

云计算下分布式数据安全读取算法研究

涂俊英,熊曾刚

(湖北工程学院 计算机与信息科学学院,湖北 孝感 432000)

摘 要: 针对传统的云计算下分布式数据安全读取算法,一直存在数据读取质量不高,读取数据重复率较高的问题,提出基于范围型属性的云计算下分布式数据安全读取算法,首先采用经验分布函数与核估计对云计算下分布式数据的分布函数进行估计;然后以分布式数据的分布函数变换进行均匀分布的假设检验;再以分布式数据分析结果为依据,利用范围型属性的值域特征,将分布式数据值域按照目标数据库的分布样本划分为多个子区间,在此基础上完成云计算下分布式数据的安全读取,实验结果分析证明,所提算法能提高分布式数据的安全读取速度和读取正确率,降低计算复杂度,适用范围广。

关键词: 云计算下;分布式数据;安全读取;经验分布函数;范围型属性

中图分类号: TP391

文献标识码: A

文章编号: 1000—7180(2017)10—0127—04

Cloud Computing Distributed Data Security Read Algorithm Research

TU Jun-ying, XIONG Zeng-gang

(School of Computer and Information Science, Hubei Engineering University, Xiaogan 432000, China)

Abstract: In view of the traditional distributed data security algorithm under cloud computing, there has been a problem that the quality of data reading is not high, and the repetition rate is high. The range of attribute based cloud computing distributed data security read algorithm, using the empirical distribution function and kernel estimation of distribution function under distributed cloud computing data are estimated; then to transform the distribution function of distributed data assuming uniform distribution of inspection inspection; then distributed data analysis based on the results, the range of feature range type attribute, the distributed data domain into multiple sub intervals according to distribution of samples into the target database, complete cloud computing security read based on distributed data. The experimental results show that the proposed algorithm can improve the security of the distributed data reading speed and reading accuracy, reduce the computational complexity, and the application scope is wide.

Key words: under the cloud computing; distributed data; security read; empirical distribution function; range type attribute

1 引言

随着科学技术的飞跃式发展,云计算下的分布式数据读取算法研究有了推动性进展^[1]. 所谓云计算下分布式数据读取就是利用云计算从分布式数据库中找寻所需知识的过程^[2-3]. 分布式数据安全读取算法研究之所以受到重视,归根结底是由于,在现实环境中,大多数的大规模数据库都是以分布式的形式存在的,因此,提出云计算下分布式数据安全读取算法系统的体系结构是非常必要的^[4-5];在数据读取过程中,经常需

要来自不同站点的数据库中的数据,这就使得数据读取必须具有分布式读取的能力,同时也需要我们根据分布式数据挖掘的特点设计出新的分布式数据读取算法^[6-7]. 随着计算机网络的快速发展与应用,互联网已然成为人类迄今为止最大的数据源. 互联网中的各类数据在以几何级数的速度疯长. 云计算环境下为了更好地利用互联网中的数据资源,确保网络传输与交换的分布式数据不会发生增加、修改或丢失,需要对分布式数据安全读取算法进行研究. 受到大部分专家学者的关注,并取得了一些成果.

文献[8]提出了基于启发式搜索的云计算下分布式数据安全读取算法,利用云计算下分布式数据文本分块技术与特征读取规则对分布式数据进行预处理分析;结合条件随分布式数据机场概率模型利用启发式搜索算法完成对分布式数据的安全读取.该算法存在数据读取质量不高且读取数据重复率较高的问题.文献[9]基于支撑矢量机的云计算下分布式数据安全读取算法,采用自组织特征映射神经网络对数据进行分类;利用支撑矢量机方法对其进行分类识别;依据分类识别结果完成分布式数据安全读取,但存在读取效率不高的问题.文献[10]提出了一种基于 RSBoost 的云计算下分布式数据安全读取算法,首先采用 SMOTE 方法对云计算下分布式数据少数类进行过采样处理;在采样处理结果的基础上结合 Boosting 算法对分布式数据进行聚类;最后依据聚类结果完成云计算下分布式数据的安全读取.此方法存在已读取数据覆盖程度较差、读取不完全及读取效率不高的问题.

针对上述算法在实际应用中产生的问题,提出一种基于范围属性的云计算下分布式数据安全读取算法.实验结果分析证明,所提算法可以提高分布式数据的安全读取速度、正确率且适应性范围广,有良好的应用价值.

