DEIM Forum 2009 D4-6

ブルームフィルタを用いたプライバシ保護検索における攻撃モデルとデータ撹乱法の一検討

渡辺知恵美† 新井裕子† 天笠俊之‡

†お茶の水女子大学大学院 人間文化創成科学研究科 〒112-8610 東京都文京区大塚 2-1-1

‡筑波大学 大学院システム情報工学研究科 〒305-8573 つくば市天王台 1-1-1

E-mail: †{chiemi,aria.yuko}@is.ocha.ac.jp, ‡amagasa@cs.tsukuba.ac.jp

あらまし クラウドコンピューティングにおいて注目されている Database as a Service(DAS)ではデータベースを暗号化し，サーバ管理者に対してもプライバシを保護しながら問合せを行うプライバシ保護検索が注目されている．著者らはこれまで文字列検索および数値の範囲検索を扱うことのできる安全性の高い検索手法を提案してきたが，サーバでの検索処理時間がかかることが課題であった．本稿では，ブルームフィルタにノイズを混入させることにより攻撃者の推測を防ぐ手法を提案する．本手法ではブルームフィルタのビットパターンから攻撃者がテーブルの特徴を推測するための攻撃モデルを定義し，その攻撃モデルに対応するビットパターン撹乱の戦略を立てる．

An Attack Model and Data Perturbation Method for Privacy Preserving Query Method using Bloom Filter

Chiemi WATANABE† Yuko ARAI† Toshiyuki AMAGASA‡‡

† Department of Information Sciences, Faculty of Science , Ochanomizu University

2-1-1 Otsuka, Bunkyo-ku, Tokyo, 112-8610 Japan

‡‡ Center for Computational Sciences, University of Tsukuba

1–1–1 Tennohdai, Tsukuba, Ibaraki 305–8573 Japan

Abstract Recently, Database-As-a-Service (DAS) has attracted considerable attention. Users require protecting sensitive data from the DAS administrators. Most of previous studies, which proposed the solutions using cryptographic techniques, assume that a large amount of data will be inserted into the database and new data will be uploaded infrequently. We propose a secure query execution model for such an environment. Our approach is to represent all schemes of each tuple in a plaintext table as one Bloom filter index, and to replace queries with keyword searches of the Bloom filter index. Same values in each tuple are transformed into same bit patterns in the bloom filter index, and adversaries may try to mining bit pattern to find out which values are included in tuples, and the value is associated to the attribute domain. Then we define the attack model for the bloom filter based privacy preserving model, and verify the effect of the attacks. Based on the verification we take an strategy by using data perturbation.

Keyword Privacy Preservation, Cloud Computing, Database Outsourcing, Bloom filter, Association Rule Hiding

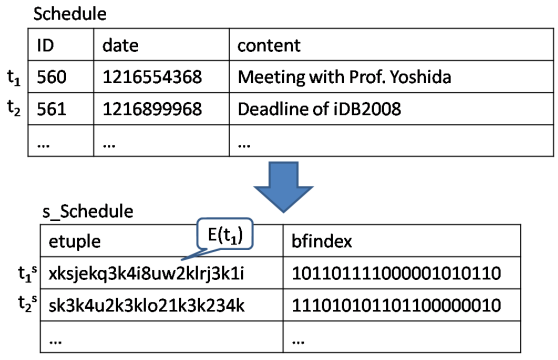
1. はじめに

近年 Database as a Service(DAS)がにわかに注目を集めている．DAS とはクラウドコンピューティング環境においてデータベース管理を請け負うサービスである．利用者はデータベースを設置・管理するための労力をかけることなく高性能なデータベース機能を利用することができるが，管理者がデータ所有者と異なる第三者であるため，データ所有者が管理者に対して機密情報を守る必要性が生じてきた．

そのための手法としてデータを暗号化した状態でデータベースに保存し，暗号化したまま問合せを施すプライバシ保護検索手法についてこれまで多くの研究がなされてきた[1][2][3][4][5][7][8]が，対象データのデータ型及び演算によって個別の手法が提案されており，実際に暗号化データベースに対する問合せを実現するためには，属性毎に別々の暗号化を施したり索引を作る必要があった．

そこで我々はブルームフィルタを用い文字列属性と数値属性を対象としたプライバシ保護検索手法を提案している[6][9][10]．本提案手法の特徴は属性毎ではなくタプル毎に索引を構成することである．また，文字列属性に対する完全一致，部分一致および数値属性に対する範囲検索をブルームフィルタによるマッチングという形で統一的に行うことにより，AND で結ばれ

た複数の検索条件を一つの検索条件に変換する．図1(a)が本提案手法におけるデータベース変換の例である．ID と日付(date)と内容(content)で構成されたテーブル Schedules が変換されて，サーバにはタプル t 全体を暗号化した t(etaple)と検索用の索引 t(bfindex)のみがアップロードされる．攻撃者はテーブル Schedules がどのような属性で構成されているかを推測することができない．



