一种面向大数据聚集分析的近似计算方法

An Approximate Calculation Method for Big Data Aggregation Analysis

顾冉

Ran Gu

北京柏睿数据技术股份有限公司　中国·北京 100020

Beijing Bairui Data Technology Co., Ltd., Beijing, 100020, China

摘　要：为促使大数据聚集分析过程更加人性化，提高其分析效率和效果，论文以“大数据聚集分析”为目标，提出一套行之有效的近似计算方法。首先，介绍大数据聚集分析相关技术。其次，从数据分割、渐进近似计算、聚集增量更新三个方面入手，完成对近似计算方法设计。最后，研究了近似计算方法优点。结果表明：在大数据聚集分析背景下，论文所提出的近似计算方法具有分析准确率高、分析速度高效、分析过程人性化等特点，完全符合电子商务、智能医疗等领域应用需求。希望通过这次研究，为相关人员提供有效的借鉴和参考。

Abstract: In order to make the process of big data aggregation analysis more humane, improve its analysis efficiency and effectiveness, this paper aims to “big data aggregation analysis” and proposes an effective approximate calculation method. Firstly, introduce the relevant technologies of big data aggregation analysis. Secondly, starting from three aspects: data segmentation, progressive approximation calculation, and aggregate incremental update, complete the design of approximation calculation methods. Finally, the advantages of approximate calculation methods were studied. The results indicate that in the context of big data aggregation analysis, the approximate calculation method proposed in this paper has the characteristics of high analysis accuracy, efficient analysis speed, and humanized analysis process, fully meeting the application needs of e-commerce, intelligent medical and other fields. I hope to provide effective reference and guidance for relevant personnel through this study.

关键词：大数据；聚集分析；近似计算；方法

Keywords: big data; aggregation analysis; approximate calculation; method

DOI: 10.12346/csai.v2i1.9109

1 引言

目前，在云计算、物联网、大数据等先进技术的不断推广和普及下，大量企业在实际管理和生产中，产生海量数据。企业要想在激烈市场竞争中立于不败之地，必须从这些海量数据中快速分析和筛选出有价值的数据，并为企业管理提出一套切实可行的决策和指导。而大数据聚集分析主要借助数据查询语句，对求和、求平均值、计数、标准差、分位数等企业数据指标的实时获取和整理，有效地提高企业重要数据分析结果的精确性和真实性。在大数据时代背景下，聚集分析操作除了分析多个数据表外，还需要连接操作多个数据表，容易导致多层嵌套查询问题的出现[1]。此时，如果使用传统分析方法，对聚集分析结果进行计算，会降低聚集分析效率和效果，同时，还会增加额外磁盘开销成本，为解决以上问题，相关人员要提出一套基于大数据聚集分析的近似计算方法。

2 相关技术概述

目前，比较常用的大数据聚集分析技术主要包含以下三种：①多维索引技术。应用该技术所构建 B- 树、网络完全满足数据分布需求。②分布式计算技术。应用该技术，所构建的基于 Hadoop 的分布式计算框架和基于 Spart 分布式计算框架，可以有效地提高大数据聚集分析速度和精确度。③

【作者简介】顾冉（1978-），男，中国北京人，硕士，工程师，从事大数据分析引擎、人工智能、隐私计算研究。

并行计算技术。应用该技术，可以对多核图形处理单元进行统一化处理。④物化视图技术。应用该技术，可以对所存储的海量中间结果进行精确化计算，使得大数据聚集分析质量和效率显著提升 [2]。

3 基于大数据聚集分析的近似计算方法

在大数据时代背景下，应用聚集分析技术，完成对如图 1 所示的近似计算流程设计，从图 1 中可以看出，近似计算流程主要包含数据分割、渐进近似计算、聚集增量更新三个环节，这些环节共同组合后，形成一套切实可行的近似计算方法 [3]。为保证大数据聚集分析结果的精确性和真实性，相关人员要严格按照近似计算流程，开展相关计算工作。

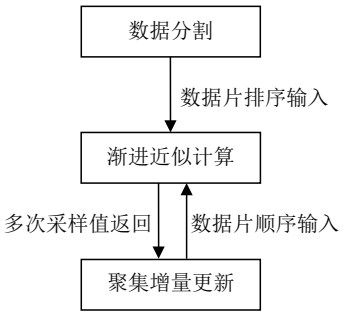

图 1 近似计算流程

3.1 数据分割

3.1.1 数据分割模块概述

数据分割模块主要是指对基础大数据顺序进行划分，使其划分为若干个数据片，这些数据片互不相交，然后，按照一定的比例，对不同数据片所对应的数据元组进行实时交换和处理 [4]。

