安全なビッグデータ分析をクラウド上で実現する秘匿分析技術

長沼 健 Naganuma Ken
吉野 雅之 Yoshino Masayuki
佐藤 尚宜 Sato Hisayoshi
佐藤 嘉則 Sato Yoshinori

大量のデータを収集分析し，新しい知識を抽出するビッグデータ分析は，購買履歴データ，医療データ，センサデータの分析などさまざまな分野で行われている。これに伴い，クラウドにデータを預託して分析するサービスも普及しつつあるが，外部のクラウドサーバ上で分析を行った場合，不正アクセスや内部犯行による情報漏えいのリスクが指摘されている。この課題の解決に向けて，暗号文のテキストマッチング処理が可能な検索可能暗号を利用し，暗号化したまま集計分析や相関ルール分析を行う秘匿分析技術を提案している。この技術を用いることで，データ分析依頼者はデータの中身を開示することなく分析を依頼することが可能となり，情報漏えいリスクを低減することができる。

1. はじめに

大量のデータを収集分析し，新しい知識を抽出するビッグデータ分析は，購買履歴データ，医療データ，センサデータなどを対象に行われている。例えば，購買履歴データの最も代表的な分析手法の一つである相関ルール分析（アソシエーションルール分析とも呼ばれる。）は，ある商品と別の商品の間にあるつながり，すなわち「おむつ」を買う人は同時に「ビール」を買うことが多いといった相関ルールを抽出し，マーケティングなどに活用する。このように大量のデータから隠れた知識を抽出するビッグデータ分析は，今日のIT（Information Technology）トレンドであり，マーケット分析などさまざまな分野で行われている。

これに伴い，自身の持つデータをクラウドサーバ上で分析するSaaS（Software as a Service）サービスも広く利用されることが予想される。しかし，外部のクラウドサーバ上で分析を行った場合，データに対する不正アクセスや内部犯行による情報漏えいリスクが指摘されており，安全にデータ分析を行う技術の開発が課題となっている。そこで，このデータ預託時に発生するセキュリティ問題の解決に向け，データを暗号化したまま，復号化することなくデータ分析を行う秘匿分析技術の研究開発を進めている。

ここでは，最も基本的な分析手法である集計分析，相関ルール分析を暗号化したまま実行する技術について述べる。この技術により，クラウドユーザーは外部クラウドにデータの中身を開示せずに暗号化データのみを預託することで集計分析や相関ルール分析を実行でき，前述の情報漏えいリスクを低減することが可能となる。また，この技術を開発するにあたっては，暗号化されたデータの安全性に加えて，大量データの分析に対応するため処理の効率性も重視した。実機を用いた実験により，10万件の暗号化データに対して約600秒（10分）で相関ルール分析が可能なこと，および中規模データに対して秘匿分析技術が実運用可能なことを確認した。

2. 秘匿分析技術

外部クラウド上で秘匿分析サービスを利用する際のシステム構成を**図1**に示す。秘匿分析システムでは，分析対象のデータを保持しているクラウドユーザー（同図の「自社組織」）がデータを自身の鍵で暗号化し，外部クラウドに暗号化データを預託する。さらに，暗号化された分析クエリ（命令）をクラウドに送信する。クラウド上では，暗号化データと暗号化クエリを用いて分析を実行し，暗号化された分析結果を算出・返信する。このとき，クラウド上ではデータ，クエリは復号化することなく，すべて暗号化状態で分析が実行される。最後にクラウドユーザーが暗号化された分析結果を復号することで，目的の結果を得る。

