4 实 验

根据上面的分析,为了实现保证数据清洗质量的同时,简化清洗流程并减小清洗代价,我们提出了如图 7 所示的混合型数据质量错误的数据清洗修复模型.

为了验证上述数据清洗修复流程的有效性,在以下 4 个数据集合上进行了实验测试,其规模见表 7.

(1) 英国大学地理信息真实数据集合(下文简写为 UD);

(2) 历史气候天气真实数据集合 Worldwide Historical Weather Data(下文简写为 WHWD);

(3) 意大利社会安全评测真实数据集合 Italian Social Security Contributors List(下文简写为 ISSCL);

(4) 学生毕业信息虚拟数据集合(下文简写为 SCD).

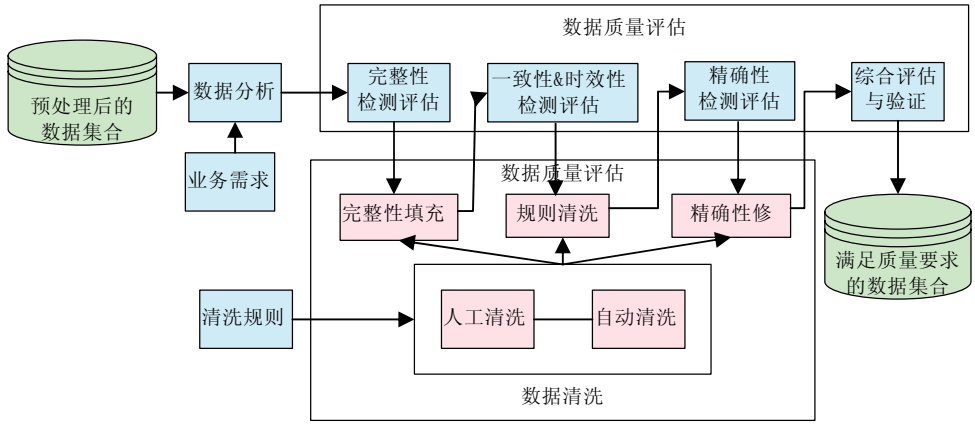


Fig.7 Comprehensive data cleaning process

图 7 综合型数据清洗流程图

Table 7 The datasets scale in experiment

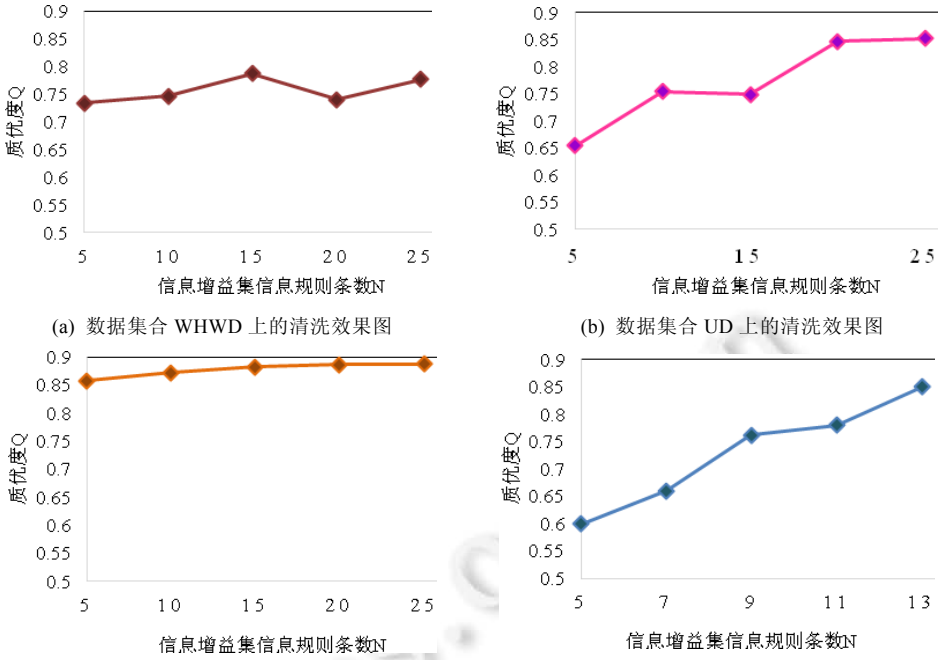
表 7 实验所用数据集合的规模示意表

	UD	WHWD	ISSCL	SCD
行数	580	20 000	1 483 712	220
列数	12	16	6	10

实验程序采用 C++编程,实验环境为内存 6GB,Inter(R) Core™ i5 CPU 处理器,64 位 Win7 操作系统.实验中,在进行实验结果分析时,我们定义了信息增益规则集合,存储针对不同性质的修复规则;采用质优度 $Q=w_1Q_{com}+w_2Q_{acc}+w_3Q_{curr}+w_4Q_{cons}$ 作为评价指标,Q 为第 3 节定义的数据质量标准系数,w 是为每种性质分配的权值.我们逐渐增加信息增益集中的规则个数进行测试,图 8 展示了在 4 个数据集合上的清洗修复效果.

在精确性方面,主要考察了错误值和孤立点的错误问题;在完整性方面,我们用空值发现来判断内容完整性.人工判断选择一组属性组作为标准属性集合,加入我们的信息增益集;在一致性和时效性方面,首先发现数据集合中的一致性规则和时效约束规则,在进行数据清洗的处理时使用这些规则,并记录修复的比例来判断一致性和时效性违反程度.

观察 4 个实验结果,发现随着清洗系统中的信息增益规则条数的增多,对数据集合的 4 个维度的修复粒度的增大,其质优度呈不同程度的增加.其中,数据集合 WHWD 上属性关联性并不强,时效性变化也并不是很大,并且数据量和属性均较为完备,因此修复效果增长较缓.在 UD 集合中,原始数据存在大量的重复、时效性低的数据,并且完备性较低,通过增加规则的个数修复,其质优度增幅较大.实验证实了本文的方法能够有效、合理地修复数据集合上的错误,并且对于原始数据集合质量较低,另外,数据集合中存在较多属性关系的数据集合处理效果更好.



