# 基于 Web 信息的关系型信息错误自动检测与修复技术研究综述

刘海龙 李战怀 陈 群 陈肇强

(西北工业大学计算机学院 西安 710072)

**摘 要** 信息质量已经成为诸多应用领域所面临的一个重要问题,自动检测和修复信息系统中的信息错误是改善信息质量的有效手段.利用 Web 对关系数据库中的信息进行扩展以助于信息错误的自动检测与修复具有对待检测与修复的信息本身依赖少、信息质量规则更灵活、适用性更广以及信息修复相对更准确等优势,可以有效克服现有的基于规则、基于扩展信息和基于人机交互的信息错误检测与修复技术的不足.文中详细分析了基于 Web 信息的信息错误自动检测与修复技术的优势及所面临的挑战,提出了基于 Web 信息的信息错误自动检测与修复技术框架.该框架包括:Web 信息自动拓展模型、基于 Web 信息的信息错误自动检测算法、基于 Web 信息的信息错误自动修复算法和基于 Web 信息的信息错误自动检测与修复算法的可靠性评估模型.基于上述框架,文中系统总结了基于 Web 信息的信息错误自动检测技术、信息错误自动修复技术以及信息扩展核心技术三个方面的研究进展,提炼出了基于 Web 信息的信息错误自动检测与修复技术需要解决的关键科学问题,对未来的研究方向进行了展望并且讨论了初步的研究思路.

**关键词** 关系数据;信息质量;错误检测;数据修复;Web

**中图法分类号** TP311 **DOI 号** 10.11897/SP.J.1016.2017.02286

# A Review on Web-Based Techniques for Automatically Detecting and Correcting Information Errors in Relational Databases

LIU Hai-Long LI Zhan-Huai CHEN Qun CHEN Zhao-Qiang

(*School of Computer Science, Northwestern Polytechnical University, Xi'an* 710072)

**Abstract** Information quality has become an important issue in many application areas. Automatically detecting and correcting information errors has proven to be an effective way to improve information quality in most information systems. Integrating information from the World Wide Web (WWW) can help us overcome the shortcomings of existing rule-based, external-information-based, human-based information error detection and correction techniques for relational databases to a great extent. The advantages of Web-based techniques include less dependence on the sufficiency of the database, more styles of constrains, wider applicability and more accurate repairs. In this review, we detail the advantages and challenges of Web-based information error detection and correction techniques. We propose a technological framework and believe it should include four components, including Web-based information expansion model, Web-based error detection algorithms, Web-based error correction algorithms and Web-based

收稿日期:2016-05-25;在线出版日期:2016-10-23.本课题得到国家"九七三"重点基础研究发展规划项目基金(2012CB316203)、国家自然科学基金(61502390,61472321)和西北工业大学基础研究基金(3102014JSJ0013,3102014JSJ0005)资助.**刘海龙**,男,1980 年生,博士,讲师,中国计算机学会(CCF)会员,主要研究方向为数据管理技术、数据质量管理技术.E-mail:liuhailong@nwpu.edu.cn.**李战怀**,男,1961 年生,博士,教授,中国计算机学会(CCF)高级会员,主要研究领域为数据存储技术、数据管理技术.**陈 群**,男,1976 年生,博士,教授,中国计算机学会(CCF)高级会员,主要研究领域为数据管理技术、数据质量管理技术.**陈肇强**,男,1988 年生,博士研究生,主要研究方向为数据质量管理技术.

evaluation models for error detection and correction algorithms. Based on the framework, we comprehensively review current research works on the topics like Web-based error detection techniques, Web-based error correction techniques and Web-based information expansion techniques. We also refine out two key scientific problems which all Web-based information error detection and correction techniques must concern. Furthermore we prospect some future research topics and ideas.

**Keywords**　relational data; information quality; error detection; data repairing; Web

## 1　引　言

信息已成为一个组织最有价值的资产之一. 对于任何数据分析和决策支持系统来说，高质量的信息都是保障获得有意义分析结果的重要前提. 而现实应用中存在严重的信息质量问题. 诸多实例①②③表明多国政府和商业机构已投入大量资金用于应对低质量的信息带来的影响. 信息质量已经成为商业、科学计算、医疗、环境和公共服务等信息技术应用领域面临的一个重要问题.

改善信息质量的途径主要有两种：一是提前预防，即在信息生命期的每个阶段都有严格的信息规划和约束来防止低质量信息的产生；二是事后诊断和修复，即事后采用特定的方法检测和修复低质量的信息. 大数据时代信息来源多产生速度快、处理过程复杂，往往无法采用提前预防的手段控制所有信息源的质量. 这样，事后检测与修复就显得尤为必要. 判断信息异常或者错误的依据一般来源于人的常识或者特定领域的专业知识. 但是，人工仅可检测和修复所熟悉领域中的少量信息中的信息错误，无法处理巨量复杂信息，对于不熟悉的领域也无能为力. 因此，用户往往希望信息系统具有自动检测和修复多领域巨量复杂信息中的信息错误的能力.

根据解决问题过程中所依赖的信息，关系数据库中信息错误自动检测与修复的相关技术可分为 3 类[1]：第 1 类只依赖数据库本身的信息，力求从数据库中挖掘信息特征来解决，其技术主要是基于规则的检测与修复技术；第 2 类需要借助数据库以外的信息，力求结合数据库及相关外部信息源发掘信息特征来解决，其技术主要是基于扩展信息的检测与修复技术；第 3 类需要辅以人的常识或特定领域的专业知识，其技术主要是以众包技术为代表的交互式检测与修复技术. 这 3 类技术互为补充，它们之间的关系如图 1 所示.

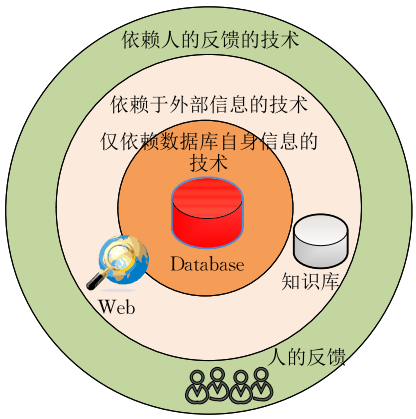


图 1　关系信息错误自动检测与修复技术分类[1]

基于规则的信息错误自动检测与修复理论及技术体系已经初具雏形，本文围绕一致性、完整性、实体同一性、时效性和精确性这 5 类错误提出了一系列的解决方案，主要包括信息质量规则描述及挖掘技术[2-13]、信息错误检测技术[5,14-16]、信息错误修复技术及信息修复的可靠性保障技术[17-22]. 但是此类技术尚存在如下不足：(1) 要求待检测与修复的关系型信息中必须包含足够的信息，否则无法发掘有效的信息质量规则或者发掘出来的信息质量规则不能够准确反映信息的真实特征，也无法借助质量规则推导出合适的候选值进行信息错误修复. 例如在表 1 中某元组的 *isbn* 号码唯一的情况下无法利用 *isbn*→*title* 这样的函数依赖推导出该元组的 *title* 值；(2) 在修复信息错误时，只是力求发掘出合适的值以确保修改后的信息与信息质量规则一致. 鉴于关系数据库中的信息是有限的，发掘出的“合适的”值不一定是真实的值.

基于扩展信息的信息错误检测与修复技术可以有效弥补基于规则的技术存在的不足，比较典型的工作有借助主信息[23]判断元组缺失等信息错误[24-25]和加入用户(领域专家)反馈以确保信息错误检测与修

① ‘Dirty Data’ is a Business Problem, Not an IT Problem, Says Gartner. http://www.gartner.com/newsroom/id/501733

② http://www-new.insightsquared.com/wp-content/uploads/2012/01/insightsquared_dq_infographic-2.png

③ Meta Group. Data Warehouse Scorecard. http://www.sfdama.org/dwscorecard.pdf, 2003

复的可靠性[26]. 这些工作中外部信息源主要是主信息. 在企业的某些关键领域(如雇员、项目和设备等),主信息通常是完全的和精确的,即封闭的. 而对于其它领域(如销售纪录和用户支持)则往往没有完整的主信息.

人工干预的方法相对来说最准确有效,但是效率较低. 另外,用户(领域专家)的知识领域有限,对于不熟悉的领域也无能为力. 采用众包策略可以有效弥补用户(领域专家)知识领域的不足,比较典型的工作有结合知识库和众包技术的数据清洗技术[27]、基于众包策略的实体识别技术[28]和多版本 Web 数据清洗技术[29]. 但是此类技术易受制于众包平台的成熟度、成本以及用户对响应时间的要求,所以更适用于较小规模信息错误的自动检查与修复.

截止到 2008 年 7 月 Google 已经在互联网上发现超过 10000 亿个独立网址,独立网页的数量还在以每天几十亿的速度持续增长①. 截止 2014 年 12 月我国网页数目达到了 1899 亿个②. Web 已经成为世界上最重要的信息源[30]. 相对于典型的知识库系统(如主信息库),Web 信息具有涵盖领域广和信息源多等优势,因此对许多应用领域的信息都可以利用 Web 对待考察的信息进行扩展以帮助进行更准确的信息错误自动检测与修复.

基于 Web 信息的信息错误自动检测与修复技术核心思想是:(1)根据关系数据库中的已知信息生成合适的查询;(2)依据查询从 Web 中获取所需的扩展信息;(3)结合扩展信息生成合适的信息质量规则或修复方案;(4)根据信息质量规则检测信息错误或根据修复方案进行信息错误修复;(5)利用 Web 校验信息错误检测与修复结果的正确性. 其示意图如图 2 所示.

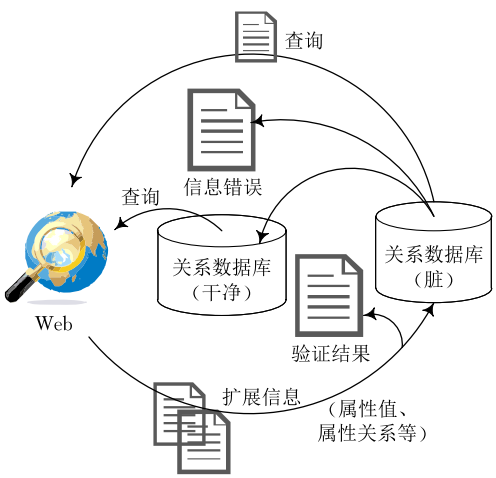


图 2 基于 Web 信息的信息错误自动检测与修复技术核心思想

本文将深入分析基于 Web 信息的关系型信息错误自动检测与修复技术的优势及面临挑战,系统总结相关技术的研究进展并对未来研究方向进行展望.

## 2 优势与挑战

基于 Web 信息的信息错误自动检测与修复技术具有如下优势:

(1)对待检测与修复的信息本身依赖少. 利用 Web 进行信息扩展,一方面可以帮助信息系统以更加全局的视野发掘信息特征从而发现更加准确的信息质量规则,另外一方面可以提供更多的参考信息帮助信息系统发现正确的信息来进行更加准确的信息修复. 所以基于 Web 信息的信息错误自动检测与修复技术可以有效克服待检测与修复的信息中信息量不足而导致的信息质量规则无法发现以及信息错误无法修复等问题. 例如:仅依赖数据库自身的信息无法为表 1 中 *isbn* 值为"978-0123814791"的元组找到一个合适的 *publisher* 值,但是可以根据该元组的已知属性值(如 *isbn* 或 *title* 值)从 Web 中获取其 *publisher* 值.

