D-023

時空間データウェアハウスにおける差分演算について
Difference Operators in Spatio-temporal Data Warehouses

趙 セイ † 石川 佳治 † 杉浦 健人 † 脇田 佑希子 †
Jing Zhao † Yoshiharu Ishikawa† Kento Sugiura† Yukiko Wakita†

1. はじめに

近年，大量のデータの高度な分析処理を行うための**データアナリティクス**（data analytics）について，研究が盛んに行われている [6]．**時空間データベース**（spatio-temporal database）においても，行動データ，移動軌跡データ，科学分野のデータなどのさまざまな領域において大規模な時空間データの分析が求められている．我々の研究グループでは，特に時空間的な大規模シミュレーションの結果を蓄積し，対話的な分析を可能とするための**時空間データウェアハウス**（spatio-temporal data warehouse）に関する研究を進めており，**シミュレーションデータウェアハウス**（simulation data warehouse）と呼んでいる [10]．

本稿では，時空間データウェアハウスにおける分析要求の一つとして考えられる**差分**（もしくは**差異**）（difference）に着目する．時空間的なデータでは，特に時間に関する変化をどのように検出するかが重要であるが，大量のデータの中から顕著な変化を検出することは容易ではない．そこで本稿では，時空間データウェアハウスから差分・差異を検出するための汎用的な演算について，その要件を分析し，提案を行う．

2. 要求の分析

ここでは，時空間的な特徴を持つデータについて考える．例として，GPS および携帯機器などにより収集できる，ユーザの各時刻における位置情報を考える．そのようなデータは，(id, x, y, t) というリレーションの形式で表すことができる．なお，対象とする 2 次元空間については，グリッドが設定されており，与えられた点 x, y に対応するグリッドセルは簡単に求められるとする．ここで，次のような要求を考える．

> 時区間 $I_1 = [t_1, t_2]$ および $I_2 = [t_3, t_4]$ において，各セルに含まれる移動ユーザ数を集計し，顕著な差異があるセルがあればそれを報告せよ

この要求では，時区間 I_1 と I_2 において，ユーザの数の分布にどのような違いがあるかを求めている．ただし，このような簡単な例においても，いくつかの考慮すべき点がある．

- どのような集計を想定しているか：最も一般的には，その時区間および対象セルに関するレコード数をカウントすることが考えられる（SQL ならば SUM 関数に対応）が，分布のパターンを見たいなら，全レコード数で割った頻度分布とすることも考えられる．また，別の対象データに関しては，AVG や MAX などの他の集計関数の利用も考えられる．
- 「顕著な違い」をどのように検出するのか：差異についての定式化が必要となる．こちらについても，対象データに応じて差分・差異に関する要求が異なる場合がある．
- これまで空間のグリッドについては所与であるとしてきたが，適切なグリッドの大きさを選択することも必要である．ユーザが必ずしも事前に適切なグリッドのサイズが分かるわけではないためである．
- 時区間の指定に関する自由度：たとえば，T_2 を時間全体とした場合，ある期間 T_1 の期間全体に対する違いを見ることができる．さらに，時区間を所与のものとはせず，「幅 τ の時区間のうち，期間全体に対してもっとも違いがあるものはどれか」といったものも考えられる．
- 報告をどのように行うか：テキストや表の形式のデータを出力するのか，何らかの可視化を行うかが考えられる．可視化を前提とするならば，それに適した差分・差異の定義が考えられる．

†名古屋大学大学院情報科学研究科
Grad. Sch. of Information Science, Nagoya University

3. 差分演算のイメージ

差分演算の事例として，ここでは単純なものを考える．図 1 にそのイメージを示す．左図は時区間 T_1 における集計結果を表している．各セルの濃淡は集計値の大きさに対応している．一方，中図は時区間 T_2 における集計結果である．両者を比較可能とするため，どちらの集計結果も，総数で割って正規化されているものとする．

