ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 37,No. 5,Sep. 2022,pp. 1049－1058
DOI:10. 16337/j. 1004-9037. 2022. 05. 010

http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax:＋86-025-84892742

基于两阶段分层抽样的近似聚合查询方法

房　俊[1,2]，赵　博[1,2]，左昌麒[1]

(1. 北方工业大学信息学院,北京 100144;2. 大规模流数据集成与分析技术北京市重点实验室(北方工业大学),北京 100144)

摘　要：以数据仓库应用为代表的交互式查询分析技术为智能决策提供了支持。随着数据规模的不断增大,准确计算聚合查询结果往往需要全局数据扫描,使得这类查询面临着实时响应能力不足的问题。基于预先抽取的样本数据,复杂聚合查询提供快速的近似答案,在许多场景下是解决该问题的可行方案。分析了分层抽样优于随机抽样的具体条件,提出了一种两阶段分层抽样方法。首先针对业务特征进行分组,每个分组中使用随机抽样方法进行随机抽样,并评估其抽样效果。再针对抽样效果较差的分组,利用自组织特征映射网络(Self-organizing feature mapping,SOM)对数值进行聚类分组,改进其近似查询效果。基于公开数据集和实际电网数据的实验结果表明:本文方法相比于随机抽样、分层随机抽样以及国会抽样算法在相同抽样率下可达到15%的性能提升;与使用K-means、基于密度的聚类算法(Density-based spatial clustering of applications with noise,DBSCAN)等聚类方法相比,自SOM具有较好的近似查询结果。

关键词：聚合查询;分层抽样;SOM聚类;预计算;近似查询

中图分类号：TP391　　**文献标志码**:A

Approximate Aggregate Query Method Based on Two-Stage Stratified Sampling

FANG Jun[1,2], ZHAO Bo[1,2], ZUO Changqi[1]

(1. College of Information, North China University of Technology, Beijing 100144, China;2. Beijing Key Laboratory on Integration and Analysis of Large-Scale Stream Data, Beijing 100144, China)

Abstract: The interactive query analysis technology represented by data warehouse application provides support for intelligent decision-making. With the continuous increase of data scale, accurate calculation of query results often requires global data scanning, which makes the group-by query face the problem of insufficient real-time response ability. Based on the pre-extracted sample data, it can provide fast approximate answers for aggregate queries, which is a feasible solution to this problem in many scenarios. This paper analyzes the specific conditions that stratified sampling is better than random sampling, and proposes a two-stage stratified sampling method. In the first stage, the sampling is grouped according to the business characteristics. In each grouping, the random sampling method is first used for random sampling, and the sampling effect is evaluated. To improve the effect of approximate query, the second stage sampling is carried out, and the self-organizing feature mapping (SOM) clustering method is used to group the values. Experimental results on the public data set and the actual power grid data show that,

基金项目:国家自然科学基金国际(地区)合作与交流项目(62061136006)。
收稿日期:2021-09-10;**修订日期**:2022-01-23

compared with random sampling, stratified random sampling and congressional sampling algorithm, performance of the proposed method can be improved by 15% at most under the same sampling rate. And SOM has better approximate query results than K-means and density-based spatial clustering of applications with noise (DBSCAN) clustering methods.

Key words: aggregate query; stratified sampling; self-organizing feature mapping clustering; pre-computing; approximate query

引　言

许多行业利用大数据、互联网和物联网等技术存储了海量数据，为数据仓库类应用带来新机遇。以国家电网谐波监测系统[1]为例，其采集的电能质量数据具有数据量大、实时性强的特点。通过时空维度、电能质量物理指标量的即时聚合统计，可实现电网谐波污染情况快速评价，保障电网安全。多维聚合查询服务能够响应业务客户的交互式查询要求，快速灵活地产生分析结果，广泛应用于故障预警、交互数据分析等场景。随着数据规模的不断增大，如何快速返回交互查询结果成为挑战。当前主要处理方式包括：(1)引入索引技术，索引可实现查询谓词快速发现，但如果谓词选择性较差，聚合响应有较长延时；(2)利用MapReduce、Spark等并行计算框架可改善查询性能，但消耗资源较多；(3)使用预计算[2]方法提前完成聚合查询。上述查询方式都产生精确结果，但消耗较多资源，许多情况下也需要全局数据扫描，使得这类查询面临着实时响应能力不足的问题，限制了其使用。

