(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 112131303 A
(43)申请公布日 2020.12.25

(21)申请号 202010988710.5

(22)申请日 2020.09.18

(71)申请人 天津大学
地址 300072 天津市南开区卫津路92号

(72)发明人 李杰 叶一舟

(74)专利代理机构 天津市北洋有限责任专利代理事务所 12201
代理人 刘子文

(51)Int.Cl.
G06F 16/26(2019.01)
G06N 3/04(2006.01)
G06N 3/06(2006.01)
G06N 3/08(2006.01)

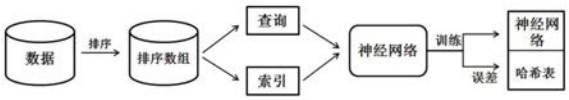
权利要求书1页 说明书6页 附图1页

(54)发明名称

基于神经网络模型的大规模数据沿袭方法

(57)摘要

本发明公开一种基于神经网络模型的大规模数据沿袭方法,包括以下步骤:(1)生成网络训练集;包括数组排序、维度标准划分和训练子集划分;根据数据集中不同维度中的值对数据集中的数据进行排序;为每个维度确定一个划分标准以解决样本穷举问题;将训练集分为许多较小的训练子集;(2)训练神经网络模型;使用分层的网络结构代替传统的神经网络结构,以解决由于样本数据差别较大造成的误差问题;分层结构具体包括网络选择器和子网两大部分;(3)可视交互与沿袭;具体包括空间散点图、时空投影视图和模式对比视图;用于对数据集进行可视化交互探索,使用可视化的方式方便用户对数据结果进行探索;并允许用户通过沿袭的方式探索数据来源。



CN 112131303 A

1.基于神经网络模型的大规模数据沿袭方法,其特征在于,包括以下步骤:

(1)生成网络训练集;包括数组排序、维度标准划分和训练子集划分;根据数据集中不同维度中的值对数据集中的数据进行排序并存储;为每个维度确定一个划分标准以解决样本穷举问题;并将训练集分为若干个子集,作为训练子集;

(2)训练神经网络模型;使用分层的网络结构代替传统的神经网络结构,以解决由于样本数据差别大造成的误差问题;分层的网络结构具体包括网络选择器和子网两大部分,使用网络选择器作为第一层,其作用是为查询找到正确的对应子网;对于每一个子网,使用训练子集分别进行训练;

(3)可视交互与沿袭;将神经网络的输出结果映射为若干种视图,具体包括空间散点图、时空投影视图和模式对比视图;用于对数据集进行可视化交互探索,使用可视化的方式方便用户对数据结果进行探索;并允许用户通过沿袭的方式探索数据来源。

基于神经网络模型的大规模数据沿袭方法

技术领域

[0001] 本专利主要涉及机器学习和数据可视化领域,具体涉及对大规模数据集的实时交互及神经网络模型优化的方法。

背景技术

[0002] 近年来研究人员面对的数据集所包含的数据量级呈指数型增长[4],这无疑给交互式可视化探索与沿袭带来了麻烦。最近提出的技术使分析人员可以实时地交互式地探索大规模数据集[5],但是这些技术忽略了人们可能关心隐藏在统计数据分布背后的真实数据[10]。我们从可视化实现了数据的反向生成,因此视觉视图将不再局限于显示数据的统计信息,它还可以用作生成更复杂的视觉视图的数据,或者探索视图子集中数据的详细分布。

