领域无关数据清洗研究综述

曹建军 刁兴春 汪 挺 王芳潇
(总参第63研究所 南京 210007)

摘 要 对领域无关数据清洗的研究进行了综述。首先阐明了全面数据质量管理、数据集成和数据清洗之间的关系，着重说明了领域无关数据清洗的特点。将领域无关数据清洗方法分为基于特征相似度的方法、基于上下文的方法和基于关系的方法分别介绍。最后对领域无关数据清洗的研究方向进行了展望。

关键词 数据质量，数据清洗，数据集成，领域无关数据清洗
中图法分类号 TP311 **文献标识码** A

Research on Domain-independent Data Cleaning: A Survey

CAO Jian-jun DIAO Xing-chun WANG Ting WANG Fang-xiao
(The 63rd Research Institute of the PLA General Staff Headquarters, Nanjing 210007, China)

Abstract Research on domain-independent data cleaning was surveyed. First, relationships among total data quality management, data integration and data cleaning were clarified, and characteristics of domain-independent data cleaning were emphasized. Then, domain-independent data cleaning was classified as feature-based similarity methods, context-based methods and relationship-based methods. They were introduced respectively. At last, the future research directions of domain-independent data cleaning were discussed.

Keywords Data quality, Data cleaning, Data integration, Domain-independent data cleaning

随着信息化进程的不断深入，人们在获得海量信息的同时，越来越被数据的质量问题所困扰，不正确或不一致数据的存在可能严重扭曲分析结果，甚至可以损失信息驱动方法的潜在效益[1]。普化永道会计事务所(Pricewaterhouse Coopers)在纽约的研究表明，75%的被调查公司存在因数据质量问题造成经济损失的现象，只有35%的被调查公司对自己的数据质量充满信心[2]；在销售自动化(sales-force automation)、直接邮寄计划(direct-mail program)和生产率提高计划(productivity improvement program)领域，缺陷数据不低于10%[3]；在数据挖掘项目中，近90%的研究者花在数据清洗和数据准备上的时间超过项目总时间的40%，25%的研究者甚至超过80%[4]。以上案例表明，数据质量问题是普遍的，造成的损失或潜在的威胁较为严重。

上世纪末，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”的研究结论，随后引入普通产品的全面质量管理思想，形成了全面数据质量管理(total data quality management)的思想[5]。数据质量提高与普通产品质量提高的思路一致，主要从两个角度来考虑[6]：一个是从预防角度出发，即在数据全生命周期[7]的各阶段防止脏数据产生；另一个是事后诊断(数据清洗)，不仅针对设计和生产阶段引入的脏数据，由于数据的演化或集成，也会有脏数据不断涌现，采取特定的算法检测和消除出现的脏数据[8]。以上二者是相互依赖的，数据质量的提高需综合利用以上两方面。

1 数据清洗及其分类

1.1 数据清洗和数据集成

数据清洗(data cleaning or data scrubbing)、数据集成(data integration)和数据ETL(data extraction, transformation and loading)及它们之间的关系在近年文献中表述不一致：文献[9]认为数据ETL是数据清洗工具；文献[10]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；文献[11]将数据集成列为了数据清洗的一个研究内容。

数据清洗和数据集成主要应用于数据仓库(data warehouse)、数据挖掘(data mining)和全面数据质量管理3个领域，随着三者应用的不断升温，数据清洗和数据集成也成为研究热点。在不同应用领域，数据清洗和数据集成的内涵有所差异，但它们的核心任务和涉及的典型技术是一致的。

在全面数据质量管理领域，数据清洗和数据集成都是提高数据质量的重要技术途径，是同等的概念范畴。数据清洗是通过检测和消除数据中的错误或不一致(脏数据)来提高数据质量的技术途径[12]；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[13,14]。尽管数据清洗和数据集成联系密切，往往相互交

