2E-07

集計データへの差分プライバシ適用における特性の一考察Ⅱ

本郷 節之 † 手塚 理貴 † 寺田 雅之 ‡ 稲垣 潤 †
北海道科学大学 † 株式会社ＮＴＴドコモ ‡

1 はじめに

プライバシ保護とデータの有効活用の両立を可能にする技術が注目を集めている．とりわけ，Dwork らが提案した差分プライバシ基準[1] は，高い安全性を保障する．しかし，データの有用性や処理効率において実用性に課題を有している．Xiao らが提案した Privelet 法[2] は，有用性の一要素である“部分和精度の劣化”の改善を実現した．この手法では，データに対して Wavelet 変換を施し，得られた Wavelet 係数に対して Laplace 摂動[1]を加えることで，差分プライバシ基準に従う秘匿を実現している．けれども，この Privelet 法を用いても，例えば人口の空間分布のような，非負の，かつ，疎な（ゼロ値が多い）データ分布に対しては，“非負制約の逸脱”，“計算量の増大”といった課題が残されていた．

寺田らは上記ふたつの問題を解決するために，この Privelet 法に対して Top-down 精緻化[3] と呼ばれる処理を導入した．Top-down 精緻化を伴う Privelet 法は，“秘匿処理後のデータに負値が存在しない”，“疎なデータの計算量を抑制できる”という特徴をもつことが，理論的に証明されている[3]．

我々は以前，Top-down 精緻化を伴う Privelet 法を人口分布データに適用し，(1)部分和精度を劣化させないこと，(2)疎なデータにおいてはむしろ部分和精度を向上させる性質があること等を示した[4]．しかし，Top-down 精緻化による計算量抑制効果の評価については，未だ実データによる比較評価は行われてはいない．

本稿では，ゼロデータの割合が異なる複数のエリアにおける人口分布データを対象に，Top-down 精緻化の計算量抑制効果の大きさを評価した結果について報告する．

A Study on Characteristics of Publishing Tabular Data with Differential Privacy II
† Hokkaido University of Science
‡ NTT DOCOMO Inc.

2 方法

本評価では，異なる複数のエリアにおける人口分布データに対して Privelet 法による秘匿処理を適用し，Top-down 精緻化処理が適用される Privelet 逆変換処理部分の処理時間を計測する．

Privelet 逆変換処理の過程で，一旦 Top-down 精緻化処理が発生すると，その片側子ノードから下は，全てのノードの値が 0 となる．Top-down 精緻化を伴う Privelet 法では，この部分の演算処理が省略できるため，計算量を大きく抑制できる．この“演算処理の省略”を，ここでは“枝刈り”と呼ぶことにする．そして，“①枝刈りあり”と“②枝刈りなし”との間の処理時間の差を“②枝刈りなし”の処理時間で正規化した値(②-①)/②をここでは“時間短縮率”と呼ぶこととし，計算量抑制効果の指標とする．

本評価では，平成 22 年度国勢調査に基づく地域メッシュ人口（1km メッシュ）のデータに対して，シンプルな一次元 Privelet 法を適用した．日本全国（$2^{11}\times2^{11}$）のデータから，(1)北海道（$2^{9}\times2^{9}$），(2)四国（$2^{8}\times2^{8}$），(3)関東（$2^{8}\times2^{8}$）の各エリアを切り出して評価用のデータとした．また，他エリアとの比較用に，隣接する縦横２×２メッシュの人口をひとまとめにした，(4)北海道 1/4（$2^{8}\times2^{8}$）も評価用データに加えた．

なお，本評価を行うにあたって，二次元上に配置された地域メッシュ人口データを，一次元化する方式による差異についても確認するべく，(a)ラスター方式，(b)ソート方式（降順），(c)Morton 方式という３方式によって一次元化を行い，各々に対して評価を行った．Morton 写像は，多次元空間から一次元空間への全単射を行うものであり，元の空間上における距離の遠近が写像先の空間における距離の遠近に反映される性質を持つ，局所性保存写像の一種である．なお，(b)ソート方式は差分プライバシを満たさないが，他２方式と比較する目的で加えている．

評価には Intel Core i7 CPU (2.93GHz)，実装メモリ 4GB のデスクトップＰＣを使用した．また，同一処理を 100 回繰り返した時間を計測して 1/100 し，計測時間の精度向上を図った．

3 結果

表１に処理時間の計測結果を示す．３方式間での時間短縮率の傾向を見ると，ソート方式において最も時間短縮率が高く，ラスター方式において最も低い．また，同一メッシュ数をもつ３エリア間での時間短縮率の傾向については，北海道 1/4 の値が最も高く，関東の値が最も低い．さらに，メッシュ数の多い北海道エリアは，著しく高い時間短縮率を示している．