[0003] 关于数据沿袭的研究已经在数据库领域进行了一段时间[7]。传统方法通过扩展基本数据模型来捕获源信息[9],由此带来的缺点是显而易见的:必须使用与实际数据不同的模型来存储访问源。Miles等人[8]提出,由数据产生的产品和描述可能隐藏结果的来源以及如何产生结果的细节,他们研究并讨论了数据来源如何可以帮助科学家进行实验。Boris Glavic等人[6]提出了使用查询重写为源元组标注结果元组的方法,并在数据库中证明了其可行性。K.Dursun等人[1]提出了一种新的中间体重用模型,该模型可缓存在查询处理过程中实现的内部物理数据结构。这项工作通过研究数据库中中间体的重用来加速分析查询的处理。R.Ikeda等人[2]的panda实现了物源捕获,存储,运算符和查询。他们将数据沿袭应用于诸如调试,审计,数据集成,安全性,迭代分析和清理之类的任务。在他们的基础上,Fotos Psallidas等人[3]提出了Smoke,这是一个内存数据库引擎,不需要牺牲沿袭捕获开销。Smoke将哈希表形式的谱系情况以哈希表的形式预先存储,以节省谱系查询带来的时间开销,可以满足实时视觉交互要求。

[0004] 上述的工作主要使用较大规模的数据集,然而这些工作都存在一些缺点和不足:首先,一些工作为每个输入创建哈希索引以加快沿袭查询,但是与此同时,随着数据大小的增加,哈希表的大小也会增加,这可能会带来诸如内存耗尽的问题。其次,最新工作使用一种方法在内存中实时实现哈希表,以加快查询速度,但即使此方法优化了实时生成哈希表的时间,它仍然带来了不可避免的存储开销和额外的查询时间。同时,上述工作无法使用查询数据再次生成可视化,它只能在多个可视化视图之间建立连接。

发明内容

[0005] 本发明的目的是为了解决现有技术中的以下问题。1.使用神经网络模型取代传统索引结构,从而减少查询带来的时间开销与存储开销。2.对于大量数据,神经网络无法很好地满足查询和索引之间的关系,因此需要使用层次结构来解决此问题。分层结构包括第一层网络选择器,用于查找查询对应的子网;以及第二层子网络,用于计算并输出查询结果。3.大规模数据集往往包含多个维度,用户可能不仅需要约束一个维度,所以要解决同时满足多维约束的沿袭查询,需要对不同的维度制定不同的划分标准,并为每一个维度分别训

练神经网络模型。因此，本发明提出了一个基于神经网络模型的框架以沿袭探索大规模数据集。首先，框架采用了一个基于神经网络模型的索引结构，满足实时交互式沿袭查询。其次，框架集成了层次结构网络模型以及哈希表，实现对误差数据的处理。最后，设计支持对该数据结果进行快速查询及交互的可视化界面。

[0006] 本发明的目的是通过以下技术方案实现的：

[0007] 基于神经网络模型的大规模数据沿袭方法，包括以下步骤：

[0008] (1)生成网络训练集；包括数组排序、维度标准划分和训练子集划分；根据数据集中不同维度中的值对数据集中的数据进行排序并存储；为每个维度确定一个划分标准以解决样本穷举问题；并将训练集分为若干个子集，作为训练子集；

[0009] (2)训练神经网络模型；使用分层的网络结构代替传统的神经网络结构，以解决由于样本数据差别大造成的误差问题；分层的网络结构具体包括网络选择器和子网两大部分，使用网络选择器作为第一层，其作用是为查询找到正确的对应子网；对于每一个子网，使用训练子集分别进行训练；

[0010] (3)可视交互与沿袭；将神经网络的输出结果映射为若干种视图，具体包括空间散点图、时空投影视图和模式对比视图；用于对数据集进行可视化交互探索，使用可视化的方式方便用户对数据结果进行探索；并允许用户通过沿袭的方式探索数据来源。

[0011] 与现有技术相比，本发明的技术方案所带来的有益效果是：

[0012] 1.可视化沿袭，这是一种基于可视化的新查询方法来查找数据，并着重于如何在人们感兴趣的可视化视图区域后面查找真实数据。据了解，现如今的方法都无法做到实时交互查找真实数据，尤其是数据中的细节。然后，可以使用这些数据生成新的可视化视图，以帮助用户进一步分析和研究。

