处理该数据， 最后创建出可以用于分析和报表过程的结构化数据。另外，许多大数据源都是半 结构化或多结构化的，而不是完全非结构化的，这些数据具有可被理解的逻辑 流程，因此可以从它们中提取出用于分析的信息。驾驭大数据的一个重要方 面是，利用这种创造性的方式将非结构化数据和半结构化数据变成可用于分 析的数据。

尽管 MapReduce 和 NoSQL 数据库给非结构化数据处理提供了新技术手段，但 在这一过程中仍有诸多不确定因素对数据质量产生影响，特别是当前数据质量控 制技术手段还集中在结构化数据上，对非结构化数据涉及较少。

**12.3.4** **结构化数据内缺少结构性**

世界各地运行着许多复杂的设备和发动机，如飞机、火车、军车、建筑设备、钻 孔设备等。因为造价昂贵，保持这些设备的稳定运转是非常重要的。从飞机发动 机到坦克等各种机器上也开始使用嵌入式传感器，目标是以秒或毫秒为单位来监 控设备的状态。物联网产生的传感器数据，已成为重要的大数据源(涂子沛， 2012)。

传感器数据给我们带来了一个非常艰巨的挑战。虽然我们收集到的大部分数 据是结构化的，每个独立的数据元素也很好理解，但元素之间的时间关系和模式却 根本无法理解。延时和无法测量的外部因素增加了问题的复杂性。如果要考虑所 有的信息识别各种数据长期的作用效果，这个过程会异常复杂。拥有结构化数据 并不一定能够保证分析方法就是高度结构化和标准化的。

第12章 大数据与大数据质量问题(293)

**12.3.5** **分** **布** **式** **数** **据** **清** **洗**

大数据算法的瓶颈是数据移动，在设计数据质量控制算法时应尽可能减少中 间结果。大数据包括海量历史数据，数据清洗是改进这部分数据质量的主要技术 手段。数据越来越多地被划分为若干片段并分布存储于不同的网络站点上。因 此，不仅需要针对集中存储的数据质量进行研究，还需要对分布式存储的数据进行 研究。目前已有的研究主要针对集中存储的数据，几乎未涉及分布式存储数据的 质量问题。研究分布式数据的质量问题面临新困难。例如，检测分布式数据中的 错误需要将一些数据传输到其他站点，保障数据传输量最小化的数据错误检测问 题在分布式环境下成为NP 完全问题(Fan et al.,2010),这就需要为分布式数据开 发全新的错误检测算法和数据纠正算法。因此，如何解决分布式环境下的数据质 量问题是一项极具挑战性的工作。

**12.3.6** **数** **据** **化** **程** **度** **不** **够**

数据代表着对某件事物的描述，数据可以记录、分析和重组它，数据化是一种 把现象转变为可制表分析的量化形式的过程，计算和记录促成了数据的诞生，它们 是数据化的根基。数字化又称电子化，指的是把模拟数据转换成用0和1表示的 二进制码，这样电脑就可以处理这些数据。

数据化与数字化存在本质差异。2004年谷歌决定将所有版权条例允许的书 籍进行数字化，让世界范围内的读者通过网络免费阅读。为此，谷歌与全球最有影 响的图书馆展开合作，还发明了一种能自动翻页的扫描仪，加速书籍扫描过程。起 初，谷歌做的是数字化文本，扫描每一页并存入服务器，书籍内容转化成了网络上 的数字图像，供全球读者查阅。但是，除非读者知道自己要的内容在哪本书上，否 则必须在浩瀚的电子书页中寻找所要的片段。至此，谷歌拥有的只是大量图像，只 能通过人的阅读才能转化为有用信息。接下来，谷歌对这些图像继续做了数据化 工作，通过识别数字图像的光学识别软件来识别文本的字、词、句和段落，将数字化 图像转换成数据化文本。随后，不仅人可以使用这些文本信息，计算机也可以处理 和分析它们。通过检索和查询，可以进行多种分析，既可以分析出某个词或词组第 一次出现的时间及其成为流行词的时间，又可以发现数百年间人类思维发展和思 想传播的轨迹，这些分析还可以支持多种语言。