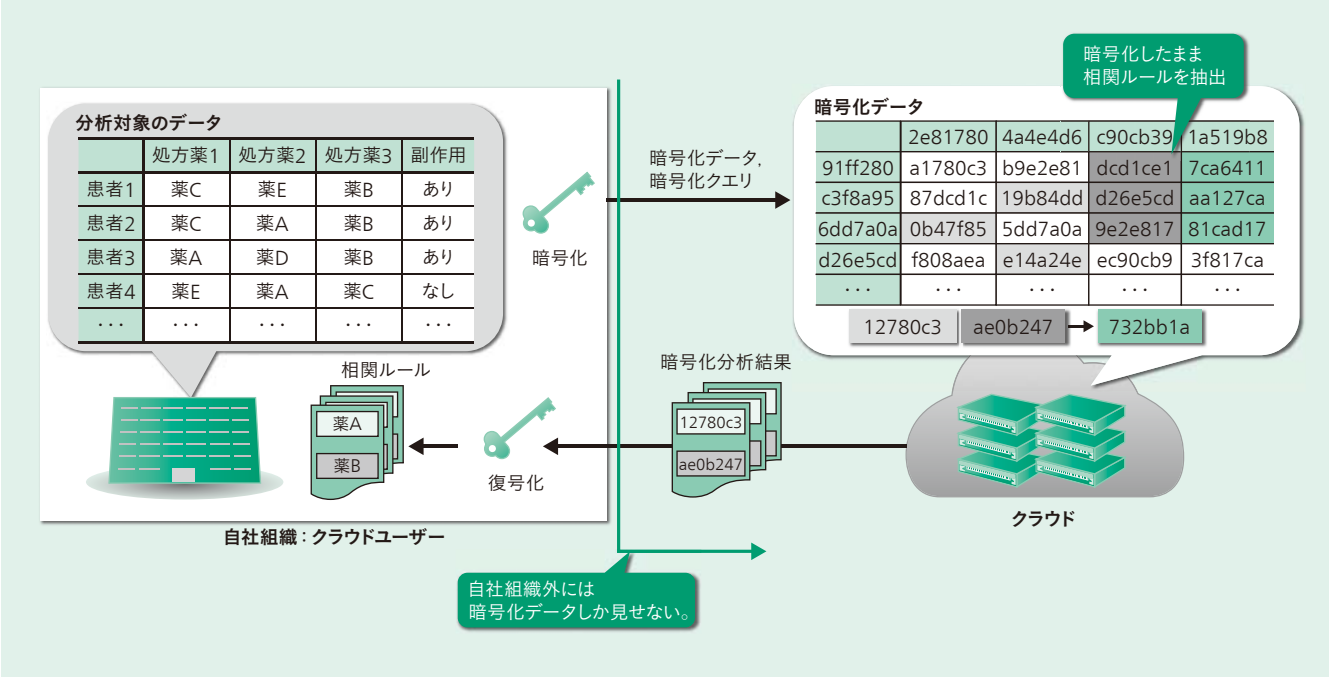


図1｜秘匿分析のシステム構成

クラウドに暗号化データを預託し，暗号化された分析クエリをクラウドに渡して分析を実行する。クラウド上ではデータが暗号化されているため，情報漏えいが発生した場合のリスクを低減することができる。

従来，分析を実行する際には，クラウド上で暗号化データを一度，復号してから処理を行う必要があったため，前述の漏えいリスクが高まる。一方，秘匿分析技術では，クラウド上でデータは暗号化されているため，情報漏えいが発生した場合のリスクを低減することができる。

以下，暗号化したままテキストマッチング（平文と平文クエリがテキストとして完全一致しているかどうか判定）処理が可能な検索可能暗号を利用し，暗号化データ上で集計分析，相関ルール分析を実行する方法を述べる。今回，集計分析，相関ルール分析に分析アプリケーションを絞った理由は，データマイニングにおいてこれらが最も基本的な分析となるからである。