[0013] 2.分层的神经网络结构。它可以有效地减少每个神经网络需要预测的值的范围，帮助控制神经元的数量，较少的神经元数量可以使得网络易于观察和调整，同时解决了网络的更新问题。该结构可以有效控制误差，一个仅需要适应数千个数据样本和标签之间关系的神经网络，可以简单地将最大允许误差控制在一个合理的值。

[0014] 3.一种基于神经网络模型的方法，它使用哈希表和分层神经网络的复合结构来支持以交互速率对大型数据进行沿袭查询。它花费很少的额外存储开销，并实现了细粒度的沿袭查询。该结构实现实时交互探索的同时，花费的存储开销比现存的所有技术都要少。该结构支持更新，从而解决了神经网络结构的常见问题。

附图说明

[0015] 图1为提出方法的总流程图。

[0016] 图2为网络训练集生成图。

[0017] 图3为神经网络结构图。该图中：a表示用户输入的查询，b表示查询通过网络选择器输出对应子网络，c表示查询对应的子网络输出对应的数据所在位置。

具体实施方式

[0018] 以下结合附图和具体实施例对本发明作进一步详细说明。应当理解，此处所描述的具体实施例仅仅用以解释本发明，并不用于限定本发明。

[0019] 本发明提出了一个基于神经网络模型的框架以沿袭探索大规模数据集。首先，框架采用了一个基于神经网络模型的索引结构，对于数据集中的每一个维度，将所有查询条件作为样本，查询结果作为标签，训练神经网络模型，满足实时交互式沿袭查询。其次，层次结构网络模型以及哈希表，通过将网络分为网络选择器与子网络两层，有效减少每个神经网络需要预测的值的范围，以及使用哈希表记录误差较大的数据，实现对误差数据的处理。具体而言，如图1所示，主要包含以下步骤：

[0020] 步骤一：网络训练集生成(图2)。具体操作包括数组排序、维度标准划分和训练子集划分。据数据集中不同维度中的值对数据集中的数据进行排序，然后存储所有这些顺序；为每个维度确定一个划分标准以解决样本穷举问题；将训练样本分为许多较小的子集以克服神经网络不能很好地适应细微的误差。

[0021] 对数组进行排序是数据结构中非常重要的组成部分。简单来说，对数组进行排序意味着根据不同维度中的值对数据集中的数据进行排序，然后存储所有这些顺序。训练神经网络需要输入和输出之间的某种功能关系，而顺序排序可以在查询和输出之间建立某种线性关系。对于数据集中的多个维度(经度，纬度，小时，天)，将为每个维度存储一个排序后的数组，排序依据是每个数据在那个维度上的值。像数据库一样，为每个数据分配唯一的主键，因此每个排序的数组仅需要存储主键，这将大大减少存储空间的开销。通常，每条数据只会带来最低的数据类型存储消耗，即4个字节。

[0022] 训练样本的生成需要分别为每个维度生成训练样本。为了使训练样本详尽无遗，需要为每个维度确定一个划分标准，即样本基于此准则。对于具有明确标准(例如天)的尺寸，将一天作为划分标准；对于没有明确标准(例如经度或纬度)的尺寸，将它们尽可能细地分开。训练样本分为输入和输出(标签)。输入是查询条件，即一个分辨率大小从零开始增加，直到达到维度的最大值；输出是每个输入对应的查询结果。输出是查询结果在与每个输入相对应的每个已排序数组中的索引位置，即最后一个数据的索引，该索引的值小于某个维的已排序数组中的输入。在这种情况下，输入和输出单调增加，因此它们之间可能存在一定的线性关系。由于神经网络不能很好地适应细微的误差，随着训练样本范围的扩大，这个问题将被放大，最终导致无法接受的误差。因此将训练样本分为许多较小的子集，然后使用它们来训练不同的神经网络。对于每个子集，使用标准化过程来简化神经网络的训练。