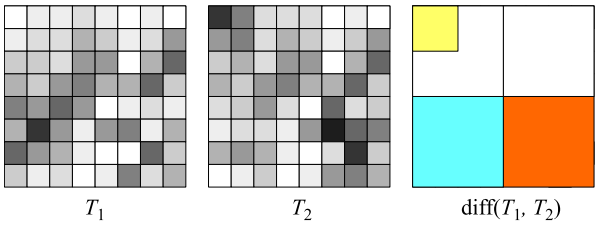

図 1: diff 演算のイメージ

このような入力に対し，右図は T_1 の集計結果に対する T_2 の集計結果の差分を近似的に表した結果である．ヒートマップ表現を用い，赤い色が強い領域ほど T_2 の時区間において T_1 時区間と比較して集計値が大きいことを示す．一方，青い色が強いほど，逆の傾向があることを意味する．また，大まかな差分の傾向を表現するため，差分のトレンドが同じようなセルが隣接する場合にはそれらをまとめて一つのセルとして表現する．ここでは四分木（quadtree）のような空間分割を想定し，セルの辺の大きさが 2 のべき乗の長さになるようにしている．

このような出力結果の提示により，ユーザは，二つの時区間における変化を容易に把握することができる．なお，上記の説明では時区間 T_1 および T_2 における集計結果をまず生成してから差分を計算すると説明したが，実装においては必ずしもこのように段階を踏む必要はないため，効率よい実装を考える余地がある．

4. 差分演算のアイデア

実装においては，**空間ヒストグラム**（spatial histogram）[1, 2] の考え方を用いる．ヒストグラムはデータベースの問合せ最適化などのために広く用いられているが [5]，これを空間データベースに拡張したものが空間ヒストグラムであり，空間データの空間的な頻度を適切に近似しようというものである．

素朴な手順としては以下が考えられる．

1. 空間を最粒度で何分割するかというパラメータ n が与えられているものとする．図 1 の例では $n = 3$ であり，最粒度の集計では，空間が $2^n \times 2^n = 64$ 分割されている．
2. 与えられた時区間 T_1, T_2 に対してそれぞれ集計を行い，さらに総数でそれぞれ正規化し，図 1 の左図，中図のような中間結果を得る．

3. 2 つの集計結果の差をとる．最粒度セルのそれぞれに対して，差分の値が計算される．
4. 差分のデータをもとに b 個のセルを用いた空間ヒストグラムを構築する．b は $b < 2^n \times 2^n$ を満たす定数であり，ユーザにより与えられる．b が小さいほど大まかな近似となる．b 個のセルの選び方の組合せの中から，最もよいものを選択することになる．

ステップ 4 における空間ヒストグラムの作り方にはさまざまなアプローチが考えられる．特にポイントとなるのは，空間をどのように分割するかという考え方である．一つには，図 1 の右図に示したように，$2^m \times 2^m$ $(m < n)$ のサイズのヒストグラムのセル（ただし四分木の境界に合わせる）を考えるアプローチがある．他のアイデアとしては，STHoles 法 [2] に見られるように，ヒストグラムのセルが空間的にオーバーラップしてよいというものもある．

これらについては，精度だけでなく，可視化における有効性などについて得失があると考えられる．今後，詳細に検討したい．

5. 差分演算の拡張

前節で述べた差分演算では，2 つの時区間 T_1, T_2 がユーザにより指定されることを想定していた．しかし，このようなアプローチは，事前にユーザが着目すべき時区間についてすでにアイデアを有しているときにしか使えない．より有用性があるものとして以下にアイデアを示す．

1) 集計の単位としての時区間の幅 τ のみがユーザにより指定されるとする．τ の時間間隔ごとにデータセットを集計し，$T_1, T_2, \ldots, T_m$ という集計結果のシーケンスを構築する．そして，T_{i+1} と T_i $(1 \leq i \leq m-1)$ の差分を求め，差分が大きいペアから順にトップ k 件を選択する．すなわち，ユーザが着目すべき差異の順に k 件を選び出すことで，ユーザの分析の負荷を減らすという効果が期待できる．
2) 1) のアイデアと似ているが，比較対象を前後の時区間とするのではなく，各時区間 T_i $(1 \leq i \leq m)$ を全体の時区間の傾向と比較してトップ k 件を選ぶことが考えられる．すなわち，ある時点における変化ではなく，一般的なトレンドと比べて乖離している度合いにより選択することになる．
3) これらと直交するが，着目する領域を合わせて指定することも考えられる．たとえば 1) と組み合わせた場合には，「この領域において大きい変化が生じた順に k 件の時区間のペアを求めよ」となる．