2.1　検索可能暗号

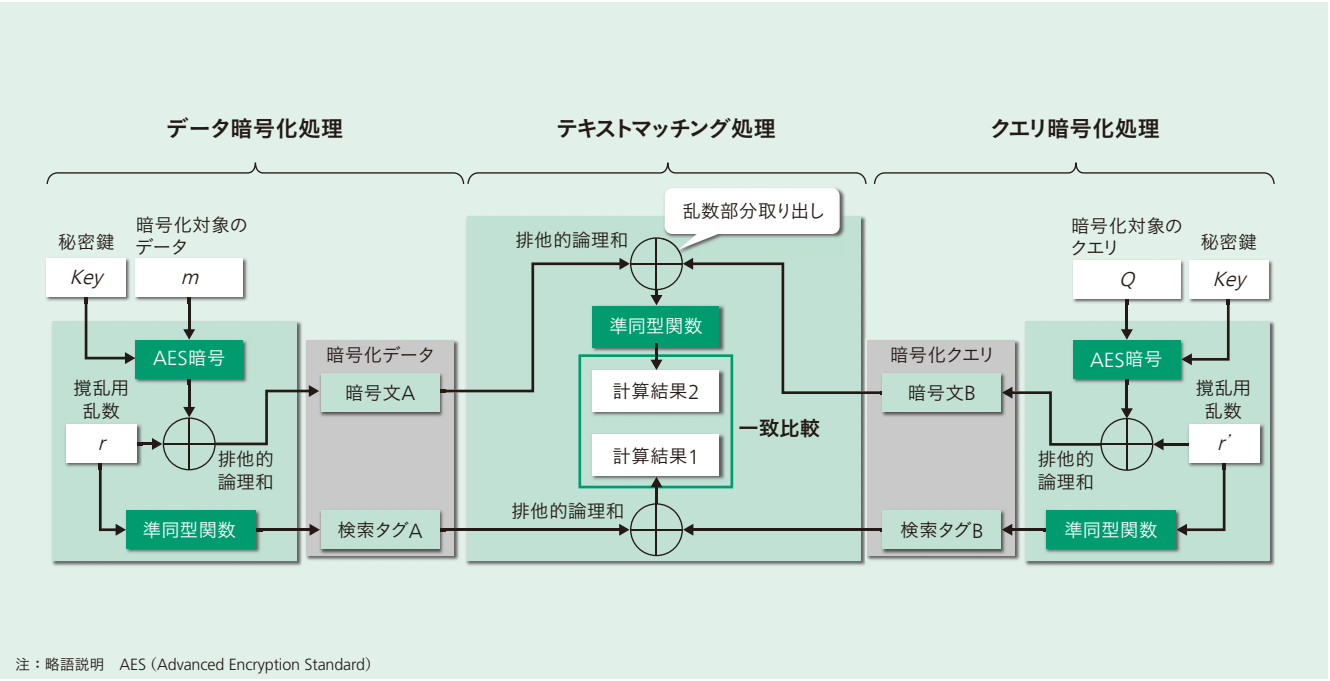
検索可能暗号とは，通常の暗号化，復号化機能に加えて，暗号文と暗号化クエリに対してテキストマッチング処理が可能な暗号方式の総称である。また，暗号化，復号化はそれぞれ暗号化鍵，復号化鍵を必要とするが，テキストマッチング処理においては特別な情報は必要なく，鍵を持たないクラウド管理者でも実行可能である。ただし，方式によっては，正当な権限者のみにテキストマッチングが実行できるよう，マッチング処理用の秘密鍵を持つものもある。

具体的な検索可能暗号方式としていくつかの方式が知られているが[1), 2), 3), 4), 5)]，これらの方式は，共通鍵方式と公開鍵方式に大別することができる。共通鍵方式は，暗号化と復号化の鍵が同一であり，公開鍵方式に比べて処理効率が高く，大容量データの暗号化に適している。公開鍵方式は，暗号化と復号化の鍵が異なり，暗号化鍵を広く公開しても安全性が確保できるが，共通鍵暗号よりも処理が複雑になる傾向があり，暗号化と復号化により多くの計算資源を必要とする。