[0023] 步骤二：训练神经网络模型(图3)。使用分层的网络结构代替传统的神经网络结构，以解决由于样本数据差别较大造成的误差问题；分层结构具体包括网络选择器和子网两大部分，使用网络选择器作为第一层，其作用是为查询找到正确的对应子网；对于每一个子网，使用训练子集分别进行训练。

[0024] 有了训练样本，传统的方法是使用它来训练神经网络，然后保存网络。在本实施例的过程中，由于样本范围的较大范围，神经网络的细微偏差被放大了许多倍，因此误差是不可接受的。因此决定使用分层结构代替传统的神经网络。层次结构主要由两部分组成：网络选择器和子网。使用网络选择器作为第一层，其作用是为查询找到正确的对应子网。根据一定数量均匀地分配训练样本，因此在输入查询时，可以轻松地通过一次计算来计算出它属于哪个子网。然后对查询进行规范化并将其输入到子网中，然后子网在排序数组中输出与查询相对应的数据的索引。应当注意此处输出的索引不是完全准确的值。这是由于神经网络的性质所致，这意味着除非神经元的数量足够大，否则它无法完全适合每条数据。因此为

神经网络的输出设置一个误差值。只要输出在误差范围内，就认为神经网络的预测是成功的。使用分层结构而不是整个神经网络的好处如下：

[0025] (1)它可以有效地减少每个神经网络需要预测的值的范围。对于大型数据集，查询和索引之间的关系总体上是线性的，但是如果放大单个记录，它将显示越来越多的不规则性。但是，如果单个神经网络只需要满足数千个查询和索引之间的关系(极少数情况除外)，则神经网络具有完美的功能。

[0026] (2)它可以帮助控制神经元的数量。例如，对于具有数千万个样本的训练集，很难确定单个神经网络中神经元的数量。但是，如果对于仅包含数千个样本的训练集，只需要几个神经元即可取得非常好的结果。另外，神经元较少的神经网络可以将较长的训练时间分成单独的小部分，这更便于观察和调整。

[0027] (3)它可以帮助设置误差。如果单个神经网络需要适合数百万甚至数千万个数据样本和标签之间的关系，则很难控制预测值和真实值之间的允许误差。但是相比之下，如果单个神经网络只需要适应数千个数据样本和标签之间的关系，那么可以简单地将最大允许误差控制在一个合理的值。

[0028] 即使将数据分成许多小部分，对于神经网络，仍然只有很少的数据无法将预测值和真实值之间的误差控制在最大允许范围内。对于神经网络“异常”的这部分数据，将使用哈希表来存储它们。因为它们的数量最少，并且哈希查询的时间复杂度仅为O(1)，所以哈希表带来的时间开销和存储开销几乎可以忽略不计。因此最后，使用预排序的数据，自适应单层神经网络和其他哈希表作为最终的索引结构。之后，在不同的维度上，使用神经网络或哈希表的输出来查找原始数据中与查询相对应的位置，然后获得不同维度的结果，对这些结果进行合取运算，并根据总结用户提供的条件，然后将结果传递到前端界面，然后完成前端界面以进行视觉呈现。

[0029] 步骤三：可视交互与沿袭。对数据集提供可视化交互沿袭。具体包括空间散点图、时空投影视图和模式对比视图；用于对数据集进行可视化交互探索，使用可视化的方式方便用户对数据结果进行探索；并允许用户通过沿袭的方式探索数据来源。

[0030] 地图空间散点图。

[0031] 可视化界面主要通过散点图反映数据在地图上的分布。用户可以根据地图上每个点的突出程度来判断该区域中的数据分布量。同时，散点图可以约束其他四个视觉视图。用户可以在散点图上选择一个框，以进一步探索和分析局部区域而不是整个地图中的数据。使用散点图上的用户选择框作为查询，返回查询的沿袭结果，将结果存储在内存中，并通过即时计算更新其他四个视图。散点图支持缩放，因此用户可以在很小的区域(例如仅包含少量数据的街道)中浏览单个数据。散点图会在其他四个视觉视图中响应用户的操作，然后在用户感兴趣的特定时间段内在地图上显示数据分布。

