(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 116662454 A
(43)申请公布日 2023.08.29

(21)申请号 202310401084.9

(22)申请日 2023.04.15

(71)申请人 复旦大学
地址 200433 上海市杨浦区邯郸路220号
申请人 星环信息科技(上海)股份有限公司

(72)发明人 荆一楠 乔冀瑜 张寒冰 徐伟
陈振强 何震瀛 王晓阳

(74)专利代理机构 上海正旦专利代理有限公司 31200
专利代理师 陆飞 陆尤

(51)Int.Cl.
G06F 16/28(2019.01)
G06F 16/2453(2019.01)
G06F 18/23213(2023.01)
G06N 3/04(2023.01)
G06N 3/08(2023.01)

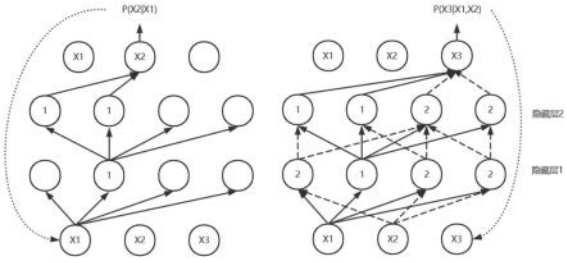
权利要求书3页 说明书6页 附图3页

(54)发明名称

基于分组掩码自回归模型的查询基数估计方法

(57)摘要

本发明属于数据库查询技术领域,具体为一种基于分组掩码自回归模型的查询基数估计方法。本发明包括列数据分组预处理、分组掩码神经网络基数估计模型训练;列数据分组预处理将单表数据进行分组排序,使模型更好的学习到其中的联合概率分布;分组掩码神经网络基数估计模型通过位置编码以不同次序学习部分列的分布,模型训练分为数据训练、混合掩码训练、直接查询训练三个阶段且混合训练,组合不同的数据集对模型进行训练,降低模型的训练时间;从数据库系统执行日志中记录误差较大的查询信息并加入三阶段训练的数据中,提高模型的训练效率。本发明可以减少传统自回归模型进行基数估计时的查询误差,减少时耗,使得模型可以更快、更稳定用于基数估计。



CN 116662454 A

1.一种基于分组掩码自回归模型的查询基数估计方法,其特征在于,具体步骤包含:列数据分组预处理、分组掩码神经网络基数估计模型的训练;

步骤(一)列数据分组预处理;

具体是将单表数据以列为单位,按照信息熵从小到大进行分组;分组后在组内按列间相关性与相对熵即KL散度的加权平均进行组内排序;排序后的数据按照分组附加不同顺序的掩码输入分组掩码神经网络基数估计模型,并进行训练;经过分组预处理的单表数据有利于使用自回归模型进行预测,降低自回归模型的预测误差,且之后的神经网络按分组对数据附加掩码进行训练;

(二)分组掩码神经网络基数估计模型的训练;

所述分组掩码神经网络基数估计模型采用自回归神经网络模型,以下简称模型,包括使用原始数据与在原始数据上的查询与查询结果,对模型进行混合训练学习,获取原始数据中不同列不同取值的联合概率分布,且可以使用掩码机制针对步骤(一)中生成的数据分组与重点查询的概率分布进行优化学习;

模型训练分为数据训练、混合掩码训练、直接查询训练三个阶段;在使用数据进行训练时,首先对数据进行编码,对于值域较小的列采用独热编码,对于值域较长的列使用词向量编码;同时需要在每条数据前附加当前预测位置的编码;其中:

数据训练阶段,将长度为n数据序列扩展为n×n的矩阵数据,对相同的数据矩阵按照分组附加不同的掩码矩阵以进行不同顺序的训练学习,其中一数据中顺序对应一种n×n大小的掩码矩阵,掩码按分组为单位进行施加,掩码矩阵与数据矩阵相乘得到训练数据,将训练数据输入神经网络进行训练;

混合掩码训练阶段,选取测试查询中误差大于设定阈值η_1的查询,根据查询所涉及的列条件顺序生成对应掩码矩阵,掩码矩阵与数据矩阵相乘得到训练数据,训练数据输入神经网络进行训练;

直接查询训练阶段,需要选择经过前两步训练后误差大于定阈值η_2的查询,将其加入数据集中,利用重参数技巧,或者拆分为子查询,直接对模型进行训练,这些查询可以来自预先设置的测试集,也可以来自在实际使用中日志记录查询。

2.根据权利要求1所述的基于分组掩码自回归模型的查询基数估计方法,其特征在于,步骤(一)中所述列数据分组预处理,具体流程为:

计算所有数据列的信息熵,按信息熵从小到大进行排序,排序前列为$\{X_1,X_2...X_n\}$,排序后为$\{X_1',X_2'...X_n'\}$,这一过程表示为:

$$\{X_1',X_2'...X_n'\}=\mathrm{Sort}(H(X_1),H(X_2)...H(X_n)), \quad (1)$$

排序完成后,按照信息熵对数据列进行聚类,聚类方法使用K-means方法,设μ_j为聚类过程中第j类的聚类中心,聚类的目标函数表示为:

$$\min \sum_{i=1}^{n} \min_{j=1,2...k} ||H(X_i')-\mu_j||^2, \quad (2)$$

完成数据列分组后进行组内排序,设组内存在k列,排序前列为$\{X_1,X_2...X_k\}$,排序后为$\{X_1',X_2'...X_k'\}$,每列数据i计算其与组内其他列j的关系值R(i,j),R(i,j)由相关性r(i,j)与相对熵$D_{KL}(i,j)$进行加权平均产生:

$$R(i,j)=w_1*r(X_i,X_j)-w_2*D_{KL}(X_i,X_j), \quad (3)$$

列间相关性计算使用成对数据最小方法或皮尔逊积矩相关系数;以原先的组内第一列

作为开始，后续选取未参与排序且与已经完成排序的前一列之间关系值R(i,j)最大的列作为组内排序的下一个列，列选取为：

$$X'_{i+1}=X_j=\max_{j\in\{i\}[\,]\{j\}}R(i,j), \quad (4)$$

其中，第一列为$X_1'=X_1$。

3.根据权利要求2所述的基于分组掩码自回归模型的查询基数估计方法，其特征在于，步骤(二)中所述分组掩码神经网络基数估计模型的训练，具体步骤为：

(1)对于一组长度为n的存储数据，首先按照列数据分组预处理方法分组后得到k组数据，k的值视列数量与聚类情况确定，按组为单位生成A_k^k种随机排列；

(2)数据训练阶段，对于步骤(1)中生成的随机排列，根据排列顺序生成对应的掩码矩阵M_1与位置序列P_1；将掩码矩阵M_1与m个数据序列扩展n*n数据矩阵进行相乘，得到掩码后的数据矩阵，数据矩阵左方附加位置序列；随后将数据矩阵输入模型，模型输出在当前序列下需要预测变量的条件概率分布，将该条件概率分布与输入数据计算交叉熵，作为模型损失$loss_1$，并进行反向传播训练模型，设定$epoch_0$次训练周期进行训练；

(3)混合掩码训练阶段，设定误差阈值η_1，选择经过数据训练阶段后误差q_{error}大于η_1的查询条件生成掩码样本，这些查询来自预先设定好的训练集与实际使用时的查询日志之中；设查询$Q=\{X_1\in R_1,X_2\in R_2...X_n\in R_n\}$，设$X_i$值域为$R_i'$，如果$R_i\neq R_i'$，则认为查询Q在i列上存在条件；根据查询是否在i列上存在条件，从查询中生成如下序列$S=\{f_1,f_2...f_n\}$，掩码矩阵长度L=sum(S)，其中：

$$f_i=\begin{cases}1 & i列上存在条件\\ 0 & i列上不存在条件\end{cases}, \quad (5)$$

根据序列S生成掩码矩阵与位置序列，矩阵大小为n*L，

将数据单条数据扩展为对应L*n大小的数据矩阵D_2，将掩码矩阵乘以数据矩阵并拼接位置序列得到该次训练数据D_2'；

将该数据与第一阶段的训练数据一起输入模型进行训练，计算模型输出，与原始数据对应位置分布的交叉熵作为损失$loss_2$，并于第一阶段的$loss_1$一同进行反向传播对模型进行训练；

(4)直接查询训练阶段，挑选出经过前两阶段训练后误差值大于设定阈值η_2的查询Q，使用查询条件与其基数值对模型进行训练；这些查询来自预先设置的测试集，或者来自在实际使用中日志记录查询；直接查询阶段使用重参数技巧直接对所有类型的查询进行训练，或者将范围查询分解为多个对应的点查询重新执行；对于范围查询，利用重参数技巧对模型输出结果与实际的label计算q_{error}作为损失$loss_3$，结合前两个阶段的$loss_1$与$loss_2$对模型进行反向传播训练；对于只包含等于条件的查询Q，将符合查询的序列直接输入模型得到估计值sel(q)'，与查询q的选择率sel(q)计算q_{error}作为损失进行对分组掩码神经网络基数估计模型反向传播。

4.根据权利要求3所述的基于分组掩码自回归模型的查询基数估计方法，其特征在于，关于模型的三个阶段的训练采用混合训练的方式，即首先使用经过预处理后的数据对模型进行数据训练阶段的训练，属于初始化训练；随后统计训练集查询与实际查询的查询信息与结果，用于混合掩码训练与直接查询训练阶段，当对应阶段的超过误差阈值的查询出现