今回，大量データの分析に対応するため，処理効率の観点から，日立が提案した共通鍵検索可能暗号方式[5)]をベースとして秘匿分析技術を開発した。この検索可能暗号アルゴリズムは，データ暗号化とクエリ暗号化のそれぞれに撹（かく）乱用乱数を利用することで，同一の平文（もしくは平文クエリ）に対しても暗号化のたびに異なる暗号化データ（もしくは暗号化クエリ）が生成されるため，暗号アルゴリズムの安全性が高い（**図2**参照）。また，アルゴリズム中の各処理は，最も標準的な共通鍵暗号方式であるAES（Advanced Encryption Standard）暗号などの高速な暗号プリミティブを用いて設計されており，公開鍵暗号技術をベースとする検索可能暗号アルゴリズムに比べて1,000倍程度高速に検索処理を実行することが可能である[5)]。

2.2　検索可能暗号を利用した秘匿分析

前述したとおり，検索可能暗号のテキストマッチング処理機能を用いれば，クラウド管理者（以下，「クラウド」と記す。）は暗号化データを復号することなく，データベース内に暗号化クエリがいくつ出現するかといった出現頻度を求めることができる。これを応用すれば，クラウドユーザーが適当な暗号化クエリをクラウドに提供することで，



注：略語説明　AES（Advanced Encryption Standard）

図2｜日立提案方式の検索可能暗号のデータ暗号化，クエリ暗号化，テキストマッチング処理の概要

データ暗号化とクエリ暗号化のそれぞれに撹（かく）乱用乱数を利用することで，暗号アルゴリズムの安全性を高めている。また，各処理は，高速な暗号プリミティブを用いて設計されており，公開鍵をベースとする検索可能暗号アルゴリズムに比べて1,000倍程度高速に検索を実行することが可能である。

アイテムの出現頻度情報のみから計算可能なデータマイニングアルゴリズムを暗号文上で実行することができる。アイテムの出現頻度情報のみで実行可能なデータマイニングアルゴリズムとしては，簡単な集計分析や，後述する相関ルール分析などが挙げられる。また，このような検索可能暗号を利用した分析の前提として，クラウドには，平文情報は開示されないものの，暗号化データの出現頻度情報は提供されている点に注意する。一方，暗号化クエリを持たない第三者に対しては，平文情報と暗号化データの出現頻度情報も秘匿されている。

2.3　相関ルール分析

相関ルール分析とは，購買履歴データなどの表形式トランザクションデータから，ある事象が発生すると（条件部）別の事象が発生する確率が高い（結論部）といった事象間の関係を抽出するデータ分析手法である[6]。以下，トランザクションデータの患者と処方薬品表に従い，相関ルール分析の手法を簡単に述べる（**表1**参照）。

同表の各行（トランザクション）は，各患者に処方された薬品を表している。例えば，患者1には「薬A」,「薬B」,「薬C」が処方されたことを示している。このとき，ある薬品Aを処方された（条件部）患者は，別の薬品Bも同時に処方されることが多い（結論部）といったルールを見つけたいとする。以下このことを，「相関ルールA⇒B」と略記する。各相関ルールは「支持度」,「信頼度」,「リフト値」の3つの指標を基に評価される。それぞれの指標の定義は以下のとおりである。

相関ルールA⇒Bの支持度（support）とは

$$\mathrm{Supp}(A \Rightarrow B)=\frac{\text{AとBを含むトランザクションの総数}}{\text{トランザクションの総数}}$$

相関ルールA⇒Bの信頼度（confidence）とは

$$\mathrm{Conf}(A \Rightarrow B)=\frac{\text{AとBを含むトランザクションの総数}}{\text{Aを含むトランザクションの総数}}$$

相関ルールA⇒Bのリフト値（lift）とは

$$\mathrm{Lift}(A \Rightarrow B)=\frac{\mathrm{conf}(A \Rightarrow B)}{\mathrm{supp}(B)}$$

