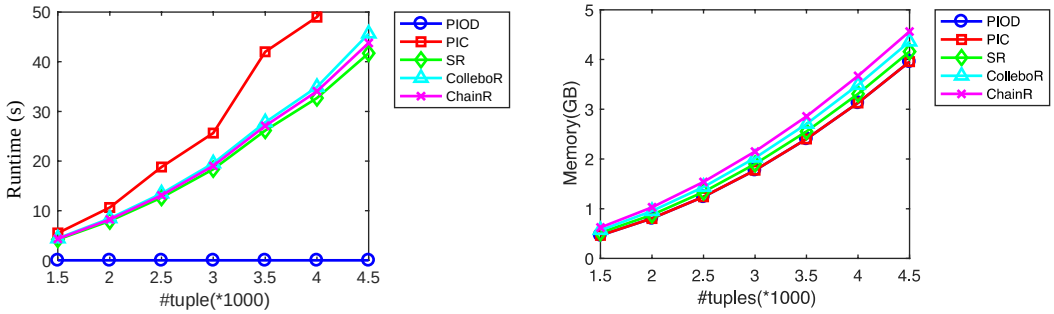
从图 3 中可以看出:从原始数据进行概率推理的方法一直没有明显效果.而从图 3(c)、图 3(d)两张图中可以看出:依据两两比较结果进行概率推理的方法在 top-50 和 top-100 的排序质量有轻微随着元组数的增加而下降的趋势.但是图 3(e)中明显可以得到下降的趋势.此实验结果与随着数据量增多,稀疏性增加,简单概率推理的效果下降的理论相符.

而对于 3 种与回归结合的概率推理方法,首先可以看到它们在 5 张排序质量图显示的结果上,因为 3 种与回归结合的概率推理方法充分考虑了属性之间的关系,所以得到的排序质量都优于两种概率推理方法,并且得到的排序质量一直很稳定,几乎不受数据量增加的影响.

- 运行时间

所有方法在不同元组数条件下的运行时间和内存消耗如图 4 所示,可以清楚地看到:从原始数据进行概率推理的方法仍然是最快的,但也是效果最差的.而 3 种与回归结合的方法的图像表明,3 种算法是随着数据集中元组数的增加呈线性增长的.依据两两比较结果进行概率推理的方法在元组数不断增加的条件下,运行时间呈非线性增长.若在相同的未知属性条件下,3 种与回归结合的方法在时间和效果上都明显优于依据两两比较结果进行概率推理的方法.而在运行时的内存消耗方面,可以清楚地看到:从原始数据进行概率推理和根据两两比较进行概率推理的方法所占内存基本相同,而 3 种与回归结合的方法的运行内存消耗顺序从小到大分别为简单回归方法、联合回归方法和链式回归方法.这种顺序是合理的,因为从简单回归方法到链式回归方法,属性之间的关系被更加详细地考虑,因此运行时所占内存联合回归方法和链式回归方法会大于简单回归方法.在元组数不断增加的条件下,5 种方法的内存消耗也呈近平方式增长,与本文所使用的两两比较方式内存的增长速度相符.



(a) 运行时间 (b) 内存消耗

Fig.4 Runtime and memory of all methods

图 4 所有方法在不同元组数条件下的运行时间和内存消耗

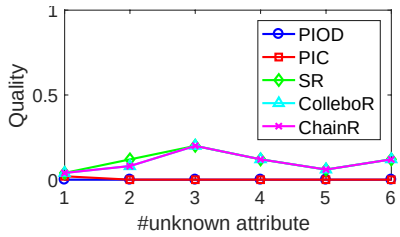
6.4.3 列扩展性

- 排序质量

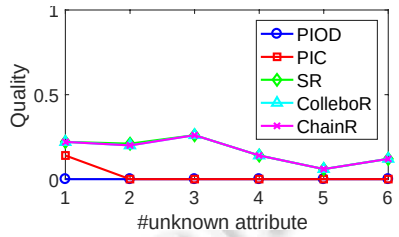
为了减少偶然性并全面观察各个算法排序质量随着未知属性数量的变化,在此实验中分别测量了在不同元组数条件下的 top-1 排序质量、top-10 排序质量、top-50 排序质量、top-100 排序质量和 top-300 排序质量,具体结果如图 5 所示.