数字化带来了数据化，但数字化无法取代数据化。将数字化等同于数据化，是 和将计算机化等同于信息化同等的理念误区，当前仍阻碍着多个领域的信息化 进程。

数据质量导论

**12.3.7** **数据稀缺**

尽管当今的数据总量以及增长速度已到了令人始料未及的状态(曹建军，等， 2012),但在某些领域，数据仍然弥足珍贵，甚至可能在未来相当长的时期内与大 数据绝缘。数据化试图要从一切事物中汲取信息，甚至包括很多以前和“信息”无 关的事物，然而，在数据稀缺领域，这一雄心壮志显得苍白无力。“巧妇难为无米 之炊”,没有数据，大数据应用以及相关技术手段也就根本无从谈起。这些领域包 括对敌海军装备的侦听水声数据、战略导弹的发射数据、运载火箭的故障数据、人 造卫星的失控逃逸数据，等等。

**12.3.8** **数据冗余**

更准确地讲，全体数据应该指携带全信息的最小数据集，而不是一开始所获取 的全部数据。事实上，在全部数据中存在大量数据冗余，这些冗余的存在会增加存 储成本，并给数据应用带来不利影响。

事实上，如果在数据处理的最后阶段大部分数据都要被丢弃，那么耗费大量的 时间和存储空间，把全部的原始数据都加载到企业级数据仓库中是完全没有意义。 如果只在一个很短的时间内需要这些数据，那么,把这些数据加载到一些长期保留 信息的地方(如数据仓库)就完全没有必要。在这种情况下，最适合使用 MapRe- duce,在进入数据库之前，把数据中的多余部分剔除掉(Bill,2013)。

确定携带全信息的最小数据，首先要明确所有可能的数据应用，对大数据而言 这是不可能的，因此，数据冗余问题现阶段不可能完全解决。

**12.3.9** **数据对实际需求的适用性**

许多组织致力于建立尽可能多并且可能覆盖所有主题的报表。发生这种 情况可能出于以下原因：业务人员提交了所有可能的需求，而不是真正需要并 且会使用的需求。结果业务人员收到多种多样的报表，可还是得不到想要的 信息。

好像埋藏在500 份报表中的东西才是业务人员所需要的，但报表多达500 份 时，他们很难从中找出自己想要的。另外，任何两个人都希望用不同的想法看待同 一件事情。每一个业务人员都希望在报表上有一个额外的度量标准，或者用一种 不同的方式去组织报表。或许会有500份报表摆在那里，但没有一份是使用人员 想要的。

第12章大数据与大数据质量问题 (**295**

重点是得到的报表要适用于实际需求，包含的数据不多也不少，且组织方式要 符合业务人员的偏好。但事实上，认为报表越多越好的观点却普遍存在。

**12.3.10** **人为选择导致的信息失真**

最为典型的是出于某些复杂的原因人为选择分析结果，当分析结果支持既定 方案时，保留分析结果，当分析结果与既定方案冲突时，则舍弃分析结果，更有甚 者，根据需要对分析结果进行人为调整。最终，尽管分析结果已偏离真实情况，但 相关组织还声称分析结果验证了方案是合理的。这种情况下，数据的应用链被人 为阻断了，根本没有产生应有价值；相反，这种注定不会产生任何效益的数据分析 工作耗费了大量成本。当分析结果被二次利用时，会继续造成损失。因为是对已 得到认可的工作的分析结果，这种信息失真相当隐蔽，除非相关人员自己承认，或 重复分析过程(一般很少有这样做的动机),否则很难被发现。

**12.4** **数据治理**

数据质量控制是数据治理的基础，也是数据治理的目标。本节在分析数据治 理动机的基础上，介绍数据治理的一般性流程；针对国内信息环境特点，设计了基 于相对不变过程的数据治理系统框架，借助相对稳定的数据生命周期和项目周期 实现了项目数据的单独治理，以及单位或业务领域数据的统一治理，为进一步的数 据治理研究和系统实现提供基础。

**12.4.1** **数据治理的出发点**

数据治理以从数据中获取最大价值为目标，将数据视为资产对数据进行优化、 保护和利用，涉及对组织人员、流程、技术和策略的优化调整。数据治理的重点在 于协调不同的、孤立的且常常冲突的可能导致数据异常的策略(Alex et al.,2011; Sunil,2014;Sunil,2010)。