(2)信息质量规则更灵活. 借助 Web 数据挖掘技术可以从 Web 中发现信息实体之间的关联关系,利用这些关联关系可以帮助用户判断数据库中的信息是否正确. 采用这种方式确定的信息质量的规则会更加灵活、适用性更好. 例如:如果 Web 中"Data Mining: Concepts and Techniques, Third Edition"与"Morgan Kaufmann"共同出现的次数明显多于"Data Mining: Concepts and Techniques, Third Edition"和"Springer"共同出现的次数,就可以判定:若该书的出版社是"Springer",则该条信息中存在错误的可能性较大. 在这种情况下,就可以使用 Web 中信息的分布情况作为一种有效的信息质量检测规则.

(3)适用性更广. 关系型信息质量主要包括完整性、一致性、实体同一性、精确性和时效性这 5 类问题[31]. 基于 Web 信息的信息错误自动检测与修复技术对这 5 类问题都适用. 根据关系数据库中的已知信息可以从 Web 中获取关系表中的缺失属性值以及描述同一个实体的不同时间、多种精度和多个

① http://www.google.cn/intl/zh-CN/about/company/history/

② 中国互联网络信息中心. 第 35 次中国互联网络发展状况统计报告. http://www.cnnic.net.cn/hlwfzyj/hlwxzbg/hlwtjbg/201502/P020150203548852631921.pdf

《计算机学报》

副本的值，因此理论上讲基于 Web 信息的方法可以有效地处理完整性、精确性和时效性问题，例如：借助 Web 可以获取表 1 中的 *isbn* 值为“978-0123814791”的元组的 *publisher* 值、可获取不同精度的中国 2015 年人口数量以及不同时期的美国总统姓名. 另外，Web 中存在大量对同一个实体的不同表述以及隐藏了多个实体之间的多种关系，可以借助这些信息有效应对一致性和实体同一性问题，例如：若在 Web 中“gem”和“jewel”与“precious stone”同时出现的频率都很高，则可以判定它们指代同一种物体的可能性比较大.

(4) 信息修复相对更准确. 传统的数值型数据错误修复策略一般取用平均值、中位数、和频繁值等近似值进行修复，不能保证修复结果的准确性. 另外，这种修复策略忽略了信息的特异性，对于信息个体来说可能导致修复结果错误. 例如：在表 1 中，“Springer”出现的频率最高，但是其并非为 *isbn* 值为“978-0123814791”的元组的 *publisher*，该元组的 *publisher* 值应为“Morgan Kaufmann”. 而基于规则的信息修复技术的原则是修复后的信息与规则匹配，并不着眼于发掘信息的真实值. 所以这两类修复方法相对来说都不能确保修复的准确性. 从理论上讲，若信息的真实值存在于 Web 中并且能够被正确发掘，则可以做到对错误信息进行比较准确的修复.

**表 1　图书记录**

| isbn | title | author | publisher | year |
|---|---|---|---|---|
| 978-0-387-30768-8 | Encyclopedia of Machine Learning | Claude Sammut | Springer | 2010 |
| 1-930708-38-6 | Database Integrity: Challenges and Solutions | Jorge Horacio Doorn | Idea Group | 2002 |
| 978-0-521-56876-0 | Computational principles of mobile robotics | Michael R. M. Jenkin | Cambridge University Press | 2000 |
| 978-0123814791 | Data Mining: Concepts and Techniques, Third Edition | Jiawei Han | | 2011 |

尽管具有上述优势，由于 Web 信息具有海量以及不一致等特征以及 Web 信息实体获取技术尚在发展，基于 Web 信息的信息错误自动检测与修复技术还面临诸多挑战，主要表现如下：

(1) Web 信息本身存在质量问题. 不同于知识库系统，Web 信息本身具有海量、信息源自治、数据动态变化、数据不规范和数据不一致等特征[32]. 这些特征使得从 Web 中获取相对正确的信息本身就变成一件非常有挑战性的工作. 例如：“谣言”会导致 Web 中存在“基督教”和“穆斯林”这两类相互冲突的对于奥巴马的宗教信仰的描述[33]，需要对相关的信息进行甄别、筛选.

(2) 缺乏适用的基于 Web 信息的信息实体扩展技术. 作为主流的 Web 信息获取工具，目前搜索引擎只能返回相关文档而无法直接返回相关信息实体，例如：利用表 1 中 *isbn* 号 “978-0123814791”或该元组的其他属性值作为搜索引擎关键词搜索时，只能得到一系列的文档而不能直接获取到“Morgan Kaufmann”. 很明显，现有的搜索引擎无法直接满足信息错误自动检测与修复平台的需求.

(3) 关系型信息的特征会给利用 Web 进行信息扩展带来困难. 关系型信息具有如下特征：无法通过关系模式获取属性之间的语义上下文关系；数据库中属性间的约束关系不等同于 Web 中这些属性之间的约束关系；数据库中对于实体的属性的描述可能与 Web 中不一致. 这些特征都会给基于 Web 信息的信息扩展技术的研究带来困难，例如：在获取表 1 中 *isbn* 值为“978-0123814791”的元组的 publisher 值时，尽管 *isbn* 值可以作为关系表的主键，但是其不一定是最佳的搜索关键词.

(4) 信息质量问题种类多样. 完整性、一致性、实体同一性、精确性和时效性这 5 类问题每类问题具有不同的特点，例如：完整性错误的修复关键在于寻找合适的缺失值，而时效性错误关键在于借助辅助时间信息判定信息的有效性. 另外，这 5 类问题常常相互交杂，例如：合适的值可能也有不同的时效和不同的精度. 因此，为这 5 类问题建立相对统一的信息扩展平台并且结合扩展的 Web 信息和关系数据中的信息进行准确的信息错误检测与修复是一件非常有挑战性的工作.

## 3　研究框架

根据图 2 可知，基于 Web 信息的关系型信息错误自动检测与修复技术需要解决如下问题：(1) 如何从 Web 中获取改善关系数据库信息质量所必需的信息？(2) 如何结合所获取的 Web 信息进行有效的信息错误检测与修复？(3) 如何自动验证信息错误自动检测与修复算法的可靠性？

针对上述问题，基于 Web 信息的关系型信息错误自动检测与修复技术体系可划分为如下 4 个方面：Web 信息自动拓展模型、基于 Web 信息的信息

《计算机学报》

错误自动检测算法、基于 Web 信息的信息错误自动修复算法及基于 Web 信息的信息错误自动检测与修复可靠性评估方法. 其中,Web 信息自动拓展模型是其它几个方面的基础技术. 各内容之间的支撑关系如图 3 所示. 各部分具体的研究内容如下:

(1) Web 信息自动拓展模型. 该部分研究利用 Web 自动判定关系型信息属性间的上下文语义关系的方法;研究利用所获得的属性值之间的上下文关系及已知信息与拓展信息之间的数据依赖关系等进行准确的信息实体获取的方法与技术.

(2) 基于 Web 信息的信息错误自动检测算法. 该部分研究基于 Web 信息的信息相对和绝对时效性的判定算法、信息绝对精确性的判定算法、属性值同一性的判定算法;研究基于 Web 信息的数据依赖规则的挖掘算法及基于相关规则的元组同一性判定算法、信息一致性的判定算法.

(3) 基于 Web 信息的信息错误自动修复算法. 该部分研究基于 Web 信息的属性值缺失错误的自动修复算法、信息相对和绝对时效性的修复算法、信息绝对精确性的修复算法、信息一致性的修复算法.

(4) 基于 Web 信息的信息错误自动检测与修复算法的可靠性评估模型. 该部分研究基于 Web 信息的信息完整性、一致性、精确性、时效性、实体同一性错误的自动检测与修复算法的可靠性评估模型,并且建立有效的机制对信息错误自动检测与修复算法的可靠性进行自动评估.

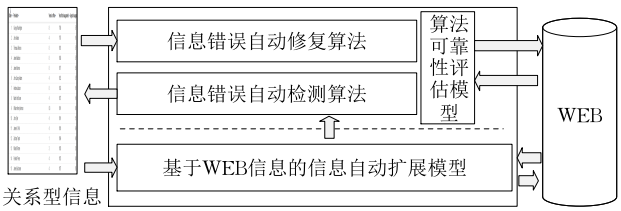


图 3 基于 Web 信息的关系型信息错误自动检测与修复技术框架

## 4 相关研究进展

目前国际上在基于 Web 信息的信息错误自动检测技术、基于 Web 信息的信息错误自动修复技术和基于 Web 信息的信息扩展核心技术三个方面已开展了初步的研究并取得了一些研究成果. 本节将详细分析相关研究进展.

### 4.1 基于 Web 信息的信息错误自动检测技术

信息错误检测的本质是对信息的正确性进行验证. 根据信息之间是否存在明确的语义关联关系,信息可分为两类. 一类信息之间存在明确的语义关联关系,称之为"连续型"信息. 例如:"Barack Obama is a Muslim"这条信息中"Barack Obama"和"Muslim"之间存在明确的语义上下文关系,所以此类陈述性信息为"连续型"信息. 另一类信息之间没有明确的语义关联关系,称之为"离散型"信息,例如:表 1 中属性 *isbn* 和 *publisher* 之间语义关联关系并不明确或者不易被自动识别,所以此类信息为"离散型"信息. 据此,本文将信息正确性验证技术划分为"连续型"信息的正确性验证技术和针对"离散型"信息的正确性验证技术.

#### 4.1.1 "连续型"信息的正确性判定技术

文献[33]研究了利用 Web 对陈述性声明信息进行正确性验证的方法. 其中将陈述性信息划分为主题单元(Topic Units)和质疑单元(Doubt Units). 例如:对于信息"Barack Obama is a Muslim","Muslim"是质疑单元,"Barack Obama"为主题单元. 文献[33]提出了一个两阶段的方法:(1)首先根据所质疑的陈述的主题单元从 Web 中获取多个质疑单元的候选值并根据这些候选值构成关于同一话题多个候选陈述,例如"Barack Obama is a Muslim"、"Barack Obama is a Christian";(2)根据候选陈述构造 Web 查询,然后根据 Web 搜索结果对所有候选陈述进行排序,最后从这些候选陈述中选择正确的可能性最高的陈述,从而完成对相关陈述的正确性验证. 例如,对"Barack Obama is a Muslim"进行正确性验证时,可分别将候选陈述"Barack Obama is a Muslim"、"Barack Obama is a Christian"作为查询,利用搜索引擎获取 Snippets,然后根据所获取的 Snippets 信息对这两个候选陈述进行排序. 本例中"Barack Obama is a Christian"的排序高于"Barack Obama is a Muslim",所以可以判定"Barack Obama is a Christian"比"Barack Obama is a Muslim"正确的可能性更大,从而可以判定"Barack Obama is a Muslim"是错误的. 在获取候选陈述和对候选陈述进行排序时,文献[33]考虑了结果的覆盖率、结果与查询的相关性、与主题单元的距离、类型的匹配程度、语义的相关度、返回网页的个数等多种因素,并且利用 TREC-8 和 TREC-9① 作为测试集验证了其方法的有效性. 该方法只支持对一个质疑单元的验证,而实际应用中可能存在多个质疑单元. 另外,由于主题单元信息的有限性(例如没有时限、精度限制

① http://trec.nist.gov/data/qamain.html

等),正确的值可能有多种,例如 1992 年诺贝尔医学奖是由两个人共同获得的、多个不同的歌手都演唱过同一首歌曲、多个人曾在不同的时期担任过美国总统等.对于此类情况文献[33]中都没有涉及.