从图 5 中可以看出,从原始数据进行概率推理的方法一直没有明显效果.而依据两两比较结果进行概率推理的方法随着未知属性数的增加,下降的趋势非常强烈;而在未知属性数超过 4 之后,此方法在 top-300 的质量也为 0 了.

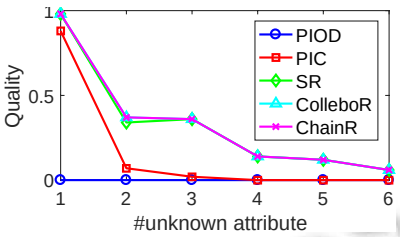
而对于 3 种与回归结合的概率推理方法,首先可以看到:随着未知属性的增加,排序质量呈下降趋势.这是由于随着未知属性数量的增加,与回归结合的概率推理方法可以学习到的信息变少的缘故.但是由于充分考虑属性之间的关系,它们在 top-1,top-10,top-50,top-100,top-300 的排序质量上都优于两种概率推理方法.



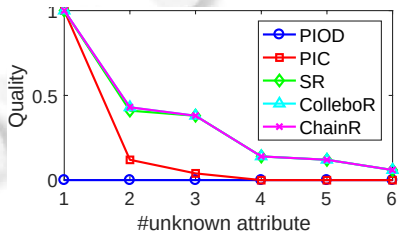
(a) 不同未知属性数条件下的 Top-1 质量



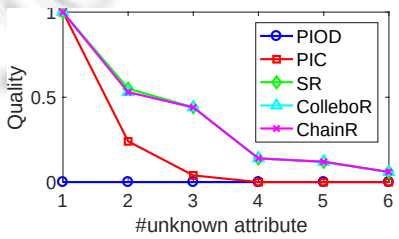
(b) 不同未知属性数条件下的 Top-10 质量



(c) 不同未知属性数条件下的 Top-50 质量



(d) 不同未知属性数条件下的 Top-100 质量



(e) 不同未知属性数条件下的 Top-300 质量

Fig.5 Top-1, Top-10, Top-50, Top-100, Top-300 quality under different number of unknown attributes

图 5 不同未知属性数条件下的 Top-1,Top-10,Top-50,Top-100,Top-300 质量

- 运行时间

所有方法在不同元组数条件下的运行时间如图 6 所示,可以清楚地看到,从原始数据进行概率推理的方法仍然是最快的,因为从原始数据进行概率推理的方法在计算时不需要进行两两比较等一系列步骤,但是也是效果最差的.而 3 种与回归结合的方法的图像表明,3 种算法是随着测试数据集中未知属性数的增加呈线性增长的.依据两两比较结果进行概率推理的方法,在未知属性数不断增加的条件下,运行时间仍然稳定,几乎不受未知属性数的影响.因为在两两比较时,未知属性数的数量相对于元组数量对运行时间的影响较小.

