L-007

非負集計データのための部分和精度に優れた差分プライバシー適用手法二次元化の一考察Ⅴ
A study of two-dimensional differential privacy application method providing high-accuracy answers for range count queries Ⅴ

本郷 節之 † 石井 貴己 ‡ 寺田 雅之 * 杉尾 信行 † 鈴木 昭弘 † 稲垣 潤 †
Sadayuki Hongo Yoshiki Ishii Masayuki Terada Nobuyuki Sugio Akihiro Suzuki Jun Inagaki

1. はじめに

本研究では，元のデータベースに含まれる個々のデータの集合体（個票）から，何らかの条件を満たすデータの個数を数えた数値データの集合体であり，さらに，全体的に疎な分布をとるような集計データを対象とする．集計データに対するプライバシー保護に関しては古くから検討されて来ているが，近年，Dwork らが提案した差分プライバシー基準[1] が，高い安全性を実現するための基準として注目を集めている．差分プライバシー基準は，データベースへの問い合わせを行った際に，「ある特定のデータがデータベースに含まれているか否かを問い合わせ結果から判別することが困難である」ことを安全性の根拠とするプライバシー保護基準である．この差分プライバシー基準を満たす代表的な手法に Laplace メカニズムがある．この手法は，データベースへの問い合わせ結果に対して，平均値が 0 の Laplace ノイズ（Laplace 分布に従う独立な乱数）を付加するものである．たとえば，構成する部分集合が互いに素であるとき，集計データの各セルに確率密度が

$$\ell = \frac{\epsilon}{2} e^{-\epsilon x}$$

に従う Laplace ノイズを加えることで差分プライバシーを満たすことができる（ϵ はパラメータ）．

しかし，この Laplace メカニズムを大規模集計データに適用すると，「非負制約の逸脱」「部分和精度の劣化」「疎データの密度急増」といった問題への対処が必要となる．そこで，これら 3 点の課題を同時に解消・改善する手法として，我々は「非負精緻化を伴う Privelet 法」を提案した[2]．これは，Xiao らによって提案された Privelet 法[3] が有する，部分和精度が高いという性質を維持しつつも，「非負制約の逸脱」に対する回避と，「疎データの密度急増」の抑制を同時に実現する手法である．

非負精緻化を伴う Privelet 法は，一次元データ列を対象としたものであり，二次元データに適用する際には，一旦一次元データ配列に変換を行った上でプライバシー保護処理を適用し，その上で，処理された一次元データを，改めて二次元データへ戻す処理が必要となっていた．しかし，上述した通り，地理的に分布した集計データは二次元状に分布していることから，そのプライバシー保護処理においても，二次元データに直接適用できる手法の開発が望まれる．そこで我々は現在，その二次元化手法の開発を進めている．これまで，提案した二次元方式の精度について論じて来たが，本稿では，演算速度について議論する．

† 北海道科学大学 Hokkaido University of Science
‡ SOC 株式会社 SOC Corporation
＃株式会社 NTT ドコモ NTT DOCOMO Inc.

2. 方法

我々は既に非負精緻化を伴う Privelet 法を二次元化する基本アルゴリズムの提案を行った[4]．これは， Xiao らが提案している標準分解に基づく多次元化手法[3]よりも秘匿に要するノイズ量が少ない手法である[2]．いま，図 1 に示す 2×2 基本構造からなる二次元 Haar Wavelet を採用すると，図 2 に示す二次元ツリー構造が構成される．ここでリーフ層には，はじめ，秘匿対象となる集計データ$v_{x,y}$が格納されており，一方，ノード層には，Wavelet 係数($cA_{h,x,y}$または$cD_{h,x,y}$)が格納される（hは階層番号）．