(a)データベース変換例

```
select *
from schedules
where date < 1216800000
   and contains(content,'Yoshida')
```

```
select etuple
from s_schedules
where smatch(etuple,bfindex,
  'skejcke','ekcj23k','45kjfk45')
```

(b) 問合せ文変換例

図 1：提案手法によるデータベースおよび問合せの変換例

図 1(b)は本提案手法による問合せ変換例である．date 属性に対する範囲検索条件や content 属性に対するテキスト検索が，索引 bfindex に対するマッチング関数 smatch に置き換えられている．そのため攻撃者は何の属性値を用いてどのような検索条件を指定したかを読み取ることができない．

先行研究[6][9][10]では本提案手法の基礎となるアルゴリズムを提案し，さらにブルームフィルタのビットパターンからタプルの特徴を推測されないよう 2 段階でハッシュ関数を適用する手法を提案し，複数のタプルがある属性において同じ値である場合にもビットパターンが異なるよう対処していた．この手法は攻撃者に対する機密性は高いものの，サーバ側でタプル毎に数回のハッシュ関数の適用が必要となるため，検索時間がかかるという問題があった．

そこで我々はブルームフィルタにノイズを混入させることにより攻撃者の推測を防ぐ手法を考案することとした．本稿ではノイズの混入における撹乱法を提案する準備段階として，先行研究で作成されるブルームフィルタのビットパターンに対する攻撃モデルについて述べ，その効果を検証した上でノイズ付与の戦略を立てる．

ビットパターンを利用した攻撃モデルとして相関ルールマイニングによる推測攻撃を想定する．まず頻出アイテム集合の抽出によりブルームフィルタに含まれる語の発見，相関ルールマイニングによる同一属性に対する語集合の抽出，数値属性の抽出などが想定される．我々はそれらの攻撃モデルに対する防御方法として，ダミービットを戦略的に混入する等のビットパターン撹乱を行う．本稿では 2 節にて先行研究にて述べたのち，第 3 節にてビットパターンに対する攻撃モデルについて述べる．第 4 節ではビットパターン撹乱をする前のデータに対して攻撃モデルを検証し,ビットパターン撹乱法について検討する．

2. 先行研究

2.1 DAS におけるプライバシ保護検索

本節では，DAS における一般的なプライバシ保護検索手法について述べる．図 2 に DAS における暗号化データベース検索の流れを示す．DAS ではデータベース管理者が攻撃者となりうるという点が大きな特徴となる．データベース管理者による攻撃モデルは以下の 3 種類があると考えられており[4]，暗号化データに対する問合せ手法は以下の攻撃に対して安全でなければならない．

(1) Direct Attack：データベース上のデータを直接盗みとる

(2) Indirect Attack：ログやシステムカタログから統計情報を盗み取る

(3) Memory Attack：サーバのメモリにあるデータを盗み取る

これらの攻撃に対してデータを守るため，信頼できるサーバもしくはデータベースを利用するクライアントを通して問合せ及びデータを暗号化したうえでサーバに送信する．サーバでは暗号化したデータに対して問合せを行うため，通常の問合せと同様に正確な（不正解データが含まれない）結果を返すことは難しい．そのため，サーバで一度問合せをして仮の結果を求め，信頼できるサーバまたはクライアントにて復号化をしたのちに再度問合せを行うことによって正確な回答を得る．

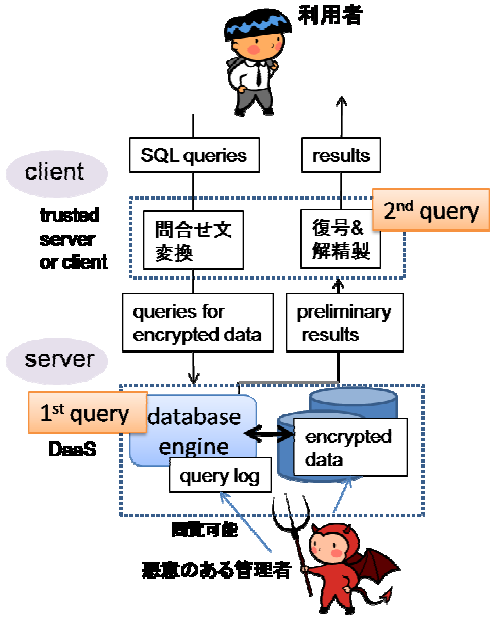


図 2：プライバシ保護検索の流れ

これまで提案されてきたプライバシ保護検索手法は，(1)文字列属性や暗号化文書に対するキーワード検索に対する手法と(2)数値属性に対する比較演算に対する手法に分けることができる．