在数据资源建设与利用过程中，不同的数据角色反复困扰于以下问题：

(1)主题专家(Subject Matter Expert):因 为IT 使用的定义同业务不一致，常 常难以理解业务需求。

(2)数据架构师(Data Architect):现有数据不能满足业务需求。

(3)业务分析人员(Business Analyst):常常需要花数小时甚至几天找到所需 信息 。

(4)数据分析人员(Data Analyst):尽管信息量很大，但是不知道哪些数据是

)数据质量导论

可靠的，能用来做决策。

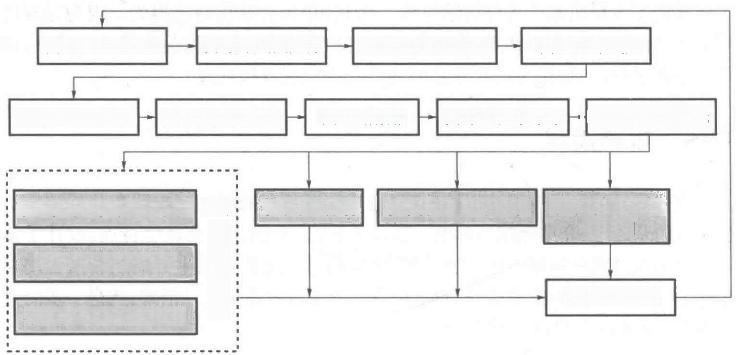
( 5 ) 项 目 执 行 者(Project Executive):因为缺乏有效的沟通，项目常常超出 预算。

(6)数据管家(Data Steward):是否存在数据风险?能否审计信息来源?

数据治理的出发点正是通过解决以上困扰，进而获取更高的数据应用价值。

**12.4.2** **数据治理的一般流程**

对一个具体组织而言，需要明确的可执行步骤来治理数据，图12-1给了一般 性的数据治理流程(Sunil,2010)。



(1)定义业务问题 (2)获取高层支持 (3)执行成熟度评价 (4)建立路线图

(5)建立组织蓝图 (6)创建数据字典 (7)理解数据 (8)创建元数据库 (9)定义质量指标

(11)治理分析 (12)管理安全和隐私 (13)管理信息生

命周期

(10.2)治理数据质量

(14)度量结果

(10.3)实现主数据治理

(10.1)指定数据治理职位 (10):

主数据治理

图12- 1 数据治理的一般性流程框图

如图12-1所示，数据治理的一般性流程包括14个主要步骤，其中10个是必 需步骤，4个是(深色部分)可选步骤。组织用10个必需步骤和部分或全部可选步 骤建立自己的数据治理流程。

对图12-1中的步骤，简要介绍如下：

(1)定义业务问题。围绕某个特定的业务问题定义数据治理计划的初始 范围。

(2)获取高层支持。得到关键IT 和业务高层的支持是数据治理计划得以执 行的必要条件，获得支持的最佳方式是以业务案例的形式实现数据治理的价值，直 观说明计划的必要性；任命兼职或全职的数据治理负责人，并保证其在组织数据治 理相关决策中的足够发言权，确保计划顺利推进。

(3)执行成熟度评价。评估组织的当前数据治理成熟度水平以及拟达到的成 熟度水平。

第12章大数据与大数据质量问题(297)

(4)建立路线图。给出每一个数据治理成熟度维度从当前状态达到目标状态 的具体措施。

(5)建立组织蓝图。建立具有足够成熟度的数据治理组织结构，并且制定日 常管理章程。组织结构分为三层：顶层是由依赖组织数据资产的主要职能和业务 领导组成的数据治理委员会；中间层是日常交往密切的中层管理者；底层是负责日 常数据质量的数据统管社区。

(6)创建数据字典。数据字典或业务术语库是一个关键术语定义的数据库， 作用是实现技术和业务之间的一致性。数据字典将业务术语通过元数据与技术术 语关联，保证对术语的一致理解。