次数大于预先设定的阈值时，会触发对应训练阶段完成对(二)中神经网络模型的训练；

模型混合训练中，需设定误差阈值η_1与η_2，设定查询数量统计阈值$N_{\eta 1}$与$N_{\eta 2}$，两阶段训练轮次$epoch_1$与$epoch_2$，记录在测试阶段与实际使用阶段误差分别大于阈值的查询Q与对应的出现次数N_{Q1}与N_{Q2}，当对应阶段的超过误差阈值的查询出现次数N_Q大于N_η时，触发对应的混合掩码训练或直接查询训练，对模型进行轮次为$epoch_{1or2}$的训练；

该模型混合训练在使用与测试时，将误差超过η_2的查询Q拆分为只包含点查询的子查询q并在系统空余时刻自动执行得到选择率sel(q)；子查询结果记录在日志中，其中产生较大误差的子查询加入直接查询训练阶段的训练数据中；

拆分查询表达为：

$split(Q)=[Q_1,Q_2,...Q_n]$，

设$Q=\{X_1\in R_1,X_3\in R_3,X_4\in R_4\}$，

则有：$split(Q)=[q_1=\{X_1=x_1\},q_3=\{X_1=x_1,X_3=x_3\},q_4=\{X_1=x_1,X_3=x_3,X_4=x_4\}]$，其中，$x_i\in R_i$，$x_i$为$R_i$范围内的随机取值；该查询总共拆分为$R_1*R_3*R_4$个子查询，从查询中进行随机取值，取其中产生较大误差的子查询加入数据集中。

基于分组掩码自回归模型的查询基数估计方法

技术领域

[0001] 本发明属于数据库查询技术领域，具体涉及一种查询基数估计方法。

背景技术

[0002] 伴随着信息时代的高速发展，数据库中的数据存量增长迅速，数据库关系模式也日趋复杂，想要在这种情况下进行查询优化，查询优化器需要能在复杂模式下具有更高准确性和稳定性的基数估计模型。传统基数估计方法大致包含直方图、数据画像与采样三种方法，但是在如今的复杂条件下它们的准确性都难以满足的查询优化的需求。

[0003] 由上，基于机器学习的基数估计方法逐渐走进了人们的视野，针对训练对象的不同，现有的基于机器学习的基数估计方法大致可分为由查询驱动的机器学习基数估计方法和数据驱动的机器学习基数估计方法。查询驱动的方式利用神经网络模型去学习查询条件与对应基数的映射关系，数据驱动的方式使用和积网络、自回归模型等模型学习数据的联合概率分布。这些方法相比于传统基数估计防范都具有更高的准确性。

[0004] 目前，基于自回归模型的数据驱动方法能适应更多的负载场景，且具有更高的准确性。但是，自回归模型难以拟合某些数据列的联合概率分布，导致涉及这些列的查询出现较大误差，进一步导致查询优化器选择错误的计划；同时，也有方案利用查询条件与其基数结果和数据去优化自回归模型，使得该方法同时由数据和查询驱动，具有更高的准确性，但是代价是大幅增加的训练时间，由此导致模型实时性降低，难以满足数据库查询优化需求。

发明内容

[0005] 本发明的目的在于提供一种具有高准确性，又满足数据库查询优化需求的基于分组掩码自回归模型的查询基数估计方法。

[0006] 本发明提出的基于分组掩码自回归模型的查询基数估计方法，具体步骤包含：列数据分组预处理、分组掩码神经网络基数估计模型的训练。

[0007] 步骤(一)列数据分组预处理；

[0008] 具体是将单表数据以列为单位，按照信息熵从小到大进行分组；分组后在组内按列间相关性与相对熵(KL散度)的加权平均进行组内排序。排序后的数据按照分组附加不同顺序的掩码输入分组掩码神经网络基数估计模型，并进行训练。经过分组预处理的单表数据更有利于使用自回归模型进行预测，降低自回归模型的预测误差，且之后的神经网络按分组对数据附加掩码进行训练。

[0009] 进一步地：

[0010] 所述数据列分组阶段，计算所有数据列的信息熵，按信息熵从小到大进行排序，排序前列为$\{X_1,X_2...X_n\}$，排序后为$\{X_1',X_2'...X_n'\}$，这一过程表示为：

[0011] $$\{X_1',X_2'...X_n'\}=\mathrm{Sort}(H(X_1),H(X_2)...H(X_n)), \quad (1)$$

[0012] 排序完成后，按照信息熵对数据列进行聚类，聚类方法可以使用K-means等聚类方法，μ_j为聚类过程中第j类的聚类中心，聚类的目标函数可以表示为：

[0013] $$\min \sum_{i=1}^{n} \min_{j=1,2\ldots k} ||H(X_i') - \mu_j||^2 , \quad (2)$$