近似查询处理[3]模式通过对数据规模进行合理规约，容忍一定范围的查询误差，以较少的资源消耗实现较快的查询响应，在探索式数据分析、故障实时预警以及机器学习等前沿应用领域获得了较多关注和研究。基于样本的近似查询方法[3]可以分为在线抽样方法和离线预计算方法。在线抽样方法根据用户需求即时生成并使用样本完成近似查询，为了提升近似查询精度，需要生成较多的样本，此时必然需要较多的抽样时间，这就会影响近似查询的时间性能。离线预计算方法则预先进行抽样，查询时不需要花费抽样时间，故查询时间性能较优，但这种方法一般需要知道数据分布状态以及查询负载的类型，通用性较差。一个近似查询处理方案可以从4个维度进行评价[4]：所支持查询的通用性，查询准确性，查询时延以及预计算的额外开销。这4个维度不是独立的，到目前为止所有的方法都是面向其中某些维度的优化方案。数据抽样具有易于理解、实现简单、普适性好以及理论支撑鲜明等特点，是数据规约[5]较多采用的技术方法。此外，与直方图、小波采样和草图等方法相比，不存在维度爆炸问题[6]。常见的数据抽样方法包括简单随机抽样和分层随机抽样等[7]。简单随机抽样方法实现比较简单，但是在面对聚合查询，特别是分组内部数据倾斜比较严重时，存在分组丢失或者分组近似查询误差大的问题[8]。一些工作[8-10]采用了分层随机抽样，并且重点考虑了不同分组样本数以及抽样率的方案设计。如国会抽样[8]根据分组属性对总体样本进行切分，每个分组根据理论分析得出的样本量进行随机抽样。CVOPT[10]利用方差系数作为度量和优化指标，得出每个分层的样本量。本文方法与上述思路都不一样，考虑的是在每个分组内部可以进一步采用聚类等方法对数值进行分层抽样，改进层内的查询精度。SSAQP方法[11]考虑数值聚类特性，将总体样本根据待聚合数值列的大小划分为3个类别，分别代表大值类、小值类和常值类，在此基础上对每个类别使用分层抽样方法构建样本集。该方法针对极端值情况可以取得较好效果，但本质上组内仍然是随机抽样。本文方法使用SOM聚类方法确定聚类结果，相比于SSAQP方法通用性更强，也更加合理。除了基于抽样的方法外，直方图[12]、小波采样和草图等方法也被利用来解决近似查询问题，这些方法在处理某些特定问题，如百分位数查询，具有优势，并且查询时延更小。但这些方法通用性一般较差，此外直方图、小波等方法面临维度爆炸问题[6]，上述不足也

限制了这些方法的应用。近年来,除了分层抽样研究之外,还出现了一些近似查询的新技术,如基于最大熵原理[13]规约数据,特别是基于人工智能的新方法具有通用性强、准确性高和时延短等特点,得到了本领域研究者普遍关注,也取得了初步研究成果。如通过模型预测近似查询结果[14-17]或者快速生成抽样样本[18],这些方法为解决近似数据查询问题带来了新思路。但正如文献[17]中所述,基于人工智能的方法目前还比较适合于作为已有近似查询方法的有益补充。

1　问题分析

1.1　基本定义

定义1　聚合查询。给定一个关系数据集合R,$D=|R|$为关系条目数,$C=\{c_i|i\in[1,m]\}$为其属性集合,m为属性数量,给定聚合列$C'\subseteq C$,其分组表示为$\{g_i|i\in[1,n]\}$,n为分组个数。定义聚合查询Q为$R\rightarrow R'$的映射,其中$R'=\{(g_i,Q(g_i))\}$。根据一个或多个属性的值将输入关系划分为多个组,使用聚合函数作用在上述分组上。为简化起见,本文暂不考虑查询谓词条件。如下例所示,由空气质量表airQuality(location, parameter, value)构成的关系中,查询每个城市PM2.5的平均值:select locationId, avg(value) from airQuality where parameter = 'pm25' group by locationId。

定义2　基于数据抽样的近似聚合查询。给定关系数据集合R和聚合查询Q,函数$Q'\times S$称为基于数据抽样的近似聚合查询。其中S是抽样函数,$S:R\rightarrow r$,$r\subset R$;函数Q'满足$\left|Q'(r)-Q(R)\right|<\theta$,$\theta$为近似误差的阈值。

近似聚合查询包括2个步骤:首先将数据集R规约为规模较小的数据集r;然后对查询Q进行适当改造,生成可作用于r的查询Q',使得$Q'(r)$逼近$Q(R)$。基于数据抽样的查询近似需要在查询时延和查询误差之间取得平衡。

定义3　聚合查询误差。Q是一个聚合查询,对关系R上的属性集C进行聚合操作。设$\{g_i|i\in[1,n]\}$为查询结果中出现的所有组的集合,n为出现的所有组个数,$Q(g_i)$和$Q'(g_i)$是分组g_i中Q查询的精确和近似聚合值,将误差ϵ_i定义为近似值和精确值间的相对误差,有

