

引用格式:王 玥. 关系数据库中分布式大数据的集成冲突消解算法[J]. 科学技术与工程, 2018, 18(3): 63—67
Wang Yue. integrated conflict resolution algorithm for distributed large data in relational databases[J]. Science Technology and Engineering, 2018, 18(3): 63—67

关系数据库中分布式大数据的集成冲突消解算法

王 玥

(南阳理工学院软件学院, 南阳 473000)

摘 要 针对现有关系数据库中分布式大数据集成冲突消解研究的不足,提出一种新的集成冲突消解算法。依据关系数据库中分布式大数据的集成过程对冲突进行分类,将其划分成语义冲突、模式冲突以及实例冲突。针对语义冲突,通过句法融合、逻辑树融合和频率融合法实现冲突消解。通过属性有向图对关系数据库中模式数据和实例数据的属性进行描述。从属性关系参与分布式大数据集成冲突的状态分析,通过关系的权重值对属性关系的重要程度进行量化处理。通过有向图全部关系的权重和对所有属性有向图的重要程度进行描述。综合分析冲突数与权重定义代价函数,在此基础上给出关系数据库分布式大数据集成冲突消解详细过程。实验结果表明,所提算法冲突识别和消解性能高。

关键词 关系数据库 分布式 大数据 集成 冲突消解

中图分类号 TP391; **文献标志码** A

因为关系数据库中分布式大数据信息集成要求若干分布式数据源集成至一个全局模式中,所以不同数据源之间一定有很多互相影响的关系,不同数据源对相同事物的表达方式、实现过程等也存在差异,上述因素一定会造成关系数据库中分布式大数据集成时冲突的出现^[1,2]。所以,冲突为关系数据库中分布式大数据集成的特点。

在分布式大数据集成时,冲突在很大程度上会出现在不同阶段。从某种程度上来说,关系数据库中分布式大数据的集成过程实质上就是冲突产生与消解过程^[3]。所以,冲突识别和消解变成关系数据库中分布式大数据集成中非常关键的技术^[4]。

当前,数据冲突消解为分布式大数据集成的研究重点与难点。通过分析关系数据库中分布式大数据集成冲突消解过程,提出解决算法,为分布式大数据的集成提供有效依据。

1 关系数据库中分布式大数据的集成冲突消解算法

1.1 分布式大数据集成冲突分类

一个关系数据库中分布式大数据集成系统可通过三元组 $I = \langle G, S, M_{GS} \rangle$ 进行描述,其中 G 为全局模式,主要通过定义在字母表 A_c 上的语言 L_c 体现,

G 中的所有元素均为通过字母表 A_c 组成的符号; S 为局部模式,其实际上是一个集合 $|S_1, S_2, \dots, S_n|$, 其中 $S_1, S_2, \dots, S_n$ 分别为不同分布式大数据的局部模式。所有局部模式 S_i 均可通过定义在字母表 A_{S_i} 的语言 L_{S_i} 进行描述, S_i 中的所有元素均是通过字母表 A_c 组成的符号; M_{GS} 为 G 和 S 间的映射,可通过形如 $R_c - R_s$ 的规则进行描述。

在关系数据库分布式大数据集成过程中,导致冲突的原因是多方面的,所有数据源均有其独立的结构和知识库,这对分布式大数据的独立性产生直接影响^[5]。因为所有数据源的信息均受到限制,所以关系数据库中的全局数据模式需和所有分布式大数据进行转换才可实现集成,也就是所有分布式大数据间都有依赖性。独立性与依赖性两个相互矛盾的因素一定会引起冲突产生。

现依据关系数据库中分布式大数据的集成过程对冲突进行分类^[6]。依据数据间存在的冲突问题,将数据冲突划分成语义冲突、模式冲突以及实例冲突。详细情况如图 1 所示。

1.2 语义冲突消解算法

1.2.1 句法融合

句法融合即对不同知识元的术语集与谓词集进行逻辑加获取的结果。下面分别对术语集合 Δ 与谓词集合 Σ 的句法融合进行定义:

$$S_{\text{yntax}}(\Delta) = S_{\text{yntax}}(s_1, s_2, \dots, s_n) \tag{1}$$

$$S_{\text{yntax}}(\Sigma) = S_{\text{yntax}}(g_1, g_2, \dots, g_m) \tag{2}$$

式中, $s_1, s_2, \dots, s_n$ 为不同术语; $g_1, g_2, \dots, g_m$ 为谓词

2017年6月22日收到
作者简介:王 玥(1978—),女,汉族,河南南阳人,硕士,讲师。研究方向:数据库技术。E-mail:wangyuewang41245@sina.com。