例えば，**表1**において相関ルール「薬A⇒副作用あり」の支持度，信頼度，リフト値はそれぞれ次のとおりとなる。

Supp（薬A⇒副作用あり）$=\frac{4}{8}=0.5$

Conf（薬A⇒副作用あり）$=\frac{4}{5}=0.8$

Lift（薬A⇒副作用あり）$=\frac{0.8}{0.625}=1.28$

次に，3つの指標が意味するところを述べる。

表1｜トランザクションデータ

各行は患者ごとに処方された薬品と副作用の発現あり，もしくは発現なしを示している。

	処方薬1	処方薬2	処方薬3	副作用
患者1	薬A	薬B	薬C	あり
患者2	薬B	薬A	薬F	あり
患者3	薬B	薬D	薬E	なし
患者4	薬C	薬E	薬F	なし
患者5	薬E	薬A	薬B	あり
患者6	薬A	薬D	薬E	なし
患者7	薬C	薬B	薬A	あり
患者8	薬C	薬E	薬F	あり

相関ルール分析は，トランザクションテーブル中に多頻出する事象間の関係を抽出することが目標であり，支持度は出現確率を表す指標である。通常は，この値が高い事象に絞って分析を行う。信頼度は，条件部の事象が発生した条件の下で，結論部の事象が発生する確率を表す。よって，この値は条件付き確率と見なすことができる。また，リフト値は，この条件付き確率と結論部の発生確率の割合であり，この値が1より真に大きければ，条件部と結論部の相関が強いと言える。上述の例では「薬A⇒副作用あり」のリフト値が1.28であるため，条件なしの通常の患者より，「薬A」を処方された患者の方が「副作用あり」の確率が1.28倍高いと言える。実際の分析では，支持度，信頼度，リフト値の高い相関ルールを抽出し，なぜその事象間に相関があるのかを別途分析してマーケティングなどにフィードバックする。

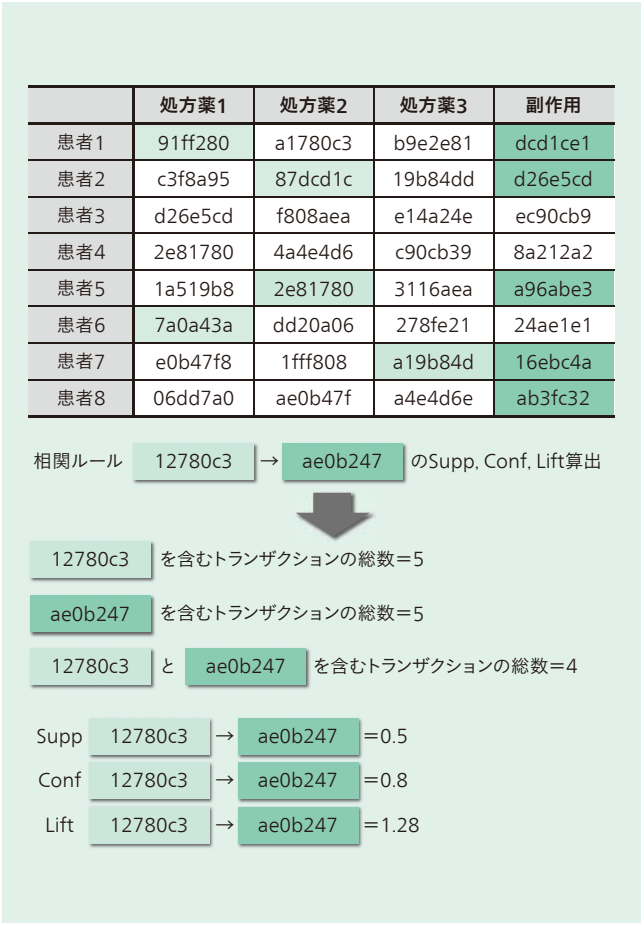
相関ルール分析を行う際の重要なポイントとして，支持度，信頼度，リフト値は，それらの定義から，特定のアイテムを含むトランザクションがいくつ存在するかといった出現頻度情報から算出が可能なことが挙げられる。このことは，検索可能暗号を用いて暗号化データ上で相関ルール分析を実行する際のキーとなる。

