时空聚集计算研究进展*)

包 磊 秦小麟
(南京航空航天大学信息科学与技术学院 南京 210016)

摘 要 时空数据库要处理大量的数据。相对于单个时空数据来说,大量数据的聚集计算结果更有信息量。本文综述了时态聚集、空间聚集和时空聚集计算领域的研究现状,着重分析了各类时空聚集算法的研究进展。讨论了目前时空聚集计算存在的问题,并指出了今后的发展方向。

关键词 时空数据库,聚集查询,聚集函数

Research Progress in Spatiotemporal Aggregation Computation

BAO Lei QIN Xiao-Lin
(College of Information Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 210016)

Abstract Spatiotemporal databases need to process vast amounts of data. In such cases, generating aggregate information from the data set is more useful than individually analyzing every entry. In this paper, we study the most relevant techniques for the evaluation of aggregate queries on spatial, temporal, and spatiotemporal data. The research progress of spatio-temporal aggregation is summarized. The problems of current research are discussed and the future directions are pointed out.

Keywords Spatiotemporal database, Aggregate query, Aggregate function

聚集计算是数据库领域的重要研究内容之一。随着近年来时空数据库研究的逐渐深入,时空聚集计算成为该领域内的研究热点。时空数据库每日要处理海量的数据,从查询需求上看,用户一般不会对大量数据中的某个单个信息感兴趣,而是需要获得大量数据的总体变化趋势或者总体统计信息。例如,交通管制系统中,一般需要查询某个路口每小时内的车辆流量而不是某个车辆具体的位置。对于时空数据库来说,提供强大的聚集计算能力非常重要。

与传统关系数据库的聚集计算相比,时空聚集计算有其特殊性。传统关系数据聚集函数针对关系中的某个或者某几个显式属性,各个元组可以抽象为数据集合中的一个独立的点,而时空对象本身在时间和空间上占据一定的范围,时空聚集不仅发生在某些显式属性上,还会在对象所占据的时间或者空间区域上进行。由于时空信息互相依赖、互为存在条件以及时空的连续性、无限性等特点,时空聚集的计算比较困难,需要深入细致的研究,包括对时空聚集的准确描述和时空聚集算法的优化。

当前对时空聚集计算的研究已经展开,包括时态聚集计算、空间聚集计算和时空聚集计算研究,已有成果包括了对各类聚集函数的描述和分类以及聚集算法的研究。有关聚集算法的研究又包括对聚集算法的优化和聚集结果的近似计算等方面,其目的都是力图以较高的性能快速获得所需的计算结果。本文对时空聚集相关领域的研究成果进行介绍和分析,其中第1节介绍聚集函数描述和定义。第2、3、4节分别介绍时态聚集、空间聚集和时空聚集查询的描述、分类以及相关的算法实现。最后对目前时空聚集研究中存在的问题以及未来的发展方向进行了总结。

1 聚集函数及其形式化定义

数据库中的聚集计算一般通过聚集函数来实现。聚集函数考察数据库中的一组元组,概括分析其总体信息,并以一个返回值反映分析结果。SQL92标准规定了5个聚集函数,分别是Max,Min,AVG,Sum和Count。SQL99在此基础上增加了Every,Some和Any三个函数。SQL_OLAP中又规定了18个新的聚集函数[1]。

聚集函数的形式化定义为:

设有关系R,其模式为$\{A_1,A_2,\cdots,A_n\}$,A_i对应的定义域为D_i,可数集合$Agg=\{f_1,f_2,\cdots,f_n\}$是$R$的聚集函数集合,对任一$f_i\in Agg$,$f_i:D_1\times D_2\times\cdots\times D_n\rightarrow D_{agg}$,其中$D_{agg}$是聚集函数的值域。对于某聚集查询,首先通过谓词$SP$对$R$的元组列表$A$进行划分,再将相应聚集函数作用于划分元组$\gamma(A)$,最终获得结果,即:

$$Agg_{f_i,SP}(R)=\{s\circ f_i(\gamma(A))\mid s\in\pi_A(R)\}$$

在对某组元组应用聚集函数进行计算之前,一般首先对元组列表进行划分,然后对获得的划分执行聚集函数。划分方法包括分组(Group)、分区(Partition)和滑动窗口(Slide Window)等。