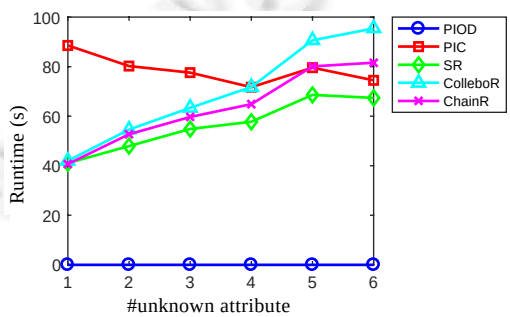


Fig.6 Runtime of all methods under different number of unknown attributes

图 6 所有方法在不同未知属性数条件下的运行时间

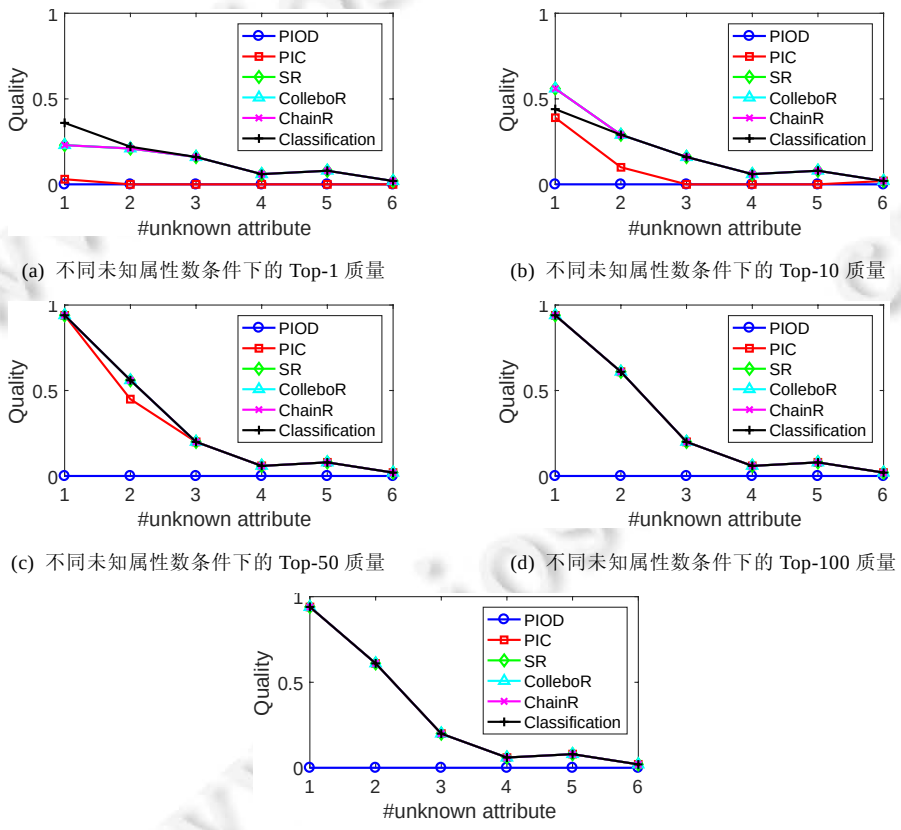
6.4.4 列扩展性(含与分类结合方法)

为了比较与分类结合的方法和其他方法在不同未知属性数条件下的效果,我们利用 WineQuality 数据集的子数据集 WineQuality-300 来对 6 种方法做不同的对比实验.WineQuality-300 中含有 300 条数据,用如此少量数据的原因,是因为多分类方法的时间效率过慢,用太大的数据集无法在可观时间内得到结果.

在此列扩展性实验中,仍然将排序质量和运行时间作为两个重要的指标.为了更详细地看到每个算法在 top-50 之前和 top-100 之前的变化趋势,此实验中还画出了在不同未知属性数条件下,top-50 前和 top-100 前的趋势图.

- 排序质量

所有方法在不同未知属性数条件下,Top-1,Top-10,Top-50,Top-100,Top-300 的排序质量图如图 7 所示(含与分类结合的方法).在图中可以清楚看到:当在 Top-1,未知属性个数为 1 时,如图 7(a)所示,与分类结合的方法比其他的方法得到的结果要好一些;但在其他情况下,如图 7(b)~图 7(e)所示,与分类结合的方法得到的排序质量并没有相对其他的方法有显著的提升.这是由于在未知属性个数为 1 时,分类时问题为二分类;而当未知属性个数为 2 时,分类问题变为四分类,以此类推.所以当未知属性个数增多时,由于类别个数呈指数增长,与分类结合的方法效果并不会相对其他方法有显著的提高.