2.4　秘匿集計分析，相関ルール分析

検索可能暗号方式を用いて，暗号化データ上で集計分析，相関ルール分析を行う方法を述べる。

図3上部の表は，**表1**の各セルを日立提案の検索可能暗号方式で暗号化したトランザクションデータである。例えば**表1**の患者1—処方薬1の「薬A」は「91ff280」と暗号化されている。また，同表の患者2—処方薬2も同じく「薬A」であるが，暗号化後のテーブルでは，「87dcd1c」と異なる暗号文に変換されている。このように，日立提案の検索可能暗号方式では同一の平文でも異なる暗号文に暗号化され，これら暗号文単独では乱数列と識別困難である。よって，暗号化トランザクションデータが漏えいして第三者の手に渡ったとしても，この暗号化トランザクションデータから平文の情報を読み取ることはできない。

今，クラウドユーザーが，事象Aと事象Bに対し，相関ルールA⇒Bの支持度，信頼度，リフト値の算出をクラウド上で実行する場合を考える。前述したとおり，相関ルールA⇒Bの支持度，信頼度，リフト値を算出するためには，トランザクションテーブル中の「事象Aを含むトランザクションの総数」，「事象Bを含むトランザクションの総数」，「事象AとBの両方を含むトランザクションの総数」，「全トランザクションの数」の4つの値が分かればよい点に注意する。全トランザクションの総数はテーブルの大きさから既知のものとする。



	処方薬1	処方薬2	処方薬3	副作用
患者1	91ff280	a1780c3	b9e2e81	dcd1ce1
患者2	c3f8a95	87dcd1c	19b84dd	d26e5cd
患者3	d26e5cd	f808aea	e14a24e	ec90cb9
患者4	2e81780	4a4e4d6	c90cb39	8a212a2
患者5	1a519b8	2e81780	3116aea	a96abe3
患者6	7a0a43a	dd20a06	278fe21	24ae1e1
患者7	e0b47f8	1fff808	a19b84d	16ebc4a
患者8	06dd7a0	ae0b47f	a4e4d6e	ab3fc32

図3｜暗号化トランザクションデータ上での相関ルール分析
暗号化クエリ「12780c3」と「ae0b247」に対し，「支持度」，「信頼度」，「リフト値」を算出している。

まず，クラウドユーザーは，事象A，事象Bを検索可能暗号で暗号化し，暗号化クエリQuery（A），Query（B）を生成する。次に，この暗号化クエリをクラウドに渡し，暗号化トランザクションデータと暗号化クエリに対して検索可能暗号のテキストマッチング機能を用いて暗号化データ上で，事象Aを含むトランザクションの総数，事象Bを含むトランザクションの総数，事象AとBの両方を含むトランザクションの総数，全トランザクションの数を算出し，支持度，信頼度，リフト値を求める。つまり，通常のテキストマッチング処理を検索可能暗号のテキストマッチング処理に変更することで，暗号化された相関ルールQuery（A）⇒Query（B）に対して支持度，信頼度，リフト値を求める。このテキストマッチング機能を用いれば，集計分析も同様の方法で可能である。

図3では，暗号化されたクエリQuery（薬A）＝「12780c3」，Query（副作用あり）＝「ae0b247」に対して，それぞれの暗号化クエリを含むトランザクションの総数を検索可能暗号のテキストマッチング機能を用いて算出し，相関ルールQuery（薬A）⇒Query（副作用あり）の支持度，信頼度，

リフト値を求めている。このとき，これらの処理はすべて暗号化データ上で行われており，平文情報はクラウドに渡す必要がない点に注意する。また，通常は相関ルール分析を行う際に，単独の相関ルールA⇒Bに対して支持度，信頼度，リフト値の算出をするのではなく，トランザクションデータ中に存在する支持度，信頼度，リフト値が，ある閾（しきい）値以上になる相関ルールをすべて抽出する。この場合，クラウドユーザーはすべての事象A，B，C，…に対して暗号化クエリQuery（A），Query（B），Query（C），…を生成し，クラウドに渡して分析実行の依頼をする必要がある。