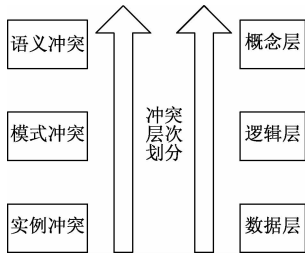


图1 分布式大数据集成中冲突分类

Fig.1 Conflict classification in distributed large data integration

集合。句法融合主要是为了滤除术语集合和谓词集合中的冗余数据^[7]。所以句法融合可实现语义冲突的初步消解,必要时可将句法融合与其余融合方法共同使用,增强准确性。

1.2.2 逻辑树融合法

在关系数据库中分布式大数据集成过程中,一条知识元的语义集合 K_s 可划分成术语集合 Δ 和谓词集合 Σ ,也就是

$$K_s = \{\Delta, \Sigma\} \quad (3)$$

若两个术语中其中一个在概念上的概括性更强,则该关系被称作类属关系或包含关系。

假设术语集合 Δ 的逻辑树集合用 $\text{Logic}(\Delta, T, N)$ 进行描述,其中 Δ 为分布式大数据中知识元的术语集合, $\Delta = \{s_1, s_2, \dots, s_n\}$; T 为 Δ 的逻辑树; N 为逻辑树节点。如果 $s_i \subset s_j$, 则 $N = s_j$; 如果 $s_i \Leftrightarrow s_j$, 则 $N = s_i$; 如果 $s_i = s_j$, 则 $N = s_k$, $N = s_k$ 代表 s_i 的父类。

1.2.3 频率融合法

针对知识元的术语集合 $\Delta = \{s_1, s_2, \dots, s_n\}$, 在术语项 s_i 出现术语冲突的情况下, 可将使用频率最高的看作融合结果, 也就是频率融合法^[8]。假设术语项 s_i 的使用频率为 f_i , 则术语集合 $\Delta = \{s_1, s_2, \dots, s_n\}$ 的频率融合可描述成 $F_{\text{frequency}}(\Delta, f, X)$ 。如果 $f_i > f_j$, 则 $X = s_i$; 如果 $f_i < f_j$, 则 $X = s_j$; 如果 $f_i = f_j$, 则 $X = s_i \text{ or } s_j$ 。

1.3 属性有向图及冲突消解算法

1.3.1 属性有向图

现通过属性有向图对关系数据库中模式数据和实例数据的属性进行描述。将约束属性取值范围称作约束条件, 前提属性取值称作前提条件^[9,10]。

通过 a_{ufun} 形式令前提关系中的前提属性和约束属性符合约束条件, 一个属性 o 中的两个 a_{ufun} 约束 i 和 j 可同时成立记作 $S_{\text{atisfy}}(a_{\text{ufun}_i}, a_{\text{ufun}_j})$ ^[11]。特别地, 在 a_{ufun} 函数前添加符号前缀 C 、 R , 确定前提属性和约束属性符合的条件。通过属性 i 对属性 j 组成一个属性关系是 $P_{i,j}(C_{\text{at}} - a_{\text{ufun}_i}, R_{\text{a}_{\text{ufun}_j}})$ 。

将属性集看作顶点集合 V , 属性间的关系集是边的集合 E , 可获取一个有向图 $G(V, E)$ ^[12,13]。

需要注意的是, 若针对 $0 \leq i \leq k-2$, 均存在 $P_i(v_i, v_{i+1}), P_{i+1}(v_{i+1}, v_{i+2}) \in E$, 同时 $S_{\text{atisfy}_i} + 1(R_{\text{attfun}_{i,i+1}}, C_{\text{attfun}_{i+1,i+2}})$, 则将 $v_0, v_1, \dots, v_k$ 称作 G 的一条长是 k 的有向路^[14]。在 $v_0 = v_k$ 的情况下, $v_0, v_1, \dots, v_k$ 被称作有向环。