(a) 数据集合 WHWD 上的清洗效果图 (b) 数据集合 UD 上的清洗效果图

(c) 数据集合 ISSCL 上的清洗效果图 (d) 数据集合 SCD 上的清洗效果图

Fig.8 Cleaning effect of different datasets

图 8 在数据集合上的清洗效果图

5 总结和展望

本文通过对数据质量的多种性质的研究,建立了数据质量多种性质模型,详细阐述并总结了数据质量在 4 种重要性质上存在的实际问题,并对这 4 个性质:完整性、精确性、一致性、时效性的违反模式给出定义,理论证明了在数据修复背景下它们之间的关联关系,基于此制定了混合型错误情况下的数据清洗修复策略,并通过实验验证了其有效性和合理性.

四维数据质量关系模型有助于数据质量的综合评估,这对于数据集合上开展数据清洗有着重要的影响力.综合评估对于数据清洗策略和具体步骤的制定在效果、效率等方面均有指导意义.我们今后的研究重点将放在四维数据质量关系模型在混合型错误的数据集合上的清洗实现,整合并优化数据清洗算法.此外,我们也将更为全面地研究关联关系理论在不同数据类型(如非结构化数据和半结构化数据)中的应用.

References:

[1] Mayer-Schonberger V, Cukier K. Big Data: A Revolution That Will Transform How We Live, Work, and Think. London: Houghton Mifflin Harcourt, 2013. 19−31.

[2] Sidi F, Shariat PPH, Affendey LS, Jabar MA, Ibrahim H, Mustapha A. Data quality: A survey of data quality dimensions. In: Proc. of the 2012 Int'l Conf. on Information Retrieval & Knowledge Management. IEEE, 2012. 300−304. [doi: 10.1109/InfRKM.2012.6204995]

[3] Guo ZM, Zhou AY. Research on data quality and data cleaning: A survey. Ruan Jian Xue Bao/Journal of Software, 2002, 13(11):2076−2082 (in Chinese with English abstract). http://www.jos.org.cn/ch/reader/view_abstract.aspx?flag=1&file_no=20021103&journal_id=jos