2.5　秘匿分析の処理フロー

前述した秘匿相関ルール分析をクラウドユーザーとクラウドの二者間で実行する際のデータ処理フローを**図4**に示す。クラウドには暗号化データのみが提供されている点に注意する。

(1) クラウドユーザーは，自身の持つトランザクションデータTの各セルを共通鍵検索可能暗号で暗号化し，暗号化されたトランザクションデータE（T）をクラウドデータベースに預託する。

(2) クラウドユーザーは，トランザクションデータ内の相関ルール分析（もしくは集計分析）対象アイテム集合{A，B，C，…}に対して，検索可能暗号の暗号化クエリ{Q（A），Q（B），Q（C），…}を生成し，生成した{Q（A），Q（B），Q（C），…}を分析依頼クエリとしてクラウドに送信する。また，このときクラウドユーザーは支持度，信頼度，リフト値の閾値の3組（a，b，c）を併せて送信する。

(3) クラウドは，検索可能暗号のテキストマッチング機能を利用した暗号文上での相関ルール分析手法を用いて受信したクエリ集合{Q（A），Q（B），Q（C），…}上の相関ルールで，支持度がa以上で信頼度がb以上，かつリフト値がc以上のものを求め，そのルール集合{Q（*）⇒Q（*），…}を分析結果としてクラウドユーザーに送信する。集計分析を行う場合も同様である。

(4) クラウドユーザーは，アイテム集合{A，B，C，…}の各アイテムと，その暗号化クエリ集合{Q（A），Q（B），Q（C），…}の各クエリの対応を知っているため受信したルール集合{Q（*）⇒Q（*），…}を復号し，目的の分析結果，つまり支持度，信頼度，リフト値がそれぞれ（a，b，c）以上の相関ルールを得る。

以上の手続きにより，クラウドユーザーは，クラウドに平文データを開示することなく，相関ルール分析もしくは集計分析を行うことができる。

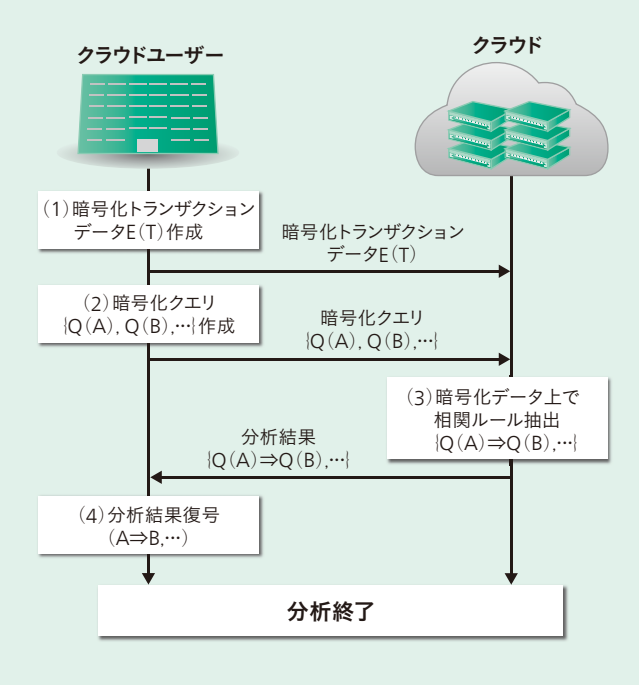


図4｜秘匿分析の処理フロー

(1) クラウドユーザーはトランザクションデータを暗号化し，クラウドに預託する。(2) 相関ルール分析対象アイテム集合の暗号化クエリをクラウドに送信する。(3)クラウドは暗号化したまま相関ルールを抽出し，結果をクラウドユーザーに返信する。(4) クラウドユーザーは分析結果を復号し，相関ルールを得る。