文字列属性に対する検索手法は[5][8]などが挙げられる．これらの手法は基本的に暗号化された文書に対してあるキーワードが含まれているかを確認することができる手法である．これはたとえば外部サーバでのメール文書のフィルタリング等では有効に働くが，データベースの文字列属性に対して適用する場合はタプル毎の認証処理が必要となるため検索時間がかかるという問題がある．

数値属性に対する比較演算に対する手法は，数値をバケットに分割しバケット番号による問合せに書き換える手法[1][3]と属性値の大小関係を維持するがその値の分散を元のデータと全く異なるものに変換する手法[2]が提案されている．

これまで提案されてきたリレーショナルデータベースに対するプライバシ保護検索では属性毎に上記のいずれかの手法でデータを暗号化する．そのため攻撃者からテーブルのスキーマ構成や問合せ条件が対象としている属性や検索のタイプを隠すことができない．一方で我々の提案手法は索引や問合せログから，スキーマ構成や問合せの傾向を読み取ることができない．

2.2 ブルームフィルタによるプライバシ保護検索

2.2.1 基本手法

本提案手法において，タプルからブルームフィルタ索引を生成する流れを図 3 に示す．まずタプル t から語の集合 Wt={w0,…,wn}を生成する．各語 wi は属性名と属性値からなる．属性値が文字列である場合は，部分一致用に単語毎や n-gram に従って分割し属性名を合わせて語とする．数値属性から語を生成する方法については 2.2.2 項にて述べる．これらの語に対して HMAC（鍵付きハッシュによるメッセージ認証関数）を複数（図 3 では r 個としている）適用し，その値に基づいてブルームフィルタにビットを立てる．

	ID	date	content
t	e560	1216554368	Meeting with Prof. Yoshida

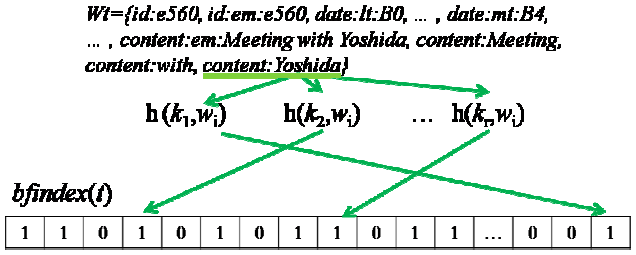


図 3：タプル t から索引を生成する流れ

2.2.2 数値からの語の生成

我々は先行研究[6]にて数値から語を生成する手法を提案している．基本的には数値属性のドメインを複数のバケットに分割し，該当するバケットの名前を属性名と合わせて語とする．図 4 は 374，671，23 とそれぞれに対する語の集合を表したものである．すべてのバケット B={B1,…,Ba}に対して，バケットの上限が値 v より小さい場合「<属性名>:lt:<バケット名>」，バケットの下限が値 v より大きな場合「<属性名>:mt:<バケット名>」それ以外の場合は「<属性名>:em:<バケット名>」という語を追加する．

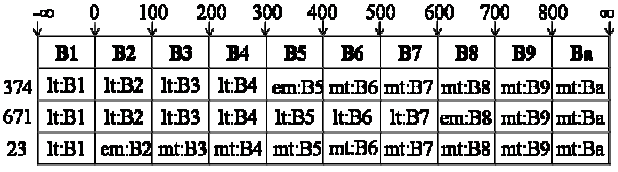


図 4：数値から語を求める．

これを用いて値の代表比較をする場合，例えば 620 より大きな値を調べたい場合は，620 が含まれるバケット(B7)の左隣(B6)に注目し，<属性名>lt:B6 という語を持つタプルを探せばよい．逆に 420 より小さな値を探す場合には，右隣のバケット(B7)に注目し，<属性名>:mt:B7 という語を持つタプルを探す．このようにして数値の比較演算を文字列のマッチングと同様に扱う．

2.3 2 段階ハッシュによるより安全な索引

基本手法で索引を生成した場合，同じ語が含まれている複数のタプルはすべて同じ位置にビットが立っているためそこからデータの傾向などを推測される可能性がある．そこで先行研究[9][10]では基本手法のよう

に一度ハッシュ関数を求めた後，タプルを暗号化した値（etuple）を鍵として再び HMAC を適用する．これにより，複数のタプルが同じ値を持っていた場合も 2 回目のハッシュ関数適用により異なる場所にビットが立ち，攻撃者はビットパターンから元データの特徴を推測することができない．この手法に関する問合せの手順を図 5 に示す．

まずクライアントで検索条件文から検索用の語を生成し，1 回目のハッシュ関数を適用しサーバに送る．サーバ側ではタプル毎に etuple を鍵にしてハッシュ関数を適用し，その値でブルームフィルタへのマッチングを行う．