3.1.2 数据分割模块具体实施

数据分割模块在具体实施中，首先，用“A”表示基础大数据，并采用分割的方式，将基础大数据（A）划分成 A_1 数据片、A_2 数据片、A_3 数据片…A_i 数据片共 i 个数据片，其中将 i 值控制为偶数，然后，将 i 份数据片划分成 $\frac{i}{2}$ 组，从而获得 [A_1，A_i]，[A_2，A_{i-1}]…[$A_{\frac{i}{2}}$，$A_{\frac{i}{2}+1}$]，同时，从每组数据片中，分别挑选出 0~100 个百分点数据元组，并对其进行相互交换，从而获得新的数据片 B_1 数据片、B_2 数据片、B_3 数据片…B_i 数据片，从而得出以下计算公式：

$$A=\bigcup_{i-1}^{i}A_i=\bigcup_{i-1}^{i}B_i$$

3.2 渐进近似计算

3.2.1 渐进近似计算模块概述

渐进近似计算模块主要是指从首个数据片入手，对数据分割模块所提交的数据片进行依次接收，并将这些数据片安全、可靠地传输至聚集增量更新模块中，由聚集增量更新模块统一计算最终聚类分析结果，并返回相应的采样值，并精确地计算出这些采样值的平均值、相对标准差、置信区间，并将三个计算值直接发送和传输至终端用户。如果终端用户成功接收到相对标准偏差、置信区间，此时，整个分析过程立即结束；反之，如果终端用户没有接收到相对标准偏差、置信区间，需要借助数据分割模块，再次输入数据片，并进行后续操作 [5]。

3.2.2 渐进近似计算模块具体实施

渐进近似计算模块在具体实施期间，主要包含以下两个环节：①将申请数据片请求发送到数据分割模块中，并接收数据分割所发送的数据片，然后，从首个数据片 B_1 开始。②当向聚集增量更新模块传输所接收的 B_1 数据片，由聚集增量更新模块对这些数据片进行统一化处理，当这些数据片处理拒收后，获得抗 n 个聚集分析结果采样值，即 $V_1^{(1)},V_2^{(1)},\cdots,V_n^{(1)}$。③对 n 个聚集分析结果值进行求平均值，$av^{(i)}=\sum_{n-1}^{n}V_i^{(i)}/n$。④计算 n 个聚集分析结果值的相对标准偏差：$rsd^{(i)}=\frac{\sqrt{\sum_{n-1}^{n}(V_i^{(i)}-av^{(i)})/n-1}}{av^{(i)}}$。⑤计算 n 个聚集分析结果值的置信区间为 $[\sum_{n-1}^{n}V_i^{(i)}/n,\sum_{n-1}^{n}(V_i^{(i)}-av^{(i)})/n-1]$。⑥将平均值、相对标准差、置信区间三个计算结果直接传输和发送至终端用户，如果终端用户接收和认可这三个计算结果值，那么大数据聚集分析过程立即结束；反之，如果终端用户不认可这三个计算结果值，那么需要返回到步骤 1，并进行后期操作。

3.3 聚集增量更新

3.3.1 聚集增量更新模块概述

当渐进近似计算模块发送的数据片传输到聚集增量更新模块后，由聚集增量更新模块合并处理这些数据片以及过去处理过的各个数据片，从而形成全局数据片，然后，借助该全局数据片，严格按照所规定的次数，进行放回采样。当每次放回采样结束后，借助数据样本集，利用聚集增量更新模块，对最终聚集分析结果进行精确计算和更新，确保所获得的聚集分析结果采样值完全满足样本集使用需求。当各个样本集处理结束后，需要向渐进近似计算模块传输最终聚集分析结果采样值 [6]。

3.3.2 聚集增量更新模块具体实施

聚集增量更新模块在具体实施时，主要包含以下三个环节：①接收由渐进近似计算模块所发送的数据片 B_m（$1\leqslant m\leqslant i$）。②如果 m=1，需要从 B_m 中，采用有放回的方式完成对 $S_1^{(1)},S_2^{(1)},\cdots,S_n^{(1)}$ 共 n 个样本集，这些样本集与 B_m 完全相同，接着，采用全量聚集分析法，精确地计算这些样本集，从而获得内部控制审计个聚集分析结果的采样值 $V_1^{(1)},V_2^{(1)},\cdots,V_n^{(1)}$，然后，向渐进近似计算模块传输和发送 n 个采样值。此外，还要结合聚集增量更新模块中的动态属性，将其采样值划分为 z 个区间，并将 B_m 中的所有数据元组所对应的动态属性值划分到指定的区间内，并将指针绑定到该区间内。在进行聚集增量更新期间，为保证最终聚集分