[0032] 视图组件

[0033] 可视化界面主要使用折线图，条形图和热图来反映数据的时间分布。用户可以通过限制散点图来执行沿袭查询。通过即时计算以热图，折线图和条形图的形式绘制查询结果，使用户可以轻松地探索和分析本地数据的时间分布。如果用户不执行约束，则视觉视图将反映整个数据集的时间段信息。用户可以在除热图之外的三个视图组件中选择框，以约束所有其他视图(包括可视图)反映的数据内容。用户可以根据感兴趣的时间段来构图视图

组件。例如，要浏览和分析周末数据，用户需要对星期投影直方图执行框选，而要在夜晚浏览和分析数据，用户需要对小时直方图执行框选。小时投影直方图反映了24小时内的数据分布，星期投影直方图反映了一周7天之内的数据分布，而天分布直方图则反映了开始日期和结束日期之间的数据分布。天分布直方图中添加了一个额外的摘要视图，该视图以细线图的形式出现，反映了时间段内整个数据的分布。用户可以根据自己的兴趣在摘要视图上选择感兴趣的时间范围，并将天分布直方图用作详细视图以可视化查询结果。

[0034] 资源占用。

[0035] 该数据结构的存储开销主要来自神经网络和预分类。表1中展示了一些细节。预排序所占用的存储空间与沿袭查询的维数有关。沿袭查询的每个维度都将使数据结构使用相同数量的原始数据再存储一个排序后的数组，尽管排序后的数组中的每个数字都是最小的4字节。存储开销的另一部分来自神经网络的参数。神经网络的总参数与神经网络的数量和每个神经网络的神经元数量有关。数据结构中神经网络的数量由不同维度的分辨率决定。在时间维度中，沿袭查询的分辨率为一小时，然后将两个时间戳为10h15m和10h20m的对象划分为同一标签，即10h。因此，定义越高，神经网络的数量或每个神经网络的神经元的数量就会越多。实现数据结构时更倾向于增加神经网络的数量，这是因为神经网络可以很好拟合的数据具有一定的规律性，并且其范围不应超过特定范围。

[0036] 在训练过程中，增加神经网络的数量并减少每个神经网络需要拟合的数据数量可以有效地提高准确性，这将导致神经网络占用的存储开销是恒定的，因为插入新数据时新分辨率不会改变，以相应的分辨率更改索引值即可完成结构的快速更新。对于新数据的每个维度添加一个索引值，该索引值与其在相应维度的排序数组中的位置相对应，该索引值的大小也为4个字节。输入对象的数量，训练时间和存储开销在前三列中报告。输入对象的数量是指数据集中有效数据的数量。训练时间是训练所有需要训练的网络所花费的时间的总和。网络列表示每个数据集使用的子网数。对象(网络)列表示与每个子网相对应的数据对象的数量。由于数据结构从原始数据中丢弃了一些有价值的信息。因此此处的存储开销可能不一定大于数据集本身的大小。

[0037] 该数据结构可以处理非常细致的数据沿袭，例如一位数的数据沿袭，精度为1.0。同时，在处理大规模数据沿袭时，它仍然具有良好的性能，较低的时间开销和存储开销。对于包含数百万或数千万数据的数据集，很难在如此多的数据中仅找到其中的几个。该数据结构能够以良好的性能实现对大规模数据集的沿袭查询的细粒度可视化。

[0038] 表1资源占用

[0039]

Dataset	Objects	Time	Size	Nets	Objects(net)
Brightkite	4.5M	2.5h	223MB	20K	1678
Crime	6.5M	3.0h	272MB	20K	1678
Taxi	14.0M	2.5h	587MB	20K	1678

[0040] 本发明并不限于上文描述的实施方式。以上对具体实施方式的描述旨在描述和说明本发明的技术方案，上述的具体实施方式仅仅是示意性的，并不是限制性的。在不脱离本发明宗旨和权利要求所保护的范围情况下，本领域的普通技术人员在本发明的启示下还可做出很多形式的具体变换，这些均属于本发明的保护范围之内。

[0041] 参考文献：

[0042] [1]K.Dursun,C.Binnig,U.Cetintemel,and T.Kraska.Revisiting reuse in main memory database systems.arXiv preprint arXiv:1608.05678,2016.