[0014] 完成数据列分组后进行组内排序，设组内存在k列，排序前列为$\{X_1, X_2 \ldots X_k\}$，排序后为$\{X_1', X_2' \ldots X_k'\}$，每列数据i计算其与组内其他列j的关系值R(i,j)，R(i,j)可以由相关性r(i,j)与相对熵(KL散度)$D_{KL}(i,j)$进行加权平均产生：

[0015] $$R(i,j) = w_1 * r(X_i, X_j) - w_2 * D_{KL}(X_i, X_j), \quad (3)$$

[0016] 列间相关性计算可以使用成对数据最小方法、皮尔逊积矩相关系数等。以原先的组内第一列作为开始，后续选取未参与排序且与已经完成排序的前一列之间关系值R(i,j)最大的列作为组内排序的下一个列，列选取为：

[0017] $$X'_{i+1} = X_j = \max_{j \in \{i\}[\,]\{j\}} R(i,j), \quad (4)$$

[0018] 其中，第一列为$X_1' = X_1$。

[0019] (二)分组掩码神经网络基数估计模型的训练；

[0020] 所述分组掩码神经网络基数估计模型采用自回归神经网络模型，以下简称模型，包括使用原始数据与在原始数据上的查询与查询结果，对自回归神经网络模型进行训练学习，获取原始数据中不同列不同取值的联合概率分布，且可以使用掩码机制针对(一)中生成的数据分组与重点查询的概率分布进行优化学习。本发明可以使用多种自回归神经网络模型，当前所使用的是由Mathieu Germain在2015年ICML上发表论文《MADE:Masked Autoencoder for Distribution Estimation》中的掩码自回归模型，结构可见图1。

[0021] 模型训练分为数据训练、混合掩码训练、直接查询训练三个训练阶段。在使用数据进行训练时，首先对数据进行编码，对于不同值数量大于64的列可以采用独热编码，对于不同值数量大于64的列可以使用词向量编码。同时需要在每条数据前附加当前预测位置的编码。

[0022] 三个训练阶段总体如下：

[0023] 数据训练阶段，将长度为n数据序列扩展为n×n的矩阵数据，对相同的数据矩阵按照分组附加不同的掩码矩阵以进行不同顺序的训练学习，其中一数据中顺序对应一种n×n大小的掩码矩阵，掩码按分组为单位进行施加，掩码矩阵与数据矩阵相乘得到训练数据，将训练数据输入神经网络进行训练；

[0024] 混合掩码训练阶段，选取测试查询中误差大于设定阈值η_1的查询，根据查询所涉及的列条件顺序生成对应掩码矩阵，掩码矩阵与数据矩阵相乘得到训练数据，训练数据输入神经网络进行训练；

[0025] 直接查询训练阶段，需要选择经过前两步训练后误差大于定阈值η_2的查询，将其加入数据集中，利用重参数技巧，或者拆分为子查询，直接对模型进行训练，这些查询可以来自预先设置的测试集，也可以来自在实际使用中日志记录查询。

[0026] 具体训练过程如下：

[0027] (1)对于一组长度为n的存储数据，首先按照列数据分组预处理方法分组后得到k组数据，k的值视列数量与聚类情况确定，通常可以取[2-4]。按组为单位生成A_k^k种随机排列。

[0028] (2)数据训练阶段，对于第一步中生成的随机排列，根据排列顺序生成对应的掩码矩阵M_1与位置序列P_1。将掩码矩阵M_1与m个数据序列扩展n*n数据矩阵进行相乘，得到掩码后

的数据矩阵，数据矩阵左方附加位置序列。如果排序数量，可以根据训练集或验证集的查询条件选择部分次序进行保留。随后将数据矩阵输入神经网络模型，神经网络输出在当前序列下需要预测变量的条件概率分布。将该分布与输入数据计算交叉熵作为模型损失$loss_1$，并进行反向传播训练神经网络模型，设定$epoch_0$次训练周期进行训练。

[0029] 以6列数据为例，按分组[(1,2),(3,4),(5,6)]分为3组，产生一种组间顺序为：Sq＝(3,4)-(5,6)-(1,2)，该序列的掩码矩阵M_1为：

[0030] $$M_1=\begin{bmatrix}0&0&1&0&0&0\\0&0&1&1&0&0\\0&0&1&1&1&0\\0&0&1&1&1&1\\1&0&1&1&1&1\\1&1&1&1&1&1\end{bmatrix}$$