$$\epsilon_i=\frac{\left|Q(g_i)-Q'(g_i)\right|}{Q(g_i)} \tag{1}$$

在单组误差评价基础上,使用分组误差的均值作为聚合查询的整体误差,有

$$\epsilon=\frac{1}{n}\sum_{i=1}^{n}\epsilon_i \tag{2}$$

1.2　分层抽样优于随机抽样的条件

抽样率一定的情况下,如何分层以及样本量分配是分层抽样的关键问题。许多工作讨论了样本量的确定方法[5,19],如样本量的确定采用比例分配或奈曼分配,可取得较好的近似效果。本文采用的是比例分配的方式进行样本量确定。一般来讲,需要估计子总体参数时,会按照自然层进行划分。如上例中需要估计各个城市PM2.5的评价值,按照城市进行自然分层,再在每层内使用简单随机抽样是自然的想法。然而这种自然分层方法有时并不能带来比随机抽样更好的效果。如表1所示A、B两个城市观测到1组PM2.5数据,按照50%抽样率进行抽样,加"—"为随机抽样样本,加粗的为分组抽样样本。聚合计算结果如表2所示,此

表1　PM2.5数据

Table 1　PM2.5 data

城市A	24	$\overline{25}$	33	88	$\overline{\mathbf{86}}$	$\overline{\mathbf{97}}$	94	62	57	$\overline{60}$
城市B	$\overline{26}$	$\overline{29}$	$\overline{\mathbf{89}}$	92	**98**	**94**	**59**	$\overline{64}$	$\overline{56}$	$\overline{\mathbf{102}}$

时随机抽样误差优于分层抽样误差。

分层随机抽样和简单随机抽样构造的统计量是原统计量的无偏估计,什么情况下分层抽样一定会比随机抽样的精度高?按照抽样理论,这两种方法的效率可以通过估计量的方差大小评价,前提是两种方法采用同样的样本量。

表2 聚合查询结果

Table 2 Aggregation results

城市	全量均值	随机抽样		简单分层抽样	
		均值	相对误差/%	均值	相对误差/%
A	62.6	67	7.03	72.8	16.29
B	70.9	61	13.96	88.4	24.68

以均值估计量的方差计算说明其比较过程,详细过程可参考文献[20]。为方便表示,令简单随机抽样的方差为 $V_1=\frac{1-f}{n}S^2$,其中 $f=\frac{n}{N}$ 为抽样比。n 为样本个数,N 为总体个数,S 为样本标准差。比例分层的均值估计量方差为 $V_2=\frac{1-f}{n}\sum_{h=1}^{L}W_hS_h^2$,其中 $W_h=\frac{n_h}{N}$ 为层权重。比较 V_1 与 V_2 的差值为

$$V_1-V_2=\frac{1-f}{n}S^2-\frac{1-f}{n}\sum_{h=1}^{L}W_hS_h^2=\frac{1-f}{n(N-1)}\left[\sum_{h=1}^{L}N_h(\bar{Y}_h-\bar{Y})^2-\frac{1}{N}\sum_{h=1}^{L}(N-N_h)S_h^2\right] \tag{3}$$

故当 $\sum_{h=1}^{L}N_h(\bar{Y}_h-\bar{Y})^2>\frac{1}{N}\sum_{h=1}^{L}(N-N_h)S_h^2$ 时,分层抽样的效率低于简单随机抽样。假设层内方差都相等,式(3)可进一步简化为

$$\frac{\sum_{h=1}^{L}N_h\left(\bar{Y}_h-\bar{Y}\right)^2}{L-1}>S_h^2 \tag{4}$$

也就是当层(组)间方差小于层(组)内方差时,分层抽样效率有可能是低于简单随机抽样的。反之,只要层(组)内方差小于层(组)间方差,分层抽样的精度就会高于简单随机抽样,并且层(组)间差异越大,分层抽样效果越好,为此最大化各层差异即成为分层抽样的关键问题。

实际应用过程中,由于分层抽样的分组属性往往层内内聚,层间差异较大,满足上述条件的情况并不多见。但上述分析提供了一种改进分层随机抽样的思路,即对于随机抽样近似查询效果不好的分组,可以使用数据值聚类的方法优化近似查询准确率。

2 两阶段分层抽样方法

本节首先介绍自组织特征映射网络SOM,然后详细描述两阶段分层抽样方法,最后提出一种样本增量更新方法应对实际应用中总体规模不断增加的情形。

2.1 SOM网络

累积平方根法[21]是一种常见的确定数值聚类边界的近似方法,需要确定不同范围的频数,可行性较差。机器学习聚类算法[22]包括:基于划分的方法,如K-means方法;密度聚类算法,如DBSCAN方法;基于模型的算法,如自组织映射神经网络方法。SOM的构建分为两个阶段:竞争阶段和合作阶段[23]。在第1阶段,选择最匹配的神经元,即"赢家";在第2阶段,调整赢家及其相邻格子的权值。SOM算法的优点包括:不需要事先定下类的个数,且受初始化的影响较小;受孤立点和噪声点的影响较小;具有自稳性,适用于大样本数据。

