情報処理学会論文誌　データベース　Vol.10　No.4　31–35 (Dec. 2017)

テクニカルノート

暗号化データベースシステムにおける クエリベースのデータ販売スキーム

秋山 賢人[1,a)]　渡辺 知恵美[2,b)]　北川 博之[3,c)]

受付日 2017年6月10日, 採録日 2017年8月6日

概要：クラウドサービスの発達にともない，暗号化データベースシステムに関する研究がさかんに行われている．暗号化データベースシステムはデータを暗号化して保存し検索を行うことができるため，データ所有者はクラウドサービスのプロバイダに対してデータを秘匿することができる．一方，データ所有者はクライアントに対してもデータ開示の制御を要求する場合がある．その一例としてデータ所有者がクライアントにデータを販売し，クエリによって得られた情報に対して検索結果を提供する前に費用を要求することなどが考えられる．本稿では，暗号化データベースでの秘匿検索フレームワーク OSIT においてデータ販売をする場合のクエリマーケットスキームを提案する．クエリによってクライアントが得る情報をヒストグラムで表し，情報利得スコアをエントロピーで定義した．実験ではこのスコアが問合せによって減少し，すべてのデータがクライアントにわたったときに 0 となることを示した．

キーワード：暗号化データベースシステム，プライバシ保護，Data marketing

A Query-based Data Selling Scheme Based on the Encrypted Database System

Kento Akiyama[1,a)]　Chiemi Watanabe[2,b)]　Hiroyuki Kitagawa[3,c)]

Received: June 10, 2017, Accepted: August 6, 2017

Abstract: With the development of cloud services, privacy preserved query schemes for encrypted database systems have been proposed. In the system, queries can be processed without decryption, therefore the data owner can preserve the confidential data against the cloud service provider. On the other hand, the data owner may require the data disclosure control towards the clients in case that the data owner sells the data to the client. In this case, we consider that the data owner may require the fee according to the amount of data the client obtains before returning the result. In this paper, we propose a data marketing scheme by using the secure query processing framework OSIT on the encrypted database system. We express the information which the client obtains from a query by histogram, and we define the information gain score by conditional entropy. From experiment, we show the score decreases by a query, and the score is zero when the client obtains all attribute values in the data. We assume the contract of buying and selling data between the data owner and the client. Then, we propose a method to calculate the information gain score for the query result.

Keywords: Encrypted database systems, Privacy preservation, Data marketing

[1] 筑波大学大学院システム情報工学研究科
Graduate School of Systems and Information Engineering, University of Tsukuba, Tsukuba, Ibaraki 305–8573, Japan
[2] 産業技術大学院大学産業技術研究科
Graduate School of Industrial Technology, Advanced Institute of Industrial Technology, Shinagawa, Tokyo 140–0011, Japan
[3] 筑波大学計算科学研究センター
Center for Computational Sciences, University of Tsukuba, Tsukuba, Ibaraki 305–8573, Japan
[a)] k_akiyama@kde.cs.tsukuba.ac.jp
[b)] chiemi@acm.org
[c)] kitagawa@cs.tsukuba.ac.jp

1. はじめに

クラウドサービスの発達により，膨大なデータをクラウド上で管理することが可能になった．サービス利用者としては個人や企業等が想定でき，データ所有者がクラウドサーバに預けたデータに対してクライアントに検索権限を与えることで，データの検索サービスも実現可能となる．

しかしながら，自身の所有するデータを外部委託してサービスを提供する場合，様々な要求が考えられる．まず，データ所有者はサーバ管理者に対し，預けたデータの中身を知られたくないという要求を持つ．次に，クライアントはデータ所有者およびサーバ管理者に対し，発行したクエリの内容を知られたくないという要求を持つ．

これらの要求を満たすために，近年では暗号化データベースシステムの研究がさかんに行われている [5], [6]．暗号化データベースシステムでは，データとクエリを暗号化して機密性を保証する．暗号化に検索可能暗号を用いることで，暗号化データを復号せずに検索が可能である．篠塚らが提案した秘匿検索フレームワーク OSIT [8] では，クライアントがサーバ上の索引を探索することで効率的な検索を実現し，準同型暗号でデータとクエリを暗号化し，暗号化したまま検索を行うことで安全性を保証している．

