文章编号：2095-2163(2021)03-0061-06 中图分类号：TP391.41 文献标志码：A

大数据近似分析方法综述

张美范，王宏志
（哈尔滨工业大学 计算机科学技术学院，哈尔滨 150001）

摘 要：大数据分析旨在从大量复杂的数据中获取价值。查询驱动的数据分析是大数据分析中最主要的部分。由于数据量的庞大，在大数据上获取准确的分析结果将带来极大的存储和计算代价。为解决这一困难，大数据近似分析方法应运而生。本文将主要针对大数据近似分析中的频率估计问题、近似查询处理问题、查询选择性估计问题近十年的解决方法进行总结和归纳。不同于以往以数据库为主视角的分析方法的总结，本文中将涵盖近几年应用或结合机器学习方法来处理上述问题的新方法。

关键词：大数据分析；频率估计；近似查询处理；查询选择性估计

A review of big data approximate analytics

ZHANG Meifan, WANG Hongzhi
(School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China)

【Abstract】 Big data analytics aims to obtain value from a large amount of complex data. Query-driven data analytics is the most important part of big data analytics. Due to the huge amount of data, obtaining accurate analysis results on big data will bring great storage and calculation costs. To solve this problem, big data approximate analysis methods came into being. This article mainly summarizes the frequency estimation methods, approximate query processing methods, and query selectivity estimation methods in big data analytics in the past ten years. Different from the previous summary of analysis methods based on the database perspective only, this article will cover the new methods of applying or combining machine learning methods to deal with the above problems in recent years.

【Key words】 big data analysis; frequency estimation; approximate query processing; query selectivity estimation

0 引 言

大数据中蕴含着海量的信息和巨大的价值，然而数据的庞大复杂使得人们不能或难以直接从数据中获得有价值的信息。大数据分析就是从大量复杂的数据中有目标地获取价值的过程。传统的数据分析方法难以应对极速增长的数据量，满足快速响应的需求。为解决这一问题，一系列近似方法应运而生，本文主要针对近些年大数据近似分析中频率估计、近似查询处理、查询选择性估计这三种基础分析任务的方法进行了总结。随着机器学习、人工智能领域的不断发展，近些年研究者们尝试将机器学习方法和大数据分析相结合，利用机器学习模型的推理预测能力，提高大数据分析方法的性能。

本文归纳并总结了近些年有代表性的大数据近似分析方法，同时涵盖了近些年将机器学习方法应用到大数据分析领域的新方法。

1 大数据频率估计

数据频率是数据最基本的统计量，同时也是网络监控、异常检测中的重要指标。然而，在大规模数据上统计准确的数据频率会占用极大的空间，为此，研究者们提出通过亚线性空间的草图来近似存储和估计数据频率。数据草图具有体积小、精度高、查询高效的特征，被广泛应用于频率估计和流数据处理上。此外，一些研究者致力于频繁元素查找及频率估计方法，频繁元素查找对应大数据管理中十分重要的 Top-k 查询，同时频繁元素频率在网络安全监控中常作为某些异常事件的衡量指标。

1.1 基于数据草图的频率估计

Count-Min(CM)[1]是最广泛应用的草图，其结构为一个二维数组，二维数组每行对应一个哈希函数。CM 通过哈希函数将数据映射到每行的相应位置，并增加这些位置的计数器来记录数据的频数。基于 CM 的某元素频率估计值为该元素通过每个哈

基金项目：国家重点研发计划(2018YFB1004700)；国家自然科学基金(U1866602, 61602129, 61772157)。

作者简介：张美范(1992-)，女，博士研究生，主要研究方向：大数据分析、大数据质量；王宏志(1978-)，男，博士，教授，博士生导师，主要研究方向：大数据管理与分析、数据库、机器学习等。

收稿日期：2020-10-08

希函数映射到位置中的计数值中的最小值,根据 CM 的结构特征,其估计结果为过量估计。CM 草图同时支持插入和删除操作。CU 草图[2]和 CM 草图的结构相同,不同之处在于 CU 草图在每次插入时只增加哈希映射位置中计数值最小的计数器的值,CU 相比于 CM 草图准确性更高,但是并不支持删除操作。

