第17卷 第9期 2017年3月
1671—1815(2017)09-0239-05

科 学 技 术 与 工 程
Science Technology and Engineering

Vol. 17 No. 9 Mar. 2017

复杂属性环境下 NoSQL 分布式大数据挖掘方法研究

梅 毅 熊 婷 罗少彬
(南昌大学科学技术学院计算机系,南昌 330029)

摘 要 由于复杂属性环境下的大数据挖掘工作需要涉及到对大数据的分析、清理、转换和集成等一系列操作,导致以往提出的复杂属性环境下大数据挖掘方法无法同时拥有较强的准确性、稳定性和实用性,故提出复杂属性环境下 NoSQL 分布式大数据挖掘方法。所提方法利用 NoSQL 数据库的物理分散逻辑,在复杂属性环境下构建 NoSQL 数据库,给出挖掘条件,对数据库中大数据的特征、位置、方向和长度属性的关联性进行分布式挖掘,经由挖掘公式给出挖掘结果。利用挖掘聚类公式对大数据的特征、位置、方向和长度属性的关联性挖掘结果进行聚类,获取所提方法的最终挖掘结果。经实验分析可知,所提方法在挖掘工作中具有较强的准确性、稳定性和实用性。

关键词 复杂属性环境 NoSQL 分布式 大数据挖掘
中图法分类号 TP391.3; **文献标志码** A

近年来,随着网络资源利用率的不断提高,各行各业对大数据挖掘的重视程度也越发增强,尤其在复杂属性环境下,大数据的各类特征参数较多,影响了使用者对大数据的利用程度,为此,人们需要利用数据库对大数据进行合理规划和有效挖掘[1—3]。科研组织曾提出过一些复杂属性环境下大数据挖掘方法,但由于复杂属性环境下的挖掘工作需要涉及到数据的分析、清理、转换和集成等一系列操作,导致以往提出的方法无法在挖掘工作中同时拥有较强的准确性、稳定性和实用性,对此类方法的研究工作仍在紧张地进行当中[4—6]。

在复杂属性环境下大数据挖掘方法中,数据库的选择格外重要,性能良好的数据库能够挖掘出更为准确和稳定的大数据,如文献[7]提出复杂属性环境下 RDBMS 大数据挖掘方法,由于 RDBMS 数据库的稳定性虽强但灵活性很差,为研究工作带来了一定的难处,使得此方法无法拥有较强的挖掘准确性,且只适用于数据量较小的挖掘工作;文献[8]提出基于标签技术的复杂属性环境下分布式大数据挖掘方法,这一方法利用 OLAP 数据库将复杂属性环境下的大数据聚合成 3D 立体图像,并采用分布式理念提高方法实用性,但其挖掘准确性和稳定性仍需提高;文献[9]提出复杂属性环境下 MYSQL 大数据挖掘方法,此方法重点研究了复杂属性环境下 MYSQL 数据库的大数据查询功能,以精准的查询工作为挖掘准确性和稳定性提供保障。但由于查询工作过于繁杂,导致此方法的实用性不强;文献[10]提出基于贝叶斯模型的复杂属性环境下大数据挖掘方法,贝叶斯模型对复杂属性环境下的大数据进行分类,再将其导入 RDBMS 数据库,在一定程度上解决了 RDBMS 数据库不灵活性给挖掘工作带来的不好影响。但此方法的实用性仍不强。

根据对以上的复杂属性环境下大数据挖掘方法进行分析,提出复杂属性环境下 NoSQL 分布式大数据挖掘方法。经实验分析可知,所提方法具有较强的挖掘准确性、稳定性和实用性。

1 复杂属性环境下 NoSQL 分布式大数据挖掘描

1.1 NoSQL 分布式大数据挖掘概述

NoSQL 是一种根据物理分散逻辑进行数据规划的分布式数据库,可较好地增强挖掘方法的准确性、稳定性和实用性。NoSQL 数据库拥有灵活性佳、实用性强、价格低廉和高效等优点,可对复杂属性环境下的大数据进行分布式挖掘。

2016年8月31日收到
第一作者简介:梅 毅(1981—),男,汉族,江西抚州人,硕士,副教授。研究方向:软件工程。E-mail:meiyi5867@163.com。
引用格式:梅 毅, 熊 婷, 罗少彬. 复杂属性环境下 NoSQL 分布式大数据挖掘方法研究[J]. 科学技术与工程, 2017, 17(9): 239—243
Mei Yi, Xiong Ting, Luo Shaobin. Research on NoSQL distributed big data mining method in complex attribute environment[J]. Science Technology and Engineering, 2017, 17(9): 239—243