(a) 不同未知属性数条件下的 Top-1 质量

(b) 不同未知属性数条件下的 Top-10 质量

(c) 不同未知属性数条件下的 Top-50 质量

(d) 不同未知属性数条件下的 Top-100 质量

(e) 不同未知属性数条件下的 Top-300 质量

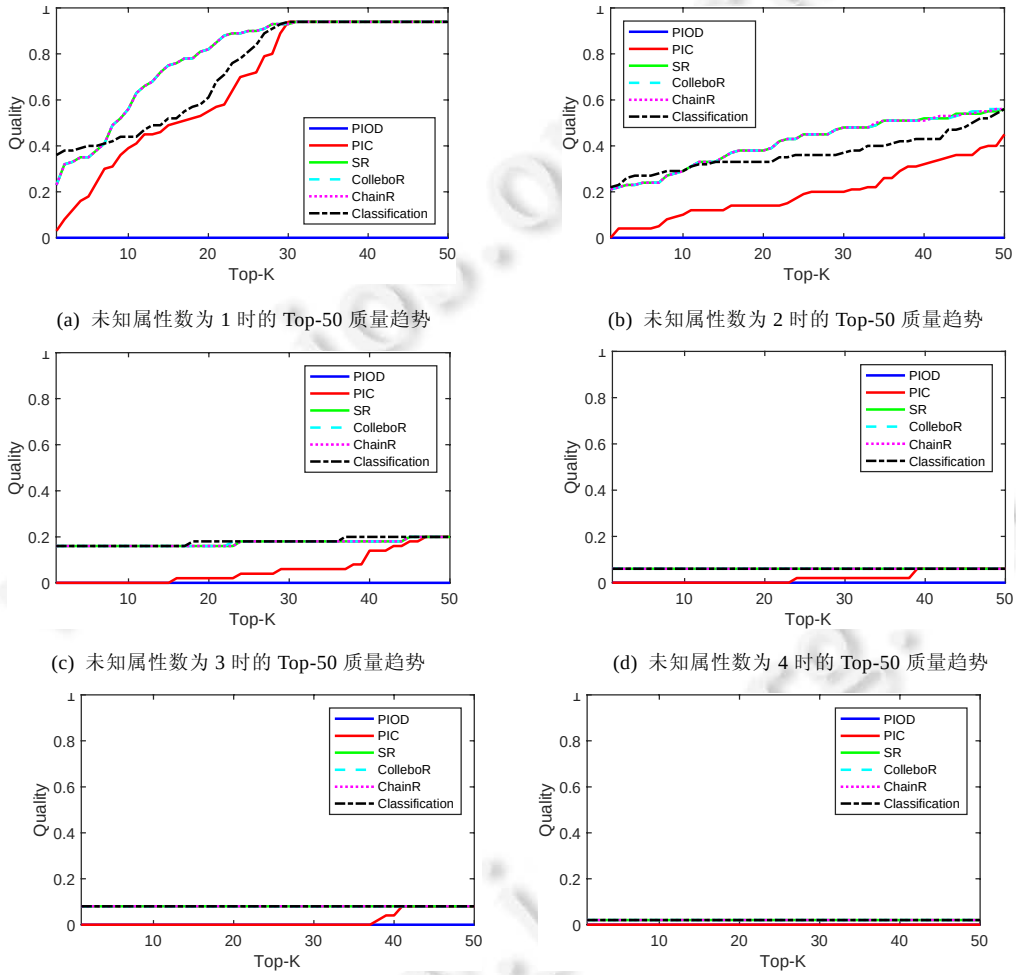
Fig.7 Top-1, Top-10, Top-50, Top-100, Top-300 quality under different number of unknown attributes

图 7 不同未知属性数条件下的 Top-1,Top-10,Top-50,Top-100,Top-300 质量

为了更加详细地研究与分类结合的方法和其他算法的效果对比,在下文中给出每个算法在 top-50 之前和 top-100 之前的变化趋势图.

- Top-50 排序质量趋势

所有算法在 top-50 之前在不同属性数条件下的变化趋势图如图 8 所示(含与分类结合的方法).图 8(a)~图 8(f)分别代表了未知属性数为 1~6 时,各个算法的 top-50 排序质量趋势.从图中可以看出,直接从源数据进行概率推理的方法仍然是最差的.而依据两两比较比较结果进行概率推理的方法在所有属性数条件下都差于与机器学习结合的方法.与分类结合的方法在前 top-10 的表现比与回归结合的方法好,而在 top-10 之后都差于或持平与回归结合的方法.