(7)理解数据。通过识别整个组织中主要数据关系，充分理解数据。

(8)创建元数据库。在数据查询时，需要从数据字典生成大量业务元数据和 技术元数据，将它们存储在元数据库中，以便共享利用。

(9)定义质量指标。选择可度量持续性能的关键指标，通过这些指标来度量 并监测数据治理进展。

(10)主数据治理。对照业务目标，定义主数据质量管理的准则、策略、流程、 业务规则和度量指标，并指定专人负责治理主数据。

(11)治理分析。优化业务用户体验，以及对分析基础架构的投资策略与 过程。

(12)管理安全和隐私。处理与数据安全和隐私相关的问题，包括敏感数据发 现、隐藏、访问权限管理等。

(13)管理信息生命周期。处理与信息生命周期各阶段相关的问题，如纸质文 档的数字化、非结构化数据的索引管理等。

(14)度量结果。依据度量指标对数据治理情况进行度量，并向IT 和业务部 门的高层领导报告。

整个数据治理流程是持续循环的，需要对治理效果进行度量，并将结果报告给 高层支持者，以便获取持续性支持。

**12.4.3** **数据治理的系统框架**

当前国内大型数据资源建设项目，是由多个建设单位分年度承担多个数据资 源建设项目来实施的，业务领域、单位项目管理规程等差异，给数据治理带来极大 挑战，其复杂性集中于基础信息环境的不稳定性。

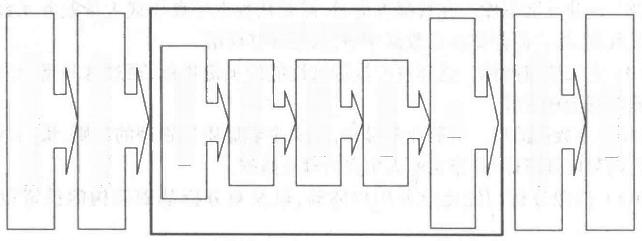
[**12.4.3.1**](12.4.3.1) **稳定信息过程和稳定信息结构**

从控制先天性数据质量问题的角度，吴建明(2004)将信息规划作为从根源上 避免问题产生的有效途径进行了讨论。传统上基于稳定信息过程的数据规划

298)数据质量导论

(Data Planning Based on Steady Information Process,DPBSIP)在发达国家获得了成 功应用，但是，业务过程不稳定所造成的信息过程不稳定是国内信息环境的基本特 征，所以 DPBSIP 在国内多数领域不适用。

为了解决 DPBSIP不适用于国内信息环境的问题，吴建明(2004)将信息结构 视为新的稳定因素提出了基于稳定信息结构的数据规划(Data Planning Based on Steady Information Structure,DPBSIS)。该数据规划方法以“信息科学方法论”为基 本思想，以“数据稳定性原理”为基本依据，以挖掘组织内稳定的信息结构为基本 目标，按照确定的步骤，遵守确定的原则，经过多次反复“收集数据一分析数据”的 循环过程，最终实现构建组织信息模型的目标。DPBSIS 的实施步骤如图12-2 所示。



完成模型

完善目标数据集

审查

任务审查

功能审查

主题审查

数据项审查

获取初始数据集

确 定 目 标 和 边 界

建立核心数据集

图12-2 基于稳定信息过程的信息规划步骤

图12-2中的步骤与 DPBSIP 需要首先确定信息处理过程不同，DPBSIS 的中 心是建立“核心数据集”,再转换成满足不同的用户需要的输出信息结构 目标 数据集，由于核心数据集的稳定性，通过更改输出信息结构即可满足不同的用户， 而输出信息结构的更改不会产生更多的“波及效应”。