到稿日期：2009-06-18 返修日期：2009-09-01 本文受江苏省博士后科研资助计划(0901014B)和国家自然科学基金(50705097)资助。
曹建军(1975－)，男，博士后，CCF会员，主要研究方向为数据质量、进化计算等，E-mail：cjj_8@163.com；刁兴春(1964－)，男，研究员，博士生导师，主要研究方向为网络及信息技术等；汪 挺(1964－)，男，高级工程师，主要研究方向为数据工程等；王芳潇(1979－)，女，工程师，主要研究方向为数据工程等。

织、互相渗透[10]，但二者从目的到典型技术的侧重点都有明显区别。比如，实现数据集成的典型技术是数据 ETL，而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)，简称为数据 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脏数据，但事实上因存在不确定性等因素，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。表 1 对数据清洗和数据集成进行了比较。

表 1　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	数据修正	数据异构性

首先，数据清洗和数据集成解决的问题不同；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。

1.2　数据清洗方法的分类

数据清洗所关注的包括缺失数据、错误数据、逻辑错误、相似重复记录(approximately duplicated records)等脏数据的检测和消除。文献[6]重点综述了针对各种脏数据的检测方法和技术。因相似重复记录的普遍性、复杂性以及对后续数据分析的影响，如何检测和消除相似重复记录一直是数据清洗研究的重点[6,9,11]。从消除层面上，相似重复记录的消除又被称为消歧(disambigualation)或实体分辨(entity resolution)。消歧一般被进一步分为对象合并(object consolidation)[1]、参照消歧(reference disambigualation)或模糊匹配(fuzzy match)[15,16]、记录链接(record linkage)[17]等。国内在相似重复记录清洗方面的研究集中在检测层面，以基于特征相似度(feature based similarity, FBS)的传统方法为主[18-19]，关于消歧的研究成果还较少见。

根据是否依赖具体业务领域知识，数据清洗分为特定领域(domain-spesific)数据清洗和领域无关(domain-independent)数据清洗。特定领域数据清洗要求参与清洗过程的人员，必须掌握仅适用于特定领域的规则和原则[20]。领域无关数据清洗具有如下特征[15]：方便与数据库管理系统(database management system, DBMS)整合；能应用于不同业务领域的数据集；具有依赖于数据集规模的合理复杂度；最低的人员参与要求，以方便普通数据库用户应用。与前者相比，领域无关数据清洗具有适用范围广、自动化程度高、对人员要求低等优点，近年受到了广泛关注。

2　领域无关数据清洗研究

数据集中有不同的数据特征信息供数据清洗使用，如属性(attributes)信息、上下文(context)信息、关系(relationships)信息等，利用以上 3 种信息的领域无关数据清洗方法分别称为 FBS 方法、基于上下文的方法(context-based methods)和基于关系的方法(relationship-based methods)。

2.1　FBS 方法

FBS 方法是指通过测量记录中各属性的相似程度进行数据清洗的方法，此类领域无关数据清洗方法的研究由来已久[21,25-30]，至今仍然是最基本的数据清洗方法。近年，对该方法的相关研究多集中于相似度函数和链接效率的改进，如文献[19]提出用机器学习技术最优化相似度函数，采用的是支持向量机(support vector machine)学习；文献[16]提出用专用索引(specialized indexes)提高链接效率。目前，最常用的是用两级相似度函数比较两条记录，首先通过比较两条记录的同一属性值，计算属性级相似度，然后联合属性级相似度计算两条记录的整体相似度。不同属性对相似重复记录检测贡献不同，文献[19]和文献[31]采用给属性加权的方法提高 FBS 方法的检测精度。