ところで，実際にクラウド上でのデータの検索サービスを考えた場合，サービスプロバイダに対してデータやクエリの機密性を保証するだけでなく，クライアントに対してのデータ保護の保障が必要となる場合がある．それはデータ所有者がクライアントに対し，無制限にデータを提供するのではなく，データ所有者とクライアントとの間でデータ利用に関しての契約を取り決め，その契約の範囲内でデータを提供する場合である．データ利用契約の分かりやすい形として，データ所有者がクライアントに対してデータ販売をすることを想定すると，クライアントがクエリを発行した際問合せ結果に対応して価格を設定し，結果入手前にクライアントに支払いを要求する形式が考えられる．Deep ら [1] や Koutris ら [2], [3] は，クエリに対して価格をつける価格設定関数を提案している．

本研究では秘匿検索スキーム OSIT を用いてデータ販売プラットフォームを提供することを想定し，問合せ結果に応じた価格設定スキームを提案する．OSIT では，クラウドサーバに暗号化索引として対象とする属性値でソートされた配列を格納し（ただしその配列の順序はサーバには分からない），m 分探索することで検索を行う．クライアントは，問合せ前に対象データのレコード数のみデータ所有者から取得し，問合せをすることでデータを取得する．我々はクライアントが各属性に対して持つ情報をヒストグラムで表し，問合せ結果を得ることでヒストグラムが詳細化されることを利用し，問合せ後に得られるヒストグラムのエントロピーを価格関数とする．

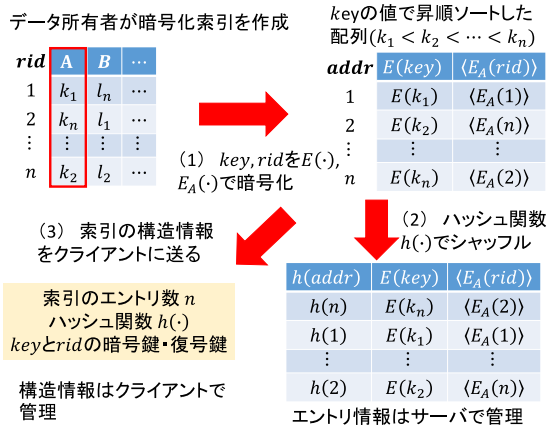


図 1　属性 A における暗号化索引の作成

Fig. 1　Construction of encrypted index on attribute A.

2. OSIT フレームワーク

我々が提案している暗号化データベースでの秘匿検索スキーム OSIT [8] について述べる．OSIT では，データ所有者，クライアント，サービスプロバイダの 3 者を想定する．データ所有者がクラウドサーバにデータを預け，サービスプロバイダが管理をする．クライアントはデータ所有者から問合せの権限を受け，サーバに問合せをして結果を受け取る．OSIT ではサーバ上に暗号化されたレコードを格納し，各属性に対し暗号化された索引を用意する．索引はクライアントとサーバで分散管理し，問合せの際にはクライアントとサーバが通信することでクライアントとサーバ双方から索引の情報を秘匿でき，高速で安全な探索ができる．

具体的な索引配置および問合せプロトコルについて述べる．データ所有者が n レコードのテーブルを持つとすると，データ所有者はテーブルをレコードごとに暗号化し，暗号化索引を作成してサーバに保管する．**図 1** に n レコードのテーブルに対する属性 A の暗号化索引作成方法を示す．

データ所有者は属性値をキーに，レコード ID を値にしたハッシュテーブルを作成し，ハッシュテーブルのレコードをキー値（key）で昇順にソートする．次に key を Paillier 暗号 [4] $E(\cdot)$ で暗号化し，ハッシュテーブルの値（rid のリスト $\langle E_A(rid)\rangle$）を加法準同型性を持つ Wong らの暗号化スキーム [6] $E_A(\cdot)$ で暗号化する（図 1 (1)）．このテーブルは key でソートされているため，属性 A に対するクエリにおいて探索ができる一方で，この状態でサーバに保管するとレコードの順序関係から属性 A の値を推測される可能性がある．そこで，データ所有者は元の索引の順序関係を秘匿するために，配列のアドレス $addr$ をハッシュ関数 $h(\cdot)$ でハッシュし，配列の要素にハッシュ値を加えたエントリ情報をサーバに保存する（図 1 (2)）．また，データ所有者はクライアントに $h(\cdot)$，key，rid の暗号鍵，復号鍵を与える（図 1 (3)）．索引を分散管理することで，サーバは索引のノード間関係を把握できず，クライアントは各ノード対