[0043] [2]R.Ikeda and J.Widom.Panda:Asystem for provenance and data.IEEE Data Eng.Bull,2010.

[0044] [3]Fotis Psallidas and Eugene Wu.Smoke:Fined-Grained Lineage Capture At Interactive Speed.Proceedings of the VLDB Endowment,2018.

[0045] [4]F.Psallidas and E.Wu.Provenance for interactive visualizations.HILDA,2018.

[0046] [5]J.Poco,J.Heer,Reverse-engineering visualizations:Recovering visual encodings from chart images.Comput.Graph.Forum,2017.

[0047] [6]B.Glavic and G.Alonso.Perm:Processing provenance and data on the same data model through query rewriting.ICDE,2009.

[0048] [7]Muniswamy-Reddy,K.-K.,Macko,P.,and Seltzer,M.Provenance for the cloud.Proceedings of the 8th USENIX Conference on File and Storage Technologies(FAST),2010.

[0049] [8]Miles,S.,Groth,P.,Deelman,E.,Vahi,K.,Mehta,G.,and Moreau,L.Provenance:The bridge between experiments and data.Comput.Sci.Engin,2008.

[0050] [9]B.Glavic.Big data provenance:Challenges and implications for benchmarking.Specifying Big Data Benchmarks-First Workshop,WBDB,2014.

[0051] [10]J.Wang,D.Crawl,S.Purawat,M.Nguyen,I.Altintas,Big data provenance:Challenges state of the art and opportunities.Big Data,2015.

[0052] 本发明并不限于上文描述的实施方式。以上对具体实施方式的描述旨在描述和说明本发明的技术方案，上述的具体实施方式仅仅是示意性的，并不是限制性的。在不脱离本发明宗旨和权利要求所保护的范围情况下，本领域的普通技术人员在本发明的启示下还可做出很多形式的具体变换，这些均属于本发明的保护范围之内。