事实上，记录属性所提供的信息是有限的，尽管 FBS 方法在检测层面能够取得较好效果，但在消除层面上往往因信息量不足而不能做出高置信度的决策。

2.2　基于上下文的方法

为了弥补 FBS 方法的不足，出现了一些考虑上下文信息的消歧方法。这类方法不但考虑记录本身的属性，还考虑上下文的属性或来自上下文记录的属性。文献[21]对层次关系数据集用直接链接实体的相似度来消除重复记录，但不适用于一般数据集；文献[22]提出了一种对象合并方法，克服了前者的缺点，具有通用性；文献[32]通过排序先使可能相似重复的记录聚集，然后用 FBS 方法进行检测；文献[33]研究了与对象合并问题相关的硬化软数据库(hardening soft databases)问题，虽然没有考虑“关系”，但与后面的基于关系的方法 RelDC 相似的是试图寻求全局最优解，文中证明了这一过程是 NP 难问题，并提出了一种基于优先队列的优化方法。

基于上下文信息的消歧方法，在计算记录相似度时用到直接链接实体的信息，与 FBS 结合提升了消歧效果。

2.3　基于关系的方法

为了从数据集中获取更多的信息用于数据清洗，又逐渐出现了超出上下文范围的基于关系的方法。文献[23]提出了基于关联规则挖掘的利用上下文属性的相似度进行参照消歧的方法，该方法仍然属于 FBS，但讨论了与关系相关的“概念层次”(concept hierarchies)；文献[34-36]提出了基于关系的目标合并技术，两两匹配决策不仅用给定属性进行匹配判定，还触发更多属性参与决策，最终确定多种对象的聚类；文献[37]研究了对象合并和糊模聚类两类不确定问题，并讨论了在引文匹配(citation matching)中的应用，方法以一个关系概率模型为基础，能够同时对多种参照进行推理，但该方法对用户要求高，对普通用户而言应用非常复杂[35,36]；文献[38]研究了自然场景文本识别问题(robust reading)，该问题是一个与对象合并和模糊聚类对应的自然语言问题，对来源于同一时期的文档聚类时，同时考虑文档中发现的地域和组织可以改进对象合并质量。

2003 年，美国学者 Kalashnikov 和 Mehrotra 提出了一种新型领域无关数据清洗框架，被称为基于关系的数据清洗(relationship-based data cleaning, RelDC)，其基本思想是用无向图对关系数据库建模得到数据库的完整实体关系图(entity-relationship graph)，通过分析实体关系图，利用实体之间的关系提高数据清洗效果[24]。实体关系图是对完整关系数

据库进行建模所得，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，蕴含着比基于上下文的方法和其它基于关系的方法丰富得多的信息资源；对数据库建模并进行实体关系图分析，针对相同数据异常情况，其清洗过程是通用的，符合领域无关数据清洗的主要特征。随后，Kalashnikov 及其合作者于 2005 年将 RelDC 应用于对象合并[1]，2006 年将 RelDC 应用于参照消歧[2]，2007 年提出了一种自适应确定实体关系图路径权重的方法，通过对数据样本学习达到适应具体数据集的目的，解决了文献[1,2]需要人员确定路径权重的问题[17]。以上工作给出了 RelDC 的一般实现流程，初步探讨了实现过程的硬约束优化和自适应方法，并用实验初步验证了 RelDC 具有良好的数据清洗效果。但作为一种新型的领域无关数据清洗方法，RelDC 还存在以下问题：①求取实体关系图中虚拟连接图(virtual connected subgraph)的任两节点间的所有路径，是 RelDC 实现的核心过程，实体关系图的规模依赖于数据库的规模，所以一般而言，这一过程运算复杂度非常高，是 Kalashnikov 课题组一直强调的 RelDC 的瓶颈；另外，选择(相似)边的权重求解也具有较高的复杂度。②为了工程上能够实现，以路径过长没有实际意义为假定，文献[1,2,17]均采用限制发现路径长度的方法以降低实现复杂度，权重修剪阈值、路径类型限定也采用了类似的硬约束方法，这种处理的合理性、约束参数如何确定以及对数据清洗效果的影响，还需进一步研究；另外，这种简化处理可能损失一部分有用信息，也不容易对不同处理对象选择最优约束参数值。③FBS 的置信度阈值决定了 RelDC 的切入时机，因 FBS 方法的数据清洗能力有限，而 RelDC 运算代价高，RelDC 何时切入涉及效率与效果的综合最优问题，即 RelDC 与 FBS 的切换机制值得专门研究。④数据库的实体关系图蕴含了丰富的实体之间的关系，目前 RelDC 的应用尚局限于对几种相似重复记录的消除上，如何对这些关系信息进一步挖掘，即将 RelDC 用于其它数据异常的清洗，也值得进一步深入研究。⑤目前 RelDC 所验证的数据集包含的实体种类较少，相应的关系也较少，实体关系图也较简单，实际工程中数据集的实体可能有几十种或更多，实体关系图非常复杂，当前的硬约束甚至求解策略的有效性需要进一步验证。