してサーバ上のアドレスのみを知り，ノード自体を取得できない．サーバは復号鍵を持たないため，サーバのデータに対する機密性が保証される．

索引の探索処理を Algorithm 1 に示す．クライアントは検索条件に含まれる属性の索引を m 分探索する．索引のエントリ数を N，クエリ値 q とする．クライアントでは探索領域 (l, u) を分割する点を $m-1$ 個選択し，暗号化したクエリ値 $E(q)$ とハッシュ値 $h(i_1), h(i_2), \ldots, h(i_{m-1})$ をサーバに送る．サーバでは，各ハッシュ値に対応する暗号化キー $E(k)$ とクエリ値との差を暗号化したまま計算してクライアントに送る．クライアントはサーバから返った値を復号し，探索領域を更新する．このように，クライアントのみがクエリ値とエントリ値の比較結果を知ることで索引の探索を実現する．

OSIT では索引を用いない検索手法 [6] に比べ，レコード数 100,000 で約 120 倍高速に検索が可能であり [10]，範囲検索 [7]，文字列検索 [9] をサポートしている．

Algorithm 1 先行研究における索引の探索

Procedure: 索引のエントリ数 N，クエリ値 q
1: クライアントでの処理
2: 探索領域の初期化 $(l, u) \leftarrow (1, N)$
3: $E(q)$ をサーバに送る
4: **while** $l < (u-1)$ **do**
5: (l, u) を m 分割する i を $(m-1)$ 個選択
6: $h(i_1), h(i_2), \ldots, h(i_{m-1})$ をサーバに送る
7: サーバでの処理
8: **for** $j = i_1$ to i_{m-1} **do**
9: $h(j)$ に対応する $E(k_j)$ を取得
10: $E(c) = (E(k_j) \cdot E(q)^{-1})^r$ を計算し，クライアントに送る
11: **end for**
12: クライアントでの処理
13: $E(c_1), E(c_2), \ldots, E(c_{m-1})$ の復号
14: $c_1, c_2, \ldots, c_{m-1}$ に基づき (l, u) を更新
15: **end while**

3. 提案手法

クラウド上でのデータの検索サービスを考えた場合，データ所有者とクライアントの間でデータ利用に関して契約を取り決め，契約の範囲内でデータを提供する必要がある．具体的には，クライアントがクエリを発行した際問合せ結果に応じて価格を設定し，結果の入手前にクライアントに支払いを要求するという契約が考えられる．このような場合，どのように価格を決定するかが課題となる．

OSIT ではクライアントがサーバにある索引を探索して検索を行う．そのため，「SELECT b FROM R WHERE a < q」のような問合せを実行した場合，索引の探索によって q より小さなレコードの数を知ることができる．これにより，OSIT でクライアントが各属性に対して持つ情報はヒストグラムで表すことができる．例として，属性 age を含むレコード数 6 のテーブルに対して問合せを繰り返したときのクライアントが持つ情報について考える．**図 2** の分布 1 に示すように，問合せ前，クライアントが知る情報はレコード数（= 6）のみである．ここで「SELECT age FROM db WHERE age = 20」の問合せを実行すると，索

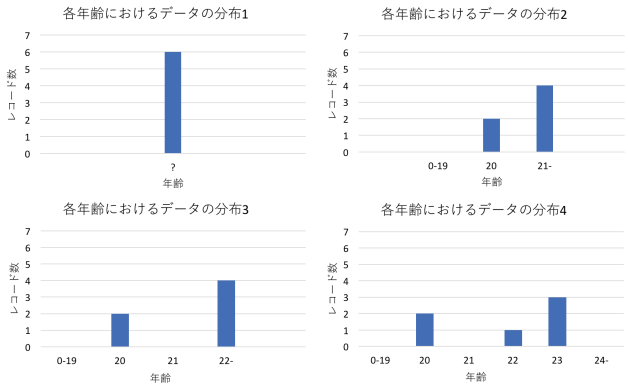


図 2　複数回の問合せにおけるヒストグラムの変化
Fig. 2　Changes of the histogram over multiple queries.

引の探索の結果 age = 20 を満たすレコードは 2 レコードあることが分かり，図 2 の分布 2 のようにヒストグラムを分割できる．分布 3 は age = 21，分布 4 は age = 22，23 の順に問合せた結果によるヒストグラムの変化を示しており，問合せを行うごとに分布が詳細になっていくことが分かる．

我々はヒストグラムにおけるデータの分布確率を利用したエントロピー計算によりクエリ結果の価格を設定する．エントロピーを利用する理由としては，データの分布が詳細になるほど値が減少していくという性質を示すからである．したがって，クライアントに開示される情報が多いほどエントロピーの値は小さくなると考えられる．