[0031] 该序列产生的位置序列P_1为：

[0032] $$P_1=\begin{bmatrix}3\\4\\5\\6\\1\\2\end{bmatrix}$$

[0033] 将数据单条数据扩展为n*n大小的数据矩阵D_1，将掩码矩阵乘以数据矩阵并拼接位置序列得到该次训练数据D_1'：

[0034] $$D_1'=[P_1|D_1*mask(M_1)]=\begin{bmatrix}3&mask&mask&X_3&mask&mask&mask\\4&mask&mask&X_3&X_4&mask&mask\\5&mask&mask&X_3&X_4&X_5&mask\\6&mask&mask&X_3&X_4&X_5&X_6\\1&X_1&mask&X_3&X_4&X_5&X_6\\2&X_1&X_2&X_3&X_4&X_5&X_6\end{bmatrix}$$

[0035] 此时产生的掩码矩阵附加与对应训练的条件概率分布见图2所示，图中实线位置代表掩码矩阵值为1模型可见该位置值；虚线位置代表附加遮蔽，掩码值为0模型不可见此处数据值。

[0036] 当分组数量超过4导致排序数量超过6时，或者后两阶段的训练数据集增长至严重影响模型训练时间时，可以根据训练集或验证集的查询条件选择部分次序进行保留，以减少这一阶段的模型训练时间。统计训练集或验证集的联合列条件分布情况，例如同时涉及{4,6}两列的查询较多时，可以保留Sq＝(3,4)-(5,6)-(1,2)这一次序使得模型可以跳过部分列而学习{4,6}两列的联合概率分布。

[0037] (3)混合掩码训练阶段，设定误差阈值η_1，阈值通常设置为1.5，也可以根据模型的实际精度需求自行确定。选择经过数据训练阶段后误差q_{error}大于η_1的查询条件生成掩码样本，这些查询可以来自预先设定好的训练集与实际使用时的查询日志之中，阈值越接近1时会收集更多的查询，模型会更加精确。设查询Q＝{$X_1\in R_1$,$X_2\in R_2$...$X_n\in R_n$}，设X_i值域为R_i'，如果$R_i\neq R_i'$，则认为查询Q在i列上存在条件。根据查询是否在i列上存在条件，可以从查询中生成如下序列S＝{f_1,f_2...f_n}，掩码矩阵长度L＝sum(S)，其中：

[0038] $$f_i=\begin{cases}1 & i列上存在条件\\ 0 & i列上不存在条件\end{cases},\qquad(5)$$

[0039] 根据序列S生成掩码矩阵与位置序列，矩阵大小为n*L，以6列数据为例，设Q＝{$X_1\in R_1$，$X_3\in R_3$，$X_4\in R_4$}为例，可以得到S＝{1,0,1,1,0,0}，掩码矩阵M_2为：

[0040] $$M_2=\begin{bmatrix}1&0&0&0&0&0\\1&0&1&0&0&0\\1&0&1&1&0&0\end{bmatrix}$$

[0041] 位置序列P_2为：

[0042] $$P_2=\begin{bmatrix}1\\3\\4\end{bmatrix}$$

[0043] 将数据单条数据扩展为对应L*n大小的数据矩阵D_2，将掩码矩阵乘以数据矩阵并拼接位置序列得到该次训练数据D_2'：

[0044] $$D_2'=[P_2|D_2*mask(M_2)]=\begin{bmatrix}1&X_1&mask&mask&mask&mask&mask0\\3&X_1&mask&X_3&mask&mask&mask\\4&X_1&mask&X_3&X_4&mask0&mask0\end{bmatrix}$$

[0045] 将该数据与第一阶段的训练数据一起输入模型进行训练，计算模型输出，与原始数据对应位置分布的交叉熵作为损失$loss_2$，并于第一阶段的$loss_1$一同进行反向传播对模型进行训练。

[0046] (4)直接查询训练阶段，挑选出经过前两阶段训练后误差值大于设定阈值η_2的查询Q，使用查询条件与其基数值对模型进行训练，阈值η_2通常设置为1.8，也可以根据模型的实际精度需求自行确定，通常需要大于η_1。这些查询可以来自预先设置的测试集，也可以来自在实际使用中日志记录查询。直接查询阶段可以使用重参数技巧直接对所有类型的查询进行训练，也可以将范围查询分解为多个对应的点查询重新执行。对于范围查询，利用重参数技巧对模型输出结果与实际的label计算q_{error}作为损失$loss_3$，结合前两个阶段的$loss_1$与$loss_2$对神经网络模型进行反向传播训练。对于只包含等于条件的查询q，可以将符合查询的序列直接输入模型得到估计值sel(q)′，与查询q的选择率sel(q)计算q_{error}作为损失进行对分组掩码神经网络基数估计模型反向传播。

[0047] 由上可见，本发明关于基数估计模型的三个阶段的训练采用混合训练的方式，即首先使用经过预处理后的数据对模型进行数据训练阶段的训练，属于初始化训练；随后统计训练集查询与实际查询的查询信息与结果，用于混合掩码训练与直接查询训练阶段，当对应阶段的超过误差阈值的查询出现次数大于预先设定的阈值时，会触发对应训练阶段完成对(二)中神经网络模型的训练。