Kalashnikov 为了寻求解决 RelDC 中问题的新思路，在文献[2]中专门将 RelDC 与文献[22]中采用了启发式算法的 BANKS(Browsing ANd Keyword Searching)系统做了比较，强调二者在建模、求解过程上的相似性，给粒子群算法[39]、蚁群算法[40]等先进启发式算法在 RelDC 中的应用研究提供了依据。

结束语 在全面数据质量管理领域，数据清洗和数据集成是数据质量提高的两个重要技术途径。领域无关数据清洗因其独特优势近年发展迅速，其发展呈现出以下趋势：(1) RelDC 作为一种新的领域无关数据清洗方法，优势明显，但其理论及实现方案将得到进一步关注和完善；(2)将先进的优化算法应用于 RelDC，以解决当前 RelDC 实现复杂度高的问题；(3)融合 FBS 方法、基于上下文的方法和基于关系的方法的优点，综合考虑效率和效果，制订出更好的数据清洗方案，是一个具有应用价值的研究方向。因此，随着领域无关数据清洗方法研究的不断推进，它们在数据质量提高中发挥的作用也将日渐明显。

参 考 文 献

[1] Chen Zhaoqi, Kalashnikov D V, Mehrotra S. Exploiting Relationships for Object Consolidation[C]// Proceedings of the IQIS Workshop at ACM SIGMOD Conference. Baltimore, MD, 2005

[2] Eppler M J, Algesheimer R, Dimpfel M. Quality Criteria of Content-driven Websites and Their Influence on Customer Satisfaction and Loyalty: an Empirical Test of an Information Quality Framework[C]// 8th International Conference on Information Quality (IQ 2003). November 2003: 108-120

[3] The MIT Total Data Quality Management[OL]. http://web.mit.edu/tdqm/www/ about. shtml, 2009-2

[4] KDnuggets Polls. Data Preparation Part in Data Mining Projects [OL]. http://www.kdnuggets.com/polls/2003/data_preparation.htm, Sep.-Oct. 2003

[5] Wang R Y. A Product Perspective on Total Data Quality Management[J]. Communications of the ACM, 1998, 41(2): 58-65

[6] 韩京宇，徐立臻，董逸生. 数据质量研究综述[J]. 计算机科学，2008, 35(2): 1-5, 12

[7] 曹建军，刁兴春，汪挺，等. 数据全生命周期过程模型与质量控制[J]. 现代军事通信，2009, 17(2): 38-42

[8] Dasu T, Johnson T. Exploratory Data Mining and Data Cleaning [M]. John Wiley, 2003

[9] 郭志懋，周傲英. 数据质量和数据清洗研究综述[J]. 软件学报，2002, 13(11): 2076-2082

[10] 管丽娟. 数据 ETL 研究与展望[J]. 电脑知识与技术(学术交流)，2007(6): 1512-1514

[11] 王曰芬，章成志，张蓓蓓，等. 数据清洗研究综述[J]. 现代图书情报技术，2007(12): 50-56

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

[13] Maurizio L. Data Integration: a Theoretical Perspective[C]// ACM SIGACT-SIGMOD- SIGART Symposium on Principles of Database Systems. 2002