図 1　2×2 基本構造

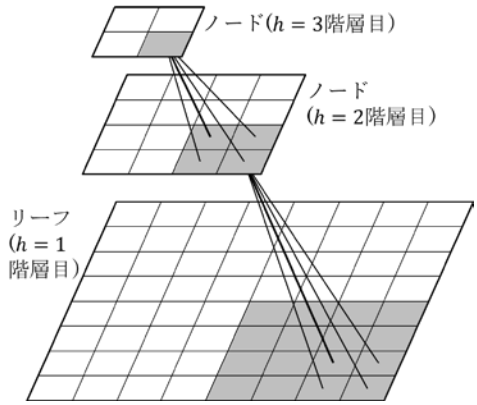

図 2　二次元ツリー構造

2.1 Wavelet 変換および逆 Wavelet 変換

いま，階層番号が$h = (1,2,3,\cdots,H)$で表され，1 層のリーフ$(h = 1)$と$H-1$層のノード$(1 < h \leq H)$ からなる二次元ツリーを考える．最初の Wavelet 変換(第 1 階層)を行った際の近似係数$cA_{h,x,y}$および詳細係数$cD_{h,x,y}$の値は，次式で求められる．ここで，$x = (0,1,2,\cdots.X-1)$ならびに$y = (0,1,2,\cdots.Y-1)$は各階層における二次元座標を表す．

$$\begin{aligned} cA_{1,0,0} &= \frac{v_{0,0}+v_{1,0}+v_{0,1}+v_{1,1}}{4} \\ cD_{1,1,0} &= \frac{v_{0,0}-v_{1,0}+v_{0,1}-v_{1,1}}{4} \\ cD_{1,0,1} &= \frac{v_{0,0}+v_{1,0}-v_{0,1}-v_{1,1}}{4} \\ cA_{1,1,1} &= \frac{v_{0,0}-v_{1,0}-v_{0,1}+v_{1,1}}{4} \end{aligned} \quad (1)$$

続いて次階層の変換(第 2 階層)を行うのに先立って，$cA_{1,x,y}$の値を 1 階層上のノード$(2, x/2, y/2)$へコピーする．その上で，第2階層の各ノードに対して，上記式(1)に準ずる処理を行う．以上の処理を最上位層まで再帰的に繰り返すことで二次元 Wavelet 変換を実現できる．

一方，逆 Wavelet 変換の処理は，Wavelet 変換処理のプロセスを逆にたどる．これにより改めて第 1 階層においてリーフ値として秘匿された数値を得ることができる．

2.2 ノイズ付加

提案手法では，Wavelet 変換（順変換）の後，詳細係数$cD_{h,x,y}$(最上位層のみ近似係数$cA_{H,0,0}$含む)に対して確率分布$\ell(x:\lambda'(h)) = (1/2\lambda'(h))e^{(-x/\lambda'(h))}$に従う Laplace ノイズを付加する．階層$h$におけるノイズ強度は$\lambda'(h) = \lambda/4^h$とする．ここで$\ell, x$および$\lambda = H/\epsilon$はそれぞれ確率密度，確率変数，ノイズ強度を表す．

2.3 非負精緻化

非負精緻化処理は，逆 Wavelet 変換処理の過程で負値の発生を排除する処理である．いま，h層での逆 Wavelet 変換処理の結果得られた 4 変数$cA^*_{h-1,2x,2y}$, $cA^*_{h-1,2x+2,2y}$, $cA^*_{h-1,2x,2y+2}$, $cA^*_{h-1,2x+2,2y+2}$のうちのいずれか（複数もあり得る)に負の値が現れたら，次式に従って，3 つの変数$cD^*_{h,x+1,y}, cD^*_{h,x,y+1}, cD^*_{h,x+1,y+1}$に対して非負精緻化処理を施し，精緻化後の$cD$値を用いて改めて逆 Wavelet 変換処理を行う(ノイズ付加の結果得られた値は*を付して，また，非負精緻化の結果得られた値は+を付して表している).

$$cD^+_{h,x+1,y} = \beta \cdot cD^*_{h,x+1,y}, \quad cD^+_{h,x,y+1} = \beta \cdot cD^*_{h,x,y+1},$$
$$cD^+_{h,x+1,y+1} = \beta \cdot cD^*_{h,x+1,y+1},$$
$$\beta = \frac{cA^*_{h,x,y}}{\left|Min\left(cA^*_{h-1,2x,2y}, cA^*_{h-1,2x+2,2y}, cA^*_{h-1,2x,2y+2}, cA^*_{h-1,2x+2,2y+2}\right) - cA^*_{h,x,y}\right|}$$

2.4 枝刈り

枝刈りは，非負精緻化を伴う Privelet 法の演算処理を効率化する実装法である[5]．非負精緻化を伴う Privelet 法では，逆 Wavelet 変換の際に，あるノードの値が0になると，そこに連結した下層のノード値は全て0となるため演算を省略できる．この“演算の省略”により効率化を 図る手法が枝刈りである．枝刈りの手法として，水平型実装法と垂直型実装法が提案されているが，ここでは，より効率性の高い垂直型実装法を採用する．