以单个二维数组为结构的 CM 和 CU 草图中的估计误差来自于哈希冲突。当不同数据被映射到同一位置,其频数在计数器中累加,造成极大的估计误差。这种冲突可以通过增加二维数组的大小来缓解,然而,这样会增大草图的空间代价,同时降低查询效率。为提高准确性和查询效率,多层结构的草图被提出。ASketch[3]在 CM 的基础上增加了一个过滤器,用于存储高频数据的频数,将高频数据和低频数据分开处理,提高了高频数据频率估计的准确性。SF-sketch [4]建立了 2 层分别称为 Slim 层和 Fat 层的结构,Fat 层为 CM 草图,Slim 层为一个尺寸更小的草图,每次插入时根据 Fat 层的观察结果更新 Slim 层,只有当 Slim 层中的计数小于 Fat 层估计值时才在计数器上加 1。由于 SF-sketch 的查询只在尺寸小的 Slim 层上进行,因此查询效率很高。由于数据频率分布不均匀,数据中大部分数据为低频数据,少量数据频率极高,因此,相同的计数器位数会造成极大的空间浪费或不准确的频率估计。为此,研究者们提出了可调节位数的计数器。Pyramid Sketch[5]为一种金字塔形状的多层结构草图,主要思想是随着频数的增长增加计数器的位数,采用共享计数器高位技术降低空间代价。ABC [6]在计数器位数溢出时通过向相邻的计数器借用位数的方法利用大量的小计数器来完成频率估计。

这里将近些年有代表性的频率估计草图的结构和支持的操作做了全面的总结详见表 1,这些草图能够估计存在于数据集或不存在于数据集中任意元素的频率,且为避免错过频繁元素,这些草图的估计误差均为单向误差,即估计结果不小于真实结果。

表 1 频率估计草图及特征总结

Tab. 1 Frequency estimation sketches and the characteristics

方法	结构	支持操作
CM[1]	单层:二维数组	插入、删除
CU[2]	单层:二维数组	插入
ASketch[3]	双层:Filter + CM	插入、删除
SF-Sketch[4]	双层:Slim 层 + Fat 层	插入、删除
Pyramid Sketch[5]	多层金字塔:计数器高位共享	插入、删除

1.2 基于计数器的频繁数据检测及频率估计

频繁数据检测在数据管理、推荐系统、网络安全监控等领域都有重要意义。为此,一些研究者将关注点放在频繁热点数据上,提出了一些只针对高频热点数据的频率估计方法,不同于数据草图能够估计所有数据的频率,这些方法只能估计频繁元素的频率,但这些方法对频繁元素估计的准确性和效率通常优于数据草图对频繁数据的估计。

其中,Space-Saving[7]是最有代表性的方法,通过有限个计数器来寻找 Top-k 个频繁数据并估计其频率,当新元素到来且能够存储的元素已满时将新元素和最小计数器对应的元素交换并增加计数器中频数。该方法能够在 O(1)时间内完成插入和查询。由 Space-Saving 衍生出 Compact Space-Saving (CSS)[8]、Scoreboard Space-Saving(SSS)[9]等变种。CCS 提出了一种较 Space-Saving 更紧凑的结构,能够避免使用大量的指针,节省空间。SSS 利用计数布隆过滤器预测一个数据是否为频繁数据,根据预测值将数据分为高频数据、潜在高频和低频三种,并只向 Space-Saving 中存储高频数据,避免了将低频数据存入 Space-Saving 结构引起的误差。

1.3 基于机器学习模型的频率估计

近些年,研究者们尝试将机器学习方法和数据库领域相结合,利用机器学习模型的推理、预测能力改进数据分析方法。文献[10]提出了一种基于机器学习模型改进的布隆过滤器,是在传统布隆过滤器的基础上增加了一个机器学习模型来预测数据是否存在在集合中,对于被预测不在集合中的元素通过插入传统的布隆分类器进行判断。以此为启发,研究者们将机器学习模型和数据草图相结合提出了学习草图。文献[11]利用机器学习模型分类高频和低频数据,为高频数据分配单独的存储单位,将低频数据存入 CM 草图中。文献[12]利用机器学习模型从历史数据中学习高频数据的频率,利用 CM 草图估计低频数据的频率,能够有效提高轻量级草图的准确性。

2 大数据近似查询处理

大数据近似查询处理的目标在于高效获取查询目标的近似结果。研究将大数据近似查询处理的方法分为在线查询和线下查询两类。其中,线上近似查询处理主要基于在线样本进行,线下近似查询处理则基于线下样本、直方图、草图等数据概要进行。近年来,研究者们将机器学习方法应用到大数据的近似查询处理中,以提高大数据近似查询处理的性能。本文将近些年有代表性的大数据近似查询处理