2.2 样本生成方法

给定关系 R 上的聚合查询 Q,C' 为聚合查询作用的属性。记数据量为 $D=|R|$,在给定某时间约束及物理资源约束条件下,可通过特定方法(如机器学习中的回归方法)近似得到查询所能承载的最大样本量,记为 D'。由于样本量越大,抽样精度越高,故使用 D' 作为目标样本量。给定分组聚合结果的相对误差阈值 $\{\theta_i\}$。

(1)基于分组属性 C' 的分层抽样

扫描一遍数据,将数据按照分组属性C'进行分组$\{g_i|i\in[1,n]\}$,并按照比例分配的方式确定每个分组的样本数d_{g_i},在每个分组中进行简单随机抽样;生成样本。

(2)抽样结果评价

首先对全量数据执行查询Q得到每个分组的准确查询结果值,再对样本数据执行查询Q'得到近似查询结果。接下来对抽样结果进行检验,将近似查询结果与准确值进行比较,按照式(1)确定其相对误差。如果$\epsilon_i\leqslant\theta_i$,则表示分组$g_i$采用简单随机抽样即可满足性能要求,放入候选组集合$\{g_{i1}\}$,否则将该分组放入需2次分层抽样候选分组中$\{g_{i2}\}$。

(3)基于数据值的分层抽样

再次扫描一遍数据,对属于候选分组的数据采用基于值的分层方法进行分组,在值分组完毕后同样基于比例分配的方式进行随机抽样。

(4)抽样结果归并

归并$\{g_{i1}\}$、$\{g_{i2}\}$中的分组抽样成为有效抽样结果,作为预计算的样本结果,保存为新的关系R'。算法伪代码如表3所示。

表3　样本生成算法

Table 3　Sample generation algorithm

步骤	伪代码	描述
Begin		
1	D' = getMaxSampleSize(D);	
2	$d[i]$ = getGroupSampleSize(D');	//计算各分组样本量
3	For each data in dataSet	//扫描数据
4	i = getWhichGroup(data);	//确定数据分组
5	$g[i]$.reservoirSampling(data, $d[i]$);	//层内随机抽样
6	End for	
7	accResultSet = calculate(data, Q);	//对全量数据执行查询
8	appResultSet = calculate(g, Q');	//对样本执行近似查询
9	For each g[i] in g	
10	If getError(accResultSet[i], appResultSet [i]) $<$ $\theta[i]$	//对每个分组判断误差是否小于阈值
11	$g1$.add($g[i]$);	//小于则分入g_1分组
12	Else $g2$.add($g[i]$);	//大于则分入g_2分组
13	End if	
14	End for	
15	For each g in $g2$	
16	SOM(valueGroups, g);	//使用SOM算法聚类
17	End for	
18	R' = $g1$.addAll(valueGroups);	//归并为新的关系R'
End		

算法1依然属于分层抽样算法。聚合查询精度方面,对于某些随机抽样效果不好的分组,第15步开始使用SOM数据聚类方法进行值聚类,根据式(3),该步骤会减少基于抽样结果的估计量方差,故算法在精度上优于分层随机抽样算法。查询时间方面,算法需至少扫描两次数据集,且需多次执行SOM

聚类算法,其时间消耗会多于分层随机抽样算法。但由于本文方法为预计算抽样,增加抽样时间以得到更准确的结果在许多应用中是可以接受的。

2.3 样本增量更新方法

将一段时间内新增总体记为R^+。二级分层抽样的样本维护方法包括以下步骤:

(1) 求取$\{g_{i2}\}$中各组数据的均值作为该组的重心,记为P_i。

(2) 增量数据分组。对滑动窗口内的增量数据按照C'进行分组。对于属于$\{g_{i1}\}$的候选数据,直接将分组后的增量数据归并到对应$\{g_{i1}\}$的数据存储空间中即可,对属于$\{g_{i2}\}$的候选数据,比较该数据与分组重心P_i,将增量数据归并到距离最近的g_{i2}数据存储空间中。

(3) 历史样本替换。经过上面步骤后,样本容量增大,查询的时间性能受到影响。当查询时间性能不满足约束要求时,可采用样本替换的方法,即在g_i中通过增量数据随机替换历史样本的方法达到层内数据的随机取样。常用的样本替换方法包括蓄水池抽样[23]等,该方法常见于流计算窗口数据规约。对于已经存储在各组数据存储空间的增量数据,在分组g_i内通过蓄水池抽样方法随机替换各组内部的历史样本数据$\{R_i'|i\in[1,n]\}$,蓄水池大小为g_i数据量大小,以此保证层内样本的随机性。

(4) 更新结果评价。依据一定更新频率,对查询结果进行评估。如果查询误差比较稳定,则表示原有聚类结果依然有效,如果查询误差有增大趋势,则需要采用2.2节的方法重新进行样本生成。

3 验证与评价

本节通过实验对两阶段抽样方法的先进性进行了验证,包括本文方法中使用SOM算法与K-means、DBSCAN等聚类算法时近似聚合查询误差对比;本文方法与随机抽样算法、分层随机抽样算法以及国会抽样算法的比较;SOM算法中网络大小参数影响分析等。

