

マーケティング情報のデータ分析を加速する暗号化データの統合・分析手法 を世界で初めて開発

～セキュリティの最高峰国際会議 ACM CCS で発表～

日本電信電話株式会社(本社:東京都千代田区、代表取締役社長:島田 明、以下「NTT」)は、暗号化したままデータ分析を行う秘密計算技術において、複数データのデータ結合およびグループ化集計(統計値計算)を行う方法を開発しました。開発した方法は 2 つの部分からなり、データ結合部分では、実測値で従来の 3.95 倍の世界最高速を達成し、グループ化集計部分では重要な統計値である中央値の計算を世界で初めて実現しました。この方法を用いることで、秘密計算による複数組織のデータの統合・分析が、これまでよりも多様な統計値を対象に、より短時間でできるようになり、秘密計算によるプライバシーを保護したデータ活用の幅が広がります。本成果は NTT コミュニケーションズ株式会社(以下 NTT Com)が提供中のサービス「[析秘](#)」に、内部演算として既に実装されており、本成果を活用した分析をお客さまにご利用になれます。

なお、本成果は 2023 年 11 月 26 日から 30 日まで、デンマーク・コペンハーゲンで開催されるセキュリティ分野の最高峰国際会議 30th ACM Conference on Computer and Communications Security (ACM CCS 2023) において発表されます。

1. 研究の背景

データの活用が拡大している昨今、活用の対象を個人に関する情報や機密データへと広げる際に、どう安全性や安心感を確保するかが課題となっていました。NTT はこの課題に対し、データを暗号化したまま計算途中の値をも一度も元に戻すことなく分析結果を計算する秘密計算技術の研究を進めてきました。理論研究においては秘密計算で任意の処理が実現できることは証明されているものの、現実的な時間で実行できる処理は限られていました。秘密計算の効果が大きいと考えられる利用シーンとして、異業種の組織間のデータを持ち寄った分析があり、大量かつ多くの種類のデータの分析のため、秘密計算で多様な分析をより高速に実現されることが望まれていました。

2. 研究の成果

従来の技術では計算可能な統計量が平均値や最大値、最小値などに限られていましたが、当該技術により、複数の組織が持ち寄ったデータを暗号化したまま、双方のデータに共通して含まれる人物に関し名寄せしたデータを作成(データ結合)し、名寄せ後のデータに対して統計値計算(グループ化集計)を行うという一連の分析(図 1)を、これまでよりも多様な統計値を対象に、より高速に実現することができるようになりました。クレジット利用額のような極端な値が含まれるデータに対しても、楽曲データを販売している会社 A とクレジットカード会社 B が、好みの楽曲とクレジットカード利用額の関係に着目した分析(グループ化集計)を行い、双方の持つデータを誰にも明かすことなく名寄せを行い、好みの楽曲のジャンルごとのクレジット利用額の統計量を計算することができます。

データ結合

組織Aの持つ
顧客の年代データ

名前	年代
Alice	30代
Bob	20代
Carol	30代
Dave	50代

組織Bの持つ
顧客の購入額データ

名前	購入額
Alice	120円
Dave	5,300円
Alice	1,250円
Carol	3,800円
Ellen	570円
Dave	200円

組織A・Bのデータに共通の
顧客の年代と購入額のデータ

名前	年代	購入額
Alice	30代	120円
Dave	50代	5,300円
Alice	30代	1,250円
Carol	30代	3,800円
Dave	50代	200円

グループ化集計

年代と購入額のデータ

年代	購入額
30代	120円
50代	5,300円
30代	1,250円
30代	3,800円
50代	200円

年代ごとの購入額の統計量(総和、最大値、中央値)

年代	購入額 (総和)	購入額 (最大値)	購入額 (中央値)
30代	5,170円	3,800円	1,250円
50代	5,500円	5,300円	2,750円