2.6　プロト性能評価実験

2.4節で述べた手続きに従い，テストデータに対して暗号化したまま相関ルールを抽出する実験を行った。実験には参考文献[6]と同じ10万件のテストデータを用いた。このテストデータは10万件のトランザクションから成り，各トランザクションは平均10のアイテムをエントリとし，アイテムの総数は1,000種類である。

検索可能暗号は日立提案方式を用いた。この検索可能暗号方式は，公開鍵方式をベースとする検索可能暗号に対して1,000倍程度高速にテキストマッチング処理が可能である。相関ルール分析を実行する際，各アイテムを含むトランザクションの総数を数え上げる処理が実行時間の大半を占めるため，公開鍵方式の検索可能暗号を利用した場合に比べて分析処理時間を約1,000分の1に削減することができる。1台の汎用PC（Personal Computer）上で実験した結果，10万件の暗号化トランザクションデータに対して約600秒（10分）で相関ルール抽出が完了した。

この結果から，数万件程度の中規模データに対しては，秘匿分析技術が実運用可能なことが確認された。一方で，今回実験は行わなかったものの，数百万件から数億件規模の大量データを分析する際には，データ件数の増加に応じた処理時間の発生が予想されるため，さらなる高速化や並列化などの工夫が必要である。大量データへの対応は今後の課題としたい。

3. おわりに

ここでは，外部クラウドサーバ上でビッグデータ分析を行う際の情報セキュリティ対策技術として，暗号化したままで分析を実行する秘匿分析技術について述べた。

提案した手法では，暗号化したまま検索が可能な検索可能暗号をコア技術として，暗号化データ上で集計分析と相関ルール分析を実現している。この秘匿分析では，暗号化データと暗号化クエリのみを用いて分析を行うため，不正アクセスが発生した場合やデータが漏えいした際のリスクを低減することが可能である。

今後も日立は，ビッグデータの強固な保護と利活用を両立する先進セキュリティ技術の研究開発を推進し，安全性の高いソリューションを提供していく。

参考文献

1) D. Boneh, et al.: Public key encryption with keyword search, EUROCRYPT 2004, pp. 506-522 (2004)
2) D. Boneh, et al.: Conjunctive, subset, and range queries on encrypted data, TCC 2007, pp. 535-554 (2007)
3) R. Curtmola, et al.: Searchable symmetric encryption: improved definitions and efficient constructions, CCS 2006, pp. 79-88 (2006)
4) Song DX, et al.: Practical techniques for searches on encrypted data. In: Security and Privacy, 2000. S&P 2000. Proceedings. 2000 IEEE Symposium on, pp. 44-55 (2000)
5) 吉野，外：DB向け検索可能暗号方式の検討(2)，The 2011 Symposium on Cryptography and Information Security (2011)
6) R. Agrawal, et al.: Mining association rules between sets of items in large databases, Proceedings of the ACM SIGMOD International Conference on Management of Data, Washington, D.C. (1993.5)

執筆者紹介

長沼 健
日立製作所 横浜研究所 情報サービス研究センタ サービスイノベーション研究部 所属
現在，クラウドコンピューティング・ビッグデータ利活用を支えるセキュリティ技術の研究開発に従事

吉野 雅之
日立製作所 横浜研究所 情報サービス研究センタ サービスイノベーション研究部 所属
現在，クラウドコンピューティング・ビッグデータ利活用を支えるセキュリティ技術の研究開発に従事
博士（科学）
情報処理学会会員，電子情報通信学会会員

佐藤 尚宜
日立製作所 横浜研究所 情報サービス研究センタ サービスイノベーション研究部 所属
現在，クラウドコンピューティング・ビッグデータ利活用を支えるセキュリティ技術の研究開発に従事
博士（数理学）
電子情報通信学会会員

佐藤 嘉則
日立製作所 横浜研究所 情報サービス研究センタ サービスイノベーション研究部 所属
現在，クラウドコンピューティング・ビッグデータ利活用を支えるセキュリティ技術の研究開発に従事
情報処理学会会員