3.1 实验环境

近似聚合查询实验系统如图1所示。该系统主要由查询管理模块和样本管理模块组成。查询管理模块包括预处理查询请求、执行查询请求以及对近似聚合查询后的结果进行评价等功能。样本管理模块包括样本量估计、样本生成以及增量样本维护等功能,该模块还存储历史全量数据、样本数据、增量数据以及查询结果日志等相关数据。

实验所用的服务器配置为4核Inter i5 CPU,主频3.10 GHz,32 GB内存,1.2 TB SAS硬盘,实验所用的数据集包括2类真实数据:一类数据集为公开环境数据集OpenAQ[24],该数据集包括2015—2018年,每天从67个国家的数万个地点收集的空气质量量测(如PM2.5、二氧化硫等)数据,在其中随机选取了1 000多万条数据。

另一类为国家电网实测的电能质量监测数据集,该数据集包括2017年从近万个监测点采集的电能质量量测(如基波电压,各次谐波电流等)数据,从中随机抽取了1 000多万条数据,其主要数据结构为(monitor_id, index, timestamp, value)。实验使用的聚合查询为:(1)select locationId, avg(value) from airQuality group by locationId;(2) select monitor_id, index, avg(value) from powerQuality group by monitor_id,index。

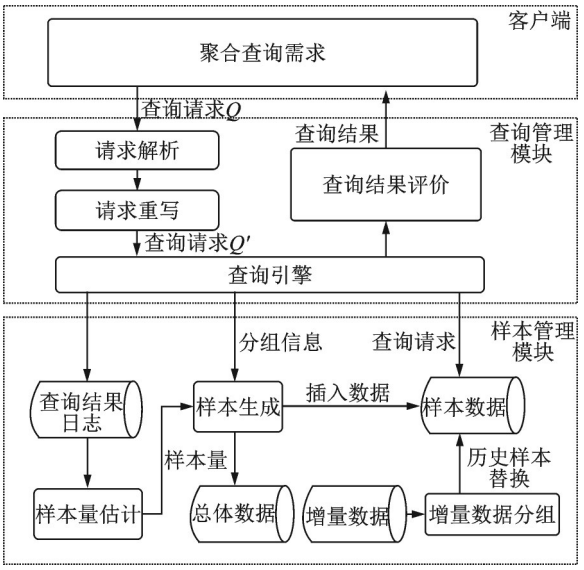


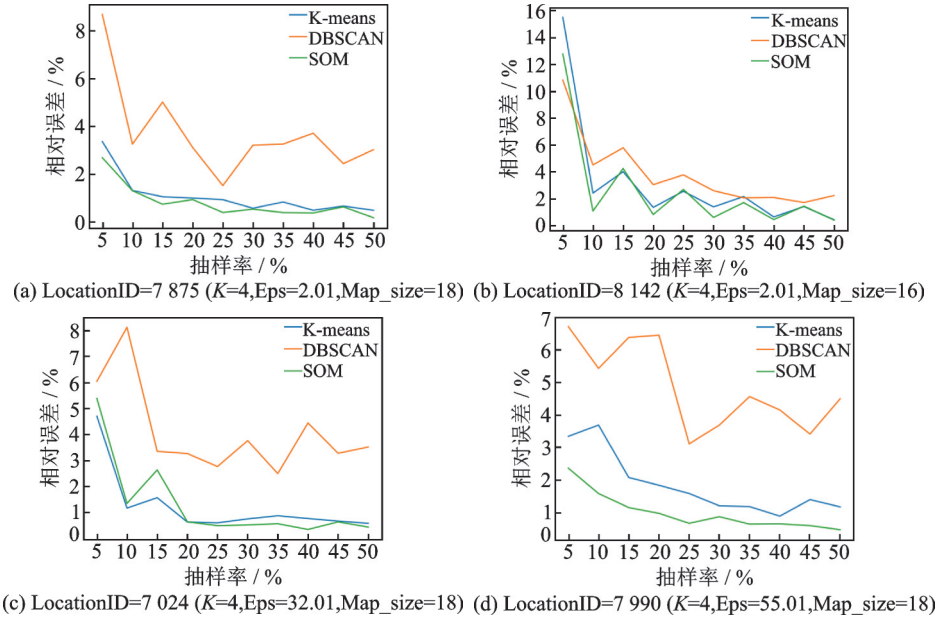
图1 系统结构

Fig.1 System architecture

3.2　聚合查询结果比较

分别将K-means、DBSCAN、SOM等3种聚类方法应用到2.2节的步骤2中,比较近似聚合查询精度差异。为了观察整体趋势,将抽样率也作为实验参数。共设置了5%～50%,步长为5%的抽样率参数。3类聚类算法用到的参数已调优,具体数值见图2。其中K是K-means中聚类个数,Eps是DBSCAN用到的邻域半径阈值,Map_size是SOM的网络边长大小。如无特殊说明,后文中性能实验都进行了10次,实验结果为10次结果的平均值。相对误差采用式(1)和式(2)计算。