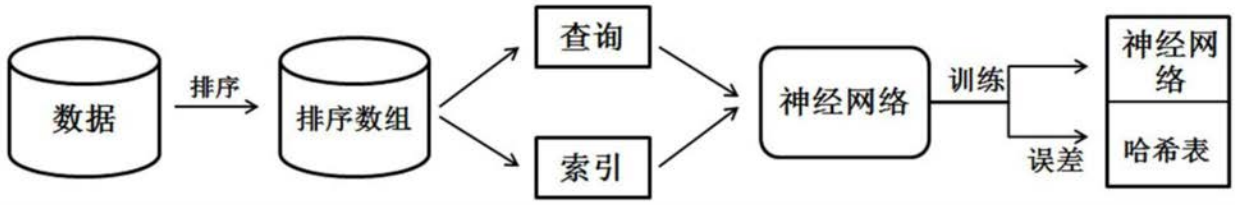

数据
排序
排序数组
查询
索引
神经网络
训练
误差
神经网络
哈希表

图1

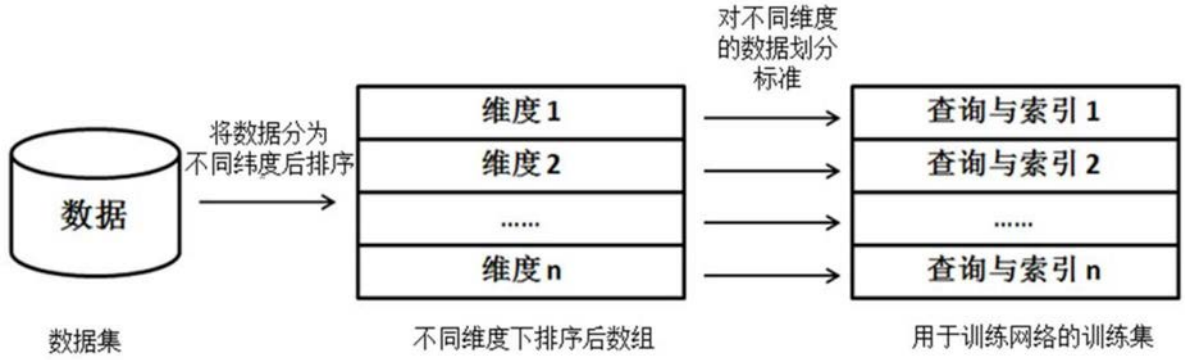

对不同维度的数据划分标准
数据
将数据分为不同纬度后排序
维度1
维度2
.....
维度n
查询与索引1
查询与索引2
.....
查询与索引n
数据集
不同维度下排序后数组
用于训练网络的训练集

图2

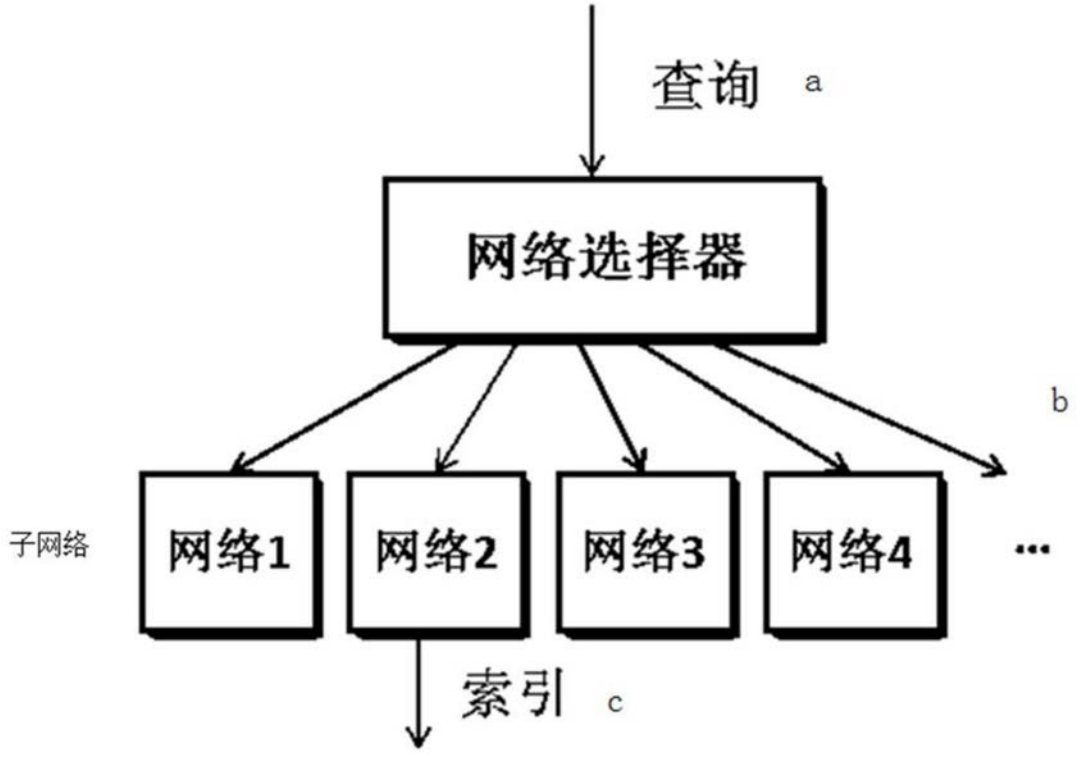

查询
a
网络选择器
b
子网络
网络1
网络2
网络3
网络4
...
索引
c

图3