DPBSIS 在应用中存在如下局限：

(1)需要原始数据集启动，不适合新领域的数据规划，仅适用于具有良好业务 实践并有一定数据积累的领域。

(2)整个实施过程是“自下而上”的，在应用复杂的领域，对人员专业性要求 高，工作量巨大，实际操作困难。

(3)未覆盖全数据生命周期过程，并且主观性步骤过多，增加了最终形成“稳 定数据结构”的难度。

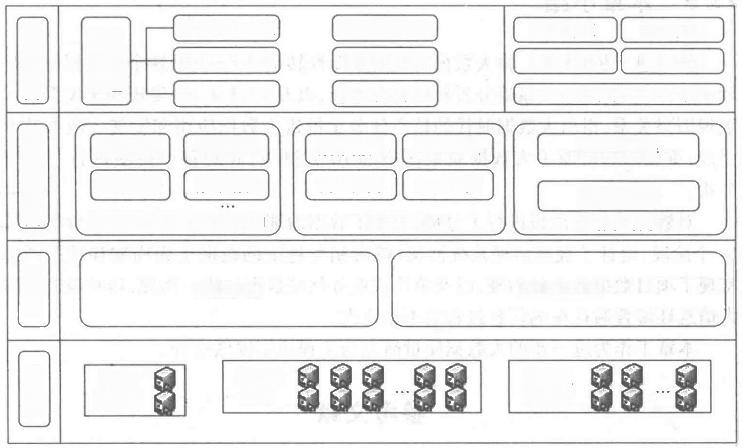
(4)以建立“核心数据集”为中心，但未阐明其与“基础数据”“主数据(Master Data)”的关系。

因此，DPBSIS只能在有限领域中应用，并且只能解决有限范围的数据质量 问题。

第12章 大数据与大数据质量问题 (**299**

[**12.4.3.2**](12.4.3.2) **基于相对不变性过程的数据治理系统**

针对国内信息环境特点，为了实现数据治理的一般性流程，设计基于相对不变 过程的数据治理系统，其系统框架如图12 - 3所示。



用户管理 角色管理 菜单管理 权限管理 系统级，模块级日志管理

流程设计器

流程解析器

工作流引擎

Hadoop hdfs

分布式

数据库

数据世系展示

元数据管理

数据活动管理

数据流程 资源管理

批量恢复 流程监控

ETL 引擎

数据周期定义

项目周期定义

数据活动定义

数据建模

数据稽核

数据采集

数据清洗

功能引擎及二次开发接口

流程管理 數 据 治 理

分布式 集群

WEB 服务器

服务及计算层

应用展现层

Oracle db

存 儲 层

设 备 层

图12 - 3 基于相对不变过程的数据治理系统框架图

如图12-3所示，系统包括应用展现层、服务及计算层、存储层、设备层四个层 次。应用展现层包含了数据周期定义、项目周期定义、数据活动定义、数据世系展 示、元数据管理、活动数据管理、用户管理、角色管理、菜单管理、权限管理以及系统 级和模块级日志管理，为本系统提供页面操作功能；服务及计算层主要提供功能引 擎、二次开发接口、ETL 引擎、工作流引擎等；存储层是数据对象及流程表单持久化 的功能层；设备层是为系统提供运行能力及数据持久化能力的硬件设备。

图12- 3中的数据治理系统的核心特点是：通过数据生命周期定义、项目周期 定义、数据活动定义，并建立数据生命周期阶段和项目周期阶段、项目周期阶段和 数据活动之间的映射来实现项目数据治理流程的实例化，而以数据生命周期为视 角，可以呈现数据资源建设单位或某业务领域的数据治理统一视图。

图12-3中的相对不变过程体现为：领域数据生命周期的统一性，单位内项目 管理周期的稳定性，数据活动的通用性与不变性，领域内数据生命周期阶段和项目 周期阶段、项目周期阶段和数据活动之间映射的稳定性，以及数据生命周期之间的 等效转换性。

图12 - 3中的数据治理系统主要解决的是当前国内信息环境下的信息过程的

数据质量导论

非稳定性带来的问题，应用两层映射，借助相对稳定的数据生命周期和项目周期实 现了项目数据的单独治理，以及单位或业务领域数据的统一治理。

**12.5** **本章小结**

从信息分析的视角，将大数据时代的革命性转变总结为：处理全体数据而不是 随机样本，分析方法从精确分析转向趋势分析，以及分析目的由发现相关关系取代 表现因果关系；指出大数据时代的核心任务是推进大数据应用和发展大数据技术 手段；重点归纳梳理了大数据质量面临的10大挑战，并对每个挑战进行了简要 分析。

对数据治理的动机进行了分析，介绍了数据治理的一般性流程，简要分析了其 14个阶段；设计了数据治理系统框架，借助相对稳定的数据生命周期和项目周期 实现了项目数据的单独治理，以及单位或业务领域数据的统一治理，巧妙规避了国 内信息环境普遍存在的信息过程的不稳定性。

本章工作为进一步的大数据质量研究与工程应用提供指导。

**参考文献**

[1]马帅，李建欣，胡春明.2012.大数据科学与工程的挑战与思考[J]. 中国计算机学会通讯，8(9):22-28. [2]李国杰.2012.大数据研究的科学价值[J]. 中国计算机学会通讯，8(9):8-15.