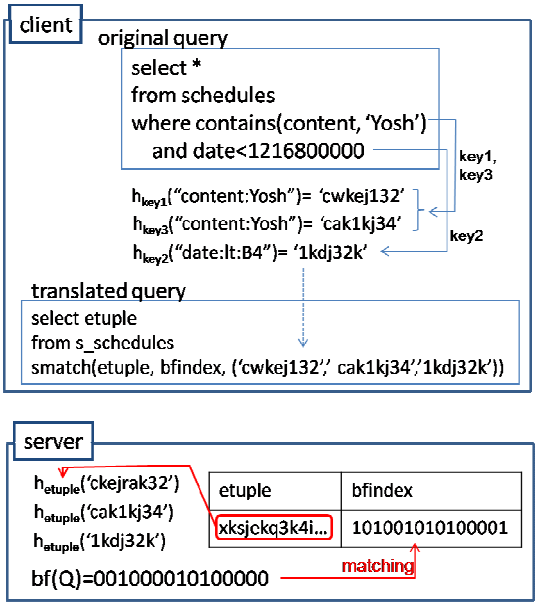


図 5：2 段階ハッシングによる問合せの手順

2.4 先行研究の問題点

先行研究における手法ではサーバ側でタプル毎にクライアントから渡されたハッシュ値の数だけハッシュ関数を適用しなければならず検索時間がかかるという問題がある．文献[10]における実験では Google App Engine でタプル数 1000，検索用のハッシュ値数 5，検索に対する正解率 60%の場合 0.4 秒かかることがわかった．その場合，タプル数が 20000 以上になるとタイムアウト(8 秒)せずに回答できる保証がなくなってしまう．対処法としてはサーバの分散化や複数行で一つの索引を作るなどが考えられるが，根本的な問題の解決とはならずスケーラビリティに問題があった．

3. ビットパターンを利用した攻撃モデル

先行研究における検索時間の問題は 2 段階ハッシングを採用していることにある．前節では，基本手法は同じ値を持つ複数のタプルは同じビットパターンを持つため，テーブルの内容を推測されやすいと延べ，その対策として 2 段階暗号を使っている．

しかし別のアプローチとして，基本手法により作成されたビットマップ bfindex に適切なノイズを混入させて攻撃を防ぐ手法が考えられる．先行研究において我々はビットパターンによる攻撃を具体的に定義してはいない．我々の提案している手法はタプルに含まれる複数の属性の値をすべてひとつの bfindex に収めているため，たとえ同じ値が同じビットパターンになるとしても，bfindex からそのパターンを推測するのは難しい可能性もある．そこで本節では攻撃者の立場になってビットパターン利用した攻撃モデルを立て，その攻撃に対しても堅強であり，かつ検索時間の早い防御法について検討する．

3.1 攻撃の目標

攻撃者にとっての目標を以下のように想定する．

目標 1：同じ値（語）を持つタプルを発見する．

同じ値を持つタプルを検出することにより，そのうちあるタプルの値が漏えいした場合に連鎖的に他のタプルの値を知ることができる．

目標 2：語が属する属性の特性を発見する．

目標 1 では発見した語が何の属性によるどんな値かまでは分からない．ここでは語が属する属性の特性（データ型，分布など）を発見することを目標とする．語が属する属性の分布を発見することにより，語を持つタプルの特性が連鎖的に推測される可能性が高まる．

3.2 想定される攻撃

攻撃 1：語を構成するビットセットの推測

基本手法ではタプルを構成するそれぞれの語に対して，k 個のハッシュ関数を適用しその値に基づいてブルームフィルタにビットを立てる．そこで攻撃者はどの k 個のビットが語に対応しているかを推測すると考えられる．基本手法ではタプル内に複数の語が含まれているため単一の bfindex では推測は難しいが，テーブル内のすべての bfindex を用いて頻出アイテム集合を求めることによりビットセットを発見し、前項にのべた目標 1 が達成される恐れがある．

攻撃 2：相関ルールマイニングで語の関係を推測する

各タプルにおける bfindex を攻撃 1 で得られた語の集合に置き換えて語の相関をマイニングすることにより前項に述べた目標 2 のうち以下の特性を発見することができる．

I. 語 A,B に対して A⇒B の確信度が 0 である場合，

A,B は同じタプルに同時に登場することがない．このような同時に登場しない語集合の支持度が 1 に近い場合はそれらの語は「カテゴリ型」属性の値であると推測される．ここでカテゴリ型属性とは例えば性別や出身地（都道府県）など単一の語で表され，カテゴリ型属性において同じ値を持つタプルを発見することにより，タプル間の相関を発見しやすくなる．推測方法としては語の非共起の関係を無向グラフで表し，その中のクリークを発見する手法が考えられる．クリークを構成する語が現れるタプルの合計数がテーブルのタプル数に近いとき，あるカテゴリ型属性の値はそれらの語のいずれかを取りえると推測できる．