上記の拡張演算については，特に 1) のアプローチについてはより大きいコストが発生しうる．一方，3) については対象データが限定されることから，効率的に処理できる可能性がある．

なお，時間間隔 τ については，ユーザに任意に選択させるのではなく，空間データと同様，たとえば $1, 2, 4, 8, \ldots$ のような 2 のべき乗の値のみを選択可能とすることが考えられる．この場合，対象のデータが静的であるならば，データウェアハウスの技術 [9] を用いて事前の集計を行っておき，差分演算の実行時の処理時間を削減することが考えられる．対話的な処理を実現する上で，現実的な解であると考える．

6. 関連研究

データウェアハウスおよび OLAP に関する研究は，元々はビジネス分野を対象になされてきたが，時空間データに関する拡張についても研究開発がなされている．[9] の 11 章や [4] に解説・サーベイがある．空間データウェアハウスに関する取り組みの多くは地図データを対象にしており，地図上での統計情報（例：人口分布）などを解析することが目的である．空間情報に加え時間情報も活用される場合もある．このような観点では，[7] などの，移動軌跡データに対するデータウェアハウスも，時空間情報を集約表現する点で本研究と技術的に関連が深い．

近年データアナリティクスが着目されているが [6]，対話的なユーザの分析を支援する上で可視化は大いに有効であり，多くの研究が行われている．データベースの観点から見たときには，大規模なデータに対する可視化を瞬時に行えることや，どのデータを可視化するかの選択技術などが重要となる．たとえば MuVE [3] では，棒グラフによる可視化を考えており，どのような観点を用いれば，指定された条件で絞り込んだデータがデータ全体に対して顕著な差異を示すかを考えている．カテゴリ属性を対象としているが，データベース可視化のための SeeDB システムも関連が深い [8]．

7. まとめと今後の課題

本稿では，時空間データベースにおける差分演算のアイデアを示した．今後はその定式化と実現手法に関する研究を進める予定である．

謝辞

本研究の一部は，科研費（16H01722, 26540043），CREST「大規模・高分解能数値シミュレーションの連携とデータ同化による革新的地震・津波減災ビッグデータ解析基盤の創出」，および文部科学省「実社会ビッグデータ利活用のためのデータ統合・解析技術の研究開発」による．

参考文献

[1] S. Acharya, V. Poosala, and S. Ramaswamy. Selectivity estimation in spatial databases. In *Proc. SIGMOD*, pp. 13–24, 1999.

[2] N. Bruno, S. Chaudhuri, and L. Gravano. STHoles: A multidimensional workload-aware histogram. In *Proc. ACM SIGMOD*, pp. 211–222, May 2001.

[3] H. Ehsan, M. A. Sharaf, and P. K. Chrysanthis. MuVE: Efficient multi-objective view recommendation for visual data exploration. In *ICDE*, pp. 731–742, 2016.

[4] L. Gómez, B. Kuijpers, and B. Moelans. A survey of spatio-temporal data warehousing. *International Journal of Data Warehousing and Mining*, 5(3):28–55, 2009.

[5] Y. Ioannidis. The history of histograms (abridged). In *Proc. VLDB*, pp. 19–30, 2003.

[6] 石川 佳治. 大規模データアナリティクスに関する研究動向と展望. 電子情報通信学会論文誌 D, J97-D(4):718–728, 2014.

[7] L. Leonardi, G. Marketos, E. Frentzos, N. Giatrakos, S. Orlando, N. Pelekis, A. Raffaetà, A. Roncato, C. Silvestri, and Y. Theodoridis. T-Warehouse: Visual OLAP analysis on trajectory data. In *Proc. ICDE*, pp. 1141–1144, 2010.

[8] A. Parameswaran, N. Polyzotis, and H. Garcia-Molina. SeeDB: Visualizing database queries efficiently. *Proceedings of the VLDB Endowment*, 7(4):325–328, 2013.

[9] A. Vaisman and E. Zimányi. *Data Warehouse Systems: Design and Implementation.* Springer, 2014.

[10] J. Zhao, K. Sugiura, Y. Wang, and Y. Ishikawa. Simulation data warehouse for integration and analysis of disaster information. *Journal of Disaster Research*, 11(2):255–264, 2016.