文献[34]介绍了两类质疑单元包含多个真值的情况.一类是相容性概念(Compatible Concepts),即一些候选值是另外一些候选值的泛化或者包含这些候选值,例如"动物"是"鱼类"的泛化,"陕西"包含"西安"等;另外一类是多值属性,即多个候选值共同构成一个完整的答案,例如 Sauvage、Stoddart 和 Feringa 共同获得 2016 年诺贝尔化学奖.针对多值问题,文献[34]提出了两类解决方案.一类是首先通过类似于文献[33]的方法获取排名 TOP $K$ 的候选值,然后利用这些候选值与排名第一的候选值之间的关系筛选出所有合适的候选值.另一类是根据候选值之间的相关性对它们进行分组,选取排序时评分总和最大的一个组作为最终的候选值.文献[34]在计算相关性时考虑了候选值之间可能存在的同义、泛化与实例、局部与整体以及相关等关系.文献[34]对来源于 TREC-8、TREC-9 和 TREC 2001 的 50 个质疑陈述进行了正确性验证,验证的准确率超过了 80%.虽然文献[34]的方法可以处理多候选值的情形,但是其必须要求事先指定质疑单元.另外,文献[34]在计算评分时并没有考虑信息源的可信度.

针对文献[33]必须指定质疑单元以及无法处理多候选值局限,文献[35]提出了一种四阶段的正确性验证模型,其核心思想是基于"可信信息源提供的信息更正确"这一假设通过衡量 Web 信息源的可信度来判定陈述性信息的正确性.4 个阶段主要包括:(1)将待验证陈述作为查询利用搜索引擎获取排名 TOP $K$ 的所有结果;(2)计算搜索结果和待验证查询之间的语义包含关系;(3)计算搜索结果的可信度;(4)根据搜索结果的可信度和对查询的支持度判定陈述的可信性.计算搜索结果 $r_i$ 和待验证陈述 $f_s$ 之间的语义包含关系时,文献[35]通过 Stanford Praser[36] 分析搜索结果中单词的语法关系,选取所需的关键词,继而计算搜索结果对待验证查询的支持度 $Sup(r_i, f_s)$.在衡量 Web 网页可信度 $Crank_i$ 时,文献[35]主要考虑了 Web 网页的重要性(Importance)和受欢迎的程度(Popularity)这两个因素.在判定待验证陈述的正确性时,首先计算相关的搜索结果 $S_{pos}$ 和中立搜索结果 $S_{neu}$ 的 $\Delta$ 函数值.其中 $\Delta_{pos}=\sum_{r_i \in S_{pos}} Sup(r_i, f_s)/Crank_i$.文献[35]坚信如果陈述为真,则 $\Delta_{pos}/\Delta_{neu}$ 一定大于某一个阈值并据此判定陈述的正确性.文献[35]将陈述作为一个整体进行考虑,所以适用于"English is the official language of Philippine"此类一个陈述有多个答案并且不指定质疑单元的场合.实验结果显示其判定准确率可以达到 80%以上.但是,文献[35]中只考虑信息源的可信度这一个因素,并没有考虑陈述性信息之间的关系及 Web 信息源所提供的信息时效性、一致性以及信息源之间引用关系等因素,信息判定的准确度还有很大的提高空间.

文献[37]提出了一种领域无关的事实陈述可信性判定模型.该模型基于"可信信息源更可能传递可信的信息"这一假设,试图通过衡量 Web 搜索引擎返回的网页的可信度来评估事实陈述的可信度.其中,首先计算事实陈述与搜索引擎返回结果中的相关文本信息之间的语义相似度,然后通过计算相关文本信息所在网页的可信度和受欢迎程度来评估相关文本信息的可信度,最后根据相关文本信息之间的语义相似度及可信度来评估事实陈述的可信度.因为将事实陈述作为一个整体来考虑,所以献[37]可以回避文献[33]必须指定质疑单元以及无法处理多候选值的限制.实验结果表明文献[37]中的方法的准确率可以达到 79%.但是因为自然语言的多样性,很难从 Web 文档中准确抽取与事实陈述语义密切相关的陈述,所以文献[37]的判定准确率并不是很高.

大多数基于 Web 的正确性验证方法的一个重要前提"由更多信息源提供的信息更准确".由于 Web 的开放性,Web 信息本身会产生冗余、不一致、不同信息源的准确度不同、相互引用和拷贝等问题.文献[38]通过对航班和股票数据的分析验证了上述结论.错误信息可能会在信息源之间拷贝以及数据会过期,这些原因导致错误的信息可能比正确的信息分布更广.针对不同信息源的影响,通过直观感觉和对真实 Web 数据的分析,文献[30]总结了 4 点启示:(1)一个对象某个属性只有一个正确的值;(2)这个正确值在不同的 Web 信息源中是相同或者相似的;(3)错误值在不同的 Web 信息源中一般是不同的;(4)在一个特定领域,如果一个网站提供的很多对象的信息都是正确的,那么它提供的其他信息正确的可能性也比较高,即如果一个网页提供的绝大多数信息可信,那么该网页更可信,可信网页提供的信息也更可信.基于上述启示,文献[30]构建了一种利用 Web 网页可信性及其与信息实体正确性之间的

计算机学报

关系从冲突的 Web 信息源中获取正确的“事实”以及识别可信网页的迭代模型. 其中,网页的可信度等于其所提供的“事实”的平均可信度. “事实”的平均可信度则由支持它的信息源以及其与其他“事实”之间的关系共同决定. 模型中,网页的可信度的初始值为服从正态分布的随机值. 文献[30]在图书作者以及人工数据集上识别准确率均达到了 90%. 该方法主要关注了信息源的质量对正确信息获取的影响,但是对于影响 Web 信息正确性判定的其他因素没有涉及.

文献[39]提出了一种半监督式的正确信息的获取方法,其策略是:首先人工判定少量的陈述,然后借助这些已判定的陈述来推导其他陈述的正确性. 在判定陈述信息的正确性时,文献[39]利用了陈述信息之间的 3 种关系:(1)相互印证,即如果一个陈述为真则另外一个也为真,例如“Google has 21000 employees”和“Google has 21500 employees”相互印证;(2)相互排斥,即如果一个陈述为真则另外一个陈述必为假,例如若“Microsoft was founded in New Mexico”和“Microsoft was founded in Washington”之中一条为正确信息则另外一条必为错误信息;(3)同源同质,即若某一个信息源提供的大多数信息是正确的,那么它提供的所有其他信息正确的可能性就比较大. 文献[39]中将信息正确性判定问题转换成了图学习的问题. 其首先将陈述信息和信息源组织成一个如图 4 所示的图结构,图中的顶点表示信息以及信息源,边表示信息之间以及信息和信息源之间的关系;然后,基于上述陈述之间的 3 种关系来推算未知信息的置信度. 例如:图 4 中,如果人工标定陈述①为真,因为陈述①和陈述⑦相互排斥,所以陈述⑦不可信,置信度需设置为−1;而陈述⑦和陈述⑥来源于同一信息源 $D_3$,所以陈述⑥也不可信. 实验表明:文献[39]在图书作者数据集上判定的准确率达到了 91%和在 Web 数据集上判定的准确率达到了 78%～96%. 上述基于图模型的方法为 Web 信息拓展提供了很好的借鉴. 但是因为同一个信息源不同部分信息质量有可能是不同的,信息可能存在多种正确的值. 这些特点文献[39]没有涉及. 该类方法的准确度尚有较大的提高空间.

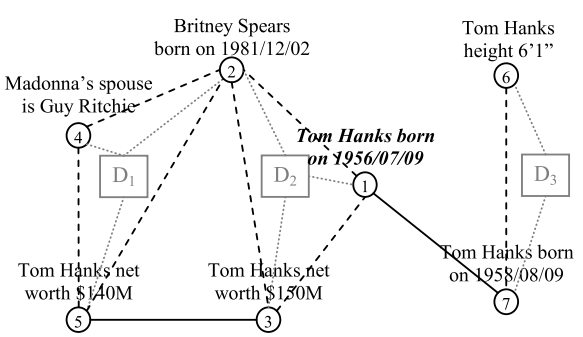


图 4 信息、信息源关系图[39]

4.1.2 “离散型”信息的正确性判定技术

关系数据是典型的“离散型”信息. 关系型信息的正确性判定主要是包含完整性、一致性、实体同一性、精确性和时效性这 5 类问题的判定. 目前,基于 Web 信息的关系型信息正确性判定技术主要集中在实体同一性判定问题上.

实体同一性判定又称重复实体检测或记录匹配. 它是信息质量方面研究最多的领域. 文献[40]将实体匹配技术划分为两大类:一类是统计学和人工智能领域所采用的分类的方法;另外一类是数据库领域所采用的基于规则的方法. 文献[40]认为分类的方法扩展性差,因为它需要将所有的元组两两配对以进行分类,不适用于大规模的数据. 基于规则的匹配方法也存在 3 个问题:一是产生匹配规则的时候会有些限制,例如必须假定某一个属性值相同或者相似;二是不能直接适用于近似匹配的场合;三是没有解决在近似匹配的场合“多相似才表示同一?”的这样一个问题. 针对第 3 个问题,文献[40]提出了一种非常有效的选择相似度函数的机制. 但是,现有的实体匹配方法中相似度函数大多是用于计算文本相似度,对于语义相似度尚少涉及. 因为自然语言灵活性,现实世界中对同一个对象存在不同的描述形式,另外缩写、简写、拼写错误使得同一性判定的问题变得更加复杂,例如“计算机”和“电脑”、“ICDE”和“International Conference on Data Engineering”以及“Jiawei Han”和“Han Jiawei”等. 此时,通过计算文本相似度来判定属性或者元组的同一性将会产生较大的偏差. 当数据库中包含的信息有限时,仅仅利用数据库中的信息很难判定语义相似度. 基于此,文献[41-45]中提出了几种基于 Web 信息的实体同一性判定方法.

文献[41]认为如果一个实体 $e$ 是另外一个实体 $e_c$ 的副本,那么它们一定会同某一组信息 $data(e_c)$ 在 Web 中频繁地一起出现. 如果“John McCarthy”和“McCarthy, N. J.”与“LISP”或“time sharing”在 Web 中共同出现的频率都非常高,则“John McCarthy”和“McCarthy, N. J.”是同一个人的可能性非常高. 基于这一前提,文献中提出了一种基于 Web 信息的实体识别方法. 该方法的步骤为:首先将 $e_c$ 作为查

询关键词获取相关文档，然后采用基于 TF-IDF 的指标体系提取 $data(e_c)$，再分别将 $e$ 和 $data(e_c)$、$e_c$ 和 $data(e_c)$ 作为查询关键词输入搜索引擎，计算两组查询返回结果的相似度. 文献[41]证实利用 Web 信息可以有效提高实体同一性判定的准确率.

单词的相似度计算是实体同一性判定的一项关键技术. 文献[42-43]提出了一种借助 Web 计算单词语义相似度的方法，其算法框架如图 5 所示. 大致过程为：首先把两个待比较的词作为查询关键词输入搜索引擎，然后根据返回的相关 Web 文档数量和文档中抽取的语法模式这两类因素计算两个单词的相似度. 文献[42-43]中提出了 WebJaccard、WebOverlap、WebDice 和 WebPMI 共 4 种基于 Web 文档数量的单词相似度的计算函数和从 snippets 中提取待比较单词之间语法模式(如：is a)及将这些语法模式分类的方法. 文献[42-43]中将所划分的 $N$ 个语法模式类和 4 种基于 Web 文档数量的相似度值作为一个 $(N+4)$ 长度的向量的维，然后用支持向量机模型将两类因素集成以判定两个单词的语义相似度. 训练支持向量机时，文献[42-43]借助 WordNet[46] 手工产生一系列同义和非同义词作为训练集. 实验结果显示：文献[42-43]中的方法较当时已有大多数方法准确率高，在两个不同数据集上分别达到了 74%和 86%.