(a) 未知属性数为 1 时的 Top-50 质量趋势

(b) 未知属性数为 2 时的 Top-50 质量趋势

(c) 未知属性数为 3 时的 Top-50 质量趋势

(d) 未知属性数为 4 时的 Top-50 质量趋势

(e) 未知属性数为 5 时的 Top-50 质量趋势

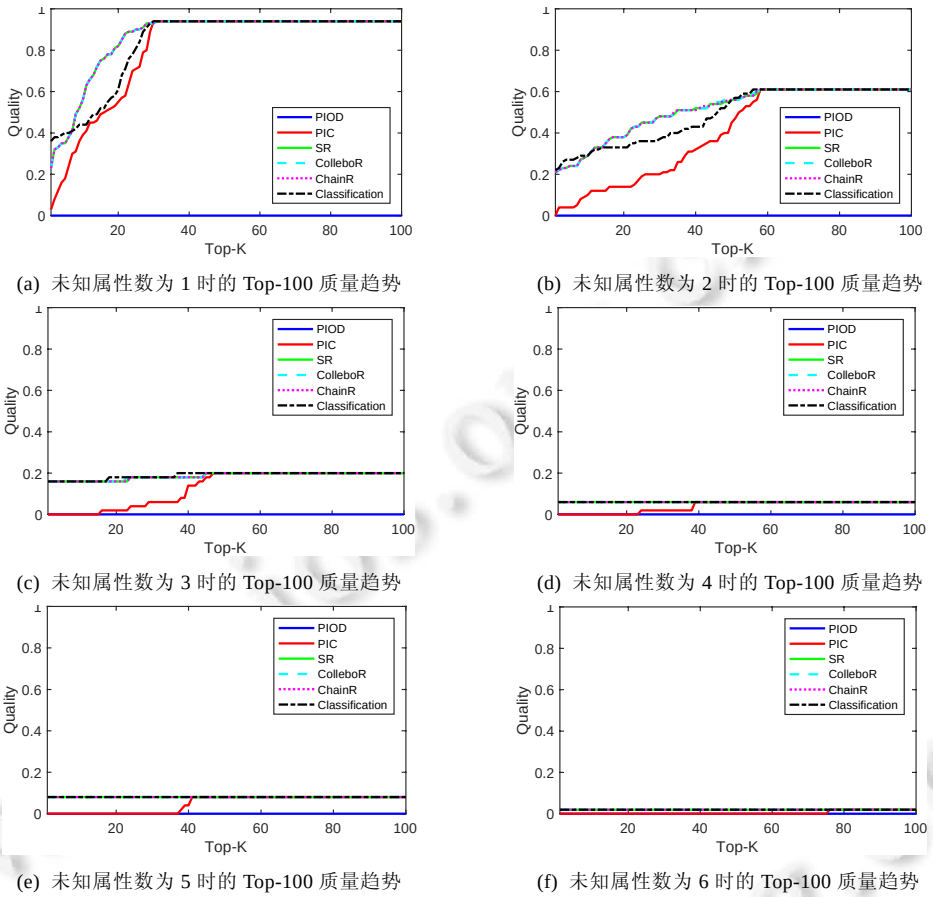
(f) 未知属性数为 6 时的 Top-50 质量趋势

Fig.8 Top-50 quality ranking trend under different number of unknown attributes

图 8 不同未知属性数条件下的 Top-50 排序质量趋势

- Top-100 排序质量趋势

所有算法在 top-100 之前在不同属性数条件下的变化趋势图如图 9 所示(含与分类结合方法).图 9(a)~图 9(f)分别代表了未知属性数为 1~6 时,各个算法的 top-100 排序质量趋势.Top-100 排序质量趋势图是对 top-50 趋势图做了扩展,以防止因为取的 top-k 数量过小而导致趋势不明.观察图 9,得到的结论与 top-50 趋势图得到的结论一致.



(a) 未知属性数为 1 时的 Top-100 质量趋势 (b) 未知属性数为 2 时的 Top-100 质量趋势

(c) 未知属性数为 3 时的 Top-100 质量趋势 (d) 未知属性数为 4 时的 Top-100 质量趋势

(e) 未知属性数为 5 时的 Top-100 质量趋势 (f) 未知属性数为 6 时的 Top-100 质量趋势

Fig.9 Top-100 quality ranking trend under different number of unknown attributes

图 9 不同未知属性数条件下的 Top-100 排序质量趋势

- 运行时间

所有方法在不同未知属性条件数下的运行时间如图 10 所示(含与分类结合方法).可以明显地看出:与分类结合的方法效率非常低,而其他方法的运行时间都明显优于与分类结合的方法.

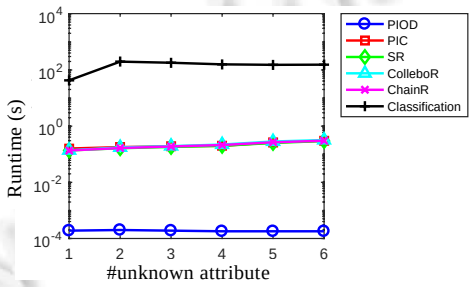


Fig.10 Runtime of all methods under different number of unknown attributes

图 10 所有方法在不同未知属性数条件下的运行时间

6.5 讨 论

在寻求 Why-not 问题解释的应用场景中,有的应用场景对响应时间要求较高,如对 NBA 球星信息的查询;有的场景对解释的准确率要求较高,如用户在查询某个公司的员工信息时,对缺失的年龄、社保号等信息的

Why-not 问题解释准确率要求较高,甚至是绝对准确.所以根据不同的应用场景,基于两两比较模型的 Why-not 问题解释及排序方法更适用于对解释准确率要求较高的场景.