[3]朱寅，杨强.2012.诺基亚移动数据挖掘竞赛[J]. 中国计算机学会通讯，8(9):67-70. [4]阳振坤，张清，王勇，等.2012.大数据的魔力[J]. 中国计算机学会通讯，8(6):17-21. [5]吴建明.2004.病态信息理论及其在装备保障中的应用[D]. 石家庄：军械工程学院.

[6]周晓方，陆嘉恒，李翠平，等.2012.从数据管理的角度看大数据挑战[J].中国计算机学会通讯，8(9):16-20. [7]城田真琴.2013.大数据的冲击[M]. 周自恒，译.北京：人民邮电出版社.

[8]涂子沛.2012.大数据[M]. 桂林：广西师范大学出版社.

[9]曹建军，刁兴春，江春，等.2013.大数据质量的10大挑战[J]. 现代军事通信，21(4):53-55,68. [10]曹建军，刁兴春，彭琮，等.2012.大数据与大数据质量问题[J]. 现代军事通信，20(4):45-48.

[11]Bill F.2013.驾驭大数据[M]. 黄海，车皓阳，王悦，译.北京：人民邮电出版社.

[12]CBINews.2012. 甲骨文为大数据时代做好准备[OE/L]<.htp://www.cbinews.com/software/news/2012>- 04-20/184085.htm.

[13]Danette M.2010. 数据质量工程实践[M]. 刁兴春，曹建军，张健美，译.北京：电子工业出版社.

[14]Sunil S.2010.IBM 数据治理统一流程[M].[S.l]:MC Press Online,LLC.

[15]Sunil S.2014.大数据治理[M].匡斌，译.北京：清华大学出版社.

[16]Viktor MS,Kenneth C.2013. 大数据时代[M]. 盛杨燕，周涛，译.杭州：浙江人民出版社.

[17]Alex B,Larry D.2011.Master Data Management and Data Governance[M].2版.[S.l]:The MeGraw Hill com- panies.

[18]Fan W,Floris Geerts,Shuai Ma,et al.2010.Detecting Inconsistencies in Distributed Data[C]//The Proceed-

第12章 大数据与大数据质量问题(301)

ings of the 26th International Conference on Data Engineering.Long Beach,Califomia,USA:IEEE Computer Society:64-75.

[19]James M,Michael C,Brad B,et al.2011.Big Data:the Next Frontier for Innovation,Competition and Productivity [R].[S.l]:McKinsey Global Institute:1-137.

[20]Laney,Douglas.2001.3-D Data Management:Controlling Data Volume,Velocity and Variety[R].META

Group Res Note:1-4.

[21]Mckinsey Global Institute.2011.Big Data:The Next Frontier for Innovation,Competition and Productivity[R].

[S.1]:McKinsey Global Institute.

[22]Merv A.2012.Big Data[JOL].TeradataMagarine.www.teradatamagarine.com/v11n01/Features/Big-Data/.

[23]Thomas C.Redman.2011.A Practitioner's View of the Really Big Data Quality(Research)Issues[J].The 16th International Conference on Information Quality,ICIQ 2011,University of South Australia,Adelaide,South

Australia:18-20.

[24]Thomas C.Redman.2008.Data Driven:Profiting from Your Most Important Business Asset[M].[S.1]:Har- vard Business Press.

**基金资助目录**

本书内容涉及的研究工作，得到了以下基金项目资助：

(1)国家自然科学基金面上项目，基于蚁群算法和云模型的领域无关数据清 洗(No.61371196),2014—2017.