II. 語 A,B に対して A⇒B の確信度が 1 である場合，語 A が登場するタプルは必ず B も登場するという包含関係が生じる．その場合，語 A,B は同じ数値属性に対する語である可能性が高い．数値属性の場合 2.2.2 項の説明からわかるように，登場するタプルに包含関係があり，例えば lt:B4 を含むタプルは必ず lt:B3 を含むという関係があるからである．そのような包含関係にある一連の語を見つけることにより，語の間の順序関係を推測することができる．推測方法としては，各語をノードとし，語 A,B が出現するタプル集合 T(A),T(B)の包含関係が T(A)⊇T(B)であるとき A から B への有向辺を張ったグラフを生成し，その中から部分連結グラフを抽出する．

III. 事前にテーブルや属性のドメインに関する情報（属性間の関係など）が攻撃者に漏れている場合，相関ルールマイニングの結果と事前情報を照らし合わせてどの語がどの属性に対応するかを推測する可能性がある．

上記のうち I,II はテーブルのドメインに特化しない共通した攻撃手法，III はドメインに特化した手法である．本稿では I,II について取り上げることとする．

4. 攻撃モデルの検証

本節では前節にて想定した攻撃モデルの検証を行う．複数のデータセット（人工データとリアルデータ）を用意し，前節に述べた攻撃 1，2，3 によってどの程度ビットパターンが推測されるかを検証する．

4.1 データセット

データセットは 2 種類の人工データと 1 種類のリアルデータを用意することとする．

TableA: 単純な構成のテーブル

属性数：3
タプル数：100,000
データ型：カテゴリ型．
全ての属性は 1 つの単語からなり，濃度は 50．
またデータの分布はすべて一様分布に従う．

TableB: 実際の構成に近いデータ

属性数：9
タプル数：100,000
データ型
・文章型（5～10 単語，単語は zipf 分布に従う）
…1 属性
・数値型（正規分布に従う．バケット分割数は 10）
…2 属性
・カテゴリ型（1 つの単語からなり，濃度は 50）
…3 属性（正規分布に従う）
＋2 属性（一様分布に従う）

TableC: リアルデータ

UCI ML リポジトリの Adult データセット[10]を使用．
属性数：14
タプル数：48842
データ型：数値型…5，カテゴリ型…9

bfindex 索引を生成するにあたり，1 語に適用するハッシュ関数の数を 5，索引のビット長を 2,048 とする．

4.2 検証 1: 頻出アイテム集合検出によるビットセットの推測

Table A, B, C から生成された bfindex 索引に対して頻出アイテムセットを抽出する．適切な最小サポート値を確かめるため，ここでは頻度分布を眺める．図 7 に TableA における頻度分布を示す．

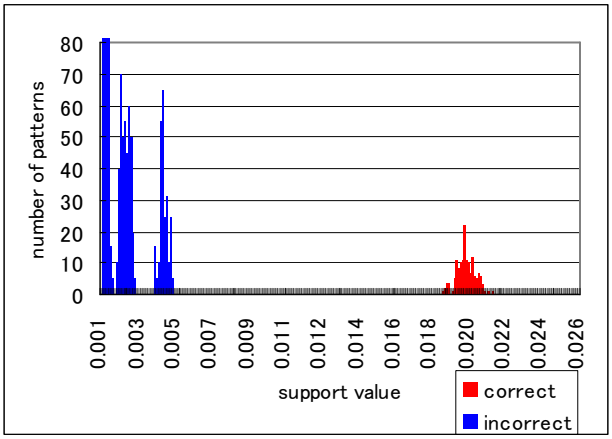


図 7：TableA（属性数 3）における頻度分布

この図では x 軸にサポート値を表し，そのサポート値を持つパターンの数を y 軸に示している．暗号前のデータベースから正解ビットパターンを抽出しその分布を correct としてプロットしている．また暗号化したデータベースの bfindex からビットが 1 となる位置をアイテムとしてリストアップし，1 タプルを 1 トランザクションとして 100000 トランザクションから抽出さ

れるアイテム集合の頻度分布を調査した．暗号化の際1つの値に対して5つのハッシュ関数を適用することを想定し，かつ攻撃者もハッシュ関数の数が事前に判明していることを想定してアイテム数を5に固定して抽出を行った．図7のincorrectはそのようにして抽出された中からcorrectに含まれていないパターン，攻撃者にとってはノイズとなるパターンである．図7を見ると，ほとんどのノイズパターンはサポート値が少なく正解パターンと分離してしまっていることがわかる．つまり最小サポート値0.019で頻出アイテム集合を抽出すると正解パターンのみ抽出できてしまうことがわかった．TableAでは属性数が少なくすべて濃度も同じ一様分布で構成されているためTableA'1，TableA'2として以下の人工データでも検証を行った．