[0048] 基数估计模型混合训练中，需设定误差阈值η_1与η_2，设定查询数量统计阈值$N_{\eta1}$与$N_{\eta2}$，两阶段训练轮次$epoch_1$与$epoch_2$，记录在测试阶段与实际使用阶段误差分别大于阈值的查询Q与对应的出现次数N_{Q1}与N_{Q2}，当对应阶段的超过误差阈值的查询出现次数N_Q大于N_η时，该训练器会触发对应的混合掩码训练或直接查询训练，对模型进行轮次为$epoch_{1or2}$的训练。

[0049] 该基数估计模型混合训练，在使用与测试时，将误差超过η_2的查询Q拆分为只包含

点查询的子查询q并在系统空余时刻自动执行得到选择率sel(Q)。子查询结果会记录在日志中，训练器将其中产生的误差大于η_2(可以默认$\eta_2=1.8$)的子查询加入直接查询训练阶段的训练数据中。

[0050] 拆分查询可以表达为：

[0051] $split(Q)=[q_1,q_2,...q_n]$，

[0052] 设$Q=\{X_1\in R_1,X_3\in R_3,X_4\in R_4\}$，

[0053] 则有：$split(Q)=[q_1=\{X_1=x_1\},Q_3=\{X_1=x_1,X_3=x_3\},Q_4=\{X_1=x_1,X_3=x_3,X_4=x_4\}]$，

[0054] 其中，$x_i\in R_i$，x_i为R_i范围内的随机取值。该查询总共可以拆分为$R_1*R_3*R_4$个子查询，从查询中进行随机取值，取其中产生较大误差的子查询加入数据集中。

[0055] 本发明的基于分组掩码自回归模型的查询基数估计方法具有以下优势：

[0056] 本发明的模型与其训练方法混合了数据驱动与查询驱动对模型进行学习，可以适应不同类型的训练数据组合情况。同时本发明提出的通过掩码由查询指导数据方面的训练方法可以减少训练时间，同时可以使模型相比于传统自回归方法可以更好地获取数据之间的联合概率分布，以提高基数估计的准确性帮助数据库生成更良好的查询计划。

[0057] 基数估计模型混合训练的流程参见图3所示。

附图说明

[0058] 图1为本发明的自回归神经网络模型(掩码自编码神经网络)图示。

[0059] 图2为数据训练阶段掩码覆盖后的训练数据与对应的条件概率。

[0060] 图3为基数估计模型混合训练流程图示。

[0061] 图4为实施方式中自回归神经网络模型的训练过程图示。

[0062] 图5为实施方式中基数估计模型在数据库中的应用图示。

具体实施方式

[0063] 下面是本发明的具体例子，进一步描述本发明图4。

[0064] Census13数据集：这是一个美国人口普查数据集，每行数据包含了年龄、工作类型、教育程度、婚姻情况等属性，总共包含5个数值型属性和8个类别型属性。本发明中使用的Census13数据集包含了48842行数据，数据集初始存储于csv文件中。对于训练和测试所需要的查询负载，则使用开源的随机查询生成器在该数据集上生成1000条查询并获取其对应的查询结果，其中500条作为训练集，500条作为测试集。

[0065] PostgreSQL数据库：这是一个开源的对象关系型数据库，使用SQL语言进行查询，具有查询优化模块和对应的基数估计模块，本发明的部署位置即在它的基数估计模块。

[0066] 首先按照列数据分组预处理方法对Census13数据集进行处理，将列数据分为3组后并进行组内排序，同时生成6种组间排序。完成预处理后进行混合三阶段训练中的数据训练，使模型初步习得数据联合概率分布并可用于查询；使用完成数据训练的模型执行500条训练集查询，计算训练集的q_{error}误差，将其中误差大于η_1($\eta_1=1.5$)的查询加入混合掩码训练部分的数据集中并再次结合原有数据进行训练；再次使用完成了以上两阶段训练的模型执行500条训练集查询，计算q_{error}误差，将其中误差大于η_2($\eta_2=1.8$)的查询加入直接查询

训练的数据集中并再次进行训练。完成训练后可用测试集测试基数估计器效果。完整训练过程见图4。至此模型已经完成了混合三阶段训练。

[0067] 将Census13数据集从csv文件中作为单表数据导入至PostgreSQL之中，并使用上述完成训练的基数模型接入至PostgreSQL的基数估计模块中用以进行查询优化，此时已经可以在该数据库中进行各项业务查询，可以用上述测试集模拟业务查询过程，查询过程可见图5输入查询、基数估计与输出结果的部分。在业务查询进行时，基数估计模块中增加日志模块记录与Census13数据表相关的查询与查询结果并返回至模型训练器中，将误差大于η_1与η_2的查询加入数据集中；如果PostgreSQL中的Census13数据表发生变化，则将其重新导出至csv文件中替代原始数据集，数据更新的过程可以见图5的数据更新连线部分。当记录的两种大于误差的查询出现次数大于查询数量统计阈值$N_{\eta 1}$与$N_{\eta 2}$时，并使用新数据集与新查询再次对模型进行训练，以提高模型查询基数估计准确度，见图5的虚线连接的数据更新与重新训练部分。