2 云计算下分布式数据安全读取算法研究

2.1 基于经验分布的分布式数据分析

首先采用经验分布函数与核估计对云计算下分布式数据的分布函数进行估计,以分布式数据的分布函数变换进行均匀分布的假设检验,以完成区间分布式数据的预处理分析.具体过程描述如下:

基于经验分布函数给出云计算下分布式数据变换公式.设 X 表示服从某一连续分布的分布式数据随机变量, $(x_1, x_2, \dots, x_n)$ 表示已得到的一组分布式数据样本,将其转化为区间分布式数据的算法是取其最大值与最小值作为分布式数据区间的两个端点,假定其他分布式数据样本在这个区间符合均匀分布.由于这一假设过于严格,若分布式数据样本符合其他分布规律,则会导致这一假设以及后续分布式数据分析结果失效.

设定 X 的分布式数据分布函数为 $F(t)$,则分布式数据的经验分布函数计算公式为

$$F_n(t) = \frac{1}{n} \sum_{i=1}^n I\{x_i \leq t\} \quad (1)$$

式中, I 表示分布式数据的示性函数; n 表示分布式数据样本数量; i 表示分布式数据样本参数值; t 表示分布式数据的分布概率; x_i 表示分布式数据的均匀分布概率.

对于给定的任意的 t ,式(1)的分布式数据经验分布函数服从二项分布,即:

$$nF_n(t) \sim B(n, F(t)) \quad (2)$$

根据式(2)可以计算得到 $F_n(t)$ 分布式数据期望与方差分别为

$$E(F_n(t)) = F(t) \quad (3)$$

$$\text{Var}(F_n(t)) = \frac{1}{n} F(t) [1 - F(t)] \quad (4)$$

则云计算下分布式数据的经验分布函数 $F_n(t)$ 二阶收敛到真实的分布式数据分布函数 $F_n(t)$.

假设分布式数据随机变量 $F(X)$ 的分布函数为 H ,则分布式数据的二阶分布函数计算表达式为

$$\begin{cases} H(y) = \Pr(F(X)) \leq y = \Pr(X \leq F^{-1}(y)) = F(F^{-1}(y)) = y \\ 0 < y < 1 \end{cases} \quad (5)$$

通过式(5)可知, $F(X)$ 服从分布式数据 $(0, 1)$ 区间上的均匀分布 $(U(0, 1))$,而分布式数据 $F_n(X_i)$ 二阶收敛到 $F(X_i)$.因此在分布式数据样本数量足够大时可以近似的认定 $F_n(X_i)$ 服从分布式数据 $(0, 1)$ 区间上的均匀分布.

由此分析可知,对于原始分布式数据随机变量 $(x_1, x_2, \dots, x_n)$ 可以通过式(6)的变换得到 $(z_1, z_2, \dots, z_n)$,转换为理论上服从 $(0, 1)$ 均匀分布的分布式数据区间,从而进行分布式数据的后续处理和分析.

$$z_i = F_n(x_i) \quad (6)$$

将分布式数据分布函数 $F(t)$ 转化为分布式数据估计密度函数,通过分布式数据积分变换将式(1)转换为以下公式计算:

$$F_n(t) = \sum_{i=1}^n \omega_i I\{x_i \leq t\} \quad (7)$$

式中, ω_i 表示基于离散均匀测度构造的分布式数据权重, $\omega_i = 1/n$,将此分布式数据权重函数进行推导可以得到分布式数据核估计.具体计算过程如下.

给出分布式数据核密度估计 $\hat{F}(t)$ 表达式为

$$\begin{cases} \hat{f}(x) = \frac{1}{nh} \sum_{i=1}^n K(x - x_i/h) \\ \hat{F}(t) = \int_{-\infty}^t \hat{f}(x) dx \end{cases} \quad (8)$$

式中, $K(x - x_i/h)$ 为分布式数据核函数; h 为分布式数据窗宽; $\hat{f}(x)$ 为分布式数据分布估计函数.通

常情况下分布式数据核函数 $K(x - x_i/h)$ 为对称函数,且满足以下条件:

$$\begin{cases} \int K(t)dt = 1 \\ \int tK(t)dt = 0 \\ \int t^2 K(t)dt = c \neq 0 \end{cases} \quad (9)$$

式中, c 表示常见的分布式数据核函数; $K(t)$ 表示分布式数据正态核. 由于式(9)中的 $\hat{f}(x)$ 范围为非负,分布式数据估计得到的核密度估计 $\hat{F}(t)$ 具有单调性特点,此分布式数据函数是可逆变换.