3. 評価と考察

図３に二次元 Privelet 法の，メッシュ人口データ（(a) 関東，(b) 四国，(c) 北海道 1/4；メッシュ数$2^8 \times 2^8$）に対する演算時間を示す．北海道 1/4 は，ほかの地域とデータサイズが同じになるように，北海道メッシュ人口データに対して2×2部分和をとって生成したものである．これら３エリアの人口分布には，(a) 関東はゼロ値含有比率が低く（61.2%），(b) 四国はゼロ値含有比率が中程度（78.7%），そして，(c) 北海道 1/4 はゼロ値含有比率が高い（90.3%）という特徴がある．

非負精緻化，枝刈りとも逆 Wavelet 変換処理内に組み込まれていることから，ここでは逆 Wavelet 変換処理に要する時間に着目する．評価には Intel Core i7-9700 CPU (3.00GHz)，実装メモリ 16.0GB 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．

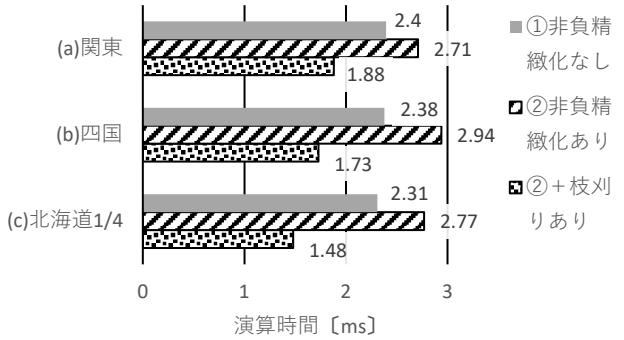

図３ 逆 Wavelet 変換時間（関東, 四国, 北海道 1/4）

図において，「①非負精緻化なし」は「非負制約の逸脱」および「疎データの密度急増」が改善されていない二次元 Privelet 法の演算時間を，「②非負精緻化あり」は，①に非負精緻化処理を組み込んだ場合の演算時間を，そして，「②＋枝刈りあり」は②に枝刈り処理を組み込んだ場合の演算時間をそれぞれ表している．

演算時間の計測結果から，各エリアとも，非負精緻化の導入による演算時間のわずかな増加，および，枝刈り処理による演算時間の顕著な短縮効果が確認できる．また，「①非負精緻化なし」においては３エリアともほぼ同程度の処理時間であること，ゼロ値含有比率が高いエリア（関東＜四国＜北海道 1/4）ほど枝刈りによる演算効率化の効果（「②非負精緻化あり」，「②＋枝刈りあり」間での処理時間の削減率）が大きい（北海道 1/4 で約 47%減）こと等が見てとれる．この結果はゼロ値含有比率が高いほど演算効率化効果が高いという，枝刈りの性質を反映している．

4. おわりに

本稿では，我々が先に提案した非負精緻化を伴う二次元 Privelet 法の演算処理速度計測結果について述べた．今後は一次元方式との比較など，より多角的な検討を進めて行く．

謝辞

本研究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：19K11970) の補助を受けて行なわれた．

参考文献

[1] Dwork C., “Differential Privacy”, Proc. 33rd Intl. Conf. Automata, Languages and Programming - Volume Part II, Bugliesi, M., Preneel, B., Sassone, V. and Wegener, I. (Eds.), Lecture Notes in Computer Science, 4052, Springer, pp. 1-12 (2006).

[2] 寺田雅之，鈴木亮平，山口高康，本郷節之, “大規模集計データへの差分プライバシの適用”，情処学論，56, No. 9, pp. 1801-1816 (2015).

[3] Xiao X., et. al., “Differential Privacy via Wavelet Transforms”, IEEE Trans. Knowledge and Data Engineering, 23, No. 8, pp. 1200-1214 (2011).

[4] 本郷，寺田，鈴木，稲垣：“非負集計データのための部分和精度に優れた差分プライバシー適用法二次元化の一考察 I ”，電気・情報関係学会北海道支部連合大会，pp.129-130 (2019).

[5] 本郷，寺田，鈴木，稲垣：“非負精緻化をともなう Privelet 法における演算効率化手法の性能向上”，情処学論，61，9，pp.1458-1471 (2020).