[14] 陈跃国，王京春. 数据集成综述[J]. 计算机科学，2004, 31(5): 48-51

[15] Kalashnikov D V, Mehrotra S. Domain-independent Data Cleaning via Analysis of Entity-relationship Graph[J]. ACM Transactions on Database Systems, 2006, 31(2): 716-767

[16] Chaudhuri S, Ganjam K, Ganti V, et al. Robust and Efficient Fuzzy Match for Online Data-cleaning[C]// Proceedings of the ACM SIGMOD Conference. San Diego, CA, 2003

[17] Christen P. Automatic Training Example Selection for Scalable Unsupervised Record Linkage[C]// Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining(PAKDD). Osaka, Japan, May 2008

[18] 韩京宇，徐立臻，董逸生. 一种大数据量的相似记录检测方法[J]. 计算机研究与发展，2005, 42(12): 2206-2212

[19] 张昌年. 一种基于 VSM 的检测相似重复记录的方法[J]. 微电子学与计算机，2008, 25(8): 184-187

[20] Chaudhuri S, Ganjam K, Ganti V, et al. Data Cleaning in Microsoft SQL Server 2005[C]// Proceedings of the ACM SIGMOD Conference. Baltimore, MD, 2005

[21] Ananthakrishna R, Chaudhuri S, Ganti V. Eliminating Fuzzy Duplicates in Data Warehouses[C]// Proceedings of the VLDB Conference. 2002

[22] Getoor L. Multi-relational Data Mining using Probabilistic Relational Models: Research Summary[C]// Proceedings of the 1st Workshop in Multi-Relational Data Mining. 2004

[23] Lee M, Hsu W, Kothari V. Cleaning the Spurious Links in Data [J]. IEEE Intell. Syst, 2004

[24] Kalashnikov D V, Mehrotra S. Exploiting Relationships for Data Cleaning[R]. TR- RESCUE-03-02. UCI Tech. Rep, 2003

[25] Newcombe H, Kennedy J, Axford S, et al. Automatic Linkage of Vital Records[J]. Science, 1959(130): 954-959

[26] Fellegi I, Sunter A. A Theory for Record Linkage[J]. J. Amer. Stat. Assoc., 1969, 64(328): 1183-1210

[27] Winkler W E, Inkler W E. Advanced Methods for Record Linkage[C]// Proceedings of the U. S. Bureau of Census. 1994

[28] Hernandez M, Stolfo S. The Merge/Purge Problem for Large Databases[C]// Proceedings of the ACM SIGMOD Conference. San Jose, CA, 1995

[29] Winkler W. The State of Record Linkage and Current Research Problems[C]// Proceedings of the U. S. Bureau of Census, TR99. 1999

[30] Mccallum A K, Nigam K, Ungar L. Efficient Clustering of High-dimensional Data Sets with Application to Reference Matching [C]// Proceedings of the ACM SIGKDD Conference. Boston, MA, 2000

[31] 陈伟, 王昊, 朱文明. 一种提高相似重复记录检测精度的方法[J]. 计算机应用软件, 2006, 23(10): 29-30, 42

[32] 俞荣华, 田增平, 周傲英. 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, 2002, 29(1): 118-119

[33] Cohen W, Kautz H, Mcallester D. Hardening Soft Information Sources[C]// Proceedings of the ACM SIGKDD Conference. Boston, MA, 2000

[34] Dong X, Halevy A Y, Madhavan J. Reference Reconciliation in Complex Information Spaces [C]// Proceedings of the ACM SIGMOD Conference. Baltimore, MD, 2005

[35] Mccallum A, Wellner B. Conditional Models of Identity Uncertainty with Application to Noun Coreference[C]// Proceedings of the NIPS. 2004