在复杂属性环境中,大数据的属性有四种,分别是特征、位置、方向和长度。大数据属性关联度是大数据在 NoSQL 数据库的挖掘聚类依据,对大数据属性关联度的准确运算,是挖掘工作对大数据进行分析、清理和转换工作的具体实现步骤。为此,所提复杂属性环境下 NoSQL 分布式大数据挖掘方法先对大数据的特征、位置、方向和长度属性的关联度进行挖掘,再对各属性的关联度进行聚类,便可获取最终的挖掘结果。

1.2 复杂属性环境下 NoSQL 数据库的挖掘条件

在复杂属性环境下 NoSQL 分布式大数据挖掘方法中,为了缩减 NoSQL 数据库运行内存、提高挖掘方法的稳定性,应对复杂属性环境下的大数据进行矩阵变换。

假设 D_j 是复杂属性环境下 NoSQL 数据库中大数据中第 j 行的单排矩阵,d_{ji} 是第 i 列、第 j 行的大数据,且 $i = 1,2,3,\cdots,m$。如果矩阵中每行共有 m 个大数据,则 $\boldsymbol{D}_j$ 可表示为

$$\boldsymbol{D}_j = (d_{j1}d_{j2}d_{j3},\cdots,d_{jm}) \tag{1}$$

如果大数据数量共有 n 个,用 T 表示矩阵的转置变换,则复杂属性环境下 NoSQL 数据库中大数据的总矩阵 $\boldsymbol{D}$ 可表示为

$$\boldsymbol{D} = (D_1,D_2,D_3,\cdots,D_n)^{\mathrm{T}} \tag{2}$$

现在对总矩阵 $\boldsymbol{D}$ 中集合 X 的大数据 x 进行挖掘,设其单属性的关联度为 $sim(X,Y)$(Y 是与 X 相对应的集合,用 y 表示 Y 中的大数据),挖掘出的样本用 s 表示,则 s 符合应下列各项条件:

$$\begin{cases} \forall x \in s \\ |Freq(x,X) - Freq(x,s)| \leqslant \theta \\ |s| \geqslant \dfrac{1}{2\theta^2}\ln\dfrac{2}{\delta} \end{cases} \tag{3}$$

式(3)中,$Freq$ 表示条件发生的频率,θ 代表所允许的最大挖掘误差,δ 为该误差的发生几率。

2 复杂属性环境下 NoSQL 分布式大数据挖掘方法研究

2.1 NoSQL 分布式大数据特征关联度挖掘

为了更好地增强所提挖掘方法的准确性和实用性,应事先对大数据特征关联度的挖掘工作进行约束,约束内容需要根据 NoSQL 数据库中给出的挖掘条件式(3)进行确定,约束内容应具有保证挖掘工作运算量小、挖掘效果强的作用。

用 $confidence(X\Rightarrow Y)$ 表示特征集合 X 中涵盖特征集合 Y 的几率,$confidence(Y\Rightarrow X)$ 与上述相反,根据式(3)的条件,对式(2)进行挖掘,则大数据特征关联度 $sim(X,Y)$ 的挖掘结果可表示为

$$sim(X,Y) = \min[confidence(X\Rightarrow Y), confidence(Y\Rightarrow X)] \tag{4}$$

由于 $confidence(X\Rightarrow Y)$ 和 $confidence(Y\Rightarrow X)$ 的取值范围为[0,1],那么,大数据特征关联度 $sim(X,Y)$ 的取值范围也为[0,1]。当 $sim(X,Y) = 0$ 时,表示复杂属性环境下大数据间的特征相互独立,此时不需要进行大数据挖掘聚类。

2.2 NoSQL 分布式大数据位置关联度挖掘

对复杂属性环境下大数据位置关联度的挖掘结果,可通过求取大数据传输通道的质心获取,将大数据集合 X 和 Y 传输通道的质心设为 c_1 和 c_2,两质心间的距离为 $|c_1c_2|$,用图 1 描述大数据位置关联度的挖掘原理。