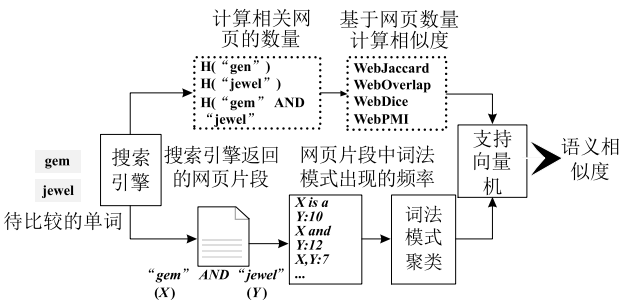


图 5　基于 Web 的单词语义同一性判定算法框架[43]

简称并没有统一的规则，例如"MOU"这个简称包括介词"of"，"MIT"却不包含介词；"ICDE"是由全称"International Conference on Data Engineering"的首字母构成，而"UbiComp"却是由"The ACM International Joint Conference on Pervasive and Ubiquitous Computing"最后两个单词前 3 个字母构成. 上述原因使得在缺乏上下文信息时，简称和全称的对应变成一个非常复杂的问题. 文献[44]提出了一种利用 Web 信息以助于简称和全称对应的方案. 其方法为：将简称或者全称作为查询关键词，然后通过分析借助搜索引擎获取的相关 Web 文档中的全称或简称出现的频率来判定简称和全称的对应关系.

文献[45]认为现有的基于语料信息(主要为 Web 信息)的实体同一性判定方法还存在如下问题：(1)受搜索引擎索引及排序机制的影响，判定的结果会存在一定的偏差；(2)一些方法，例如文献[44]，在进行相似度计算时需要和搜索引擎交互，存在比较大的通信开销，不适用于在线的应用；(3)在上下文中以词或者 N-GRAM 统计分布忽略了一个语义单位可能是由多个词构成的以及单词或者短语本身存在歧义性这一状况；(4)主要关注的是两个单词或者短语在 Web 文档中是否临近出现，这种模式更侧重于计算其语义相关性而非相似性. 文献[45]尝试了从一个源于 Web 文档的概率语义网络 Probase[47] 中直接提取上下文信息以提高处理的效率和准确度. 相对于原始的 Web 文档，Probase 已存储大量语义上下文信息. 所以相对于现有方法，文献[45]的方法效果更好，在 NASDAQ 数据集上的准确率达到了 99%左右.

在不同数据库中，姓名可以有多种不同的描述形式，例如：Han Jiawei、Jiawei Han 和 J.W. Han 等. 这些因素使得姓名的同一性判定变成一件非常困难的工作，也引起了很多学者的兴趣. 文献[48]对已取得的进展进行了概要的综述. 借助 Web 信息来进行姓名同一性判定的工作主要包括文献[49-54].

假设需要判定"Mark Orey and David Miller, Diagnostic Computer Systems for Arithmetic, Computers in the School, volume 3, #4, 1987"和"Miller, D., Atkinson, D., Wilcox, B., Mishkin, A., Autonomous Navigation and Control of a Mars Rover, Proceedings of the 11th IFAC Symposium on Automatic Control in Aerospace, pp. 127-130, Tsukuba, Japan, July 1989."这两条文献中的"David Miller"和"Miller, D."是否为同一个人. 很明显，两条文献中待判定的作者描述形式是不同的. 另外，文献中也没有其他共同的作者，论文题目和出处也不同. 这些因素使得作者的同一性判定变得非常困难. 文献[49]试图从 Web 中获取扩展信息来应对上述困难. 其中将所有的作者组织成一个图，图中的顶点代表包含待判定作者的文献引用，边的权表示两个作者同一的概率. 其核心思想为：将两篇文献的文献名组合起来构成一个查询关键词输入到搜索引擎中，然后采用两种策略来更新图中边的权值. 一种是：如果搜索引擎返回相关的结果则更新这两篇文献之间的权值. 另一种为：将返回结果中的一个 Web 网页作为新的节点加入到图中，并在两篇文献和新的节点之间建立新的边，然后借助这些新的

顶点判定作者同一的概率.第 1 种策略在 DBLP[①]、Penn 和 Rexa 等数据上判定的准确率分别达到了 90.5%、87.7%和 91.8%,第 2 种策略 DBLP 数据集上的准确率达到了 88.2%.其实验还表明:文献中的优化策略可以使得通过 1%的查询获得其中 53.5%的结果.

文献[50]尝试通过 Web 获取待判定对象的合作作者集合,然后通过判定合作作者集合的相似度来判定待判定作者的同一性.其中使用了 1999 年~2006 年韩国的一些出版文献作为测试集,结果显示其 F1 值可以达到 0.85.

文献[51]则尝试通过 Web 获取作者的个人主页和简介等信息,然后借助这些信息来进行同一性的判定.其基本思想为:如果一组待判定的文献出现在同一个人的个人主页中,那么待判定作者为同一人的概率就比较高.其处理过程为:(1)为每一个待判定文献引用构造查询提交给搜索引擎,选取排名前 n 的文档构成文档集合;(2)从文档集合中选取描述待判定作者个人信息的文档,并且根据 IMF[55] 值为这些文档赋予不同的权重;(3)借助获取的文档集合,对待判定的文献引用进行分组,使得每个聚类代表一个唯一的作者.DBLP 上的实验结果显示,文献[51]的 K、pairwise F1 和 cluster F1 的指标分别为 0.8、0.76 和 0.14.除了利用"同一个作者的文献一般在会在同一个主页中出现"这一信息外,文献[52]还假设同一个作者关注的主题是相关的,并结合了这两个因素来判定文献作者是否同一.文献[52]中判定的平均准确率可以达到 75%.

上述研究主要关注的是文献作者这一特殊属性的同一性判定问题.因为利用 Web 中可以获取合作作者、作者主页的文献列表和文献主题等诸多特定信息,所以相对来说作者的同一性比较容易判定.因为可利用的信息相对较少,所以其他类别的信息的同一性判定要困难很多,目前相关工作尚少.文献[53]关注了作者单位的同一性判定问题.文献[53]注意到:文献中作者单位的表述并不统一,包括:(1)简称全称不统一,例如"M. I. T."和"Massachusetts Institute of Technology";(2)地址的详细程度不一致,例如:有的包含院系甚至街道和邮箱号等具体的地址,有的只有机构名称等.文献[53]尝试将一组作者单位分别作为关键词输入到搜索引擎,通过计算返回结果中 URL 的相似度来判定作者单位的相似度.尽管只考虑了 URL 这一个因素,文献[53]中的方法相对于诸如编辑距离的传统字符串相似度计算方法要更有效.

上述工作都证明:利用 Web 信息可以有效的提高实体同一性判定的准确率.但是获取 Web 信息本身是比较耗时的工作,搜索引擎对返回的结果也有一定的限制.所以用尽可能少的代价获取所需的信息就变得尤为重要.文献[54]针对实体匹配问题提出了一种代价敏感的 Web 信息获取模型.该模型利用信息源之间的依赖关系来减少不必要的 Web 查询.实验证明文献[54]中的方法可以减少 90%的 Web 查询.

#### 4.1.3 小 结

总体而言,现有基于 Web 信息的信息错误检测技术相关工作主要集中在陈述性信息的正确性验证和关系型信息同一性判别.现有研究成果中对关系型信息正确性验证尚少涉及.尽管陈述性信息的正确性验证可以给关系型信息正确性验证提供很好的借鉴,但是相关方法还存在如下问题:正确性验证时没有考虑质待验证的信息时效性、精确性以及实体同一性等问题.离散型信息属性值语义同一性判别工作也只是进行了一些的尝试,其处理速度及判定的准确度都有待提高.目前尚未发现基于 Web 信息的关系数据信息的一致性、时效性和精确性检测方法.

### 4.2 基于 Web 信息的信息错误自动修复技术

#### 4.2.1 完整性错误修复

现有的大多数完整性错误修复算法仅仅利用数据库中的信息本身的分布特征来进行计算[56]或者借助数据之间的关系来推导[57]出一个近似值来进行修复,修复的目的着重于消除数据缺失带来的影响而非获取数据的真实值.当源数据中包含的信息较少时,仅利用关系表中的信息的方法将会变得失效,例如很难从表 1 推导出 *isbn* 值为"978-0123814791"的元组的 *publisher* 值.鉴于此,有学者开始关注引入外部信息来进行缺失属性值填补,代表性的工作有文献[58-66]等.

为了应对训练信息不完整对所构建的贝叶斯网络精确性的影响,文献[58]提出了一种从 Web 获取缺失属性值的方法.其处理过程为:首先人工构造查询,然后利用 Google 获取选取相关文档,最后根据已知属性和不完整属性之间的关系构建信息抽取模式利用信息抽取工具 WHISK[67]从相关文档中获取所需信息.其中主要考虑了点概率和定性的影响这

① http://dblp.uni-trier.de/xml/

两类已知属性和不完整属性之间的关系. 文献[58]认为：相对于基于统计学的方法，基于 Web 信息的属性值缺失错误修复方法可以有效应对信息缺失率较高的情况. 文献[58]中实验结果表明：对于部分参数，填补的准确率可以达到 60%以上. 该方法侧重于发现数值信息，比较适用于分类器的构建. 另外，它是一种半监督式的方法，查询关键词由领域专家人工生成，不能完全实现自动化.

文献[59]提出了一个基于 Web 信息的关系数据缺失属性值填补框架. 文献[59]认为基于 Web 信息的完整性错误解决方案需要解决 3 个核心问题：(1) 如何根据已知属性值构建查询以助于有效地获取缺失的属性值?(2) 如何从获取的候选属性值中选择一个“最合适的”值?(3) 存在多个缺失属性值时如何确定缺失属性值填补的顺序?文献[59]还认为构建查询的方法和信息抽取的方法是密切相关的，为此提出了两种 Web 信息获取的方法，分别为 P-DI 和 C-DI. P-DI 方法的核心思想为：基于完整的元组来学习语法模式，然后根据该语法模式为存在缺失属性值的元组构造查询并从查询结果中抽取候选值，例如：根据 (“Jack M. Davis”, “jdavis@mit.edu”) 或 (“Tom Smith”, “tomsmith2@cs.cmu.edu”) 构造查询，从搜索引擎返回的文档中可以抽取 Name 和 Email 之间的语法模式“send email to [NAME] (Email: [EMAIL])”，利用该模式构造查询“send email to Bill wilson (Email:”+“)”来获取 Bill Wilson 的 Email 值. 该方法的弊端为：相关的语法模式过于严格，会导致从 Web 获取候选值有限，甚至无法获取到合适的候选值. C-DI 方法的核心思想为：根据完整元组中的信息学习两个属性之间的通用的上下文关系，然后利用该通用上下文关系构造查询及获取候选信息，例如：基于 (MIT, MA)、(CMU, PA) 和 (Yale, NY) 学习 (University, Location) 之间诸如“locate at”之类的通用上下文关系. 与 P-DI 不同的是：C-DI 中只需要抽取两个属性之间最频繁出现的上下文而非严格的语法模式. 两个属性之间可能存在多种语法模式或者上下文关系，所以可能会产生多种查询. 文献[59]中采用支持度和覆盖率这两个指标对查询进行评估，并且根据查询的置信度计算候选值的置信度. 对于一个元组存在多个缺失属性值的场合，可以为每一个缺失属性值构造查询获取候选字进行填补. 但是因为已知属性值不一定是最适当的“种子”，所以这种方式不能够获取最合适的候选值. 鉴于此，文献[59]采用迭代的方法来进行缺失值的填补，其策略为：每一轮采用已知属性值构造查询来进行填补，填补结果保存在数据库的一个中间实例中；下一轮填补时，新填充的属性值中的一些置信度高的值会被用来作为“种子”来重新填补一些已经填充过的缺失值. 迭代的算法框架如图 6 所示. 文献[59]实验结果显示：相对于传统的基于规则的近似方法，文献[59]中的方法填补的准确率更高，其在 PersonInfo 和 DBLP 这两个数据集上都达到了 70%～80%的准确率.