图2是随机挑选的4个LocationID聚合查询的精度结果。绿色线条代表的SOM聚类算法在所有抽样率下都处于性能较优的位置,且误差整体比较稳定;蓝色线条代表的K-means聚类算法比SOM算法略差,而DBSCAN算法性能较差,且超过一定抽样率后出现误差上升的情形。



(a) LocationID=7 875 (K=4,Eps=2.01,Map_size=18)　(b) LocationID=8 142 (K=4,Eps=2.01,Map_size=16)

(c) LocationID=7 024 (K=4,Eps=32.01,Map_size=18)　(d) LocationID=7 990 (K=4,Eps=55.01,Map_size=18)

图2　聚合查询精度比较

Fig.2　Comparison of aggregation query precision

为了进一步验证上述结论,又使用电能质量数据集进行了实验,该数据集核密度分布如图3所示,仍然采用比例抽样方式(5%～50%,增幅5%)。使用3种算法的聚合查询精度如图4所示。从图4可见,DBSCAN的初始误差与SOM差不多,随着抽样比例的增加,呈现出快速上升趋势。而K-means算法表现较差。分析上述现象的原因,是因为电网数据集不仅数据倾斜,还存在较大数值的离群点,DBSCAN和SOM这两种聚类算法具有清除噪点的功能,在聚类运算时会筛选出噪点并排除在外,而K-means会对所有数据进行聚类。

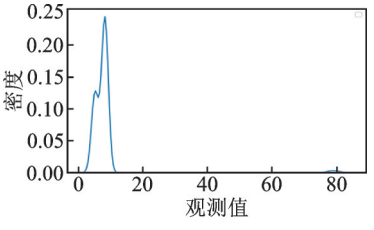


图3　电能质量数据分布

Fig.3　Example of power quality data distribution

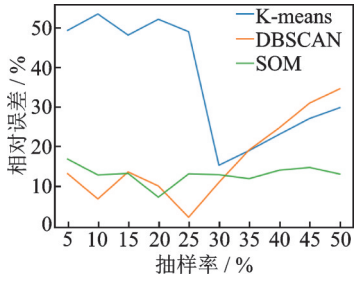


图4　聚合查询精度

Fig.4　Accuracy of aggregation query

3.3 与基准算法的比较

实验步骤如下。根据2.2节步骤(1,2)提取随机查询误差较高的分组。该数据集共217个分组,设定相对误差阈值为10%,共提取出19个误差较高的分组。从这些分组中随机选取了4个LocationId,图5是其监测数值的核密度分布图,可以发现这些数据都存在明显的数据倾斜特性。

针对上述分组,使用SOM聚类算法对数据值进行2次分层抽样,并计算查询相对误差,得出本文方法的查询精度。使用随机抽样、分层随机抽样和国会抽样等基准算法分别进行聚合查询,计算查询相对误差,结果如表4所示。

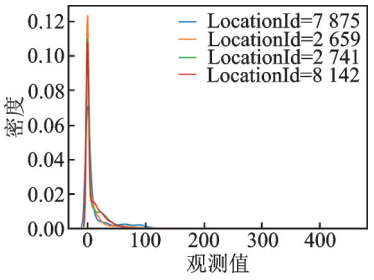


图5 OpenAQ数据分布样例

Fig.5 Example of OpenAQ data distribution

表4 聚合查询精度

Table 4 Precision of aggregate query %

LocationId	随机抽样	分层随机抽样	国会抽样	本文方法
2 536	34.333	27.462	27.404	5.234
2 659	17.263	19.128	15.105	4.224
2 728	14.999	13.077	13.011	1.131
2 741	18.014	17.127	16.011	2.002
4 212	9.325	6.904	8.336	0.649
45 008	9.674	6.471	8.222	6.736
63 094	15.495	14.359	12.706	3.314
7 024	18.241	22.076	17.047	4.616
7 440	20.979	29.485	17.705	1.925
7 674	18.847	21.341	15.886	1.814
7 871	13.029	13.531	11.063	4.204
7 875	8.497	14.532	7.157	2.517
7 983	23.862	15.076	20.775	6.812
7 986	22.578	20.565	19.642	2.526
7 988	28.526	31.882	25.586	2.473
7 989	12.871	31.161	10.677	2.562
7 990	15.291	12.337	13.381	5.754
8 142	15.012	36.061	12.371	7.892
8 172	19.496	10.735	16.951	2.526
误差ϵ	17.702	19.121	15.212	3.627

观察可见,在抽样率为5%情形下,本文方法的相对误差相比基准算法有12%~15%的性能提升。进一步地,设置5%~50%、步长为5%的抽样率,分别进行上面的实验,得到如图6所示的对比结果。可以看出,本文方法相较于基准算法都有一定性能提升,且抽样率越低,性能提升越显著。