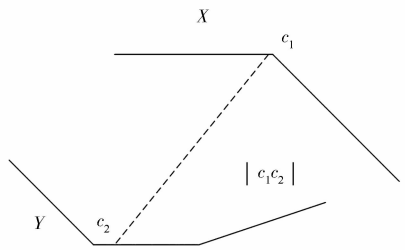

图 1 大数据位置关联度挖掘原理图

Fig. 1 Big data mining location related principle diagram

由图 1 可知,大数据位置关联度的挖掘结果等同于两质心间片段聚类 $(\overline{X},\overline{Y})$。为了求取两质心间片段聚类,需要将传输通道进行分段处理,求取每段质心,再对求取结果进行聚类,则有:

$$(\overline{X},\overline{Y}) = \left[\frac{\sum_{i=1}^{n-1}(x_{i+1}^2 - x_i^2)}{2\sum_{i=1}^{n-1}(x_{i+1}^2 - x_i^2)}, \frac{\sum_{i=1}^{n-1}(y_{i+1}^2 - y_i^2)}{2\sum_{i=1}^{n-1}(y_{i+1}^2 - y_i^2)}\right] \tag{5}$$

2.3 NoSQL 分布式大数据方向关联度挖掘

复杂属性环境下的大数据方向关联度是指大数据集合 X 和 Y 传输方向间的角度 (s_1,s_2),其余弦值可表示为

$$\cos(s_1,s_2) = \frac{s_1s_2}{|s_1||s_2|} \tag{6}$$

由式(6)可知,大数据集合 X 和 Y 传输方向间的角度 (s_1,s_2) 越大,$\cos(s_1,s_2)$ 值越小,当 (s_1,s_2) 超出 180° 后,$\cos(s_1,s_2)$ 值为负数。为了避免大数据位置关联度挖掘结果对大数据方向关联度挖掘结果产生影响,所提复杂属性环境下 NoSQL 分布式大数据挖掘方法使用 $[1-\cos(s_1,s_2)]$ 的正弦值表示方法,取代传统 $[1-\cos^2(s_1,s_2)]$ 的正弦值表示方法,令大数据方向关联度被精准挖掘出来。

基于上述方法,将大数据方向关联度的挖掘结果设为 $sim(dist)$,对大数据集合 X 和 Y 传输方向的

平均值 $avg(|s_1||s_2|)$ 进行加成运算,则有:

$$sim(dist) = avg(|s_1||s_2|)[1 - \cos(s_1, s_2)] \quad (7)$$

2.4　NoSQL 分布式大数据长度关联度挖掘

复杂属性环境下的大数据长度关联度的挖掘工作,是对大数据位置关联度挖掘原理图的异向思维运算,也是对大数据位置关联度挖掘结果的加成运算,其实质是对图 1 中两个大数据集合 X 和 Y 传输通道的长度进行求取的过程。大数据长度关联度的挖掘结果 $sim(length)$ 可表示为

$$sim(length) = (\overline{X}, \overline{Y}) \frac{|X - Y|}{\max(X, Y)} \quad (8)$$

2.5　NoSQL 分布式大数据挖掘聚类

将式(5)、式(6)、式(7)、式(8)根据式(3)给出的挖掘样本 s 条件进行聚类,确定最终的复杂属性环境下大数据挖掘结果。用 F 表示挖掘样本 s 的挖掘频率,那么,F_s 即可代表大数据挖掘聚类,也是所提方法的挖掘结果,其运算公式为

$$F_s = \frac{F_j}{F_{\max}} \lg \frac{N}{N_s} \quad (9)$$

式(9)中,F_j 代表大数据特征、位置、方向和长度属性同时出现的几率,$F_{\max}$ 代表大数据特征、位置、方向和长度属性关联度中的最大值;N 代表未进行挖掘工作前的大数据样本总数量,N_s 代表挖掘出的大数据特征、位置、方向和长度属性关联度样本的总数量。

3　实验分析

3.1　实验现场

为了准确分析本文所提方法挖掘工作的准确性、稳定性和实用性,实验对三台相同的数据挖掘设备进行了初始化,并在复杂属性环境下给予三台设备相同的初始大数据矩阵,每个矩阵中含有的四种属性的待挖掘样本数量均为 10 个,共计 40 个。使用本文方法、基于标签技术的分布式挖掘方法以及 MYSQL 挖掘方法对初始大数据矩阵进行挖掘。实验现场如图 2 所示。

实验进行 120 min,在实验进行到 60 min 时向初始大数据矩阵中加入平移、转动和收缩三种干扰因素。

3.2　实验数据处理标准

本次实验对所提方法挖掘工作的准确性、稳定性和实用性进行分析,对于方法的准确性,实验将三种方法对复杂属性环境下大数据的挖掘结果与实验给定的初始大数据矩阵的待挖掘样本参数进行对比,根据式(10)求取方法的准确性 u 。