在利用核算法对分布式数据进行估计时,分布式数据窗宽 h 的选择至关重要,因此采用基于似然函数的分布式数据交叉验证标准,其表达式如下:

$$L_{CV}(h) = \frac{1}{n} \sum_{i=1}^n \ln \hat{f}_{-i}(x_i) \quad (10)$$

式中, $\hat{f}_{-i}$ 代表剔除第 i 个分布式数据样本之后得分布式数据核密度估计. 选取基于似然函数分布式数据交叉验证标准 $L_{CV}(h)$ 的最大值 h 作为分布式数据窗宽. 对变换后的分布式数据进行是否符合均匀分布的假设检验,以便进行分布式数据的后续分析. 设定原始分布式数据矩阵如下:

$$\mathbf{I} = \begin{bmatrix} I_1 \\ I_2 \\ \vdots \\ I_n \end{bmatrix} = \begin{bmatrix} I_{11} \cdots I_{1p} \\ I_{21} \cdots I_{2p} \\ \vdots \\ I_{n1} \cdots I_{np} \end{bmatrix}_{n \times p} \quad (11)$$

式中,原始分布式数据矩阵 $\mathbf{I}$ 的每一行表示一组分布式观测数据;每一列表示一个分布式数据变量的观测值; p 表示分布式数据的样本观测数量. 设分布式数据观测值有 M 类,令 $(I_1, \dots, I_{n1}), (I_{n1} + 1, \dots, I_{n2}), \dots, (I_{n_{M-1}} + 1, \dots, I_n)$ 分别隶属不同的分布式数据类别,即分布式数据样本本身具有一定的分类结构,对变换后的分布式数据分组进行约简,整理成分布式区间数据表:

$$F_{rk}(t) = \frac{1}{n} \sum_{i=1}^n I\{I_{ik} \leq t\} \quad (12)$$

式中, I_{ik} 表示分布式数据约简参数,则分布式数据样本统计量矩阵如下:

$$\mathbf{Z} = \begin{bmatrix} F_{n1}(I_{11}) \cdots F_{np}(I_{1p}) \\ F_{n1}(I_{21}) \cdots F_{np}(I_{2p}) \\ \vdots \\ F_{n1}(I_{n1}) \cdots F_{np}(I_{np}) \end{bmatrix}_{n \times p} \quad (13)$$

通过式(12)进一步对分布式数据中每一类内部的样本进行整理可以得到分布式数据的次序统计

量为

$$\mathbf{Y} = [Y_1, Y_2, \dots, Y_M]^T \quad (14)$$

式中, T 表示分布式数据确定统计量. 完成分布式数据分析,为后续分布式数据安全读取提供依据.

2.2 基于范围型属性的分布式数据安全读取算法

以 2.1 节中的分布式数据分析结果为依据,利用范围型属性的值域特征,将分布式数据值域按照目标数据库的分布样本划分为多个子区间. 具体步骤如下.

将待划分的分布式数据的每个子区间上限值设定为 $c_i, c_o = 0$,则分布式数据中的每个子区间均可表示为 $[c_{i-1}, c_i]$;将分布式数据样本中的每个子区间上限记作 a_j ,则分布式数据样本中的每个子区间可表示为 $[a_{i-1}, a_i]$.

假设 $a_o = 0$,分布式数据样本区间 $[a_{i-1}, a_i]$ 中分布式数据记录的集合表示为 $b_1, b_2, \dots, b_j$;故分布式数据的范围属性 c_1 的计算过程如式(15):

$$\begin{cases} c_1 = (a_2 - c_o) + \frac{k - b_1 - b_2}{b_3} \times (a_3 - a_2) \\ b_1 + b_2 + b_3 \geq k \\ b_1 + b_2 < k \end{cases} \quad (15)$$

根据上述公式依此向下推导,设 $q > 2$ 是满足 $\sum_{l=1}^q b_l \geq k$ 的最小正整数, l 表示分布式数据范围属性域值, b_l 表示分布式数据查询记录属性值. 则得到 c_1 为

$$c_1 = (a_{q-1} - c_o) + \frac{k - \sum_{j=1}^q b_j}{b_q} \times (a_q - a_{q-1}) \quad (16)$$

式中, b_q 表示分布式数据中重复的数据记录; j 表示分布式数据限定性域值.

计算得到 c_1 的结果后, c_2 可以按照上述思路计算:

$$c_2 = \frac{\frac{a_m - c_1}{a_m - a_{m-1}} \times b_m + \sum_{l=m+1}^q b_l}{b_q} \times (a_{m+1} - a_m) \quad (17)$$

式中, m 表示分布式数据读取时间. 按照同样的思路即可求得 c_i . 假设 a_t 为分布式数据序列 $a_o, a_1, a_2, \dots$ 中第一个大于 c_i 分布式数据元素.

3 实验结果与分析

实验环境建立在 PC 机配置为: DDR4

2133MHz,CPU Corei7-4790 3. 4GHz,内存为 2GB, Windows7 操作系统. 实验仿真系统利用 Java 语言编写实现. 实验数据采用北京外国语大学实验室提供的分布式数据基准语料库. 选取了计算机、汉语言文学、对外贸易、国际金融等四类分布式数据作为实验测试数据. 每类分布式数据各取 500 篇共计 2 000 篇分布式数据文档进行实验. 采用 10 折交叉验证法,将整个分布式数据集随机分为 10 份,选取其中 8 份合并并作为分布式数据训练集,剩下的 2 份作为分布式数据测试集. 实验运行 6 次,选其平均值作为最后的分布式数据读取结果.