7 结论与未来工作

7.1 结 论

本文受利用两两比较方法寻找函数依赖的算法启发,提出了将两两比较方法和统计学方法以及机器学习方法进行结合的针对 Why-not 问题寻找解释并对解释进行排序的算法.在寻找解释之前,对传统解释的形式进行了重新定义.对于取值域非常大的属性,提出利用两两比较方法将属性的值域压缩到{0,1},并首先利用统计学方法对得到的两两比较结果值进行初始统计,得到一种候选解释和其概率,计算统计学特征后进行排序.之后,和机器学习中的多分类和回归方法进行结合,经过实验证明,与回归方法结合较与多分类结合方法更为有效;与多分类结合的方法比直接进行概率统计的方法更为有效.

与之前已有的成果相比,本文更多地考虑了在求解释的过程中对解释进行详细化和排序,是在以前的相关工作中并未被仔细考虑的方面.实验证明:详细化和排序确实对用户更好更快地寻找解释起到了积极作用.

7.2 未来工作展望

在后续工作中,可以从两个个方面对现有工作进行扩展.

- 首先,考虑更大规模的数据.目前,我们利用两两比较方法寻找解释并对解释进行排序的时候,考虑的是整个数据库中的对比结果.而在后续工作中,可以考虑对原始数据集进行适当采样,使得采样结果可以大致描述整个数据集的样子.然后,在采样结果上进行两两比较的操作与运算;
- 其次,可以考虑频繁更改的数据库.现有的工作只针对于确定的数据库,对于频繁更改的数据库,可以考虑在数据库的时间切片上进行采样,并利用某个时间窗内找到的可能解释,综合解释的时间戳,找到最合理的解释.

综上所述,上述两方面将是日后研究的重点.在将来的工作中,将争取对现有工作进行扩展,得到能够解决更加通用的实际问题的方法.

References:

[1] Benjelloun O, Sarma AD, Halevy A, Widom J. ULDBs: Databases with uncertainty and lineage. In: Proc. of the 32nd Int'l Conf. on Very Large Data Bases. VLDB Endowment, 2006. 953−964.

[2] Bhagwat D, Chiticariu L, Tan WC, Vijayvargiya G. An annotation management system for relational databases. The VLDB Journal, 2005,14(4):373−396.

[3] Bidoit N, Herschel M, Tzompanaki K. Query-based why-not provenance with nedexplain. In: Proc. of the Extending Database Technology (EDBT). 2014.

[4] Bohannon P, Fan W, Geerts F, Jia X, Kementsietsidis A. Conditional functional dependencies for data cleaning. In: Proc. of the Data Engineering (ICDE 2007). IEEE, 2007. 746−755.

[5] Peter B, Khanna S, Tan WC. Why and where: A characterization of data provenance. In: Proc. of the Int'l Conf. on Database Theory. Berlin, Heidelberg: Springer-Verlag, 2001. 316−330.

[6] Chapman A, Jagadish HV. Why not? In: Proc. of the 2009 ACM SIGMOD Int'l Conf. on Management of Data. ACM Press, 2009. 523−534.

[7] Cheney J, Chiticariu L, Tan WC. Provenance in databases: Why, how, and where. Foundations and Trends® in Databases, 2009, 1(4):379−474.

[8] Cormode G, Golab L, Flip K, McGregor A, Srivastava D, Zhang X. Estimating the confidence of conditional functional dependencies. In: Proc. of the 2009 ACM SIGMOD Int'l Conf. on Management of Data. ACM Press, 2009. 469−482.

[9] Cui Y, Widom J. Lineage tracing for general data warehouse transformations. The VLDB Journal—The Int'l Journal on Very Large Data Bases, 2003,12(1):41−58.

[10] Cui Y, Widom J. Practical lineage tracing in data warehouses. In: Proc. of the 16th Int'l Conf. on Data Engineering. IEEE, 2000. 367−378.

