较大,当α为一个较小值(α=0.65)时,如图 8(a)所示,更新次数也相对较多,即对于权值波动较大的数据集,α的变化对更新次数的影响相对较小.因此α在图 10(a)设置的参数范围内发生变化时,CTF-Stream 准确性的变化也较小.此外,CTF-Stream 在更新点处执行 CRH 算法.而在对静态数据集进行真值发现时,CRH 与 GTM 相比表现出很高的准确性[5].因此,即使 CTF-Stream 并没有在数据流上连续地更新数据源的权值,由于其在更新点处的准确性很高,相比于每一时刻都更新数据源权值的迭代算法 GTM,CTF-Stream 依然表现出很好的准确性.并且,在 Intel Berkeley 实验室数据集上,GTM 的均方误差随着数据集合的增大而增大,而 CTF-Stream 在大部分参数设置下其准确率都是随着数据集合的增大而减小的,这表明,CTF-Stream 在处理大规模感知数据流时具有很大优势.

上述实验在验证了 CTF-Stream 的高效性和准确性的同时,也充分说明了本文选取的两个真实数据集合具有比较显著的差异,进而说明 CTF-Stream 在处理不同类型和变化模式的数据集合时,均能表现出良好的性能.

6 结 论

真值发现作为数据集成中一种冲突消解的有效手段,在传统数据库领域已经得到了广泛的研究.但是由于时间、空间复杂度等限制,基于传统数据库的真值发现技术无法应用于一种越来越普遍的数据模型——数据流中.本文针对一种特殊的数据流——感知数据流上的连续真值发现问题进行了研究.结合感知数据自身及其应用特点,提出了一种变频评估数据源可信度的策略以平衡感知数据流真值发现的效率和准确率.本文首先定义并研究了感知数据流真值发现的相对误差和累积误差,以及它们较小时数据源在相邻时刻可信度的变化应满足的条件.进而提出一种概率模型,以预测数据源在相邻时刻可信度的变化满足该条件的概率.最后,整合上述结论,将感知数据流真值发现中的累积误差预测问题转化为一个最优化问题,在限制了累积误差的前提下,最大化数据源可信度的评估周期以提高效率.在此基础上提出了一种基于累积误差预测的数据源可信度变频更新算法——CTF-Stream,对连续到达的感知数据流进行真值发现.CTF-Stream 结合历史数据,动态地确定数据源可信度的评估周期,同时以一定概率确保真值发现结果的准确性.CTF-Stream 通过减少更新数据源可信度的次数,减少了迭代过程的执行,极大地提高了真值发现的效率.最后,本文在真实的感知数据集合上进行实验,实验结果表明,本文提出的算法在处理数据流上的真值发现问题时具有较高的准确率和效率.

References:

[1] Yin XX, Han JW, Yu PS. Truth discovery with multiple conflicting information providers on the Web. IEEE Trans. on Knowledge and Data Engineering, 2007,20(6):796−808. [doi: 10.1109/TKDE.2007.190745]

[2] Galland A, Abiteboul S, Marian A, Senellart P. Corroborating information from disagreeing views. In: Proc. of the WSDM. New York, 2010. 131−140. https://hal.inria.fr/inria-00429546/document

[3] Zhao B, Han JW. A probabilistic model for estimating real-valued truth from conflicting sources. In: Proc. of the QDB. Istanbul, 2012. http://web.engr.illinois.edu/~hanj/pdf/qdb12_bzhao.pdf

[4] Zhao B, Rubinstein BIP, Gemmell J, Han JW. A Bayesian approach to discovering truth from conflicting sources for data integration. PVLDB, 2012,5(6):550−561. [doi: 10.14778/2168651.2168656]

[5] Li Q, Li YL, Gao J, Zhao B, Fan W, Han JW. Resolving conflicts in heterogeneous data by truth discovery and source reliability estimation. In: Proc. of the SIGMOD. Snowbird, 2014. 1187−1198. http://hanj.cs.illinois.edu/pdf/sigmod14_jgao.pdf

[6] Li Q, Li YL, Gao J, Demirbas M, Zhao B, Su L, Fan W, Han JW. A confidence-aware approach for truth discovery on long-tail data. PVLDB, 2014,8(4):425−436.

[7] Dong XL, Berti-Equille L, Srivastava D. Integrating conflicting data: The role of source dependence. PVLDB, 2009,2(1):550−561.