方法分类总结如下。

2.1　基于在线样本的近似查询处理

本文将在查询时抽样的近似查询处理和在线聚集都归纳为基于在线样本的近似查询处理。在查询时抽样的近似查询处理方法在执行查询时将抽样加入查询计划中，并根据样本上的查询处理结果估计整体数据上的查询处理结果。在线样本需针对每一个查询建立，且无需预先获取数据分布等先验知识。文献[13]中对查询时抽样的方法进行了总结。

在线聚集起源于文献[14]，通过增加样本量逐步提高结果准确性，当结果准确性满足需求后，可以提前中止查询。Join[15]通过在基于连接关系建立的连接图上随机游走的方式获取多表样本进行在线聚集。此外，文献[13,16]中介绍了多种在线聚集方法。

2.2　基于线下数据概要的近似查询处理

线下数据概要形式主要包括线下样本、直方图、草图、小波、数据方块或预查询等。直方图、草图、小波多用于查询选择性估计，且近些年基于这些方法进行近似查询处理的研究不多。因此，本节着重介绍基于线下样本、数据方块和预聚集查询的近似查询处理。

线下样本可基于数据分布或统计信息建立，相较于线上样本获取耗时更多、准确性更高，且可以存储用于多个查询。通常简单的随机抽样方法并不能够为均匀分布以外的数据分布提供高质量的样本。为提高估计结果的准确性，近似查询处理系统如BlinkDB[17]、VerdictDB[18]均应用了分层抽样的方法。

文献[19]提供了一种基于学习的分层抽样方法，是从样本中学习分类器用于评价数据元组对复杂查询的贡献得分，并根据与预测得分相关的概率进行分层抽样。该研究在利用机器学习方法提高复杂查询执行效率的同时，能够和抽样方法一样为结果提供置信区间。文献[20]提出了一种通过深度生成模型学习数据分布生成样本来代替传统抽样方法进行近似查询的方法，该方法在模型训练完毕后能够实现不接触原数据进行采样，从而避免从大数据中采样的代价，提高采样效率。

数据方块或预查询通过存储预先计算的特定范围的聚集查询结果来估计未来的查询结果。文献[21]提出了一种方法，将查询结果视为变量，从而根据旧查询估计新查询，该方法能够以低误差估计稀有数据。AQP++[22]将抽样与数据方块相结合，根据预计算的聚集查询结果和由抽样估计的新旧查询的差值来估计新查询的结果。

2.3　结合机器学习模型的近似查询处理

研究中将结合机器学习模型的近似查询处理方法分为 2 类。第一类是数据驱动的机器学习模型，第二类是查询驱动的机器学习模型。

数据驱动的机器学习模型在于通过历史数据或样本数据模拟数据分布或数据之间的关系。上述基于机器模型获取样本的方法[19-20]均可归类于数据驱动的机器学习模型。此外，DBEst[23]通过样本数据建立密度模型和回归模型进行近似查询处理，但是并不能像抽样方法一样提供估计结果的置信区间。DeepDB[24]通过和积网络模型模拟数据分布概率模型。EntropyDB[25]基于最大熵模型建立数据摘要，通过在模型上进行概率推断来回答查询。查询驱动的机器学习模型在于模拟历史查询中查询和结果之间的关系。ML-AQP[26]不需要接触数据或数据样本，仅根据历史数据建立模型，能够高效估计查询结果。

文中将近些年有代表性的基于机器学习模型的近似查询处理方法及其用到的模型类别、是否提供置信区间、是否需要访问数据和历史查询这四方面特征做了总结，参见表 2。

表 2　基于机器学习模型的近似查询处理方法

Tab. 2　Approximate query processing based on machine learning models

方法	模型	置信区间	访问数据	历史查询
学习抽样[19]	分类模型	√	√	×
生成样本[20]	生成模型:变分自编码器	×	√	×
DBEst[23]	核密度估计+回归模型	×	√	×
DeepDB[24]	概率模型:关系和积网络	√	√	×
EntropyDB[25]	概率模型:最大熵模型	√	√	×
ML-AQP[26]	回归模型	×	×	√