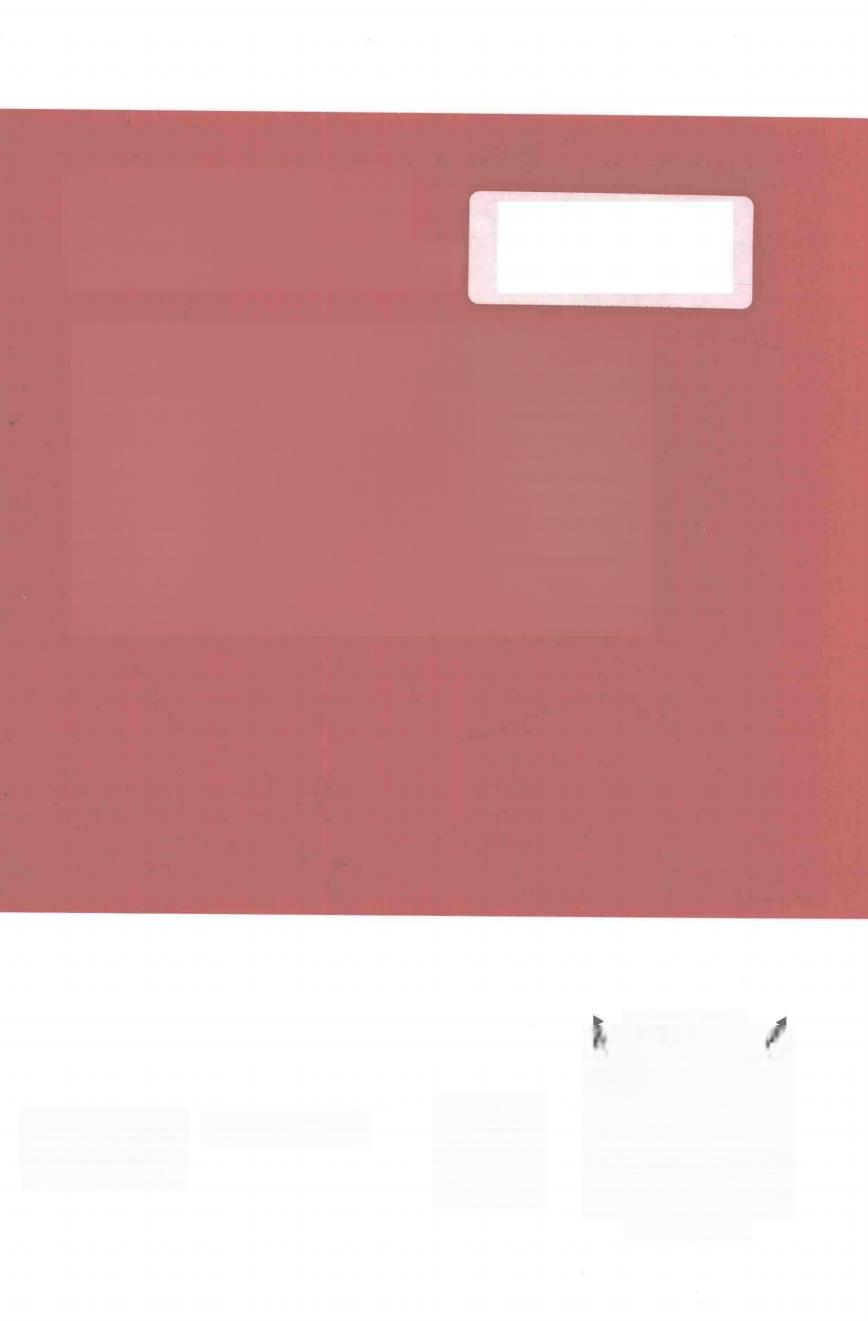
(2)中国博士后科学基金特别资助项目，××信息质量控制方法研究及应用 (No.201003797),2010—2013.

(3)中国博士后科学基金面上项目，××信息质量控制方法研究及应用 (No.20090461425),2009—2013。

(4)江苏省博士后科研资助计划项目，基于蚁群算法的领域无关数据清洗方 法研究(No.0901014B),2009—2013.

(5)解放军理工大学预研基金项目，业务领域无关的数据清洗方法研究 (No.20110604),2011.

(6)解放军理工大学预研基金项目，依赖模式挖掘及在缺失数据处理中的应 用(No.41150301),2015.

**数据质量导论**

**Introduction to Data Quality**

**大数据治理与应用丛书**

■ **数据质量导论**

· **数据质量改进实践指南**

**决胜高质量数据**

· **数据与信息质量：维度、原则与技术**

责任编辑：张冬晔

责任校对：苏向颖

**封面设计：蒋秀芹**

dyzhang@ndip.cn



上架建议：数据库



定价：79.00元