2 时态聚集计算

时态数据只在一段时间上有效,并且随时间持续变化。对于传统关系数据库,聚集函数应用于某组元组之上,针对某个或者某几个显式属性展开计算。对于时态数据库来说,时间域定义为由基本时间粒(granules)组成的全序集合,而聚集计算一般是时间域上从某个粒度向更粗粒度上的一个概括。

*)基金项目:本项目由国家自然科学基金资助(69973032),江苏省自然科学基金(BK2001045)。包 磊 讲师,博士研究生,研究方向为时空数据库,作战仿真系统;秦小麟 教授,博士生导师,研究方向为数据库技术、GIS。

例如,"计算今年每个职员的年平均工资"。

2.1 时态聚集算法研究

现有时态聚集算法可分为2类,基于时态索引的聚集计算和无索引聚集计算。基于索引的聚集算法预先对数据作局部聚集计算,并将结果存储在外存中,在需要进行聚集计算时,利用已存储的局部聚集结果快速计算出结果。无索引聚集算法通过对数据库的一次预扫描,在内存中生成一些可用于存储局部聚集结果的数据结构来进行计算。文[2]通过对时态关系的一次扫描,建立聚集树(aggregation tree),利用深度优先搜索法可快速求得聚集结果。聚集树不是平衡树,因此其节点的插入会大大影响算法整体性能。文[3]提出2-3树,利用叶节点存储聚集结果的时间信息,2-3树是平衡树,因此其搜索效率和调整效率较高,但是由于事先需要对各元组按时间排序,因此额外开销很大。文[4]提出的PA树以AVL树组织数据,存储时间戳和局部聚集值,在计算时,中序遍历PA树得到结果。文[5]提出利用meta数组将关系元组分为各个子集,采用分治法对各个子集进行聚集计算,最终结果通过对各子集结果的归并得到。文[6~8]还给出了一些并行时态聚集算法。基于时态索引的聚集计算采用存储于外存的数据结构来提高算法效率。文[9]提出SB树,树中存储了递增的局部聚集结果和其对应的时间段,聚集结果通过对SB树的深度优先搜索得到。在调整时,SB树利于新节点的插入,但是不利于节点的删除,因为对节点的删除会破坏聚集值的递增规律。与SB树不同,MVSB树[10]为一系列SB树,每个时间戳对应一个SB树,适合时态范围聚集计算的实现。文[11]基于B树和R树,给出了一种聚集结果近似计算算法,能够处理Sum,Count聚集函数的近似计算。

总体来说,非索引聚集算法需要预先对关系作一次预扫描,根据查询谓词在主存中生成相应数据结构,现有的算法都需要聚集值满足一定的递增规律,因此只适合某些有分布性的聚集函数。基于索引的时态聚集算法通过事先存储在外存中的索引来提高算法效率,不需要作预扫描,但是不考虑属性上的谓词操作,只能处理作用于单值数据上的聚集函数,如SQL92中所规定的5类聚集函数;聚集函数的最终结果都是通过对局部聚集结果的组织得到,对于无法通过局部结果组合得到的整体聚集函数也无法进行处理,如Median。

2.2 基于数据流的时态聚集算法研究

数据流是数据的有序序列。由于数据流中的每个数据都带有时间戳以记录该数据的产生、接收时间,数据流可看作某种时态数据。数据流在网络控制、通信传输、Web应用以及传感器网络等方面有较广泛的应用。由于数据流包含大量数据,一般来说数据流中的数据只使用一次就被丢弃或者存档,对数据流的聚集计算相对于某个单独数据的查询更加重要[26]。Datar等人提出了对数据流使用指数直方图来求取Count和Sum聚集函数的近似解的算法[12],Zhang提出依据粒度层次对数据流进行时态聚集的机制,基本思想是根据数据的已存储时间采用不同粒度作聚集计算,已存储时间越长,粒度越粗[13]。