3　大数据查询选择性估计

查询选择性(基数)是指满足查询谓词的元组占整体数据的比例。查询选择性估计是查询优化过程中的必要环节，查询选择性估计的准确性将影响查询计划的效率。综述[27]中介绍了基于样本、直方图、小波、草图等数据概要进行查询选择性估计的方法。直方图和抽样是近些年的查询选择性估计的主要手段，此外，近些年也提出了一些基于机器学习的新方法。本文将近些年有代表性的大数据查询选择性估计方法分类总结如下。

3.1 基于直方图的查询选择性估计

直方图是查询选择性估计最常用的手段。直方图将数据分成多个桶并存储每个桶的边界和桶内的统计信息。直方图根据不同的划分方式分为等宽直方图、等高直方图、V-optimal 直方图等等。其中,等宽直方图将数据范围等分为若干份,等高直方图每个桶内的数据量相同,V-optimal 直方图的数据分布和原始数据分布之间的 L_2 距离最接近。直方图用于近似查询处理的优势在于查询效率高;但其不足在于仅能应用于数值类型的数据,且多维数据获得合适的划分方式的难度大。一维等宽直方图、等高直方图、V-optimal 直方图等在综述[27]中均有介绍。多维直方图的构建相对于一维直方图更为复杂,原因在于随着维度的增长,划分数据的自由度增加了,不同于一维直方图只需确定一个维度上桶边界的位置,多维直方图需确定多个维度上的划分位置、数量以及不同维度的处理顺序。不同于数据驱动的直方图,查询驱动的直方图通过负载中的查询反馈自适应地建立直方图[28-29]。查询驱动的直方图不需要根据数据建立,提高了构建效率,同时对负载中查询范围内的查询的准确性较高,但无法准确估计负载查询范围以外的查询。

3.2 基于抽样的查询选择性估计

抽样作为处理大数据分析任务的重要手段之一,也被应用在查询选择性估计及连接结果大小估计问题上。数据表的连接结果大小估计可以视为一种特殊的选择性估计,其估计的是多表根据相同属性连接后得到的表的大小。文献[30-32]基于抽样对连接结果的大小进行估计。核密度估计作为一种基于样本的估计方法被许多研究者采用,文献[33]提出了一种 GPU 加速的核密度估计模型,能够自适应地处理数据及查询的变化。文献[34]提出了一种将抽样和数据概要相结合估计合取查询选择性的方法。文献[35]通过实验评估了基于抽样和基于直方图的空间大数据查询选择性估计方法。

3.3 基于机器学习的查询选择性估计

本文总结了近些年将机器学习方法应用到查询选择性估计中的有代表性的工作[36-39]。用到的机器学习的方法主要分为有监督学习和无监督学习两种。其中,有监督学习通常需要提前收集查询结果作为训练数据,无监督的方法通常需要从数据本身中学习概率分布模型。对此可做研究阐释如下。

(1)有监督学习模型:QuickSel[36]是一个查询驱动的查询选择性学习框架,相比于查询驱动的直方图效率更高。文献[37]提出了一种基于多集卷积网络(MSCN)的基数估计算法,能够解决没有样本满足查询谓词的问题,显著提高估计质量。

(2)无监督学习模型:文献[38]提出了一种基于 MADE 模型无监督学习样本的联合概率分布的方法以及一个有监督的从查询负载中获得的回归模型。文献[39]提出了一种被称为 Naru 的基于无监督深度自回归模型的查询选择性估计方法。无监督的学习方法直接从数据中学习模型,不需要像有监督的模型一样收集大量的查询结果用于训练,因此获得模型的效率更高。

4 结束语

大数据分析发展迅速,研究成果日新月异。本文对近些年大数据分析中频率估计、近似查询处理、查询选择性估计这三种重要的近似分析任务进行了归纳总结。

参考文献

[1] CORMODE G, MUTHUKRISHNAN S . An improved data stream summary: The count-min sketch and its applications[J]. J Algorithms,2005, 55(1):58-75.

[2] ESTAN C, VARGHESE G. New directions in traffc measurement and accounting: Focusing on the elephants, ignoring the mice[J]. ACM Trans. Comput. Syst., 2003,21(3):270-313.

[3] ROY P, KHAN A, ALONSO G. Augmented sketch: Faster and more accurate stream processing[M]//ÖZCAN F, KOUTRIKA G, MADDEN S. Proceedings of the 2016 International Conference on Management of Data, SIGMOD Conference 2016. San Francisco, CA, USA:ACM, 2016:1449-1463.