以上をふまえて，本研究では価格設定スキームを条件付きエントロピーで定義する．

ある属性の各属性値 y について，属性値が y であるレコード数が x である確率 $P(X = x|Y = y) = p_{x,y}$ を考える．問合せ前はすべての属性値について 0 件から n 件までの可能性があり，問合せを繰り返してヒストグラムが詳細化するにつれてその可能性は絞られてくる．これをエントロピーで表す．2 つの確率変数 X，Y，属性値を y，属性値が y であるレコード数を x，条件付きエントロピー $H(X|Y)$ とし，価格設定スキームを式 (1) で定義する．

$$H(X|Y) = \sum_y p_{x,y} H(X|Y = y) \tag{1}$$

ヒストグラムに含まれるレコード数を n，$Y = y$ における x の選び方を ${}_nC_x$，y の最大値を $M-1$，最小値を m とする．条件付き確率 $p_{x,y}$ は式 (2) で表せる．

$$p_{x,y} = {}_nC_x \frac{(M-m-1)^{n-x}}{(M-m)^n} (x = 0, 1, \ldots, n) \tag{2}$$

$(M-m)^n$ は $m \leq y \leq M-1$ におけるレコードの分布の選び方の総数，$(M-m-1)^{n-x}$ は $Y \neq y$ におけるレコードの分布の選び方の総数を表し，n，m の値は問合せを繰り返すことで更新されていく．また，図 2 の分布 2 の age = 20 である部分に着目して式 (2) を計算すると

$M = 21$，$m = 20$，$n = 2$ であり，分母がつねに 1，分子は 0^{n-x} となる．$p_{x,20}$ はすべての x で計算したときに，$x = 2$ のときのみ 1 を，それ以外では 0 をとる．これにより，問合せで分かった部分についてはエントロピー計算を行っても $1log1 = 0$ または $0log0 = 0$ となることが分かる．すなわち，$H(X = x|Y = y)$ の計算は問合せ結果以外のヒストグラムについて行えばよいことになる．また，問合せ結果以外のすべての y について考えると，x は 0 件から n 件のすべての値をとりうることから $H(X = x|Y = y)$ の値は x が決まれば y の値に関係なく同じ値をとるので，属性値 y の最大値を $M_0 - 1$，最小値を m_0 とすると，y のとりうる値の種類は $M_0 - m_0$ となり $P(Y = y) = \frac{1}{M_0 - m_0}$ である．これらの結果を式 (1) に代入すると，価格設定スキームは式 (3) のようになる．

$$H(X|Y) = \frac{-\sum_{y,x} \frac{{}_nC_x(M-m-1)^{n-x}}{(M-m)^n} log \frac{{}_nC_x(M-m-1)^{n-x}}{(M-m)^n}}{M_0 - m_0} \quad (3)$$

4. 評価実験

実験では提案した式 (3) に基づき価格設定を行えるかを確認する．提案手法のエントロピー計算を Adult Data Set [*1] を用いて行った．その際，Adult Data Set をそのまま利用すると計算に時間がかかり結果を得ることができなかったため，Adult Data Set のレコードの中から重複なく無作為に 10,000 件抽出したデータを利用した．実験では，(1)「SELECT age FROM db WHERE age = x」，(2)「SELECT hours-per-week FROM db WHERE hours-per-week = x」の 2 つの属性に対する問合せを考え，x の値を増加させ繰り返し問合せを行うことでエントロピーの値の変化を確認する．問合せ結果以外のデータは一様分布に従っていると仮定し，クライアントはレコード数，属性値の最小値，最大値をもとにエントロピーを計算する．提案したスキームは問合せを繰り返すことによりエントロピーの値が単調に減少し，すべての結果が得られた際にエントロピーの値が 0 になると考えられるので，実験により確認する．

(1)，(2) の問合せにおける実験結果を**図 3** に示す．図 3 は属性 age，hours-per-week に対して繰り返し問合せを行うことによるエントロピーの変化を示している．縦軸はエントロピーの値，横軸は問合せ実行回数を表す．図 3 からどちらの属性に対しても問合せによりクライアントに開示されるレコードが増加するにつれ，エントロピーの値が単調減少することを確認できた．また，属性ごとの分布の違いにより減少の仕方は異なるが，すべてのレコードが開示された場合にエントロピーの値が 0 になることも確認できた．

[*1] http://archive.ics.uci.edu/ml/datasets/Adult

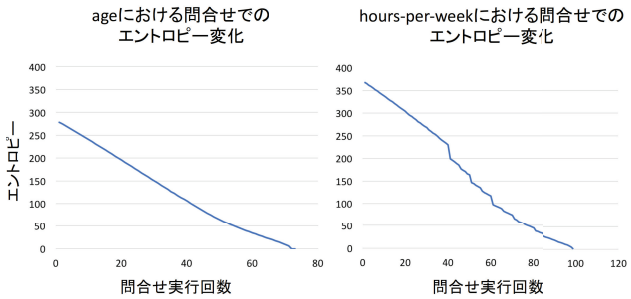


図 3　繰返し問合せによるエントロピーの変化
Fig. 3　Changes of the entropy over multiple queries.