[8] Dong XL, Berti-Equille L, Srivastava D. Truth discovery and copying detection in a dynamic world. PVLDB, 2009,2(1):562−573. [doi: 10.14778/1687627.1687691]

[9] Dong XL, Berti-Equille L, Hu YF, Srivastava D. Global detection of complex copying relationships between sources. PVLDB, 2010,3(1-2):1358−1369.

[10] Dong XL, Berti-Equille L, Hu YF, Srivastava D. Solomon: Seeking the truth via copying detection. PVLDB, 2010,3(1-2): 1617−1620. [doi: 10.1145/1966883.1966887]

[11] Dong XL, Gabrilovich E, Murphy K, Dang V, Horn W, Lugaresi C, Sun S, Zhang W. Knowledge-Based trust: Estimating the trustworthiness of Web sources. PVLDB, 2015,8(9):938−949.

[12] Pochampally R, Das-Sarma A, Dong XL, Meliou A, Srivastava D. Fusing data with correlations. In: Proc. of the SIGMOD. Snowbird, 2014. 433−444. http://lunadong.com/publication/fusionWCorr_sigmod.pdf

[13] Li X, Dong XL, Lyons K, Meng W, Srivastava D. Truth finding on the deep Web: Is the problem solved. PVLDB, 2012,6(2): 97−108.

[14] Song SX, Zhang AQ, Wang JM, Yu PS. SCREEN: Stream data cleaning under speed constraints. In: Proc. of the SIGMOD. Melbourne, 2015. 827−841. http://ise.thss.tsinghua.edu.cn/sxsong/doc/15sigmod-screen.pdf

[15] Cao L, Yang D, Wang QY, Yu YW, Wang JY, Rundensteiner EA. Scalable distance-based outlier detection over high-volume data streams. In: Proc. of the ICDE. 2014. 76−87. [doi: 10.1109/ICDE.2014.6816641]

[16] Zhao Z, Cheng J, Ng W. Truth discovery in data streams: A single-pass probabilistic approach. In: Proc. of the CIKM. Shanghai, 2014. 1589−1598. http://er2004.cse.ust.hk/~wilfred/paper/cikm14a.pdf

[17] Li JZ, Li JB, Shi SF. Concepts, issues and advance of sensor networks and data management of sensor networks. Ruan Jian Xue Bao/Journal of Software, 2003,14(10):1717−1727 (in Chinese with English abstract). http://www.jos.org.cn/ch/reader/create_pdf.aspx?file_no=20031007&journal_id=jos

[18] Zhao Z, Ng W. A model-based approach for rfid data stream cleansing. In: Proc. of the CIKM. Hawaii, 2012. 862−871. http://www.cs.ust.hk/~wilfred/paper/cikm12b.pdf

[19] Cheng SY, Li JZ, Yu L. Location aware peak value queries in sensor networks. In: Proc. of the INFOCOM. 2012. 486−494. [doi: 10.1109/INFCOM.2012.6195789]

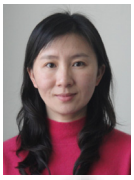
[20] Raza U, Camerra A, Murphy A, Palpanas T, Picco GP. Practical data prediction for real-world wireless sensor networks. IEEE Trans. on Knowledge and Data Engineering, 2015,PP(8):1. [doi: 10.1109/TKDE.2015.2411594]

[21] Li YL, Li Q, Gao J, Su L, Fan W, Han JW. On the discovery of evolving truth. In: Proc. of the SIGKDD. Sydney, 2015. 675−684. http://www.cse.buffalo.edu/~lusu/papers/KDD2015Yaliang.pdf

附中文参考文献:

[1] 李建中,李金宝,石胜飞.传感器网络及其数据管理的概念、问题与进展.软件学报,2003,14(10):1717−1727. http://www.jos.org.cn/ch/reader/create_pdf.aspx?file_no=20031007&journal_id=jos

李天义(1992－),女,辽宁锦州人,学士,主要研究领域为数据质量.



李芳芳(1977－),女,博士,讲师,CCF 会员,主要研究领域为数据库技术,传感器网络 CPS 数据管理.

谷峪(1981－),男,博士,副教授,CCF 高级会员,主要研究领域为图,空间数据管理.

于戈(1962－),男,博士,教授,博士生导师,CCF 会士,主要研究领域为数据管理理论与技术,分布与并行系统.

马茜(1988－),女,硕士生,CCF 学生会员,主要研究领域为感知数据管理.

© 中国科学院软件研究所 http://www.jos.org.cn