3 空间聚集计算

空间聚集根据不同空间粒度对数据元组进行聚集计算,空间数据的聚集查询一般需要先根据对象的空间范围进行分类和划分得到某个元组集合,然后再在此元组集合上执行相应的聚集函数得到最终结果。典型的空间聚集查询的例子为"计算四川省内各个县的森林覆盖面积"。其中"四川省内"为空间谓词,"各个县"指需要按每个县的空间范围进行划分。

目前已有的空间聚集算法一般都是返回某个方形区域内的所有空间对象的聚集函数值,称为Box Aggregation。其算法都是通过空间索引来进行计算。文[14]指出空间对象的拓扑关系一般不能保证满足聚集函数的分布特性,事先存储的局部聚集值不能直接用于整体聚集结果的计算,必须依照一定的规则对空间对象进行分解,才能得到可用的预聚集结果。文[15]提出aR树在R树节点的最小外接矩形MBR内标注了其所包含的所有对象的聚集值,这样在聚集计算时,可以直接引用该信息,不用对每个对象进行聚集计算。文[16]对aR树进行了优化,给出了专门用于求取Min,Max聚集函数的MR树。对于Sum,Count,Avg等聚集函数,文[18]中使用了递增的聚集索引,包括用于内存的静态ECDF树和用于外存的ECDF-B树[17],针对ECDF树节点插入成本较高的缺点,提出的BA树可以有效地对节点数目进行调整。文[19]提出的MRA树适合于多维空间中点对象的聚集计算,以节点局部聚集结果对整体结果的不确定程度的影响因子作为该树的遍历顺序,在搜索的过程中可以不断修正结果。文[20]的aP树借用时态聚集算法处理二维空间中的点对象的聚集计算,其最大优点是遍历时间与对象数目无关。

对于空间聚集来说,目前已有的研究成果都是返回某个区域内的所有空间对象的聚集函数值(Box Aggregation),所支持的空间谓词有限,在利用空间谓词对元组进行划分之后,对象的空间信息将丢弃掉,不参与聚集计算,因此仅支持具有分布特性的聚集函数的计算。

4 时空聚集计算

时空对象同时在时间上空间上占据一定的区域,对象的显式属性和空间范围都会随时间不断变化,这种变化可以是离散的,也可能是连续不断的。时空聚集根据不同粒度对数据元组进行进行分类和划分得到某个元组集合,然后再在此元组集合上执行相应的聚集函数得到最终结果。而时空粒度是时间粒度和空间粒度的结合,如"每年每省","每小时每平方米"典型的时空聚集查询的例子为"计算10年内中国各省的年平均森林覆盖面积"。在此查询中,"10年内"为时态谓词,"中国境内"为空间谓词,"各省"指需要按省进行空间划分。

针对时空聚集算法的研究目前比较少,相关的研究开始于2002年,Zhang等人在其基于数据流的聚集算法基础上作了扩充,其基本思路是对时空对象进行降维分解,其各个分解部分通过ECDF树或者BA树来完成其聚集计算[18,21]。Papadias给出了一种基于索引的时空聚集算法[22],他将空间对象按所在区域进行分组,将各个静态区域以R树组织起来,每个区域对应的时态信息以一棵B树存储,最终获得的数据结构称为aRB树,对aRB树的进一步扩充还有aHRB树和a3DRB树,它们与aRB树的区别在于用于分组的区域可以是动态划分的。Papadias的aRB树无法处理在查询区域内同一时间段内重复出现的对象,这个问题被称为"Distinct counting problem"。为了解决这个问题,Tao对aRB树进行了修改[23],针对Count和Sum聚集函数,解决了该问题。Sun针对连续运动的对象,提供了聚集函数近似结果的计算方法[24],它将二维空间划分为w×w个网格单元,每个单元内存储了当前时间内包含的对象数目,通过其提出的数据结