5. おわりに

本稿では OSIT 上でのデータ販売を想定し，価格設定スキームを提案し，問合せの際に変化するデータの分布確率に注目してエントロピーを計算することで価格の設定を行った．実験から，提案スキームは全区間において単調減少することが確認でき，エントロピーが小さいほど高い価格を設定することで価格決定が行えることを示せた．今後の課題としては，エントロピー計算をクライアント，サーバ，データ所有者のいずれが行うかを検討し，秘密計算等を用いてエントロピー計算を安全に行うことがあげられる．

謝辞　本研究の一部は NICT 高度通信・放送研究開発委託研究「欧州との連携による公共ビッグデータの利活用基盤に関する研究開発」，JSPS 科研費 JP16K00149 の助成を受けたものです．

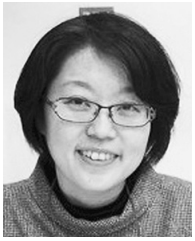
参考文献

[1] Deep, S. and Koutris, P.: The Design of Arbitrage-Free Data Pricing Schemes, *ICDT 2017*, pp.12:1–12:18 (2017).
[2] Koutris, P., Upadhyaya, P., Balazinska, M., et al.: Query-based data pricing, *Proc. PODS 2012*, pp.167–178 (2012).
[3] Koutris, P., Upadhyaya, P., Balazinska, M., et al.: Toward Practical Query Pricing with QueryMarket, *Proc. ACM SIGMOD 2013*, pp.613–624 (2013).
[4] Paillier, P.: Public-key cryptosystems based on composite degree residuosity classes, *Proc. EUROCRYPT '99*, pp.223–238 (1999).
[5] Popa, A.R., Redfield, C., Zeldovich, N., et al.: Cryptdb: processing queries on an encrypted database, *Comm. ACM*, Vol.55, No.9, pp.103–111 (2012).
[6] Wong, K.W., Cheung, W.L.D., Kao, B., et al.: Secure query processing with data interoperability in a cloud database environment, *Proc. ACM SIGMOD 2014*, pp.1395–1406 (2014).
[7] 篠塚千愛，渡辺知恵美，北川博之：DaaS 環境におけるデータとクエリ双方のプライバシ保護を実現する効率的な秘匿検索，DEIM Forum 2015 G2-6 (2015).
[8] 篠塚千愛，渡辺知恵美，北川博之：暗号化データベースにおけるデータの秘匿性を保証した検索手法，DEIM Forum 2017 H6-4 (2017).
[9] 篠塚千愛，渡辺知恵美，北川博之：クラウド環境における暗号化索引を用いた文字列属性の部分一致検索手法，

コンピュータセキュリティシンポジウム 2015 論文集, pp.979–986 (2015).
[10] 秋山賢人, 渡辺知恵美, 北川博之：秘匿検索フレームワーク OSIT における最適なクエリプラン選択法, DEIM Forum 2017 H6-5 (2017).

秋山 賢人　（学生会員）

1992 年生．2016 年筑波大学情報学群情報科学類卒業．現在，筑波大学大学院システム情報工学研究科に在学中．データベースにおけるプライバシを保護した検索技術の研究に従事．日本データベース学会学生会員．



渡辺 知恵美　（正会員）

1975 年生．産業技術大学院大学産業技術研究科准教授．2003 年お茶の水女子大学大学院人間文化研究科博士後期課程修了．同年奈良女子大学理学部情報科学科助教，2005 年お茶の水女子大学理学部情報科学科講師，2013 年筑波大学システム情報工学研究科助教，2017 年現職．データベースシステムに関する技術，特にアウトソーシングデータベースにおけるプライバシ保護検索技術，匿名化処理，プライバシリスク提示等の研究活動に従事．日本データベース学会会員．博士（理学）．

北川 博之　（正会員）

1978 年東京大学理学部物理学科卒業．1980 年同大学大学院理学系研究科修士課程修了．日本電気（株）勤務の後，筑波大学電子・情報工学系講師，同助教授を経て，現在，筑波大学計算科学研究センター教授．理学博士（東京大学）．データベース，情報統合，データマイニング，情報検索等の研究に従事．日本データベース学会前会長，情報処理学会フェロー，電子情報通信学会フェロー，ACM，IEEE，日本ソフトウェア科学会各会員．

（担当編集委員　是津 耕司）