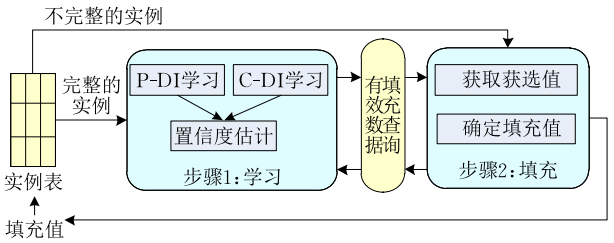


图 6　文献[59]迭代算法框架

在计算查询置信度时，文献[59]会考虑所有的已知属性值组合并且使用所有的完整元组作为训练集，这种方式代价高昂. 针对该问题，文献[60-61]提出一些优化策略用于减少训练元组的个数和每个元组所产生的查询的个数以降低估计填充查询置信度的代价. 在获取语法模式时，文献[59]需要基于完整的元组来评估所有查询的置信度. 对于每个属性来说，可以使用多种已知属性的组合来构造查询. 因此，在所有完整元组上对所有的缺失属性的所有的组合进行评估代价高昂. 文献[60]认为查询评估的代价主要由两部分构成. 一部分是元组内的代价，即为获取某一个缺失属性所构造的查询数目. 另一部分为元组间的代价，即基于完整元组计算某一个查询的支持度和覆盖率时，该查询被调用的次数. 文献[60]提出了优化策略以减少这两部分的代价，其核心思想为：通过抽样减少训练元组的个数和尽早删除置信度低的查询来提高元组内的评估代价，利用语法模式之间的包含关系、查询中所使用的属性的包含关系以及属性之间交换性等性质制定规则以减少元组间的代价. 文献[60]中的实验表明：采用元组间的优化策略，代价可以减少 45%～50%；采用元组内的优化策略，代价可以减少 80%；两种策略都采用的话，代价可以减少 85%～92%. 文献[61]中的方法则着重于减少元组间的代价，即减少查询被调用的次数. 其核心思想为：(1) 只有当一个答案比其他候选答案的置信度都高时，才能作为被选中的候选值. 因此，可以基于已调用查询的查询结果和未

计算机学报

调用查询的数量及置信度计算每一个候选答案的置信度上界和置信度下界.如果发现一个答案的置信度下界比其他所有答案的置信度上界都高的话,则可将其作为最终的候选值,其余查询不必调用;(2)针对相同的缺失属性的具有类似格式和包含相同属性值组合的查询在不同训练元组上表现是相似的,所以如果一个查询在一组训练元组上无法获得目标属性值则在其他训练元组上也无法获得,例如:如果使用"Title+Name+Zip"组合无法获得某一组训练元组的 Street 值,那么使用该查询模式获取其他训练元组 Street 值的可能性也非常低.因此可以预先识别并且剔除这些查询.实验证明,采用文献[61]中的优化策略,查询执行的次数比文献[59]减少了85%以上.

因为并非所有的缺失属性值都可以从外部信息源中获取到,所以文献[62]提出了一种混合式的方法,试图结合基于 Web 的方法和基于机器学习的方法来提高填补算法的覆盖率.但是,文献[62]中使用的 Web 数据仅仅局限于 Web 表格.因为文献中缺乏混合式的方法的实验数据,所以其效果还有待确认.

在整合不同的数据库时,需要找出两个数据库属性之间的映射关系.文献[63]提出了一种利用 Web 补充两个实体集中的实体映射关系的方法.该方法利用已知的实体映射关系作为导引借助 Web 挖掘两个实体集的映射上下文关系模式,然后根据上下文关系模式从 Web 中获取未知的实体映射关系.其中首先结合基于术语频率、位置邻近和甄别信息的方法提出了一种上下文关系抽取模型,然后基于该抽取模型提取合适的上下文关系来构建查询.构建查询时,采用了一种基于上下文关系树的查询构建方法,以达到只用调用少部分查询即可获取有效的映射关系.实验表明:在一系列真实的数据集上基于 Web 的方法相对于基于模式的方法准确率要高 50%以上.尽管不同关系表映射关系的补全与单一关系表缺失属性的填补的应用场景有所不同,相关方法还是可以为单一关系表信息完整性错误的修复提供很好的借鉴.

基于 Web 信息的缺失属性值填补技术在获取目标属性值时可能会得到多个候选属性值,合理的候选属性值的排序算法就显得尤为重要.常规的排序方法是累加多种因素的影响,如候选属性值出现的频率、候选属性值所在网页的排名、查询关键字的覆盖率、候选属性值与关键字的文本距离以及候选属性值的文本模式等.每一种因素的权重通过训练获得.此类方法容易受制于因素的选择策略,需要特定的领域专家参与.文献[64]认为,候选属性值所代表的实体与它的周边实体是相互影响的,并提出了用网页片段的信息来构建实体图,以随机游走的方式来传递实体间的影响进而对候选属性值进行排序.由于 Web 上的文档大多是非结构化的,并且通用搜索引擎只是提供相关文档,并不保证是语义匹配.针对上述问题,文献[65]提出了基于贝叶斯概率估计的候选属性值排序框架,它将多种特征以语义推理的方式融合到一个统一的概率评估框架.该排序框架包括网页影响力模型和语义匹配度模型,其中,语义匹配度模型用于评估网页片段中的候选属性值与存在缺失值的元组的语义匹配度.该模型又可细分为上下文匹配概率和属性值匹配概率的计算,计算框架分别如图 7、图 8 所示.

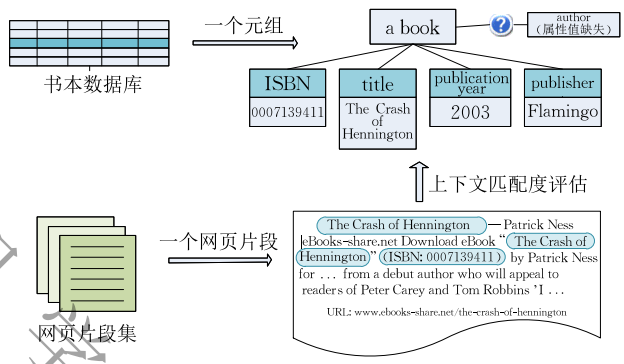


图 7 上下文匹配概率计算框架[65]

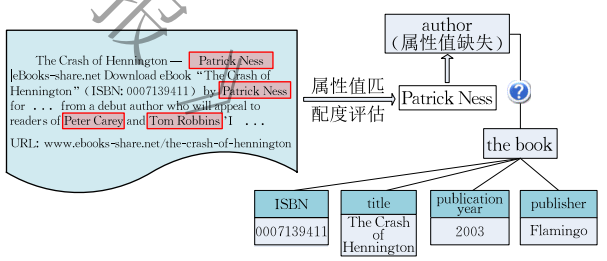


图 8 属性值匹配概率计算框架[65]

仅依赖数据库自身信息的缺失属性值填补技术的查全率(recall)比较低,基于外部信息的缺失值填补技术查全率高但是效率较低.文献[66]探讨了结合仅依赖数据库自身信息进行推导的缺失属性值填补技术和从外部信息中获取缺失属性值的填补技术以提高查全率缩短填补时间的可行性,创新性提出了一种混合式的缺失属性值填补方案.其策略是交替采用这两类方法以达到通过最少的查询来获得最高的查全率的目标.文献[66]的实验表明:相对于纯粹基于规则推导的方法,基于 Web 信息获取的方法具有更高的查全率和查准率;结合基于规则和基于

Web 信息获取的方法可以在获取高的查全率和查准率的同时提高处理效率. 尽管文献[66]中只考虑了函数依赖和条件函数依赖这两种推导方案，其中混合式的修复思路为我们开展完整性错误修复技术的研究提供很好的借鉴.

4.2.2　精确性错误修复

用户在表述地理信息时往往是不精确的，例如‘Midwest in the US’. 对于此类不精确的地理信息描述，不同的人会有不同的解读. 文献[68-71]尝试了借助 Web 信息来自动界定不精确地理区域边界.

文献[68]的策略是：利用 Web 获取候选地理位置以及这些地理位置之间的关系，然后根据这些位置点建立核密度图，从而确定不精确地理位置的范围. 其中地理位置之间的关系通过“A is a town in B”之类信息抽取模式从 Web 文档中提取.

文献[69]的思路为：(1) 首先利用搜索引擎从 Web 中获取该区域内的地理位置点，例如城市和小镇等. 这些位置点中可能包含噪声；(2) 结合地理知识库发掘这些位置点的坐标，并将其都标红色，然后根据这些位置点的坐标信息划定一个区域的长方形边界；(3) 在地理信息库中查找长方形边界内的其他位置点，这些位置点在 Web 中没有搜索到，所以标定为蓝色；(4) 寻找一个多边形使得尽可能多的红色的点在它的内部，尽可能多的蓝色点在它的区域外. 该方法中，Web 中的信息为非精确地理位置的界定提供了有力的支撑. 文献[69]在 Triger 中定义了区域和地理位置点的上下文模式，基于 Triger 中模式从 Web 中获取与区域相关的地理位置点. 其地点辨识的准确率为 Wales 地区 88%、Midlands 地区为 81%、South East 地区为 76%以及 East Anglia 地区为 59%. 文献[70]提出了两种方法以寻找一个更加合适的多边形把文献[69]中的红色点和蓝色的点准确地区分开来.

因为自然语言描述的多样性，文献[69]中所定义的 Triger 不能覆盖所有的上下文模式，可能会导致部分地理位置没有被有效获取. 文献[71]认为这种方式不够有效转而采用命名实体识别的方法来获取地理信息点. 但是因为 Web 网页中的信息存在噪声和偏差，搜索引擎返回的信息可能是不准确的. 另外，地理位置名称也存在歧义，例如：(1) 同一个名字表示不同区域中不同地理位置点；(2) 一个地理位置点有多个不同的名字；(3) 一个名字既可以表示地理位置也可以表示一个组织的名称或者人名[72]. 这些问题，文献[69,71]都没有提出通用有效的解决方案. 鉴于以上原因，上述方法判定的精度尚有较大的提高空间.

4.2.3　小　结

总的来说，现有基于 Web 信息的信息错误修复技术相关工作主要集中在信息完整性错误的修复技术和地理位置信息精确性的修复技术. 影响这些技术准确率的关键因素在于能否通过已知信息获取所需要的缺失或者辅助信息. 目前尚无较好的基于 Web 信息的信息一致性错误、时效性错误以及除地理位置以外的其他信息精确性的修复方法.

### 4.3　基于 Web 信息的信息扩展核心技术

针对 Google、Yahoo 和 Bing 等主流搜索引擎的主要目标是获取相关文档而非相关实体的现状，文献[73]提出了利用搜索引擎直接从 Web 中获取相关实体的构想并且提出了一种基于对象链接关系分析的实体排序算法——PopRank. PopRank 算法中同时考虑了对象在 Web 中的流行程度及对象之间的关系. 与 PageRank[74] 算法不同的是 PopRank 算法中认为对象之间链接关系的语义和重要度是不相同的，为此 PopRank 算法为每一个链接引入了一个流行度传播因子(popularity propagation factor)并且为不同的链接关系(如：cited-by、authored-by 和 published-by)赋予不同的传播因子. 其中 Web 对象的流行度的计算公式为

$X$ 的流行度 $=\varepsilon\times X$ 的流行度 $+(1-\varepsilon)\times$ 其它所有邻接对象对 $X$ 的影响.