图 2　实验现场图

Fig. 2　The experiment site map

$$u = \frac{F_s - \frac{10}{D}}{D} \times 100\% \quad (10)$$

对于方法的稳定性(稳定性标准有两个,分别是最低抗干扰强度 Min 和标准抗干扰强度 Mean,Min 和 Mean 的值越大,方法的稳定性越高)和实用性 M_E(M_E 的实质是干扰下的方法平均挖掘误差,该值越小,方法的实用性就越高),分别用式(11)和式(12)求取。

$$\begin{cases} \text{Min} = \min(ps_1, ps_2, \cdots, ps_n) \\ \text{Mean} = \dfrac{1}{n} \sum_{i=1}^{n} (ps_i) \end{cases} \quad (11)$$

$$M_E = \frac{1}{n} \sum_{i=1}^{k} (|clu(D)| - |clu(D^1)|) \quad (12)$$

式中,p 代表单次抗干扰强度;k 是大数据挖掘聚类的次数;$clu(D)$ 和 $clu(D^1)$ 分别代表初始大数据矩阵的待挖掘聚类结果和加入干扰后大数据矩阵的待挖掘聚类结果。

3.3　实验结果分析

根据式(12)对三种方法实验结果进行计算,挖掘准确性对比情况用表 1 描述。

由表 1 中的数据可知:在三种方法中,基于标签技术的分布式挖掘方法的挖掘准确性最低,均不高于 90%,挖掘准确性的平均值为 84.55%,仅在对大数据挖掘准确性要求不高的复杂属性环境下具有一

表 1　本文方法与其他方法准确性对比表

Table 1　Accuracy compared with other methods this method table

大数据属性	挖掘准确性 u/%		
	本文方法	基于标签技术的分布式挖掘方法	MYSQL 挖掘方法
特征	98.39	87.65	92.51
位置	98.00	81.23	96.50
方向	98.06	86.01	96.21
长度	98.74	83.29	95.24

定的利用价值;MYSQL 挖掘方法的准确性较为平庸,挖掘准确性的平均值为 95.12%;本文方法的挖掘准确性均高于 98%,其平均值高达 98.30%,可见本文方法在挖掘工作中几乎不存在错误挖掘的情况,可证明本文方法的挖掘工作具有较强的准确性。

实验向初始大数据矩阵中加入平移、转动和收缩三种干扰因素后,根据式(1)计算出三种方法的稳定性数据,用表 2、表 3 描述。

表 2 本文方法与其他方法稳定性最小值对比表

Table 2 Stability the minimum contrast table method with other methods

干扰因素	Min		
	本文方法	基于标签技术的分布式挖掘方法	MYSQL 挖掘方法
平移	0.13	0	0
转动	0.40	0.18	0.22
收缩	0.38	0.11	0.13

表 3 本文方法与其他方法稳定性标准值对比表

Table 3 Stability standard comparison method with other methods

干扰因素	Mean		
	本文方法	基于标签技术的分布式挖掘方法	MYSQL 挖掘方法
平移	0.25	0	0
转动	0.69	0.41	0.53
收缩	0.74	0.38	0.57

由表 2、表 3 可知:基于标签技术的分布式挖掘方法以及 MYSQL 挖掘方法均无法抵御复杂属性环境下大数据矩阵中的平移干扰,且两种方法对转动和收缩干扰的抵御能力也远低于本文方法。对比来看,本文方法在挖掘工作中具有较强的稳定性。

根据式(14)计算出三种方法的实用性数据,用表 4 描述。

表 4 本文方法与其他方法实用性对比表

Table 4 Method practical comparison with other methods

参数	本文方法	基于标签技术的分布式挖掘方法	MYSQL 挖掘方法
M_E	0	1.2	3.6

由表 4 中的数据可知,三种方法的实用性由低到高的排列次序依次为:MYSQL 挖掘方法、基于标签技术的分布式挖掘方法和本文方法。本文方法的 M_E 值为 0,说明复杂属性环境下的干扰因素虽然对本文方法单属性的大数据挖掘结果存在一些微小影响,但对挖掘聚类几乎无影响,证明本文方法具有较强的实用性。

4 结论

本文根据对以往提出的复杂属性环境下大数据挖掘方法进行分析,确定出数据库的合理选择对挖掘方法的重要性,为此,本文提出复杂属性环境下 NoSQL 分布式大数据挖掘方法。NoSQL 是一种根据物理分散逻辑进行数据规划的分布式数据库,拥有灵活性佳、实用性强、价格低廉和高效等优点,可增强挖掘方法的准确性、稳定性和实用性。实验通过对比方式对本文方法的准确性、稳定性和实用性进行分析。实验结果可非常直观地证明,本文方法的挖掘工作具有较强的准确性、稳定性和实用性,在复杂属性环境下的大数据挖掘工作中本文方法非常合理且有效。