表１　処理時間の計測結果

一次元化方式	エリア	メッシュ数	処理時間（100ループの平均） ①枝刈あり[ms]	処理時間（100ループの平均） ②枝刈なし[ms]	処理時間比率[%] ①／②	時間短縮率[%] (②－①)/②
ラスター方式	北海道	262,144	22.8	48.1	47.4%	52.6%
	四国	65,536	10.2	12.3	82.6%	17.4%
	関東		11.2	12.0	93.7%	6.3%
	北海道 1/4		8.4	12.0	69.7%	30.3%
ソート方式	北海道	262,144	4.3	47.9	9.0%	91.0%
	四国	65,536	3.4	11.9	28.3%	71.7%
	関東		5.4	11.7	46.0%	54.0%
	北海道 1/4		1.8	11.9	15.3%	84.7%
Morton方式	北海道	262,144	10.5	47.2	22.2%	77.8%
	四国	65,536	7.9	12.0	66.0%	34.0%
	関東		9.3	11.9	77.9%	22.1%
	北海道 1/4		5.7	11.7	48.3%	51.7%

4 考察

表２に Top-down 精緻化発生回数とゼロ値比率の値を示す．前述の通り，演算時間の短縮は，Top-down 精緻化の効果によって生ずるものであり，Top-down 精緻化の発生回数と時間短縮率の間に正の相関関係が期待できる．しかし，評価の結果，そのような関係は見られなかった．

表２　Top-down 精緻化発生回数とゼロ値比率

一次元化方式	エリア	メッシュ数	元データのゼロ値含有比率	時間短縮率[%] (②－①)/②	Top-down 精緻化発生回数	処理結果のゼロ値含有比率
ラスター方式	北海道	262,144	95.0%	52.6%	57,370	90.8%
	四国	65,536	78.7%	17.4%	23,867	68.2%
	関東		61.2%	6.3%	20,649	50.9%
	北海道 1/4		90.3%	30.3%	20,805	81.8%
ソート方式	北海道	262,144	95.0%	91.0%	6,261	96.9%
	四国	65,536	78.7%	71.7%	4,835	83.5%
	関東		61.2%	54.0%	5,640	64.9%
	北海道 1/4		90.3%	84.7%	3,033	92.8%
Morton方式	北海道	262,144	95.0%	77.8%	23,841	95.3%
	四国	65,536	78.7%	34.0%	16,622	73.9%
	関東		61.2%	22.1%	15,322	55.6%
	北海道 1/4		90.3%	51.7%	13,273	85.9%

次に，処理結果のゼロ値比率と時間短縮率の関係について評価を行った．前述の通り，Top-down 精緻化が起こると，その片側子ノードから下の値は全て 0 となる．この特性に着目すると，枝刈りによる時間短縮の効果（時間短縮率）は，処理結果のゼロ値含有比率との間に正の相関を持つことが期待できる．

図１に，今回行った評価における，処理結果のゼロ値比率と時間短縮率の関係を示す．グラフから，両者の間に高い正の相関が見て取れる．

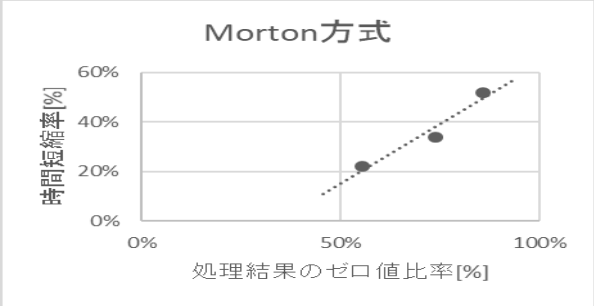

図１　処理結果のゼロ値比率と時間短縮率
（Morton 方式の例）

5 まとめ

Top-down 精緻化の計算量抑制効果の評価を行った．本評価から，元データのゼロ値含有比率が高いほど時間短縮率が高く，さらに，メッシュ数の多い北海道エリアについては，著しく高い時間短縮率を示すことがわかった．また，時間短縮率が，処理結果のゼロ値含有比率と正の相関を示すことも明らかとなった．

謝辞

本研究の一部は日本学術振興会科学研究費補助金基盤研究(C)（課題番号：15K00190）による補助を受けて行なわれた．

参考文献

[1] Barak, B., Chaudhuri, K., Dwork, C., Kale, S., McSherry, F., Talwar, K. and Berkeley, U. C.: Privacy, accuracy, and consistency too: a holistic solution to contingency table release, Proc. 26th ACM SIGMOD-SIGACT-SIGART symp. Principles of database systems - PODS '07, ACM Press, 273–282 (2007).

[2] Xiao, X., Wang, G., Gehrke, J. and Jefferson, T.: Differential Privacy via Wavelet Transforms, IEEE Trans. Knowledge and Data Engineering, 23, 8, 1200–1214 (2011).

[3] 寺田，鈴木，山口，本郷: 大規模集計データへの差分プライバシの適用，情報処理学会論文誌, 56, 9,1801–1816 (2015).

[4] 本郷，石崎，寺田，稲垣，岡崎：集計データへの差分プライバシ適用における特性の一考察Ｉ，電気・情報関係学会北海道支部連合大会，189-190 (2016).