図 1 開発技術が実現する処理

一方、NTT が今回開発したグループ化集計のアルゴリズムを使うと、中央値やパーセンタイルなどの、極端な値の影響を受けにくい重要な統計量をも世界で初めて計算できるようになり、これまでよりもデータの特徴を捉えた分析ができるようになります(図 2)。また、今回開発したデータ結合のアルゴリズムは従来の技術と比較して実測値で 3.95 倍の世界最高速を達成しています。これらのアルゴリズムを使うことにより、100 万件の顧客データ同士を名寄せしてから好みの楽曲のジャンルごとのクレジット利用額の 5 種類の統計値(件数、総和、最大値、最小値、中央値)を計算するまでの一連の処理を 9 秒で完了する実験結果を得られました。

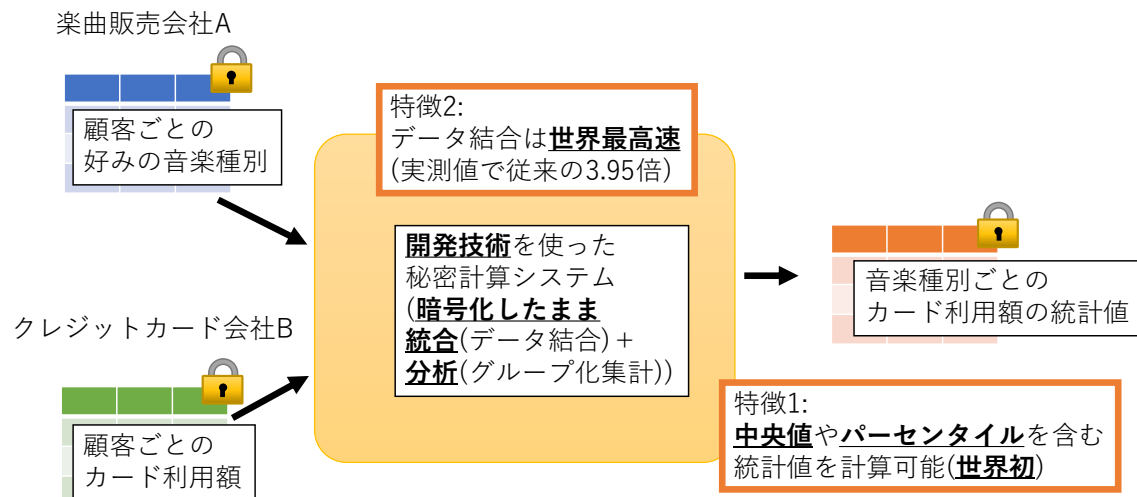


図 2 開発技術の利用イメージ

3. 技術のポイント

NTT が今回開発した方法では、従来の技術とは異なる方法でデータ結合とグループ化集計を実現しています。ポイントは大きく 2 つあります。

- 1 つめのポイントは、データ結合で累積和の計算を利用したことです。データ結合の従来手法では、集計木と呼ばれる技術を用いていましたが、集計木には前の計算に依存した計算が多く含まれるため、並列性が高くなりにくいという問題がありました。これに対し NTT が開発したデータ結合の方法は、集計木を使わずに並列性を高めやすい累積和の計算を利用することで、高い並列性を生かした高速なデータ結合を実現しました。
- 2 つめのポイントは、グループ化集計でデータの並びを効果的に利用したことです。グループ化集計の従来研究では、集計木を応用することにより、層別の総和や最大値などの統計値計算を実現できることが証明されていました。これに対し NTT が開発したグループ化集計の方法では、データを昇順に並び替え、データの並びが昇順であることを効果的に利用することによって、世界初の中央値やパーセンタイルを含む、従来手法よりも多くの種類のグループ化集計を実現しました。

4. 今後の展開

本成果は NTT Com が提供中のサービス「[析秘](#)」に実装されており、サービスを通じて複数の組織が持ち寄ったデータの安全な統計分析に貢献しています。今後はさらなる演算の拡充により、より広範囲のデータベース処理の実現を進めていきます。



〈参考〉

NTT コミュニケーションズ株式会社 秘密計算サービス 析秘(セキヒ)

<https://www.ntt.com/business/services/secihi.html>

※「析秘」はエヌ・ティ・ティ・コミュニケーションズ株式会社の商標です。

■ 本件に関する報道機関からのお問い合わせ先

日本電信電話株式会社

サービスイノベーション総合研究所

企画部広報担当

nttrd-pr@ml.ntt.com