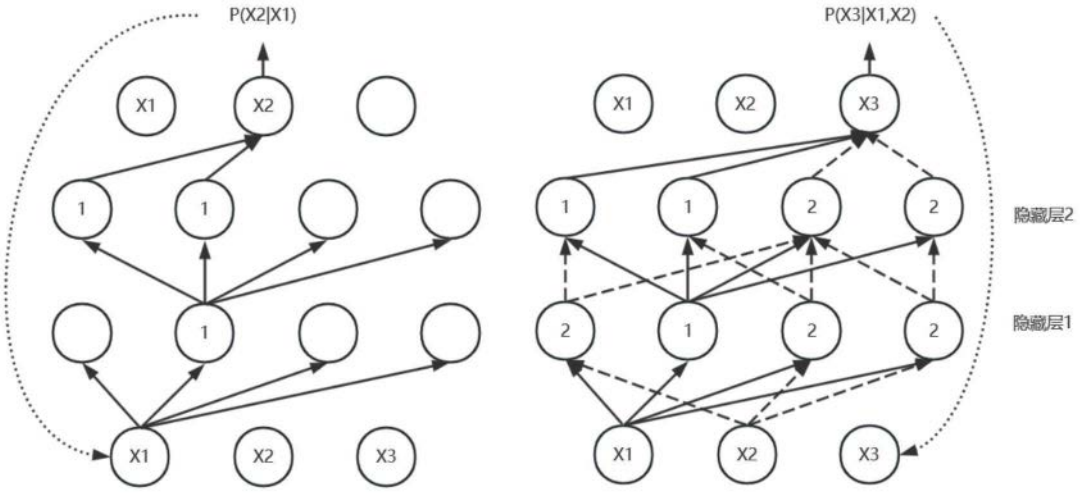

P(X2|X1)
P(X3|X1,X2)
X1
X2
X3
1
2
隐藏层2
隐藏层1

图1

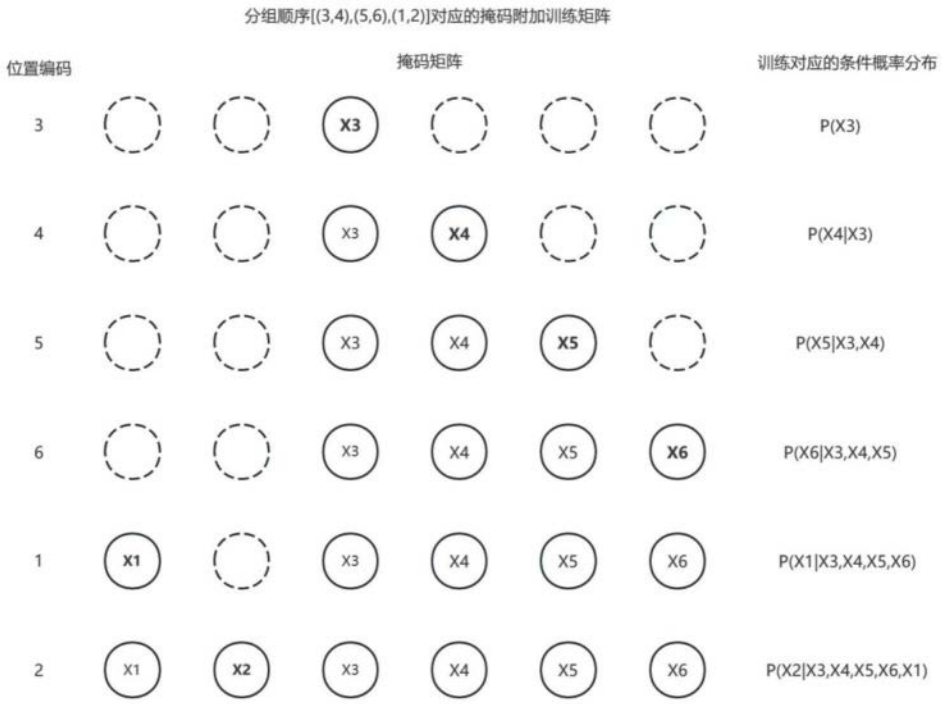

分组顺序[(3,4),(5,6),(1,2)]对应的掩码附加训练矩阵
位置编码
掩码矩阵
训练对应的条件概率分布
3
4
5
6
1
2
P(X3)
P(X4|X3)
P(X5|X3,X4)
P(X6|X3,X4,X5)
P(X1|X3,X4,X5,X6)
P(X2|X3,X4,X5,X6,X1)