在计算其他邻接对象对 $X$ 的影响时考虑了对象之间的链接的流行度传播因子. 文献[73]着重于计算实体的重要度而非实体与查询关键词的相关度. 但是，因为实体重要度可以作为实体相关度的一个重要参考依据，所以文献[73]的工作可以为本文的工作提供很好的借鉴.

文献[75]研究了根据一组关键词从一组文档中获取相关对象的方法，例如输入“lightweight”和“business use”期望从一组文档关于计算机评价的文档中获得“Dell Inspiron 700m”和“Sony VAIO”. 其中有两个关键问题需要解决：(1) 如何匹配一个对象和一组关键词，例如“Sony VAIO”与(“lightweight”，“business use”)？(2) 如何计算一个目标对象的相关度？例如：如何计算“Dell Inspiron 700m”、“Dell Latitude D810”与(“lightweight”，“business use”)的相关度？即：如何判定“Dell Inspiron 700m”与“Dell Latitude D810”哪一个与查询关键词更相关？文献[75]中假设搜索的对象为一组已建立索引

计算机学报

的特定 Web 文档或者某种全文搜索系统,首先根据查询关键词搜索出相关的文档并且计算文档对关键词的支持度;然后从文档中获取匹配实体,其核心思想为:包含该实体的文档覆盖的关键词越多、匹配度越高那么该实体的与关键词相关度就越高. 文献[75]还提出一些优化策略以减少获取排名 TOP $K$ 的对象时的代价. 作为 Web 信息拓展模型信息获取一个阶段,相关工作可以为 Web 信息拓展模型的研究提供很好的借鉴.

基于已知信息从 Web 或者知识库中获取所需信息的典型应用还包括集合自动扩展系统(Automatic Sets Expansion)和问答系统(Question & Answering). 比较典型的集合自动扩展系统有 SEAL[76-77]、NTT Set Expansion System[78]、WebSets[79] 和 Google Sets[80]. 这些系统从 Web 表格中抽取与种子实体相关的实体,然后按照相关度进行排序,得到扩展后的实体集合. 这些工作的目标只着重于发现同类事物,例如根据一组水果发现更多种类的水果,其适用的领域非常有限. 而且,截止目前,这些方法都只对一些特定类别的集合(如汽车和水果等)扩展效果较好,所以相关项目的开展都不是很成功,Google Sets 已经下马.

基于 Web 的问答系统比较典型的有卡内基梅陇大学开发的开源问答系统 Ephyra[81]. 系统自动分析用户输入的问题,然后借助 Web 自动获取答案. 为了更有效地分析问题的语义、确定备选答案的域以及抽取备选答案,Ephyra 中使用了 WordNet、维基百科和自然语言处理语料库等通用知识库以及用户自定义知识库. 鉴于自然语言处理和语义识别的复杂性,目前该问答系统的准确率还不高,有待改善. 另外利用计算机自动将“离散型”信息转化成“连续型”信息本身也是一件非常复杂的工作,所以 Ephyra 并不能被我们直接使用.

文献[82]研究了如何从大规模互联网中的详细页面中自动抽取结构化数据及计算大规模信息网络中对象之间的相似性度量方法. 文献[83]创新性地提出了一种从微博中形成基于众包技术的专家知识发现框架,并探讨了众包技术在此问题上的效用. 这些工作都只局限于 Web 信息拓展技术的一个极小的方面,例如信息的结构化和信息源的选择等,对于影响 Web 信息扩展其他诸多因素都没有涉及.

总之,现有的基于 Web 信息的信息扩展技术考虑利用信息之间的关系进行信息扩展,可以为本文相关的研究提供很好的借鉴,但是尚不成熟. 现有的基于 Web 信息的信息扩展系统只限定于特定的应用领域及场景,适用面窄、准确度不高,不能直接适用.

### 4.4 测试数据及数据生成工具

为了更好地帮助读者开展相关研究,本节将对相关工作中采用的测试数据集进行梳理并且对数据质量测试数据生成工具研究现状进行分析.

现有工作中测试数据集情况如下:

在对连续型数据的正确性验证技术进行评估时,使用频率较高的测试数据集是 TREC,例如文献[33-35]. TREC 是一个广泛应用于信息检索和自然语言处理领域的标准数据集,其中包含了很多陈述性问题以及这些问题的正确答案. 对实体同一性判定技术进行评估时,使用频率较高的数据集主要包括:DBLP、ACM Digital Library① 和 GENOMES② 等来源于专业机构的公开数据集. 因为 Web 中存在大量与这些数据集相关的数据,所以比较适合采用基于 Web 信息的技术进行处理. 对基于 Web 的完整性错误的修复技术进行评估时,使用频率较高的数据集主要包括两类:一类是公开的数据集,如 DBLP 和 Google Books③ 等;另外一类研究单位自己从公开数据源整理的数据集,例如:Multilingual Disney Cartoon Table (Disney)[59-61]、Personal Information Table (PersonInfo)[59-61] 以及 Academic Staff & University(Staff)④等. 在评估地理位置精确性修复技术时,一般选取一个特定国家,例如:UK,来构造与位置相关的查询作为测试数据集,使用维基百科⑤来校验答案的正确性.

通过上面的梳理可知:现有工作中采用的数据集比较多样. 截止目前,尚没有与数据质量相关的标准测试数据集.

现有工作中使用的数据集大多是干净的数据集,其中的错误需要注入. 而现有工作中都没有提供数据错误的注入工具,也鲜有对数据错误注入方法的介绍. 针对上述现状,文献[84]中研制了一款信息错误自动注入的工具——BART. BART 最典型的特点是可以根据用户定义的一组数据质量规则自动随机注入多种数据错误并且可以生成大规模的脏数

① http://dblp.uni-trier.de/xml/
② http://www.ornl.gov/sci/techresources/Human_Genome/acronym.shtml
③ http://books.google.com/
④ http://www.arwu.org/
⑤ https://www.wikipedia.org/

《计算机学报》

据集合.其可注入的错误包括:拼写错误、重复、数据缺失和假值等.BART 是截止目前出现的第一款开源的数据错误生成工具,相信会被广泛应用于数据质量相关技术的研究中.

### 4.5　研究进展小结

基于 Web 的信息错误自动检测与修复技术研究进展小结如表 2 所示.目前国际上基于 Web 信息的关系型信息错误自动检测和修复技术的相关研究尚少,主要集中在实体同一性的判定、完整性与精确性错误修复等方面.相关技术尚存在如下问题:(1)大多只是针对关系数据库中特定类型的数据,例如实体同一性判定技术主要针对文献中的姓名,而精确性修复技术则主要针对地址信息;(2)只是针对信息质量的部分特定问题(完整性、实体属性值同一性和精确性),对其它数据质量问题没有涉及,更没有提供系统的解决方案;(3)没有提供实用的信息错误检测准确性和错误修复可靠性自动评估方法.

**表 2　研究进展小结**

| 研究方向 | 技术 | 代表性工作 | 进展情况总结 |
|---|---|---|---|
| 基于Web信息的信息错误自动检测技术 | 正确性判定技术 | “连续型”信息:文献[30,33-35,37,39]<br>“离散型”信息:文献[41-45,48-54] | 进展:陈述性信息的正确性验证和关系型信息同一性判定.<br>不足:尚未发现基于Web信息的关系数据信息的一致性、时效性、精确性错误的检测方法. |
| 基于Web信息的信息错误自动修复技术 | 完整性错误修复 | 缺失属性值或映射关系填补:文献[58-62,64-66]<br>映射关系填补:文献[63] | 进展:关系信息完整性错误修复.<br>不足:对其他类型的信息错误极少涉及. |
| | 精确性错误修复 | 地理位置信息精确性:文献[68-71] | 进展:地理位置信息精确性修复.<br>不足:对地理位置以外的其它信息精确性极少涉及. |
| 基于Web信息的信息扩展核心技术 | 信息扩展技术 | 实体流行度排序:文献[73]<br>实体获取:文献[75]<br>结构化:文献[82]<br>知识发现:文献[83] | 进展:实体排序或根据关键词获取目标实体.<br>不足:只考虑了部分因素或只涉及信息扩展的部分阶段. |
| | 信息扩展系统 | 集合扩展:文献[76-80]<br>问答系统:文献[81] | 进展:可扩展同类信息或对问题进行回答.<br>不足:技术不成熟,适用面窄,准确度不高. |
| 测试数据集 | 数据生成器 | 文献[84] | 进展:第 1 款开源的数据生成器<br>不足:用户界面有待改善 |

## 5　研究方向及思路

本文认为若想利用 Web 信息有效改善关系数据库中的信息质量需要解决如下关键问题:

(1)基于 Web 信息的信息扩展模型的准确性保障.基于 Web 信息的信息扩展模型的准确性是保障信息错误自动检测与修复技术的有效性和可靠性的重要前提.信息扩展的关键在于信息获取.如何结合关系型信息的特征自动形成 Web 查询关键词和进行信息实体的抽取及评估是信息获取的关键所在.另外,因为目前 Web 信息获取手段的局限性,信息扩展时搜索引擎会返回大量无关信息.因此,信息扩展的边界越大,代价越大,但是并非一定意味着获取准确结果的可能性越高.针对关系数据的信息质量的特定问题扩展信息时,需要确定所扩展的信息的边界.

(2)基于 Web 信息的信息正确性评估和信息修复的可靠性评估.基于 Web 信息的信息正确性评估模型是信息错误自动检测技术的基础,基于 Web 信息的信息错误修复的可靠性评估模型是衡量基于 Web 信息的修复方法有效性和正确性的重要依据.这两种评估模型的本质都是衡量信息当前状态与真实状态的差距,所以可以统一为信息正确性评估模型.有别于传统的基于规则的正确性验证和评估系统,基于 Web 信息的关系型信息错误自动检测与修复技术解决的核心问题是:在缺乏严格规则系统的情况下如何利用关系型信息不同属性值在数据库和扩展信息中潜在的约束关系、分布特征以及来源属性等信息构筑合理的评估模型来评估信息的正确性.

鉴于此,未来需在基于 Web 信息的自动扩展模型、自动检测与修复算法以及自动检测与修复算法的可靠性模型方面开展更加深入的研究.后续小节中将讨论针对上述诸方面的基本研究思路.

### 5.1　基于 Web 信息的信息自动扩展模型

基于 Web 信息的信息扩展模型需要解决的核心问题是:如何利用关系数据库中的已知信息从 Web 中获取所需信息.目前搜索引擎是从 Web 中搜集信息的主要手段.利用搜索引擎收集信息,首先需要构筑合适的查询关键词.如何自动选择关系型信息的属性集构造 Web 查询关键词可以抽象为组合优化的 NP 难问题,对于属性个数较多的信息集采用穷举属性子集构造查询关键词的方法显然效率低下.鉴于此,需要研究利用机器学习的方法自动判定离散关系属性值之间的上下文关系,并借此自动形成查询关键词模式.在设计机器学习算法时需要考虑训练元组个数,搜索引擎返回的含有目标属性值的文档片段个数,每次考察的片段个数等因素.在

《计算机学报》

训练和调整的过程中需要关注这些因素与信息扩展边界之间的关系并根据统计数据建立相关的数学模型.