参 考 文 献

1 齐雪婷,马训鸣,刘 霞,等. 基于 CAN 的分布式顶升控制系统设计. 西安工程大学学报,2016;30(1):118—123
Qi Xueting, Ma Xunming, Liu Xia, *et al*. Design of distributed jack-up control system based on CAN bus. Journal of Xi′an Polytechnic University, 2016;30(1):118—123

2 朱建生,汪健雄,张军锋. 基于 NoSQL 数据库的大数据查询技术的研究与应用. 中国铁道科学,2014;35(1):147
Zhu Jiansheng, Wang Jianxiong, Zhang Junfeng. Research and application of large data query technology based on NoSQL database. China Railway Science, 2014;35(1):147

3 李挥剑. 大数据环境下频繁项集挖掘的研究. 青岛科技大学学报(自然科学版),2015;36(2):224—231
Li Huijian. Research on frequent itemsets mining in large data environment. Journal of Qingdao University of Science and Technology (Natural Science Edition), 2015;36(2):224—231

4 李善青,赵 辉,宋立荣. 基于大数据挖掘的科技项目查重模型研究. 图书馆论坛,2014;34(2):78—83
Li Shanqing, Zhao Hui, Song Lirong. Study on detection model of similar scientific project based on big data mining. Library Tribune, 2014;34(2):78—83

5 尤海浪,钱 锋,黄祥为,等. 基于大数据挖掘构建游戏平台个性化推荐系统的研究与实践. 电信科学,2014;30(10):27—32
You Hailang, Qian Feng, Huang Xiangwei, *et al*. Research and practice of building a personalized recommendation system for mobile game platform based on big data mining. Telecommunications Science, 2014;30(10):27—32

6 宫 宇,吕金壮. 大数据挖掘分析在电力设备状态评估中的应用. 南方电网技术,2014;8(6):74—77
Gong Yu, Lü Jinzhuang. Application of big data mining analysis in power equipment state assessment. Southern Power System Technology, 2014;8(6):74—77

7 杨 坤,李石柱. 大数据挖掘技术应用于血吸虫病监测预警研究的探讨. 中国寄生虫学与寄生虫病杂志,2015;33(6):461—465
Yang Kun, Li Shizhu. Application of big data mining technology in monitoring and early-warning of schistosomiasis. Chinese Journal of Parasitology and Parasitic Diseases, 2015;33(6):461—465

8　杨　斐,艾晓燕,张永恒,等. 大数据精准挖据处理架构及预测模型研究. 电子设计工程,2016;24(12):29—32
Yang Fei, Ai Xiaoyan, Zhang Yongheng, *et al*. New mining architecture and prediction model for bid data. Electronic Design Engineering,2016;24(12):29—32

9　高　芹,陈　亚. 数据挖掘中一种高效的聚类通用框架研究. 科学技术与工程,2014;14(16):112—118
Gao Qin, Chen Ya. Research on an efficient clustering general framework in data mining. Science Technology and Engineering, 2014;14(16):112—118

10　王秀英,张　玲,张聪聪. 探讨地震前兆观测中的大数据挖掘与应用. 震灾防御技术,2015;10(1):39—45
Wang Xiuying, Zhang Ling, Zhang Congcong. Discussion on the big data mining application on earthquake precursor observation. Technology for Earthquake Disaster Prevention, 2015; 10 (1): 39—45

Research on NoSQL Distributed Big Data Mining Method in Complex Attribute Environment

MEI Yi, XIONG Ting, LUO Shao-bin
(Department of Computer Engineering, Nanchang University College of Science and Technology, Nanchang 330029, P. R. China)

[**Abstract**] Due to the complex attribute environment of data mining work need to involve the analysis of data, cleaning, conversion and integration of a series of operation, resulting in past the complex nature of environmental data mining methods can't have good accuracy, stability and practicability, therefore, the property of the complex environment NoSQL distributed data mining method was produced. Method using NoSQL database physical distributed logic, build NoSQL database under the complex nature of the environment, the characteristics of database data, location, direction and length property of the association mining and distributed, through mining formula is given for the mining results. Using clustering formula for large data feature mining, position, direction and length property of the association mining results are clustered to obtain the final mining results. The experimental results show that the proposed method is of high accuracy, stability and practicability.

[**Key words**] complex attribute environment　NoSQL　distributed　big data mining