在实际应用过程中,云计算下分布式数据安全读取的不同算法有不同优缺点,为了进一步证明基于范围型属性云计算下分布式数据安全读取算法的有效性,引入分布式数据读取准确率(%) E 来衡量算法的有效性能.

$$E=\frac{\text{读取正确的分布式数据数目}}{\text{实际读取的分布式数据总数}}$$

(18)

图 1 中,数据集 1 代表实验选取的北京外国语大学的计算机专业分布式数据集;数据集 2 代表汉语言文学分布式数据集;数据集 3 代表对外贸易分布式数据集;数据集 4 代表国际金融分布式数据集.

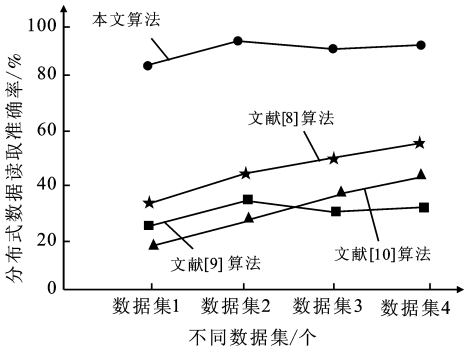


图 1 不同算法的分布式数据读取准确率

根据图 2 可知,文献[8]算法对不同数据集的分布式数据安全读取时间最长,严重影响分布式数据读取效率;文献[9]算法对不同数据集的分布式数据安全读取时间有所降低,但读取速度依然较慢;文献[10]算法对不同数据集的分布式数据安全读取时间较快,但与本文所提算法相比较读取效率仍然不高,由此可以看出,本文所提算法的分布式数据读取效率最佳.

4 结束语

针对传统方法存在计算复杂度高,读取误差大且效率低的问题. 提出基于范围型属性的云计算下

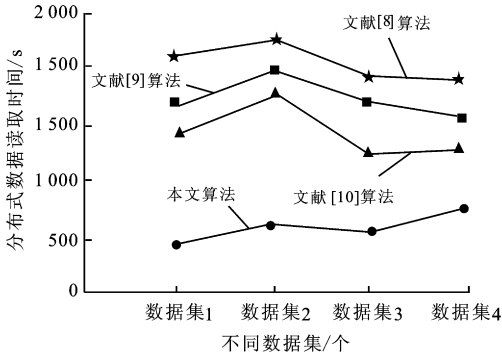


图 2 不同算法的分布式数据读取时间

分布式数据安全读取算法. 并通过实验结果分析证明,所提算法可以提高分布式数据的安全读取速度、读取正确率且适应性范围广,具有良好的应用价值.

参考文献:

[1] 王建中,杨璐. 高速实时系统数据采集与传输[J]. 计算机科学,2016,43(增刊 2):604-606.

[2] 褚征,于炯,鲁亮,等. 基于内存云的大块数据对象并行存取策略[J]. 计算机应用,2016,36(6):1526-1532.

[3] 丁有伟,秦小麟,刘亮,等. 一种异构集群中能量高效的大数据处理算法[J]. 计算机研究与发展,2015,52(2):377-390.

[4] 夏秀峰,张羽. PDM 中基于负载信息的文件读取均衡算法[J]. 计算机工程与设计,2015,36(10):2729-2732.

[5] 唐颖峰,陈世平. 一种面向分布式数据流的闭频繁模式挖掘方法[J]. 计算机应用研究,2015,32(12):3560-3564.

[6] 陈付梅,韩德志,毕坤,等. 大数据环境下的分布式数据流处理关键技术探析[J]. 计算机应用,2017,37(3):620-627.

[7] 张振华,吴开超. 一种分布式 Twitter 数据处理方案及应用[J]. 计算机应用研究,2015,32(7):2073-2077.

[8] 陈恒. 一种基于 Spark 的大规模语义数据分布式推理框架[J]. 计算机科学,2016,43(增刊 2):93-96.

[9] 于彦伟,齐建鹏,陆云辉,等. 时空轨迹大数据分布式蜂群模式挖掘算法[J]. 计算机工程与科学,2016,38(2):255-261.

[10] 胡楠,杨壮观,李峰,等. 基于 GIS 的电网地理信息分布式数据挖掘[J]. 电子设计工程,2016,24(15):20-24.

作者简介:

涂俊英 女,(1970-),硕士,副教授. 研究方向为云计算.
E-mail:tujunying@yeah.net.

熊曾刚 男,(1974-),博士,教授. 研究方向为云计算、大数据和物联网.