[11] Cui Y, Widom J, Wiener JL. Tracing the lineage of view data in a warehousing environment. ACM Trans. on Database Systems (TODS), 2000,25(2):179−227.

[12] Danaparamita J, Gatterbauer W. QueryViz: Helping users understand SQL queries and their patterns. In: Proc. of the 14th Int'l Conf. on Extending Database Technology. ACM Press, 2011. 558−561.

[13] Flach PA, Savnik I. Database dependency discovery: A machine learning approach. AI Communications, 1999,12(3):139−160.

[14] Foster I, Vockler J, Wilde M, Zhao Y. Chimera: A virtual data system for representing, querying, and automating data derivation. In: Proc. of the 14th Int'l Conf. on Scientific and Statistical Database Management. IEEE, 2002. 37−46.

[15] Grust T, Rittinger J. Observing SQL queries in their natural habitat. ACM Trans. on Database Systems (TODS), 2013,38(1):3.

[16] He Z, Lo E. Answering why-not questions on top-*k* queries. IEEE Trans. on Knowledge and Data Engineering, 2014,26(6): 1300−1315.

[17] Hernández M, Koutrika G, Krishnamurthy R, Popa L, Wisnesky R. HIL: A high-level scripting language for entity integration. In: Proc. of the 16th Int'l Conf. on Extending Database Technology. ACM Press, 2013. 549−560.

[18] Herschel M, Hernández MA. Explaining missing answers to SPJUA queries. Proc. of the VLDB Endowment, 2010,3(1-2):185−196.

[19] Huang J, Chen T, Doan A, Naughton JF. On the provenance of non-answers to queries over extracted data. Proc. of the VLDB Endowment, 2008,1(1):736−747.

[20] Islam MS, Zhou R, Liu C. On answering why-not questions in reverse skyline queries. In: Proc. of the 2013 IEEE 29th Int'l Conf. on Data Engineering (ICDE). IEEE, 2013. 973−984.

[21] Lopes S, Petit JM, Lakhal L. Efficient discovery of functional dependencies and armstrong relations. In: Proc. of the Int'l Conf. on Extending Database Technology. Berlin, Heidelberg: Springer-Verlag, 2000. 350−364.

[22] Meliou A, Gatterbauer W, Moore KF, Suciu D. The complexity of causality and responsibility for query answers and non-answers. Proc. of the VLDB Endowment, 2010,4(1):34−45.

[23] Miles S, Wong SC, Fang W, Groth P, Zauner KP, Moreau L. Provenance-based validation of e-science experiments. Web Semantics: Science, Services and Agents on the World Wide Web, 2007,5(1):28−38.

[24] Mutsuzaki M, Theobald M, De Keijzer A, Widom J, Agrawal P, Benjelloun O, Das Sarma A, Murthy R, Sugihara T. Trio-one: Layering uncertainty and lineage on a conventional DBMS. In: Proc. of the 3rd Biennial Conf. on Innovative Data Systems Research. 2007. 269−274.

[25] Qi DR. On concise explanations of non-answers over big data. In: Proc. of the 2017 ACM Int'l Conf. on Management of Data. ACM Press, 2017. 10−12.

[26] Tran QT, Chan CY. How to conquer why-not questions. In: Proc. of the 2010 ACM SIGMOD Int'l Conf. on Management of Data. ACM Press, 2010. 15−26.

[27] Zhang AQ, Song SX, Wang JM. Reducing explanations of Non-answers using data quality rules. Journal of Computer Research and Development, 2013,(zl):221−229 (in Chinese with English abstract).

[28] Wyss C, Giannella C, Robertson E. Fastfds: A heuristic-driven, depth-first algorithm for mining functional dependencies from relation instances extended abstract. In: Proc. of the Int'l Conf. on Data Warehousing and Knowledge Discovery. Berlin, Heidelberg: Springer-Verlag, 2001. 101−110.

附中文参考文献:

[27] 张奥千,宋韶旭,王建民.基于数据质量规则的缺失结果解释约减.计算机研究与发展,2013,(zl):221−229.

祁丹蕊(1997－),女,内蒙古赤峰人,硕士生,主要研究领域为数据清洗.

宋韶旭(1981－),男,博士,副教授,博士生导师,CCF 专业会员,主要研究领域为数据库.

王建民(1968－),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,工作流.