TableA'1：正規分布による属性数3のテーブル

TableA'2：属性数40のテーブル（一様分布）

また数値属性が含まれるTableBの人工データでも検証をいった。それぞれの頻度分布を図8，図9，図10に示す．図8をみると，正規分布だと頻出アイテム集合を取ることによって正解パターンは抽出されるものの，半数はノイズパターンにまぎれるため再現率は半減することがわかる．また属性数の多いTableA'2はノイズパターンが増えて正解パターンの分布が混合しており，容易に分離することが難しくなってくる．2つの数値属性を含むTableBでは大量のノイズパターンが正解パターン数を上回り抽出がかなり難しい状態になっている．

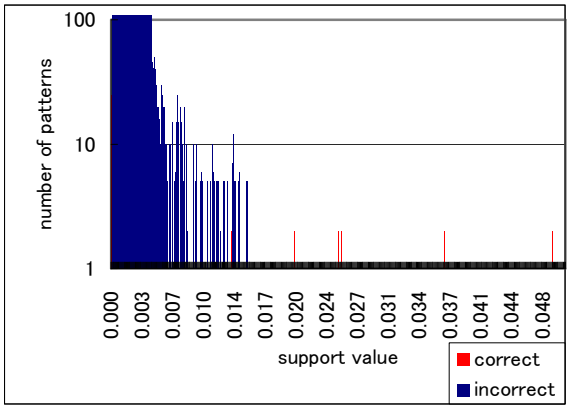


図8：TableA'1 (正規分布，属性数3)の頻度分布

これらの結果から以下のように考察する．

・一様分布の場合は属性数が少ないとノイズパターン数の頻度が少なく容易に分離できてしまうが，属性数が増えるにしたがってノイズパターンの数が正解パターンを上回るようになり、正解パターンを隠すことができる．

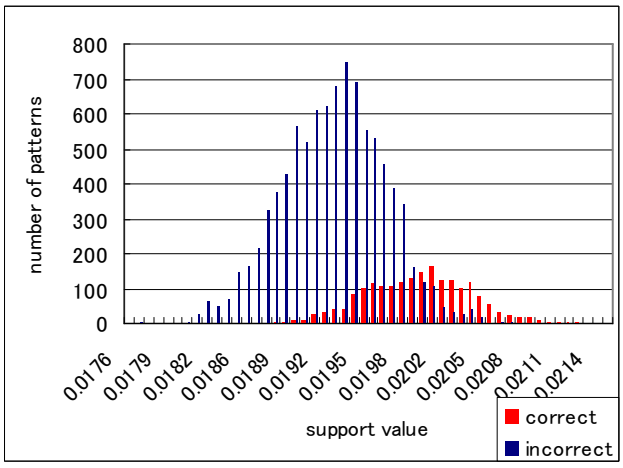


図9：TableA'2（属性数40）の頻度分布

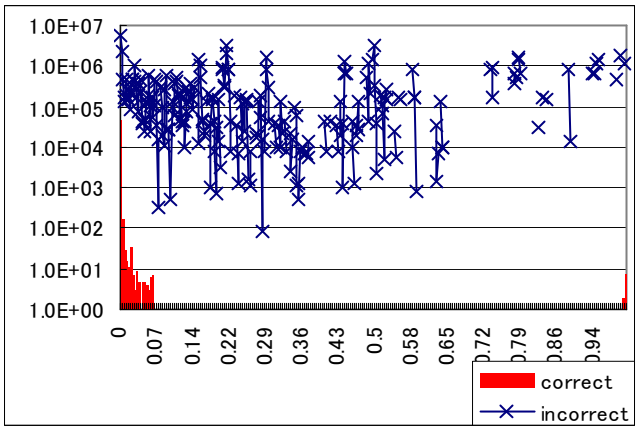


図10：TableB(数値属性を含むテーブル)の頻度分布

・正規分布（など一様分布以外）の場合は正解でも頻度の少ないものが多く，頻度が高いごく一部の正解パターン以外のほとんどはノイズパターンで隠蔽することができる．

・数値属性が含まれると多くのタプルに複数の語が含まれることが多くなり，複数の語からなるビット集合の組合せが爆発的に増えるためノイズパターンが圧倒的に多くなる．そのため数値属性が含まれたテーブルではノイズパターンが正解パターンを大きく上回り頻出アイテム集合の抽出では正解パターンを見つけることができなくなる．