析结果的精确性和真实性，需要对运算的某个属性进行全面地分析和比较，如果该属性与嵌套子查询结构进行分析和对比，所获得的属性为典型的对称属性，反之，所获得的属性为静态属性。③在 B_m 中各个数据元组内，结合动态属性值，将其划分到相应的区间内，并对相关指针进行有效地关联，便于相关人员借助指针快速查询到相应区间的数据元组。同时，还要从 $\bigcup_{i-1}^{i} B_i$ 中，采用有放回的模式，完成对 $S_1^{(1)}, S_2^{(1)}, \cdots, S_n^{(1)}$ 共 n 个样本集的采样，这些样本集与 $\bigcup_{i-1}^{i} B_i$ 完全相同。在此基础上，应用聚集增量更新模块，对嵌套子查询所对应的结果值进行实时更新，并结合各个数据元组，将其精确属性和动态属性所对应的取值划分到相应的区间中，从而获得最终聚集分析结果。当各个样本集处理结束后，将最终采样值传输和发送至渐进近似计算模块中。

4 近似计算方法优点

通过应用本文所提出的近似计算方法，可以借助渐进近似计算模块和聚集增量更新模块，实现对聚集分析结果的精确化计算和获取，将大数据聚集分析所造成开销成本降到最低 [7]。此外，应用该近似计算方法，可以与终端用户进行直接交互和互动，确保最终聚集分析结果无限趋近于大数据聚集分析结果值，促使大数据聚集分析过程变得更加简单化、人性化。另外，该近似计算方法仅仅用到三个模块，单个模块实施流程简单高效，容易操作，突破开发工具和编程语言的局限性，同时，还能直接扩展和应用到分布式开发环境中，极大地提高近似计算结果精确度和真实性 [8-10]。

5 结语

综上所述，在大数据聚集分析背景下，本文所提出的近似计算方法主要涉及数据分割模块、渐进近似计算模块、聚集增量更新模块。借助数据分割模块，可以对原始大数据顺序进行科学划分，使其被划分为若干个流式数据片，然后对该流式数据片内数据元组进行科学调整；借助渐进近似计算模块，可以从第一个数据片入手，将其准确输入和传输到各个数据片中，从而获取相应采样值，同时，以相对标准差的方式，返回最终近似分析结果。总之，该近似计算方法具有分析精确度高、分析高效、分析过程人性化等特点，完全符合实际应用需求，值得被进一步推广和应用。

参考文献

[1] 申金鑫,吴烨,陈荦,等.面向空间在线分析的并行近似聚集查询[J].计算机科学与探索,2018,12(10):1559-1570.

[2] 郑晓东,郑业爽,宋思琪.时空大数据分析在人群聚集统计中的应用[J].计算机时代,2023(4):67-71+85.

[3] 史英杰,杜方,尤亚东.MSOLA:基于多维分层采样的大数据在线聚集技术[J].计算机应用研究,2018,35(2):375-380.

[4] 王庆国,赵海,万婕.宜出行大数据支持的武汉市主城区职住特征研究[J].测绘通报,2023(3):144-149.

[5] 宋文杰,刘娟,田家兴,等.基于旅游点评大数据的传统村落文化旅游特征分析——以北京市28个传统村落为例[J].小城镇建设,2022,40(6):110-118.

[6] 宋卿清,曲婉,冯海红.基于制度分析与发展(IAD)框架的先行先试政策推广评估理论研究——以国家大数据(贵州)综合试验区为例[J].科技管理研究,2022,42(2):16-25.

[7] 肖静,王翠敏.河北省大数据产业发展现状分析[J].合作经济与科技,2021(23):46-48.

[8] 李沛林.遥感通信信息属性大数据聚类分析系统[J].电子设计工程,2021,29(15):133-136+141.

[9] 谢奇爱,李正茂.基于大数据关联规则的网络恶意行为识别检测[J].合肥学院学报(综合版),2021,38(2):85-91.

[10] 赵宇,潘青亮,王景丽.5G时代通信大数据在应急管理中的应用研究[J].中国新通信,2021,23(13):15-17.