一个属性关系所关联的属性到另一个属性关系的属性的有向图如图2所示。

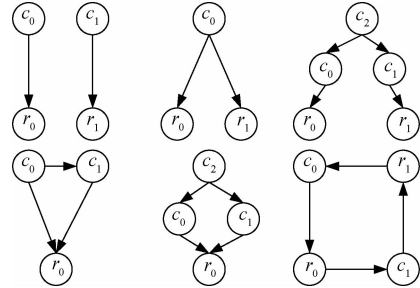


图2 两个属性关系间基本关系分类

Fig.2 Classification of basic relations between two attribute relations

1.3.2 代价函数冲突消解算法

从属性关系参与分布式大数据集成冲突的状态分析, 一个属性关系参与的冲突数越多, 则可消解的冲突越多^[15,16]。图3为多冲突场景举例图。

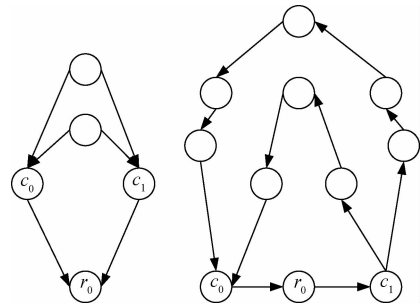


图3 多冲突场景举例

Fig.3 Examples of multi conflict scenarios

现通过关系的权重值对属性关系的重要程度进行量化, 权重越高越重要^[17]。所有属性有向图的重要程度可通过有向图全部关系的权重和进行描述^[18]。综合分析冲突数与权重表示的代价函数为

$$f_i = \frac{\alpha C_{\text{onflict}} N_{\text{um}_i}}{\beta w_i + \lambda} \quad (4)$$

式(4)中, f_i 为有向图中第 $i+1$ 个属性关系的代价, 所有属性关系均存在一个消解的冲突代价; $C_{\text{onflict}} N_{\text{um}_i}$ 为第 $i+1$ 个属性关系参与冲突数量; w_i 为第 $i+1$ 个属性关系在有向图中的权重值; α 、 β 、 λ 为代价函数中冲突数量和权重值的常数^[19,20]。

冲突消解详细过程为:

(1) 对属性有向图中的数据关系进行初始化处理, 并赋予其权重。

(2) 检测属性有向图中的冲突类型, 对所有属

性关系参与的冲突数量进行记录。

(3) 如果属性有向图中总冲突数量非零,则求出所有属性关系的代价函数值,继续进行下一步;否则,结束迭代。

(4) 选择目前代价函数值最大的属性关系,将其删除,获取新的属性有向图,重新进行步骤(2)。

2 实验验证

为了更好地验证本文算法的有效性,对本文提出消解算法进行实验说明,实验主要包括冲突识别和冲突消解两部分。

2.1 冲突识别实验验证与分析

为了更好地验证本文算法的适用性,实验开展了和概念相似度算法与规则推理算法的对比实验。分析实验结果可知,本文算法能够有效发现数据冲突,同时可指出数据类型。

评价指标选择召回率和准确率,召回率为发现冲突量占总冲突量的比例,精确率为在发现冲突为真正冲突的数量占发现冲突的比例。现对某关系数据库中随机选择分布式大数据,每次选择 500 条数据,一共选择 4 次。利用本文算法、概念相似度算法与规则推理算法对其中的集成冲突进行识别,结果如表 1 所示。

分析表 1 中的数据可以看出,不管是针对语义冲突、模式冲突还是实例冲突,本文算法召回率和准确率均高于概念相似度算法和规则推理算法,说明本文算法能够有效消解关系数据库中不同类型分布式大数据集成冲突。

为了进一步验证本文算法的有效性,单独使用概念相似度算法和规则推理算法对冲突进行识别,再将二者结合在一起对冲突进行识别,把识别过程中的召回率和准确率和本文算法进行比较,得到的结果分别见图 4 和图 5。

分析图 4 和图 5 可以看出,单独使用规则推理算法比单独使用概念相似度算法效果略好,这主要是由于在进行概念相似度计算的过程中,无法求出

关系数据库中数据相似度,导致部分矛盾冲突未被发现,然而规则推理算法由于通过约束逻辑关系进行推理,所以得到的冲突识别结果更加全面。将二者结合在一起,发挥了二者的优势,得到的结果比单独使用两种算法的召回率与准确率都高,但仍旧低于本文算法的召回率和准确率,进一步验证了本文算法在冲突识别方面的性能。

2.2 冲突消解实验

对关系数据库中分布式大数据集成冲突进行消解后,其冲突消解效果可利用冲突识别指数、冲突消解密度指数、冲突消解完备率指数三个指标进行衡量,下面给出三个指标的计算公式。