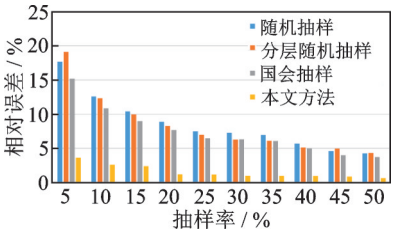


图6 不同抽样率下算法误差对比

Fig.6 Comparison of relative errors

3.4　网络大小对聚合查询结果的影响

SOM聚类算法中输出层网络的神经元数量对于聚类效果影响较大,在本方法中自然会影响查询精度,本实验使用二维输出层网络,在实验数据集上分别选择网络边长Size从10到70进行聚类,以观察其趋势。依然将抽样率也作为参数进行设置,共设置了5%～50%、步长为5%的抽样率参数,结果如图7所示。可以看到,网络边长Size为10和20时,相对误差较大,但当size增加到30以上相对误差明显变小,且趋于稳定。可见网络的大小对于最终聚类查询结果的误差有一定影响,但是当网络大小超过一定阈值后,该聚类查询的误差变化不大,趋于稳定。

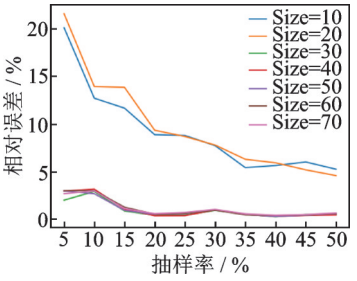


图7　不同网络边长聚合查询精度

Fig.7　Aggregate query accuracy with different sizes

3.5　样本生成时间

样本生成采用了预计算的方式,在查询前已经完成,不会影响聚合查询的时间性能,但预计算需要占用和耗费资源,故有必要评估其时间成本。分别采用不同规模数据、不同抽样率(抽样率从5%到20%,增幅5%)进行实验,记录本文方法、分层随机抽样以及简单随机抽样生成样本的时间,结果如图8所示(未考虑从数据库中装载原始数据以及将样本数据保存到样本表中的时间)。可以看出,本文的样本生成方法所花费的时间要明显高于随机抽样以及分层随机抽样。从总抽样时间角度进行比较,由于加载上述规模原始数据和保存样本数据的时间在分钟级,本文方法虽然依然劣于基准方法,但差距已不明显。

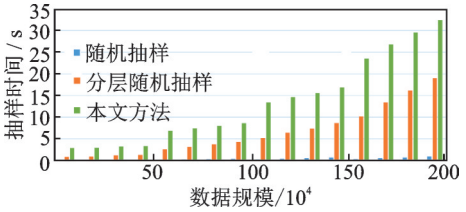


图8　样本生成时间开销

Fig.8　Sampling time

综上实验结果可得,本文提出的两阶段分层抽样方法相较于随机抽样、分层随机抽样以及国会抽样等基准方法具有较好的近似查询精度。实验所选取的聚类方法中,总体来看SOM效果较好,查询精度较高,不同抽样结果较稳定。K-means方法受异常点影响较大,DBSCAN方法参数设置困难,且稳定性较差。两阶段分层抽样所花费的时间比基准方法要多一些,但作为一种预计算方案,样本抽样时间有所上升是可以接受的。

4　结束语

近似查询技术能够在一定精度范围内快速返回查询结果。随着数据规模的不断增大,查询性能问题愈加突出,近似查询逐渐成为解决交互式数据查询与决策分析的主要手段。本文在已有的基于数据抽样的近似查询相关研究基础上,提出一种两阶段分层抽样方法及聚合查询实现架构,该方法对分组数据使用SOM算法进行值聚类,期望获得比随机抽样更好的近似查询精度。基于真实数据上的多个实验验证了该方法的有效性和先进性。未来的工作包括样本增量更新方法完善、人工智能方法的引入等。

参考文献:

[1]　WANG Tongxun, LI Yaqiong, DENG Zhanfeng, et al. Implementation of state-wide power quality monitoring and analysis system in China[C]//Proceedings of Power and Energy Society General Meeting. Porland: IEEE, 2018: 1-5.

[2]　盛家, 房俊, 郭晓乾, 等. 时序数据多维聚合查询服务的实现[J]. 重庆大学学报, 2020, 43(7): 121-128.
SHENG Jia, FANG Jun, GUO Xiaoqian, et al. Implementation of multidimensional aggregate query service for time series data[J]. Journal of Chongqing University, 2020, 43(7): 121-128.

[3]　LI Kaiyu, LI Guoliang. Approximate query processing: What is new and where to go[J]. Data Science and Engineering, 2018, 3(4): 379-397.