构"适应性多维直方图"(Adaptive Multidimensional Histogram),可快速计算 Count 聚集函数。

时空聚集算法的研究工作集中在对空间聚集算法的扩充方面,因此其算法都是基于相应的时空索引来完成,在利用时空谓词对时空关系进行了划分之后,这些算法舍弃了参与聚集计算对象的时态信息和空间信息,因此才会发生"Distinct count"之类的问题。与对应的空间聚集算法类似,现有的时空聚集算法依赖于预聚集生成的局部聚集结果,只适合几种具有分布特性的聚集函数的计算。

5 存在问题和将来研究方向

本文在简要介绍时空聚集的基础上,对近年来有代表性的时态聚集、空间聚集和时空聚集的研究工作进行了重点介绍。在时空聚集研究领域中,单独的时态聚集方面的研究已经比较充分,提出了一些很有价值的算法。但是在空间聚集和时空聚集方面的研究还很不够,尤其是时空聚集的研究还处于"萌芽"阶段。

现有的时空聚集算法分为 2 类,基于索引的聚集算法和非索引聚集算法。非索引算法要求在聚集计算之前进行预聚集,获得局部聚集结果,基于索引的聚集算法利用事先存储的索引信息进行聚集计算。通过预聚集,聚集算法可以大大提高,但是这种依赖于预处理所获得的局部聚集值的算法无法处理 Median 等非分布的聚集函数。在文[25]中,Palpanas 在研究传统关系数据库中的聚集查询时注意到了这个问题,提出了适用于所有聚集函数的通用预聚集处理方法,但是其方法只可应用于传统关系数据库中。

现有的时空聚集算法还有一个主要缺点是,在利用时空谓词对关系进行划分之后,算法舍弃了对象的时空信息进行聚集计算,这样不仅会导致一些错误的发生,而且往往这些时空信息本身才是用户所关心的。例如"返回某暴风云团的最大时速,以及达到最大时速的具体时间和地点"。

通过总结近年来时空聚集的研究进展,我们认为以下几方面是未来研究发展的方向:① 支持各类聚集函数的统一聚集算法;② 支持各类时态谓词、空间谓词和时空谓词的时空聚集查询处理;③ 针对某类或者某个聚集查询的高效算法。

参 考 文 献

1 Melton J. Advanced SQL:1999. Understanding Object-Relational and Other Advanced Features. The Morgan Kaufman Series in Data Management Systems. Morgan Kaufmann Publishers, San Francisco, CA, 2003

2 Kline N, Snodgrass R T. Computing Temporal Aggregates. In: Proc. of the Intl. Conf. on Data Engineering, Taipei, Taiwan, March 1995. 222～231

3 Kline R N. Aggregation in Temporal Databases. PhD thesis, University of Arizona, Tucson, Arizona, May 1999

4 Kim J S, Kang S T, Kim M-H. On Temporal Aggregate Processing based on Time Points. Information Processing Letters, September, 1999, 71(5-6): 213～220

5 Moon B, Vega Lopez I F, Immanuel V. Scalable Algorithms for Large Temporal Aggregation. In: Proc. of the Intl. Conf. on Data Engineering, San Diego, CA, March 2000. 145～156

6 Moon B, Vega Lopez I F, Immanuel V. Efficient Algorithms for Large-Scale Temporal Aggregation. IEEE Transactions on Knowledge and Data Engineering, May-June 2003, 15(3): 744～751

7 Gendrano J A G, Huang B C, Rodrigue J M, et al. Parallel Algorithms for Computing Temporal Aggregates. In: Proc. of the Intl. Conf. on Data Engineering, Sydney, Australia, March 1999. 418～427

8 Ye X, Keane J A. Processing Temporal Aggregates in Parallel. In: IEEE International Conference on Systems, Man, and Cybernetics, Orlando, FL, Oct. 1997. 1373～1378

9 Yang J, Widom J. Incremental Computation and Maintenance of Temporal Aggregates. In: Proc. of the Intl. Conf. on Data Engineering, Heidelberg, Germany, April 2001. 51～60