[4] Batini C, Cappiello C, Francalanci C, Maurino A. Methodologies for data quality assessment and improvement. ACM Computing Surveys, 2009,41(3):No.16. [doi: 10.1145/1541880.1541883]

[5] Wang RY, Strong DM. Beyond accuracy: What data quality means to data consumers. Journal of Management Information Systems, 1996,12(4):5−33. [doi: 10.1080/07421222.1996.11518099]

[6] Cong G, Fan W, Geerts F, Jia XB, Ma S. Improving data quality: Consistency and accuracy. In: Proc. of the 33rd Int'l Conf. on Very Large Data Bases. VLDB Endowment, 2007. 315−326. http://dl.acm.org/citation.cfm?id=1325890&preflayout=flat

[7] Bohannon P, Fan W, Geerts F, Jia XB, Kementsietsidis A. Conditional functional dependencies for data cleaning. In: Proc. of the 23rd IEEE Int'l Conf. on Data Engineering. Istanbul: IEEE, 2007. 746−755. [doi: 10.1109/ICDE.2007.367920]

[8] Fan W, Geerts F, Wijsen J. Determining the currency of data. ACM Trans. on Database Systems, 2012,37(4):25−41. [doi: 10.1145/2389241.2389244]

[9] Li MH, Li JZ, Gao H. Evaluation of data currency. Chinese Journal of Computers, 2012,35(11):2348−2360 (in Chinese with English abstract).

[10] McGilvray D. Executing Data Quality Projects: Ten Steps to Quality Data and Trusted Information. Burlington: Elsevier, 2008. 16−59.

[11] Fan W, Ma S, Tang N, Yu WY. Interaction between record matching and data repairing. Journal of Data and Information Quality, 2014,4(4):16. [doi: 10.1145/2567657]

[12] Tee SW, Bowen PL, Doyle PH. Rohde F. Factors influencing organizations to improve data quality in their information systems. Accounting & Finance, 2007,47(2):335−355. [doi: 10.1111/j.1467-629x.2006.00205.x]

[13] Eckerson W. Data quality and the bottom line, Vol.1. TDWI Report, Data Warehouse Institute, 2002. 1−31.

[14] Pipino LL, Lee YW, Wang RY. Data quality assessment. Communications of the ACM, 2002,45(4):211−218. [doi: 10.1145/505248.506010]

[15] https://en.wikipedia.org/wiki/Cronbach%27s_alpha

[16] Yue K. Data Engineering: Processing, Analysis and Service. Beijing: Tsinghua University Press, 2013. 169−180 (in Chinese).

[17] Fan W, Geerts F. Relative information completeness. ACM Trans. on Database Systems, 2010,35(4):97−106. [doi: 10.1145/1862919.1862924]

[18] Bravo L, Fan W, Ma S. Extending dependencies with conditions. In: Proc. of the 33rd Int'l Conf. on Very Large Data Bases. VLDB Endowment, 2007. 243−254. http://dl.acm.org/citation.cfm?id=1325882&CFID=627672245&CFTOKEN=70772333

附中文参考文献:

[3] 郭志懋,周傲英.数据质量和数据清洗研究综述.软件学报,2002,13(11):2076−2082. http://www.jos.org.cn/ch/reader/view_abstract.aspx?flag=1&file_no=20021103&journal_id=jos

[9] 李默涵,李建中,高宏.数据时效性判定问题的求解算法.计算机学报,2012,35(11):2348−2360.

[16] 岳昆.数据工程——处理、分析与服务.北京:清华大学出版社,2013.169−180.

丁小欧(1993－),女,黑龙江哈尔滨人,硕士,CCF 学生会员,主要研究领域为数据质量管理,数据清洗.

李建中(1950－),男,黑龙江哈尔滨人,博士,教授,博士生导师,主要研究领域为海量数据管理与计算,无线传感器网络,数据质量.

王宏志(1978－),男,博士,教授,博士生导师,CCF 高级会员,主要研究领域为数据库,大数据,数据质量.

高宏(1966－),女,博士,教授,博士生导师,CCF 高级会员,主要研究领域为海量数据计算,无线传感器网络.

张笑影(1994－),女,学士,主要研究领域为数据质量.