4.3 検証2：同一の属性に対する語の推測

前節に述べた攻撃2におけるIの成功率を検証する．本来の攻撃では抽出したビットセットを語に置き換えた上で同時に出現しない語の集合を求める．ただしここでは実験1の結果と実験2の結果を分離するため，暗号前のデータベースから求めた語を用いて実験した．

TableBでは非共起の関係を求めた結果ノード数673，

エッジ数 155919 のグラフが生成された．ノード数 673 の完全グラフの辺数の約 3 割が使われており，密なグラフである．その中から抽出されたクリークはノード数が 477 であり，予測したような同一の属性の値だけではなく，多くの値を混合していた．属性数が多い場合はこのような手法で語を属性ごとに分類するのは難しいと判断される．

TableC の場合は 1500 タプルを読み込んでグラフを作成した時点でクリークを検出したところ 929 のノードからなるクリークが検出された．これも全く異なる属性の値から構成されており，これを利用して属性の分類を行うことは難しいと判断できる．

4.4 検証 3：数値属性に対する語の推測

前節に述べた攻撃 2 における II の成功率を検証する．本実験では数値属性が含まれている TableB, TableC でのみ実験を行う．確信度が 1 である語のペアを発見し，その相関関係をグラフで表し，部分連結グラフを抽出したときにそこから数値属性が抽出できるか検証する．

TableB から抽出した部分連結グラフを図 11 に示す．TableB に含まれる数値属性に対応する語の関係が抽出できていることがわかる．

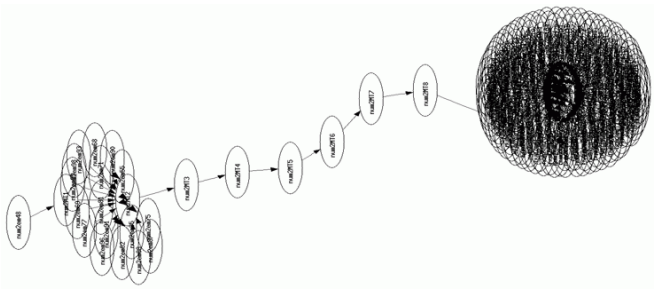


図 11　数値型の属性値を表す語の検出（TableB）

TableC（リアルデータ）による検証の結果を図 12 に示す．この場合にはさらにはっきりと，5 属性中の 3 属性が抽出されている．

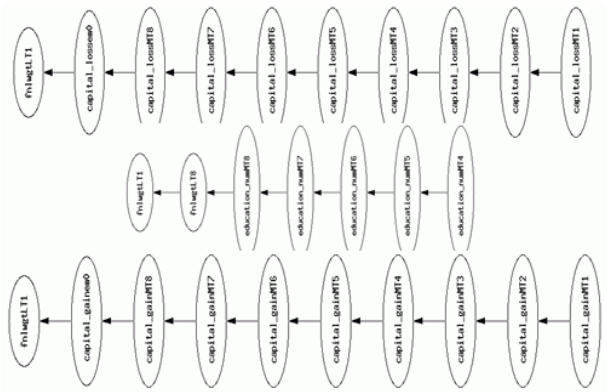


図 12　TableC から抽出された数値属性

4.4　考察

4 節における検証から以下のことが考察される．

(1) ビットパターンからの語の抽出はカテゴリ型のみから構成され属性数が少ない場合には頻出アイテム集合を使って抽出が可能である．属性数が多い場合や数値属性を含む場合はノイズパターンが多くなり抽出が難しくなる．特に数値属性は複数のパターンが多くのタプルに同時に含まれるため非常に多くのノイズパターンが抽出されるために正解パターンの推測が非常に難しくなる．

(2) 抽出された各語が何の属性に対するものかを推測する際，攻撃 2 の I で述べた非共起関係を用いてカテゴリ属性を抽出することは容易ではないが，数値属性の抽出は非常に精度高く抽出することができる．

(3) 検証ではビットパターンからの語の抽出と，語からの属性ドメイン情報の抽出を別々に検証したが，実際はビットパターンから語を抽出できないとその次に進むことは難しい．攻撃 2 では数値属性の関係が比較的容易に抽出されてしまったが，その一方で攻撃 1 では数値属性によりノイズパターンが多く含まれるため語の抽出が難しいため攻撃 2 まで進めることが難しいと推察される．

(4) しかしながら数値属性により圧倒的にノイズパターンが増えることがわかると，今度はほとんどのタプルに登場するビットを分離して頻出アイテム集合を求める手法が考えられる．

以上のことから攻撃に対する防御法として以下の手法を提案する．

手法 1：ダミー属性の追加

数値属性がテーブルに含まれることにより頻出アイテム集合抽出によって語を推測することが難しくなる．そこで数値属性が含まれないテーブルに対して数値属性をダミーとして混入することによって語の推測を防御する．しかしながら数値属性のみを分離される可能性もあるため，文字列型・カテゴリ型のノイズ用属性も混在させることを考える．