从搜索引擎返回的文档中自动抽取相关实体的准确率是影响基于 Web 信息的信息扩展模型准确率的一个关键因素.因为自然语言的灵活性,现有基于语言模式的抽取方法准确率不高.关系数据库同一个表中每个元组及其属性之间的关系都是同质的,因此可以选取训练样例从中获取属性及其相关扩展实体之间的多种关系(上下文和语义)关系模式并结合 Web 信息的特征来进行信息抽取.

### 5.2 基于 Web 信息的信息错误自动检测算法

对于信息相对时效性判定问题,其核心是确定信息之间的时效偏序关系.待确认的信息的时效由其所依附的主属性的时效来确定.而主属性不同时间的版本隐藏于 Web 中.例如,图 9(a)上下文信息中隐藏了 3 位美国总统在职的时间顺序,图 9(b)中网页的发布时间隐含了两位西北工业大学校长在职的时间顺序.信息的绝对时效性判定问题,其核心是确定主属性及其在不同时期版本之间的相对时效性.这两类问题都归结于确定 Web 中两个信息的相对时效性.所以可以考虑利用信息在 Web 网页中的上下文及语义关系、信息关联的时间戳、Web 网页更新的时间戳、链接关系以及 PageRank 值(相对权威性)等因素构建时效性评估模型.

(a)维基百科中克林顿总统信息片段 (b)校园网中校长相关的不同新闻片段

图 9 利用 Web 信息判定信息时序关系实例

文本信息的精确性涉及到自然语言的语义,判定比较困难,所以可以首先关注可度量型信息绝对精确性判定方法.令 $D$ 表示数据集合,$t$ 表示 $D$ 中的一个元组,$v(t)$表示 $t$ 的当前值,$v'(t)$表示 $t$ 的真实值,$dis(v(t),v'(t))$表示由于上述因素所导致的元组 $t$ 的精确性错误.基于 Web 信息的方法的目标是从 Web 中获取可以降低$\sum_{t\in D} dis(v(t),v'(t))$的 $v(t)$. $v'(t)$通常是未知的.最理想的情况是借助 Web 确定 $v'(t)$,然后利用 $v'(t)$来进行信息错误修复.可以充分利用 Web 信息源的相对精确性、不同版本信息值的数量及分布特征等信息来建模以期提高 $v'(t)$的准确度.首先借助关联属性设法从 Web 中获取同一个信息不同版本或者来源的值,然后判定这些值的相对精确性,从中发现 $v'(t)$.

对于属性值同一性判定问题,可以使用图模型描述从 Web 中发掘的相关实体之间的关系,利用现有的图聚类方法聚焦与待检测与修复的信息聚合度高的扩展信息,通过比较高聚合信息的相似性来判定属性值的同一性及元组的同一性.图 10 描述了利用扩展信息的关系图的相似度判定 ICDE 和 International Conference on Data Engineering 在语义上是否同一的思路.

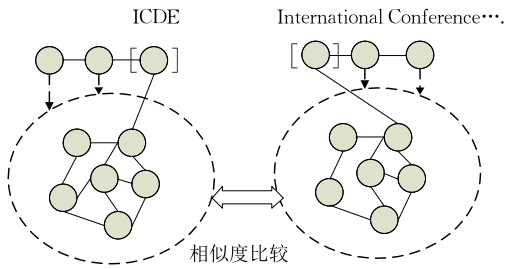


图 10 基于扩展信息的实体同一性判定模型

对于一致性规则的发现问题,在待检测与修复的信息中信息量不足时可借助 Web 进行信息扩展以发现与待检测与修复信息聚合度高的扩展信息,在高聚合信息的基础上拓展现有的函数(条件函数依赖)挖掘算法来发现数据依赖规则;对于一致性错误的检测,可充分利用现有的算法结合数据依赖规则的特征来判定信息的一致性.

### 5.3 基于 Web 信息的信息错误自动修复算法

对于属性值缺失错误的自动修复问题,可将其转换为根据已知属性从 Web 获取未知属性的问题.其关键在于如何根据已知属性及属性之间的上下文关系形成查询关键词和如何结合已知信息对从 Web 中获取的候选实体进行评估和筛选.

对于时效性和精确性错误的修复问题,因为在进行相对时效性和精确性判定的同时我们确定了信息之间的时效偏序关系和精确性的偏序关系,所以选择相对来说最新或者最精确的值进行修复即可.

对于一致性错误的修复问题,因为数据依赖规则主要依据待检测与修复的信息及其扩展信息获得,所以可以利用这些依赖规则形成查询关键词和实体抽取模型从 Web 中获取相对一致的信息进行

信息错误修复.

多种信息错误并存时，需根据错误的类型、粒度（属性级、元组级以及表级）对规则和错误结果进行聚类，采用启发式算法构建的最小代价修复模型以确定合理的错误结果输出和修复方案.

### 5.4 基于 Web 信息的信息错误自动检测与修复算法可靠性评估模型

基于 Web 信息的信息错误自动检测与修复算法可靠性评估的主要目的是借助 Web 对检测和修复结果进行再验证以确定检测算法及修复算法的可靠性，所以可以将信息错误自动检测与修复算法的可靠性评估问题转换为利用 Web 进行关系型信息正确性评估的问题. 在进行正确性评估时，可充分利用 T-verifier[33] 和 Ephyra[81] 等正确性验证和问答系统已有成果，构筑合理的模型利用网页的时间戳信息、PageRank 值、链接关系、信息来源权威性以及信息源的数量等信息来确定信息的正确性.

## 6 结　论

引入 Web 信息可以有效改善关系数据库中的信息质量，但是极具挑战性. 本文讨论了基于 Web 信息的关系型信息错误自动检测与修复技术的优势及面临的挑战，系统总结和展望了相关研究工作. 期望本文能为研究者开展相关研究提供借鉴.

## 参 考 文 献


[1] Liu H, Li Z, Jin C, Chen Q. Web-based techniques for automatically detecting and correcting information errors in a database//Proceedings of the 2016 International Conference on Big Data and Smart Computing (BigComp). Hong Kong, China, 2016: 261-264

[2] Fan W. Dependencies revisited for improving data quality//Proceedings of the 27th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems. Vancouver, Canada, 2008: 159-170

[3] Fan W, Geerts F, Li J, Xiong M. Discovering conditional functional dependencies. IEEE Transactions on Knowledge and Data Engineering, 2011, 23(5): 683-698

[4] Chiang F, Miller R. Discovering data quality rules. Proceedings of the VLDB Endowment, 2008, 1(1): 1166-1177

[5] Fan W, Geerts F, Jia X, Kementsietsidis A. Conditional functional dependencies for capturing data inconsistencies. ACM Transactions on Database Systems, 2008, 33(2): 6

[6] Bravo L, Fan W, Ma S. Extending dependencies with conditions//Proceedings of the 33rd International Conference on Very Large Data Bases. Vienna, Austria, 2007: 243-254

[7] Fan W, Gao H, Jia X, et al. Dynamic constraints for record matching. The VLDB Journal, 2011, 20(4): 495-520

[8] Song S, Chen L. Discovering matching dependencies//Proceedings of the 18th ACM Conference on Information and Knowledge Management. Hong Kong, China, 2009: 1421-1424

[9] Fan W, Geerts F, Wijsen J. Determining the currency of data. ACM Transactions on Database Systems, 2012, 37(4): 25

[10] Cao Y, Fan W, Yu W. Determining the relative accuracy of attributes//Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data. New York, USA, 2013: 565-576

[11] Fan W, Jia X, Li J, Ma S. Reasoning about record matching rules. Proceedings of the VLDB Endowment, 2009, 2(1): 407-418

[12] Caruccio L, Deufemia V, Polese G. Relaxed functional dependencies—A survey of approaches. IEEE Transactions on Knowledge and Data Engineering, 2016, 28(1): 147-165

[13] Chen W, Fan W, Ma S. Incorporating cardinality constraints and synonym rules into conditional functional dependencies. Information Processing Letters, 2009, 109(14): 783-789

[14] Fan W, Geerts F, Ma S, Müller H. Detecting inconsistencies in distributed data//Proceedings of the 2010 IEEE 26th International Conference on Data Engineering (ICDE 2010). California, USA, 2010: 64-75

[15] Fan W, Li J, Tang N, Wenyuan Yu. Incremental detection of inconsistencies in distributed data. IEEE Transactions on Knowledge and Data Engineering, 2014, 26(6): 1367-1383

[16] Fan G, Fan W, Geerts F. Detecting errors in numeric attributes//Proceedings of the International Conference on Web-Age Information Management. Macau, China, 2014: 125-137

[17] Fan W, Li J, Ma S, et al. Towards certain fixes with editing rules and master data. Proceedings of the VLDB Endowment, 2010, 3(1-2): 173-184

[18] Cong G, Fan W, Geerts F, et al. Improving data quality: Consistency and accuracy//Proceedings of the 33rd International Conference on Very Largedata Bases. Vienna, Austria, 2007: 315-326

[19] Fan W, Geerts F, Tang N, Yu W. Inferring data currency and consistency for conflict resolution//Proceedings of the 2013 IEEE 29th International Conference on Data Engineering (ICDE). Brisbane, Australia, 2013: 470-481

[20] Fan W, Li J, Ma S, et al. Interaction between record matching and data repairing. Journal of Data and Information Quality, 2014, 4(4): 16

[21] Bohannon P, Flaster M, Fan W, Rastogi R. A cost-based model and effective heuristic for repairing constraints by value modification//Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data. Maryland, USA, 2005: 143-154


计算机学报

[22] Chiang F, Miller R. A unified model for data and constraint repair//Proceedings of the 2011 IEEE 27th International Conference on Data Engineering. Hannover, Germany, 2011: 446-457

[23] Loshin D. Master Data Management. San Francisco, USA: Morgan Kaufmann Publishers Inc., 2008

[24] Fan W, Geerts F. Capturing missing tuples and missing values//Proceedings of the 29th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems. Indianapolis, USA, 2010: 169-178

[25] Fan F, Geerts F. Relative information completeness. ACM Transactions on Database Systems, 2010, 35(4): 27

[26] Yakout M, Elmagarmid A, Neville J, et al. Guided data repair. Proceedings of the VLDB Endowment, 2011, 4(5): 279-289

[27] Chu X, Morcos J, Ilyas I, et al. KATARA: A data cleaning system powered by knowledge bases and crowdsourcing// Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data. Melbourne, Australia, 2015: 1247-1261

[28] Vesdapunt N, Bellare K, Dalvi N. Crowdsourcing algorithms for entity resolution. Proceedings of the VLDB Endowment, 2014, 7(12): 1071-1082

[29] Tong Y, Cao C, Zhang C, et al. Crowdcleaner: Data cleaning for multi-version data on the Web via crowdsourcing// Proceedings of the 2014 IEEE 30th International Conference on Data Engineering. Chicago, USA, 2014: 1182-1185

[30] Yin X, Han J, Yu P. Truth discovery with multiple conflicting information providers on the Web. IEEE Transactions on Knowledge and Data Engineering, 2008, 20(6): 796-808

[31] Fan W, Geerts F. Foundations of data quality management. Synthesis Lectures on Data Management, 2012, 4(5): 1-217

[32] Wan Chang-Xuan, Deng Song, Liu Xi-Ping, et al. Web data source selection technologies. Journal of Software, 2013, 24(4): 781-797(in Chinese)
(万常选, 邓松, 刘喜平等. Web 数据源选择技术. 软件学报, 2013, 24(4): 781-797)