[36] Singla P, Domingos P. Multi - relational Record Linkage [C]// Proceedings of the MRDM Workshop. 2004

[37] Pasula H, Marthi B, Milch B, et al. Identity Uncertainty and Citation Matching[C]// Proceedings of the NIPS Conference. 2002

[38] Li X, Morie P, Roth D. Identification and Tracing of Ambiguous Names: Discriminative and Generative Approaches [C]// Proceedings of the AAAI. 2004

[39] Lin Jing, Sun Jun, Xu Wenbo. Quantum-behaved Particle Swarm Optimization with Adaptive Mutation Operator [C]// ICNC 2006, Part I, LNCS 4221. Heidelberg: Springer-Verlag Berlin, 2006: 959-976

[40] 曹建军, 张培林, 王艳霞, 等. 一种求解子集问题的基于图的蚂蚁系统[J]. 系统仿真学报, 2008, 20(22): 6146-6153, 6157

(上接第 14 页)

[28] Qin X, Lee W. Attack plan recognition and prediction using causal networks[C]// Proc. of ACSAC 2004. Washington DC: IEEE Computer Society Press, 2004: 370-379

[29] Liu P, Zang W Y, Yu M. Incentive-based modeling and inference of attacker intent, objectives, and strategies[J]. ACM Transactions on Information and System Security (TISSEC), 2005, 8 (1): 78-118

[30] Cabrera J B D, Lewis L, Qin X, et al. Proactive intrusion detection-a study on temporal data mining[C]// Barbara D, Jajodia S, eds. Applications of data mining in computer security. Berlin: Springer-Verlag, 2002: 195-227

[31] Vel O de, Anderson A, Corney M, et al. E-mail authorship attribution for computer forensics[C]// Barbara D, Jajodia S, eds. Applications of data mining in computer security. Berlin: Springer-Verlag, 2002: 229-250

[32] Julisch K, Dacier M. Mining intrusion detection alarms for actionable knowledge[C]// Proc. of KDD-2002. New York: ACM Press, 2002: 366-375

[33] Viinikka J, Debar H, Mé L, et al. Time series modeling for IDS alert management [C]// Proc. of AsiaCCS 2006. New York: ACM Press, 2006: 102-113

[34] Pietraszek T. Using adaptive alert classification to reduce false positives in intrusion detection[C]// Proc. of RAID 2004. Heidelberg: Springer Berlin, 2004: 102-124

[35] Porras P A, Fong M W, Valdes A. A Mission-Impact-based approach to INFOSEC alarm correlation [C]// Proc. of RAID 2002. Heidelberg: Springer Berlin, 2002: 95-113

[36] Xiao Min, Xiao Debao. Alert verification based on attack classification in collaborative intrusion detection[C]// Proc. of SNPD 2007. Washington DC: IEEE Computer Society Press, 2007: 739-744

[37] Wang L, Liu A, Jajoda S. An efficient and unified approach to correlating, hypothesizing, and predicting intrusion alerts[C]// Proc. of ESORICS 2005. Heidelberg: Springer Berlin, 2005: 247-266

[38] Cuppens F, Miège A. Alert correlation in a cooperative intrusion detection framework[C]// Proc. of the 2002 IEEE Symposium on Security and Privacy. Washington DC: IEEE Computer Society Press, 2002: 202-215

[39] Zhai Y, Peng Ning, Xu J. Integrating IDS alert correlation and OS-level dependency tracking[C]// Proc. of ISI 2006. Heidelberg: Springer Berlin, 2006: 272-284

[40] Tedesco G, Aickelin U. Data reduction in intrusion alert correlation[J]. WSEAS Transactions on Computers, 2006, 5(1): 1-8

[41] Lincoln P, Porras P, Shmatikov V. Privacy-preserving sharing and correlation of security alerts[C]// Proc. of USENIX-Security '04. San Jose: USENIX, 2004: 17-32