手法 2：複数タプルに対して 1 つの bfindex を生成

攻撃モデルはタプル毎の bfindex のパターンをマイニングすることによりテーブルの特徴を推測してきた．そこで複数タプルをまとめて暗号化し，それに対して一つの bfindex 索引を作ることによりビットパターンを撹乱させる．この手法は適用が容易である上にある一定の効果が見込めるが，複数のタプルのうち一つでも問合せの条件に該当するものがあった場合全てのデータをクライアントに送らなければいけないため，いわゆる擬陽性が高くなる恐

れがある．

手法 3：数値属性の語の推測を防ぐダミー語の混入

4.3 節にて検証した結果により，語の集合から数値属性を抽出するのは容易にできてしまうことがわかった．そこで手法 1 で混入させるダミー語を決定する際にタプルの包含関係の抽出を妨害するための混入戦略を立てる．

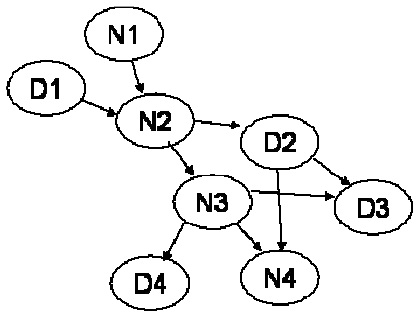


図 13：数値属性候補の撹乱法の一案

現在検討している手法では図 13 に示すように本来の数値属性の語（N1,N2,N3,N4）の周辺にダミー語を混入させ，数値属性候補を複数抽出させることで本来の数値属性を判別しにくくする．その際，少ないダミー語で効率的に候補を生成する手法を考案する必要があると考えている．

5. まとめと今後の課題

DAS モデルにおけるプライバシ保護検索として提案しているブルームフィルタによる検索手法について，先行研究で用いている 2 段階ハッシュ関数適用を使わずにビットパターン撹乱手法を用いる方法を検討することを目指し，その手始めとしてビットパターンによる攻撃モデルの方針を立て，その検証を行った．その検証の結果からダミー属性を混入させて推測を撹乱させる防御法の戦略を立てることができた．

今後はビットパターン撹乱法を提案し，その検証を行う必要がある．その際ビットパターンを混入させることによって擬陽性が生じることも考慮しなければならない．本研究ではサーバ側の問合せで擬陽性が生じることはすなわちサーバからクライアントへの結果送信，クライアントでの復号化，クライアントでの 2 回目の問合せのコストに影響が出ることになる．これらのコストとサーバ側での問合せのコストを総合的に考えた上で安全かつ高速なプライバシ保護検索を実現するための適切な撹乱法を確立していきたい．

文　献

[1] Hacigumus H., Iyer B. R., Li C. and Mehrotra S.: Executing sql over encrypted data in the database-service-provider model, Proceedings of the 2002 SIGMOD International Conference, pp. 216–227 (2002).

[2] Agrawal R., Kiernan J., Srikant R. and Xu Y.: Order preserving encryption for numerical data, Proceedings of the 2004 SIGMOD International Conference, pp. 563 – 574 (2004).

[3] Hore B, Mehrotra S. and Tsudik G.: A privacy-preserving index for range queries, Proceedings of the 30th International Conference on Very Large Data Bases, pp.720—731 (2004).

[4] Ting Yu and Shushil Jajodia: Secure Data Management in Decentralized Systems, Springer-Verlag New York Inc, p.462 (2006).

[5] Boneh D, Crescenzo G.D., Ostrovsky R. and Persiano G, Public Key Encryption with Keyword Search, Proceedings of EUROCRYPT ’04, vol.3027 LNCS, pp. 506—522 (2004)

[6] 新井裕子，渡辺知恵美: データベースアウトソーシングにおけるプライバシ保護に考慮した範囲検索法，電子情報通信学会 第 19 回データ工学ワークショップ，C1-1 (2008).

[7] Tingjian Ge, Stanley B. Zdonik: Answering Aggregation Queries in a Secure System Model. , Proceedings of VLDB 2007, pp.519-530 (2007).

[8] Bellovin S. and Cheswick W. : Privacy-enhanced searches using encrypted bloom filters” (2004).

[9] 渡辺知恵美，新井裕子：DAS におけるスキーマ情報と複合的な検索条件を隠ぺいしたプライバシ保護検索手法，情報処理学会研究会報告，No.2008-DBS-146，Vol.2008，No.88，pp.163—168 (2008).

[10] Watanabe C. and Arai Y.: Privacy-Preserving Queries for a DAS model using Two-Phase Encrypted Bloomfilter, Proc. of International Conference on Database Systems for Advanced Applications (2009) (to appear)