[33] Li X, Meng W, Yu C. T-verifier: Verifying truthfulness of fact statements//Proceedings of the 2011 IEEE 27th International Conference on Data Engineering. New York, USA, 2011: 63-74

[34] Li X, Meng W, Yu C. Verification of fact statements with multiple truthful alternatives//Proceedings of the 12th International Conference on Web Information Systems and Technologies. Roma, Italy, 2016: 87-97

[35] Wang T, Zhu Q, Wang S. Multi-verifier: A novel method for fact statement verification. World Wide Web, 2015, 18(5): 1463-1480

[36] De Marneffe M, MacCartney B, Manning C. Generating typed dependency parses from phrase structure parses// Proceedings of LREC. Portorož, Slovenia, 2006, 6(2006): 449-454

[37] Wang Teng, Zhu Qing, Wang Shan. Facts statements verification based on semantic similarity. Chinese Journal of Computers, 2013, 36(8): 1668-1681(in Chinese)
(王腾, 朱青, 王珊. 基于语义相似度 Web 信息可信分析. 计算机学报, 2013, 36(8): 1668-1681)

[38] Li X, Dong X, Lyons K, et al. Truth finding on the deep Web: Is the problem solved? Proceedings of the VLDB Endowment, 2012, 6(2): 97-108

[39] Yin X, Tan W. Semi-supervised truth discovery//Proceedings of the 20th International Conference on World Wide Web. Hyderabad, India, 2011: 217-226

[40] Wang J, Li G, Yu J, et al. Entity matching: How similar is similar. Proceedings of the VLDB Endowment, 2011, 4(10): 622-633

[41] Elmacioglu E, Kan M Y, Lee D, et al. Web based linkage// Proceedings of the 9th Annual ACM International Workshop on Web Information and Data Management. Lisboa, Portugal, 2007: 121-128

[42] Bollegala D, Matsuo Y, Ishizuka M. Measuring semantic similarity between words using web search engines//Proceedings of the 16th International Conferencer on World Wide Web. Banff, Canada, 2007: 757-766

[43] Bollegala D, Matsuo Y, Ishizuka M. A Web search engine-based approach to measure semantic similarity between words. IEEE Transactions on Knowledge and Data Engineering, 2011, 23(7): 977-990

[44] Tan Y, Elmacioglu E, Kan Y, et al. Efficient Web-based linkage of short to long forms//Proceedings of the International Workshop on the Web and Databases (2008). Vancouver, Canada, 2008

[45] Li P, Wang H, Zhu K, et al. Computing term similarity by large probabilistic isA knowledge//Proceedings of the 22nd ACM International Conference on Information & Knowledge Management. San Francisco, USA, 2013: 1401-1410

[46] Miller G. WordNet: A lexical database for English. Communications of the ACM, 1995, 38(11): 39-41

[47] Wu W, Li H, Wang H, et al. Probase: A probabilistic taxonomy for text understanding//Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data. Scottsdale, USA, 2012: 481-492

[48] Ferreira A, Gonçalves M, Laender A. A brief survey of automatic methods for author name disambiguation. ACM SIGMOD Record, 2012, 41(2): 15-26

[49] Kanani P, McCallum A, Pal C. Improving author coreference by resource-bounded information gathering from the Web// Proceedings of the 20th International Joint Conference on Artificial Intelligence. Hyderabad, India, 2007: 429-434

[50] Kang I, Na S, Lee S, et al. On co-authorship for author disambiguation. Information Processing & Management, 2009, 45(1): 84-97

[51] Pereira D, Ribeiro-Neto B, Ziviani N, et al. Using Web information for author name disambiguation//Proceedings of the 9th ACM/IEEE-CS Joint Conference on Digital Libraries. Austin, USA, 2009: 49-58

[52] Yang K, Peng H, Jiang J, et al. Author name disambiguation for citations using topic and Web correlation//Proceedings of the 12th International Conference on Theory and Practice of Digital Libraries. Aarhus, Denmark, 2008: 185-196

[53] Aumueller D, Rahm E. Web-based affiliation matching//Proceedings of the 14th International Conference on Information Quality. Potsdam, Germany, 2009: 246-256

[54] Tan Y, Kan M. Hierarchical cost-sensitive Web resource acquisition for record matching//Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology. Toronto, Canada, 2010: 382-389

[55] Tan Y, Kan M, Lee D. Search engine driven author disambiguation//Proceedings of the 6th ACM/IEEE-CS Joint Conference on Digital Libraries. Chapel Hill, USA, 2006: 314-315

[56] Zhu X, Zhang S, Jin Z, et al. Missing value estimation for mixed-attribute data sets. IEEE Transactions on Knowledge and Data Engineering, 2011, 23(1): 110-121

[57] Song S, Zhang A, Chen L, et al. Enriching data imputation with extensive similarity neighbors. Proceedings of the VLDB Endowment, 2015, 8(11): 1286-1297

[58] Tang N, Vemuri V. Web-based knowledge acquisition to impute missing values for classification//Proceedings of the 2004 IEEE/WIC/ACM International Conference on Web Intelligence. Beijing, China, 2004: 124-130

[59] Li Z, Sharaf M, Sitbon L, et al. Webput: Efficient Web-based data imputation//Proceedings of the International Conference on Web Information Systems Engineering. Paphos, Cyprus, 2012: 243-256

[60] Li Z, Sharaf M, Sitbon L, et al. A Web-based approach to data imputation. World Wide Web, 2014, 17(5): 873-897

[61] Li Z, Shang S, Xie Q, et al. Cost reduction for Web-based data imputation//Proceedings of the 19th International Conference on Database Systems for Advanced Applications. Bali, Indonesia, 2014: 438-452

[62] Ahmadov A, Thiele M, Eberius J, et al. Towards a hybrid imputation approach using Web tables//Proceedings of the 2015 IEEE/ACM 2nd International Symposium on Big Data Computing. Limassol, Cyprus, 2015: 21-30

[63] Li Z, Sharaf M, Sitbon L, et al. Core: A context-aware relation extraction method for relation completion. IEEE Transactions on Knowledge and Data Engineering, 2014, 26(4): 836-849

[64] Chen Zhao-Qiang, Li Jia-Jun, Jiang Chuan, et al. A context-aware entity ranking method for Web-based data imputation. Chinese Journal of Computers, 2015, 38(9): 1755-1766(in Chinese)
(陈肇强，李佳俊，蒋川等. 基于上下文感知实体排序的缺失数据修复方法. 计算机学报，2015，38(9)：1755-1766)

[65] Chen Z, Chen Q, Li J, et al. A probabilistic ranking framework for Web-based relational data imputation. Information Sciences, 2016, 355: 152-168

[66] Li Z, Qin L, Cheng H, et al. TRIP: An interactive retrieving-inferring data imputation approach. IEEE Transactions on Knowledge and Data Engineering, 2015, 27(9): 2550-2563

[67] Soderland S. Learning information extraction rules for semi-structured and free text. Machine Learning, 1999, 34(1-3): 233-272

[68] Purves R, Clough P, Joho H. Identifying imprecise regions for geographic information retrieval using the Web//Proceedings of the 13th Annual GIS Research UK Conference. Glasgow, UK, 2005: 313-318

[69] Arampatzis A, Van Kreveld M, Reinbacher I, et al. Web-based delineation of imprecise regions. Computers, Environment and Urban Systems, 2006, 30(4): 436-459

[70] Reinbacher I, Benkert M, Van Kreveld M, et al. Delineating boundaries for imprecise regions//Proceedings of the European Symposium on Algorithms. Mallorca, Spain, 2005: 143-154

[71] Jones C, Purves R, Clough P, et al. Modelling vague places with knowledge from the Web. International Journal of Geographical Information Science, 2008, 22(10): 1045-1065

[72] Smith D, Mann G. Bootstrapping toponym classifiers//Proceedings of the HLT-NAACL 2003 Workshop on Analysis of Geographic References. Edmonton, Canada, 2003: 45-49

[73] Nie Z, Zhang Y, Wen J, et al. Object-level ranking: Bringing order to Web objects//Proceedings of the 14th International Conference on World Wide Web. Chiba, Japan, 2005: 567-574

[74] Page L, Brin S, Motwani R, et al. The PageRank citation ranking: Bringing order to the Web//Proceedings of the 7th International World Wide Web Conference. Brisbane, Australia, 1998: 161-172

[75] Chakrabarti K, Ganti V, Han J, et al. Ranking objects based on relationships//Proceedings of the 2006 ACM SIGMOD Iconference on Management of Data. Chicago, USA, 2006: 371-382

[76] Wang R, Cohen W. Language-independent set expansion of named entities using the Web//Proceedings of the 7th IEEE International Conference on Data Mining. Omaha, USA, 2007: 342-350

[77] Wang R, Cohen W. Iterative set expansion of named entities using the Web//Proceedings of the 2008 Eighth IEEE International Conference on Data Mining. Leipzig, Germany, 2008: 1091-1096

[78] Sadamitsu K, Saito K, Imamura K, et al. Entity set expansion using topic information//Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. Portland, USA, 2011: 726-731

[79] Dalvi B, Cohen W, Callan J. Websets: Extracting sets of entities from the Web using unsupervised information extraction//Proceedings of the 5th ACM International Conference on Web Search and Data Mining. Seattle, USA, 2012: 243-252

《计算机学报》

[80] Tong S, Dean J. Automatically creating lists from existing lists: U.S. Patent 8, 429, 164. USA, 2013-4-23

[81] Schlaefer N, Ko J, Betteridge J, et al. Semantic Extensions of the Ephyra QA System for TREC 2007//Proceedings of the 16th Text REtrieval Conference. Gaithersburg, USA, 2007, 1(2): 2

[82] Liu H, He J, Zhu D, et al. Measuring similarity based on link information: A comparative study. IEEE Transactions on Knowledge and Data Engineering, 2013, 25(12): 2823-2840

[83] Cao C, She J, Tong Y, et al. Whom to ask?: Jury selection for decision making tasks on micro-blog services. Proceedings of the VLDB Endowment, 2012, 5(11): 1495-1506

[84] Arocena P, Glavic B, Mecca G, et al. Messing up with BART: Error generation for evaluating data-cleaning algorithms. Proceedings of the VLDB Endowment, 2015, 9(2): 36-47

**LIU Hai-Long**, born in 1980, Ph.D., lecturer. His research interests include data management technologies and data quality management technologies.

**LI Zhan-Huai**, born in 1961, Ph.D., professor. His research interests include data storage technologies and data management technologies.

**CHEN Qun**, born in 1976, Ph.D., professor. His research interests include data management technologies and data quality management technologies.

**CHEN Zhao-Qiang**, born in 1988, Ph.D. candidate. His current research interest is data quality management.

《计算机学报》

**Background**

Information quality has become an important issue in many application areas. Automatically detecting and correcting information errors has proven to be an effective way for improving information quality in most information systems. Supported by the Major State Basic Research Development Program (973 Program) of China (No. 2012CB316203), we have made some progress on rule-based error detecting and correcting technologies. We find that rule-based techniques always require adequate well-structured data in a database, or else they will not work properly. Furthermore, integrating information from external sources, like the World Wide Web (WWW), can help us overcome the shortcomings of existing techniques to a great extent. Supported by the National Natural Science Foundation of China (No. 61502390) and the Basic Research Fund of Northwestern Polytechnical University (No. 3102014JSJ0013), we try to use Web data augment the quality of databases.

As the start, this review analyzes the shortcomings of existing rule-based and human-based techniques and details the advantages and challenges of Web-based techniques. Based on the framework of the research topics on Web-based techniques, this paper reviews current research works in detail, furthermore prospect some future research topics and ideas.