冲突识别密度指数即识别冲突总量在总数据样本规模中的分布情况,用 χ 进行描述,计算公式为

$$\chi = \log_{s_{ize}} v \tag{5}$$

式(5)中, $s_{ize} > 0$ 为关系数据库中分布式大数据规模值,常被称作规模参数。

冲突识别密度指数越高,则冲突消解的次数越多。

冲突消解密度指数即消解冲突总量占数据样本规模的相对分布。用 γ 进行描述,计算公式为

$$\gamma = s_{ize}^{\rho} \tag{6}$$

式(6)中, ρ 为冲突消解次数; γ 值越高,认为冲突消解密度越高。

冲突消解完备率即冲突消解密度指数占冲突识别密度指数的相对分布,用 Z 进行描述,计算公式如下

$$Z = s_{ize}^{\frac{\chi}{\gamma}} \tag{7}$$

冲突消解完备率是冲突识别情况和冲突消解情况一致性的体现,二者越一致, Z 越接近 1, 否则 Z 越偏离 1。

通过本文提出的算法对随机选择的四组分布式大数据进行冲突消解。对其中冲突识别密度指数和冲突消解密度指数进行统计,可获取图 6 所示的结果,对其中的冲突消解完备率进行统计可获取图 7 所示的结果。

表 1 三种算法集成冲突识别结果比较

Table 1 Comparison of conflict resolution results between three algorithms

数据选择 次数及指标	本文算法			概念相似度算法			规则推理算法		
	语义冲突/个	模式冲突/个	实例冲突/个	语义冲突/个	模式冲突/个	实例冲突/个	语义冲突/个	模式冲突/个	实例冲突/个
①	15	8	13	8	12	8	10	11	11
②	17	12	15	10	13	13	9	9	18
③	15	11	19	9	10	14	10	14	12
④	11	16	28	10	9	24	10	13	16
召回量/个	58	47	75	37	44	59	39	47	57
确认量/个	55	43	74	30	38	51	31	40	49
召回率/%	89.96	93.81	94.53	72.36	80.27	85.32	76	81.3	85.6
准确度/%	94.93	91.49	98.67	81.08	86.36	86.44	82.49	87.11	87.96

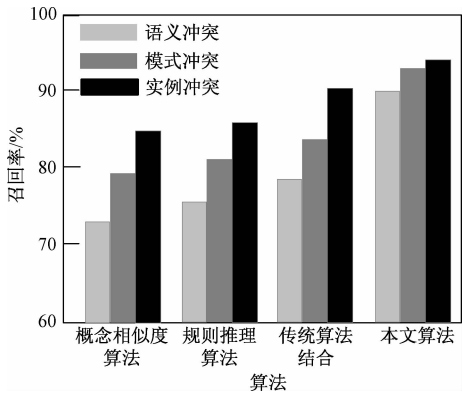


图4 传统两种算法结合识别召回率与本文算法比较结果
Fig. 4 Traditional two algorithms combined with the recognition recall rate and the results of this algorithm

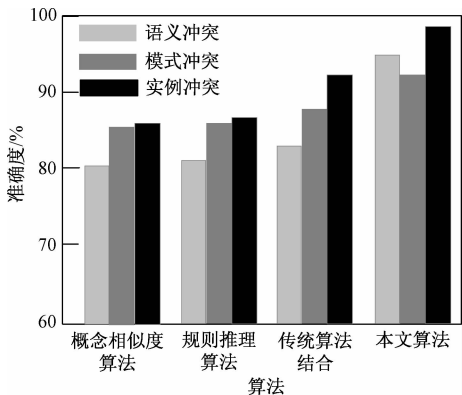


图5 传统两种算法结合识别准确度与本文算法比较结果
Fig. 5 Traditional two algorithms combine the recognition accuracy and the results of this algorithm

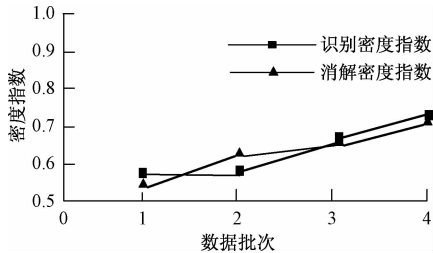


图6 冲突密度指数和冲突消解指数统计结果
Fig. 6 Statistical results of conflict density index and conflict resolution index