10 Zhang D, Markowetz A, Tsotras V J, et al. Efficient Computation of Temporal Aggregates with Range Predicates. In: Proc. of the ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, Santa Barbara, CA, May 2001. 237～245

11 Tao Y, Papadias D, Faloutsos C. Approximate Temporal Aggregation. In: Proc. of the Intl. Conf. on Data Engineering, Boston, USA, March 30-April 2, 2004

12 Datar M, Gionis A, Indyk P, Motwani R. Maintaining Streams Statistics over Sliding Windows (Extended Abstract). In: Proc. of the annual ACM-SIAM Symposium on Discrete Algorithms, San Francisco, CA, January 2002. 635～644

13 Zhang D, Gunopulos D, Tsotras V J, Seeger B. Temporal Aggregation over Data Streams using Multiple Granularities. In: Proc. of the Conf. on Extending Database Technology, Prague, Czhech Republic, March 25-27 2002. 646～663

14 Pedersen T B, Tryfona N. Pre-aggregation in Spatial Data Warehouses. In: Proc. of the International Symposium on Advances in Spatial and Temporal Databases, Redondo Beach, CA, July 12-15 2001. 460～480

15 Papadias D, Kalnis P, Zhang J, Tao Y. Efficient OLAP Operations in Spatial Data Warehouses. In: Proc. of the Intl. Symposium on Advances in Spatial and Temporal Databases, Redondo Beach, CA, July 12-15 2001. 443～459

16 Zhang D, Tsotras V J. Improving Min/Max Aggregation Over Spatial Objects. In: Proc. of the ACM-GIS Conference, Atlanta, GA, November 2001. 88～93

17 Bentley J L. Multidimensional Divide-and-Conquer. Communications of the ACM, 1980, 23(4): 214～229

18 Zhang D, Tsotras V J, Gunopulos D. Efficient Aggregation Over Objects with Extent. In: Proc. of the ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, Madison, WI, June 2002. 121～132

19 Lazaridis I, Mehrotra S. Progressive Approximate Aggregate Queries with a Multi- Resolution Tree Structure. In: Proc. of the ACM-SIGMOD Conference, Santa Barbara, CA, May 2001. 401～412

20 Tao Y, Papadias D, Zhang J. Aggregate Processing of Planar Points. In: Proc. of the Conf. on Extending Database Technology, Prague, Czhech Republic, March 25-27 2002. 682～700

21 Zhang D. Aggregation Computation over Complex Objects: [PhD thesis]. University of California, Riverside, August 2002

22 Papadias D, Tao Y, Kalnis P, Zhang J. Indexing Spatio-Temporal Data Warehouses. In: Proc. of the Intl. Conf. on Data Engineering, San Jose, CA, February 26-March 1 2002. 166～175

23 Tao Y, Kollios G, Considine J, Li F, Papadias D. Spatio-Temporal Aggregation Using Sketches. In: Proc. of the Intl. Conf. on Data Engineering, Boston, USA, March 30-April 2, 2004

24 Sun J, Papadias D, Tao Y, Liu B. Querying about the Past, the Present, and the Future in Spatio-Temporal Databases. In: Proc. of the Intl. Conf. on Data Engineering, Boston, USA, March 30-April 2, 2004

25 Palpanas T, Sidle R, Cochrane R, Pirahesh H. Incremental Maintenance for Non-Distributive Aggregate Functions. In: Proc. of the VLDB Conf., Hong Kong, China, August 2002. 802～813

26 Qiao L, Agrawal D, Abbadi A E. RHist: Adaptive Summarization over Continuous Data Streams. In: Proc. of the ACM-CIKM Conference, McLean VA, November 2002. 469～476

27 Tao Y, Papadias D. The MV3R-Tree: A Spatio-Temporal Access Method for Timestamp and Interval Queries. In: Proc. of the VLDB Conf., Roma, Italy, Sep. 11-14, 2001. 431～440