[4]　CHAUDHURI S, DING Bolin, KANDULA S. Approximate query processing: No silver bullet[C]//Proceedings of the 2017

ACM International Conference on Management of Data.[S.l.]: ACM, 2017: 511-519.
[5] FAN Wenfei. Making big data small[EB/OL]. (2019-05-08)[2020-11-01]. https://royalsocietypublishing. org/doi/10.1098/rspa.2019.0034.
[6] CORMODE G, GAROFALAKIS M, HAAS P J, et al. Synopses for massive data: Samples, histograms, wavelets, sketches [J]. Foundations and Trends in Databases, 2012, 4(1/2/3): 1-294.
[7] COCHRAN W G . Sampling techniques [M].3rd ed. Hoboken: John Wiley & Sons, 1977.
[8] ACHARYA S, GIBBONS P B, POOSALA V. Congressional samples for approximate answering of group-by queries[C]// Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data.[S.l.]: ACM, 2000: 487-498.
[9] PENG Jinglin, ZHANG Dongxiang, WANG Jiannan, et al. AQP++ connecting approximate query processing with aggregate precomputation for interactive analytics[C]//Proceedings of the 2018 International Conference on Management of Data. New York: ACM, 2018: 1477-1492.
[10] NGUYEN T D, SHIH M H, PARVATHANENI S S, et al. Random sampling for group-by queries[C]//Proceedings of 2020 IEEE 36th International Conference on Data Engineering (ICDE). [S.l.]: IEEE, 2020: 541-552.
[11] 谢金星, 李晖, 陈梅, 等. CSSAQP: 一种基于聚类的分层抽样近似查询处理算法[J]. 计算机与数字工程, 2017, 45(6): 1121-1126.
XIE Jinxing, LI Hui, CHEN Mei, et al. CSSAQP:An approximate query algorithm based on clustering stratified samping[J]. Computer & Digital Engineering, 2017, 45(6): 1121-1126.
[12] ZHANG Meifan, WANG Hongzhi, LI Jianzhong, et al. SUM-optimal histograms for approximate query processing[J]. Knowledge and Information Systems, 2020, 62(8): 3155-3180.
[13] PARK Y, MOZAFARI B, SORENSON J, et al. Verdictdb: Universalizing approximate query processing[C]//Proceedings of the 2018 International Conference on Management of Data. New York: ACM, 2018: 1461-1476.
[14] THIRUMURUGANATHAN S, HASAN S, KOUDAS N, et al. Approximate query processing for data exploration using deep generative models[C]//Proceedings of 2020 IEEE 36th International Conference on Data Engineering (ICDE). [S.l.]: IEEE, 2020: 1309-1320.
[15] ZHANG Meifan, WANG Hongzhi. LAQP: Learning-based approximate query processing[J]. Information Sciences, 2021, 546: 1113-1134.
[16] SAVVA F, ANAGNOSTOPOULOS C, TRIANTAFILLOU P.ML-AQP: Query-driven approximate query processing based on machine learning[J]. CoRRabs, 2020. DOI:10.48550/arXiv.2003.06613.
[17] MA Qingzhi, TRIANTAFILLOU P. DBEST: Revisiting approximate query processing engines with machine learning models [C]//Proceedings of the 2019 International Conference on Management of Data.New York: ACM, 2019: 1553-1570.
[18] ZHANG Meifan, WANG Hongzhi. Approximate query processing for group-by queries based on conditional generative models [EB/OL]. (2021-01-01)[2021-08-20]. http;//zrxiv.org/abs/2101.02914.
[19] RÖSCH P, LEHNER W. Sample synopses for approximate answering of group-by queries[C]//Proceedings of the 12th International Conference on Extending Database Technology: Advances in Database Technology. New York: ACM, 2009: 403-414.
[20] 金勇进. 抽样: 理论与应用[M]. 北京: 高等教育出版社, 2016.
JIN Yongjin. Sampling: Theory and application[M]. Beijing: Higher Education Press, 2016.
[21] DALENIUS T, HODGES J R J L. Minimum variance stratification[J]. Journal of the American Statistical Association, 1959, 54(285): 88-101.
[22] XU Dongkuan, TIAN Yingjie. A comprehensive survey of clustering algorithms[J]. Annals of Data Science, 2015, 2(2): 165-193.
[23] VAN HULLE M M. Self-organizing maps[J]. Handbook of Natural Computing, 2012(1): 585-622.
[24] GitHub, Inc.OpenAQ data[EB/OL].(2018-08-02)[2020-11-01]. http://openaq.org.

作者简介:

房俊(1976-),**通信作者**,男,博士,副研究员,研究方向:大数据查询处理、行业大数据分析、服务计算,E-mail: fangjun@ncut. edu. cn。

赵博(1996-),男,硕士研究生,研究方向:交互式数据查询,E-mail:1422108622@qq.com。

左昌麒(1999-),男,本科生,研究方向:大数据查询处理,E-mail:1192936732@qq.com。

(编辑:刘彦东)