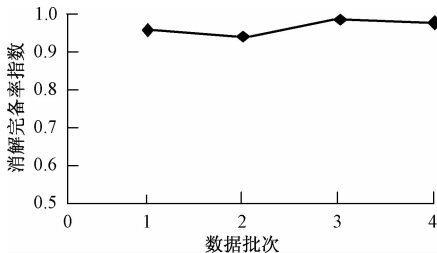


图7 冲突消解完备率统计结果
Fig. 7 Conflict resolution complete rate statistic results

分析图6和图7可以看出,本文算法冲突识别密度指数稳定性较高,消解密度指数也同样有很高的稳定性。利用消解完备性指数对冲突识别性能进行衡量有很高的效果。除此之外还可以看出,本文算法的冲突密度指数和冲突消解指数值均较高,说明本文算法的冲突消解次数较多,能够有效实现冲突消解。且本文算法处理下冲突消解完备率均趋近于1,说明本文算法处理下冲突识别情况和冲突消解情况一致性较高,进一步验证了本文算法的有效性。

3 结论

提出一种新的关系数据库中分布式大数据集成冲突消解算法。针对语义冲突、模式冲突以及实例冲突进行冲突消解。经实验验证,所提算法冲突识别和消解性能高。

参 考 文 献

- 1 李卫榜,李战怀,陈 群,等. 分布式大数据函数依赖发现. 计算机研究与发展,2015;52(2):282—294
Li Weibang, Li Zhanhuai, Chen Qun, et al. Functional dependencies discovering in distributed big data. Journal of Computer Research and Development, 2015;52(2):282—294
- 2 姚 刚. 基于 BIM 的工业化住宅协同设计中的冲突消解方法研究. 施工技术,2016;45(18):43—47
Yao Gang. Research on BIM-based method of conflict resolution to the collaborative design for the industrialized housing. Construction Technology, 2016;45(18):43—47
- 3 孙彩虹. 网络舆情之于司法审判:冲突与优化. 河南大学学报(哲学社会科学版),2015;55(5):28—35
Sun Caihong. From network public opinion to judicial trials: conflict and optimization. Journal of Henan University (Social Science), 2015;55(5):28—35
- 4 熊 嵩,周 军,呼卫军. 分布式攻防对抗仿真负载分析与平衡算法研究. 计算机仿真,2014;31(4):42—45
Xiong Song, Zhou Jun, Hu Weijun. Load analysis and balancing solution design for distributed confrontation simulation. Computer Simulation, 2014;31(4):42—45
- 5 Jenie Y I, Kampen E J V, Ellerbroek J, et al. Taxonomy of conflict detection and resolution approaches for unmanned aerial vehicle in an integrated airspace. IEEE Transactions on Intelligent Transportation Systems, 2017;18(3):558—567
- 6 李木国,王延国,孙慧涛. 基于 EtherCAT 总线的串联型分布式数据采集系统设计. 计算机测量与控制,2016;24(6):195—198
Li Muguo, Wang Yanguo, Sun Huitao. Design of series-connection distributed data acquisition system based on etherCAT bus. Computer Measurement & Control, 2016;24(6):195—198
- 7 潘 理,郑 红,刘显明,等. 基于 Petri 网局部性的极大冲突集枚举算法. 电子学报,2016;44(8):1858—1863
Pan Li, Zheng Hong, Liu Xianming, et al. Maximal conflict set enumeration algorithm based on locality of Petri nets. Acta Electronica Sinica, 2016;44(8):1858—1863
- 8 何洪雨,林志文,魏国珩,等. 基于多 Agent 的分布式信号流模型