图2

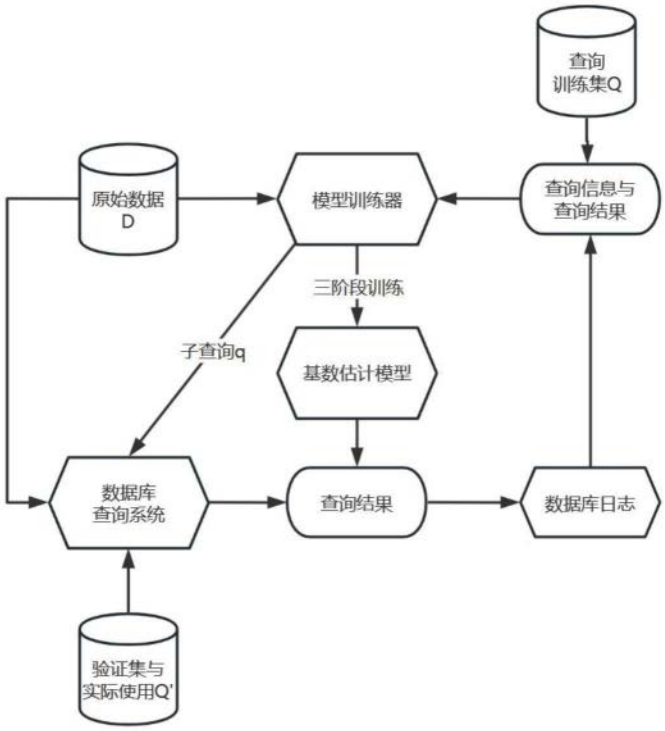

查询
训练集Q
原始数据
D
模型训练器
查询信息与
查询结果
三阶段训练
子查询q
基数估计模型
数据库
查询系统
查询结果
数据库日志
验证集与
实际使用Q'

图3

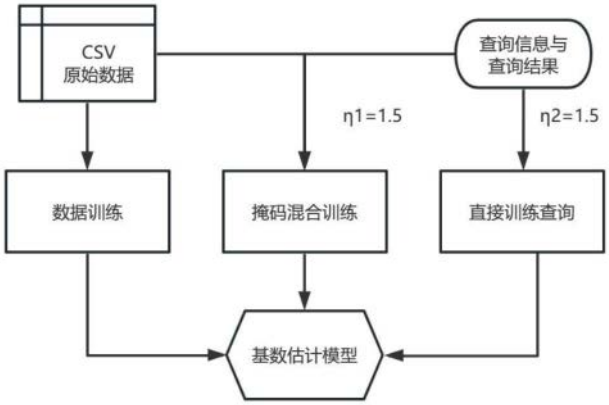

CSV
原始数据
查询信息与
查询结果
η1=1.5
η2=1.5
数据训练
掩码混合训练
直接训练查询
基数估计模型

图4

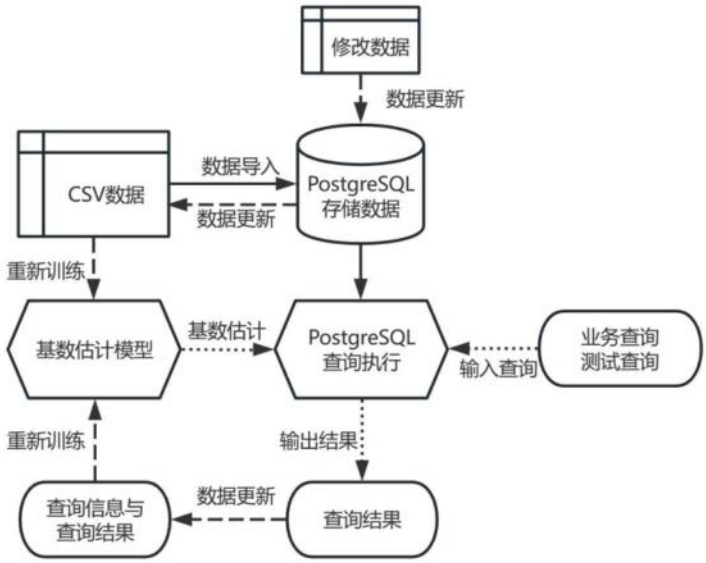

修改数据
数据更新
CSV数据
数据导入
数据更新
PostgreSQL
存储数据
重新训练
基数估计模型
基数估计
PostgreSQL
查询执行
输入查询
业务查询
测试查询
重新训练
输出结果
查询信息与
查询结果
数据更新
查询结果

图5