[4] LIU Lingtong, SHEN Yulong, YAN Yibo ,et al. Sf-sketch: A two-stage sketch for data streams[J]. IEEE Trans. Parallel Distrib. Syst., 2020,31(10):2263-2276.

[5] YANG Tong, ZHOU Yang, JIN Hao, et al. Pyramid sketch: A sketch framework for frequency estimation of data streams[J]. Proc. VLDB Endow., 2017,10(11):1442-1453.

[6] GONG Junzhi, YANG Tong, ZHOU Yang, et al. ABC: A practicable sketch framework for non-uniform multisets[C]//NIE Jianyun, OBRADOVIC Z, SUZUMURA T, et al. 2017 IEEE International Conference on Big Data, BigData 2017. Boston, MA, USA: IEEE Computer Society,2017:2380-2389.

[7] METWALLY A, AGRAWAL D, ABBADI A El. An integrated effcient solution for computing frequent and top-k elements in data streams[J]. ACM Trans. Database Syst., 2006, 31(3):1095-1133.

[8] BEN-BASAT R, EINZIGER G, FRIEDMAN R, et al. Heavy hitters in streams and sliding windows[C]// 35th Annual IEEE International Conference on Computer Communications, INFOCOM 2016. San Francisco, CA, USA:IEEE, 2016:1-9.

[9] GONG Junzhi, TIAN Deyu, YANG Dongsheng, et al. SSS: An accurate and fast algorithm for fnding top-k hot items in data streams[C]// 2018 IEEE International Conference on Big Data