- 故障诊断方法研究. 舰船电子工程, 2014; 34(8): 145—147
- He Hongyu, Lin Zhiwen, Wei Guoheng, *et al.* Distributed fault diagnosis method based on multi-agent signal flow model. Ship Electronic Engineering, 2014; 34(8): 145—147
- 9 宋顶利, 张 昕, 于复兴. 分布式优化 Apriori 算法的交通运行状态数据分析模型. 科技通报, 2016; 32(10): 202—206
- Song Dingli, Zhang Xin, Yu Fuxing. Data analysis model of traffic running based on the distributed optimal apriori algorithm. Bulletin of Science and Technology, 2016; 32(10): 202—206
- 10 武玉英, 李 豪, 蒋国瑞. 基于 RBF 神经网络和强化学习算法的供应链产销协同计划冲突消解研究. 计算机应用研究, 2015; 16(5): 1335—1338
- Wu Yuying, Li Hao, Jiang Guorui. Research on conflict resolution of supply chain production-marketing collaborative planning based on RBF and Q-reinforcement. Application Research of Computers, 2015; 16(5): 1335—1338
- 11 李文昊, 李海芳. 确定性分布式数据库中长事务处理方法研究. 科学技术与工程, 2016; 16(13): 92—95
- Li Wenhao, Li Haifang. Research on the method of long transaction in deterministic distributed database. Science Technology and Engineering, 2016; 16(13): 92—95
- 12 沈 虎, 吕绍和, 王晓东, 等. 一种应用干扰消除进行冲突消解的分布式无线 MAC 协议. 计算机科学, 2014; 41(12): 60—66
- Shen Hu, Lü Shaohe, Wang Xiaodong, *et al.* Distributed collision-resolvable MAC protocol for wireless LANs with interference cancellation support. Computer Science, 2014; 41(12): 60—66
- 13 Delhibabu R, Behrend A. A new rational algorithm for view updating in relational databases. Applied Intelligence, 2015; 42(3): 1—15
- 14 Khazalah F, Malik Z, Rezgui A. Automated conflict resolution in collaborative data sharing systems using community feedbacks. Information Sciences, 2015; 29(8): 407—424
- 15 高丽萍, 郭淑娴, 张玉本, 等. 依赖图文档模型下协同冲突消解研究. 小型微型计算机系统, 2015; 36(12): 2639—2643
- Gao Liping, Guo Shuxian, Zhang Yuben, *et al.* Research on collaborative conflict resolution in dependency graph document model. Journal of Chinese Computer Systems, 2015; 36(12): 2639—2643
- 16 杨 沁, 唐 伟, 李建国. 产品定制中客户需求冲突的高效消解研究. 机械科学与技术, 2015; 34(1): 94—98
- Yang Qin, Tang Wei, Li Jianguo. Study on the efficiently solution of customer requirements conflicts in product customization. Mechanical Science and Technology for Aerospace Engineering, 2015; 34(1): 94—98
- 17 Wei F, Ito K, Sakata K, *et al.* Pretreatment and integrated analysis of spectral data reveal seaweed similarities based on chemical diversity. Analytical Chemistry, 2016; 87(5): 281—296
- 18 孟秀丽, 王海燕, 唐 润, 等. 基于协商视角的食品质量链冲突消解策略. 系统工程理论与实践, 2014; 34(12): 3130—3137
- Meng Xiuli, Wang Haiyan, Tang Run, *et al.* Strategy of conflict resolution for food quality chain based on negotiation. Systems Engineering-Theory & Practice, 2014; 34(12): 3130—3137
- 19 陈锐忠, 魏理豪, 梁哲恒, 等. 基于 Hadoop 的海量数据处理模型研究和应用. 电子设计工程, 2016; 24(14): 101—103
- Chen Ruizhong, Wei Lihao, Liang Zheheng, *et al.* Research and application of mass data processing model based on Hadoop. Electronic Design Engineering, 2016; 24(14): 101—103
- 20 朴成日, 沈治河. 联合机动编队兵力配置冲突消解. 舰船科学技术, 2015; 37(2): 120—123
- Piao Chengri, Shen Zhihe. Conflict resolution on disposition of forces in joint maneuver formation. Ship Science and Technology, 2015; 37(2): 120—123

Integrated Conflict Resolution Algorithm for Distributed Large Data in Relational Databases

WANG Yue

(Software College, Nanyang Institute of Technology, Nanyang 473000, China)

[**Abstract**] A new integrated conflict resolution algorithm is proposed for the research on the conflict resolution of distributed large data integration in existing relational databases. According to the integration process of distributed large data in relational database, conflicts are classified and classified into semantic conflict, model conflict and instance conflict. Aiming at semantic conflict, conflict resolution is realized by syntax fusion, logical tree fusion and frequency fusion. The properties of schema data and instance data in relational databases are described by attribute directed graphs. From the relational analysis of attribute relation to the conflict of distributed large data integration, the importance of attribute relation is quantified by the weight value of relation. The weight of all relations and the importance of the directed graph for all attributes are described by means of directed graphs. Based on the analysis of the conflict number and the weight, the cost function is defined, and the detailed process of distributed data integration and conflict resolution in relational database is given. Experimental results show that the proposed algorithm has high performance in conflict recognition and resolution.

[**Key words**] relational database distributed big data integration conflict resolution