and Smart Computing, BigComp 2018. Shanghai, China: IEEE, 2018: 106-113.
[10] KRASKA T, BEUTEL A, CHI E H, et al. The case for learned index structures[M]// DAS G, JERMAINE C M, BERNSTEIN P A. Proceedings of the 2018 International Conference on Management of Data, SIGMOD Conference 2018. Houston, TX, USA: ACM, 2018: 489-504.
[11] HSU C Y, INDYK P, KATABI D, et al. Learning-based frequency estimation algorithms[C]// 7th International Conference on Learning Representations, ICLR 2019. New Orleans, LA, USA: OpenReview.net, 2019.
[12] ZHANG Meifan, WANG Hongzhi, LI Jianzhong, et al. Learned sketches for frequency estimation[J]. Inf. Sci., 2020, 507: 365-385.
[13] CHAUDHURI S, DING Bolin, KANDULA S. Approximate query processing: No silver bullet[M]//SALIHOGLU S, ZHOU Wenchao, CHIRKOVA R, et al. Proceedings of the 2017 ACM International Conference on Management of Data, SIGMOD Conference 2017. Chicago, IL, USA: ACM, 2017: 511-519.
[14] HELLERSTEIN J M, HAAS P J, WANG H J. Online aggregation [M]// PECKHAM J. SIGMOD 1997, Proceedings ACM SIGMOD International Conference on Management of Data. Tucson, Arizona, USA: ACM, 1997: 171-182.
[15] LI Feifei, WU Bin, YI Ke, et al. Wander join: Online aggregation via random walks[C]// ÖZCAN F, KOUTRIKA G, MADDEN S. Proceedings of the 2016 International Conference on Management of Data, SIGMOD Conference 2016. San Francisco, CA, USA: ACM, 2016: 615-629.
[16] LI Kaiyu, LI Guoliang. Approximate query processing: What is new and where to go? - A survey on approximate query processing [J]. Data Science and Engineering, 2018, 3(4): 379-397.
[17] AGARWAL S, MOZAFARI B, PANDA A, et al. Blinkdb: queries with bounded errors and bounded response times on very large data[C]// Eighth Eurosys Conference 2013, EuroSys '13. Prague, Czech Republic: dblp, 2013: 29-42.
[18] PARK Y, MOZAFARI B, SORENSON J, et al. Verdictdb: Universalizing approximate query processing [M]// DAS G, JERMAINE C M, BERNSTEIN P A. Proceedings of the 2018 International Conference on Management of Data, SIGMOD Conference 2018. Houston, TX, USA: ACM, 2018: 1461-1476.
[19] WALENZ B, SINTOS S, ROY S, et al. Learning to sample: Counting with complex queries[J]. Proc. VLDB Endow., 2019, 13(3): 390-402.
[20] THIRUMURUGANATHAN S, HASAN S, KOUDAS N, et al. Approximate query processing for data exploration using deep generative models [C]//36th IEEE International Conference on Data Engineering, ICDE 2020. Dallas, TX, USA: IEEE, 2020: 1309-1320.
[21] GALAKATOS A, CROTTY A, ZGRAGGEN E, et al. Revisiting reuse for approximate query processing[J]. Proc. VLDB Endow., 2017, 10(10): 1142-1153.
[22] 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, SIGMOD Conference 2018. Houston, TX, USA: ACM, 2018: 1477-1492.
[23] MA Qingzhi, TRIANTAFLOU P. Dbest: Revisiting approximate query processing engines with machine learning models [M]// BONCZ P A, MANEGOLD S, AILAMAKI A, et al. Proceedings of the 2019 International Conference on Management of Data, SIGMOD Conference 2019. Amsterdam, The Netherlands: ACM, 2019: 1553-1570.
[24] HILPRECHT B, SCHMIDT A, KULESSA M. Deepdb: Learn from data, not from queries! [J]. Proc. VLDB Endow., 2020, 13(7): 992-1005.
[25] ORR L J, BALAZINSKA M, SUCACMIU D. Entropydb: A probabilistic approach to approximate query processing[J]. VLDB J., 2020, 29(1): 539-567.
[26] SAVVA F, ANAGNOSTOPOULOS C, TRIANTAFLLOU P. ML-AQP: Query-driven approximate query processing based on machine learning[J]. CoRR, abs/2003.06613, 2020.
[27] CORMODE G, GAROFALAKIS M N, HAAS P J, et al. Synopses for massive data: Samples, histograms, wavelets, sketches. Foundations and Trends in Databases, 2012, 4(1-3): 1-294.
[28] BRUNO N, CHAUDHURI S, GRAVANO L. Stholes: A multidimensional workload-aware histogram[C]//Proceedings of the 2001 ACM SIGMOD International Conference on Management of Data. Santa Barbara, CA, USA: ACM, 2001: 211-222.
[29] KHACHATRYAN A, MÜLLER E, BÖHM K, et al. Improving accuracy and robustness of self-tuning histograms by subspace clustering [C]// 32nd IEEE International Conference on Data Engineering, ICDE 2016. Helsinki, Finland: IEEE, 2016: 1544-1545.
[30] VENGEROV D, MENCK A C, ZAÏT M, et al. Join size estimation subject to flter conditions[J]. Proc. VLDB Endow., 2015, 8(12): 1530-1541.
[31] CHEN Yu, YI Ke. Two-level sampling for join size estimation [M]// SALIHOGLU S, ZHOU Wenchao, CHIRKOVA R, et al. Proceedings of the 2017 ACM International Conference on Management of Data, SIGMOD Conference 2017. Chicago, IL, USA: ACM, 2017: 759-774.
[32] WANG Taining, CHAN C Y. Improved correlated sampling for join size estimation[C]//36th IEEE International Conference on Data Engineering(ICDE 2020). Dallas, TX, USA: IEEE, 2020: 325-336.
[33] HEIMEL M, KIEFER M, MARKL V. Self-tuning, gpu-accelerated kernel density models for multidimensional selectivity estimation[M]// SELLIS T K, DAVIDSON S B, IVES Z G. Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data. Melbourne, Victoria, Australia: ACM, 2015: 1477-1492.
[34] MÜLLER M, MOERKOTTE G, KOLB O. Improved selectivity estimation by combining knowledge from sampling and synopses [J]. Proc. VLDB Endow., 2018, 11(9): 1016-1028.
[35] CHASPARIS H, ELDAWY A. Experimental evaluation of selectivity estimation on big spatial data [M]//BOUROS P, SARWAT M. Proceedings of the Fourth International ACM Workshop on Managing and Mining Enriched Geo-Spatial Data. Chicago, IL, USA: ACM, 2017: 8:1-8:6.
[36] PARK Y, ZHONG Shucheng, MOZAFARI B. Quicksel: Quick selectivity learning with mixture models [M]// MAIER D, POTTINGER R, DOAN A, et al. Ngo, editors, Proceedings of the 2020 International Conference on Management of Data, SIGMOD Conference 2020. Portland, OR, USA: ACM, 2020: 1017-1033.

(下转第 69 页)