

# 暗号化データベースにおける 統計情報に強いノイズ付きブルームフィルタ索引の生成

柿澤美穂 (指導教員：渡辺知恵美)

## 1 はじめに

データベースの機能をネットワーク経由で利用できる DBaaS (DataBase as a Service) は、クラウド上の外部の管理者にデータベースサーバの管理・運用を委託する。利用者はこの管理者からもプライバシーを保護し、機密情報を安全に保存・検索したいと要求する。この問題に対する解決策として、データを暗号化した状態でデータベースに保存し、暗号化されたまま問合せを行う暗号化データベース管理システム (EDBMS) [1][2] がある。我々は EDBMS で用いられる索引に、複数の属性から作られる一つの多属性索引を用いる手法を提案してきた。先行研究 [3] では、多属性索引の一つとしてブルームフィルタを用いたプライバシー保護検索手法を提案した。しかし、カテゴリ型の属性かつ属性値の頻度分布に偏りがあるといった特定の場合に、ブルームフィルタ索引から属性の頻度情報が明らかになり、元の値を推測されてしまう恐れがある。

そこで我々は、頻度情報を用いた攻撃モデルとその安全性の指標を定義し、その指標を満たすためのノイズ付与戦略を提案してきた。本研究では、文献 [3] に基づいたノイズ付きブルームフィルタ索引の生成システムを提案する。

## 2 ブルームフィルタを用いた集合索引

### 2.1 基本概念

我々は、複数属性を対象とした多属性索引の一例としてブルームフィルタ索引を提案してきた。属性名と属性値の組に対する複数のハッシュ値からビットパターンを生成し、各タプルのビットパターンの論理和をとったものを、ブルームフィルタ索引として用いる。この索引を用いることで、属性値に対するキーワード検索が可能となる。またタプルに含まれる属性値の情報を隠しているため、統計情報による攻撃に強い。

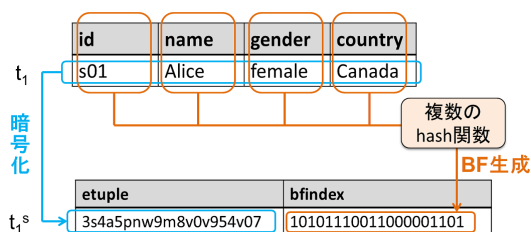


図 1: ブルームフィルタ索引の生成例

### 2.2 安全性に関する問題点

多属性索引であるブルームフィルタ索引は、単属性索引を用いる手法よりも統計情報による攻撃に強い。しかし、カテゴリ型の属性や値の頻度分布に特徴がある属性に関しては、元の値を隠しきれず安全性に問題が出る場合がある。データベース管理者が保存されているテーブルの統計情報を知っていて、生成されたブルーム

フィルタ索引のビットパターンから属性値の頻度情報を得ることができる場合、統計情報と頻度情報を比較することで値の特定が行われてしまう。

### 2.3 攻撃モデル

多属性索引は、ある程度統計情報による攻撃に強いが、完全に攻撃を防ぐことが出来る手法ではない。我々は、攻撃者による統計情報を用いた攻撃に強い手法を提案するため、カテゴリ属性に対する攻撃モデルを定義し、その攻撃を防ぐためのプライバシー保護指標を提案した。

攻撃モデルとして、攻撃者が元の値を推測する過程には 2 段階ある。(1) ビットパターンの候補を列挙する。(2) 持っている統計情報とビットパターンの割合を比較することで属性に対応するビットパターンを特定し、さらに索引値に対応する元の値を特定する。

この攻撃モデルに対し、我々はビットパターン集合を利用した安全性の指標を定義している [3]。暗号化されたリレーションの多属性索引と攻撃者が持つ統計情報を用いて、属性に対するビットパターン候補を求めた時、必ず  $\gamma$  個以上のビットパターン集合候補が求められるようにする。攻撃者は、自分が持っている統計情報を用いてビットパターン集合候補を  $\gamma$  個以下に絞り込むことが出来ない。そこで本研究では、与えられたテーブルの統計情報をとって、それに応じて最適な  $\gamma$  の値になるようノイズを付与するための戦略を提案する。

## 3 ノイズ付きブルームフィルタ索引生成システム

上の攻撃モデルに対して定義されたプライバシー保護指標を用いて、ビットパターン集合候補を最適な個数にするための戦略を提案する。戦略を適用してビットパターン集合の候補を増やすことで、攻撃者はどのビットパターン集合が正解か判断することが難しくなり、元の値を特定できない。

**戦略 1** 頻度の類似した属性集合で索引を作ることによって、攻撃者が推測するビットパターン集合候補を増やす。

**戦略 2** 頻度情報が等しい  $\gamma - 1$  個の擬似属性を多属性索引に追加し、ビットパターン集合候補を増やす。

**戦略 3** 誤検出率を増加させ、多属性索引から抽出される特徴の頻度を増加させる。

**戦略 4** 頻度の高い属性値を分割し、属性集合の頻度を少なく見せかける。

利用者の用意するテーブルの情報に応じて、適切な戦略を選んで適用する。戦略の詳しい適用方法は以下の例によって示す。

例えばここに、属性 id, name, gender, country を持つ学生情報の入ったテーブルがあるとします。システム利用者がこのテーブルをシステムに渡すと、システムは属性毎に統計情報を作成する。id や name は全て同じ頻度であり索引から特徴をつかむことは難しいため、gender と country の 2 属性を選択して索引を生成する。

しかし、ただ属性毎に統計情報を作成するだけでは値の分布に特徴があるため、ブルームフィルタ索引を生成しても攻撃者から元の値を推測されてしまう可能性がある。そこで、先程の戦略を適用する。まず戦略 1 を適用し、頻度の高い順に並べ替え同じような頻度分布をグループ化する (図 2)。そして、戦略 2 を適用し、各属性に同じ数の擬似属性を混ぜる。ここで、属性 male のような突出した属性があると特定されてしまうおそれがある。そこで他の属性と同程度の頻度にするため戦略 4 を利用する (図 3)。図 3 のように、ブルームフィルタを作成するのに最適な頻度分布であると利用者が判断すると、システムはブルームフィルタの長さの見積もり式からブルームフィルタ索引の長さを算出し、ブルームフィルタ索引が生成され利用者に渡される。

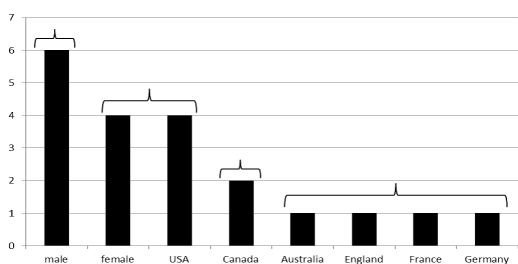


図 2: 戦略 1 適用後

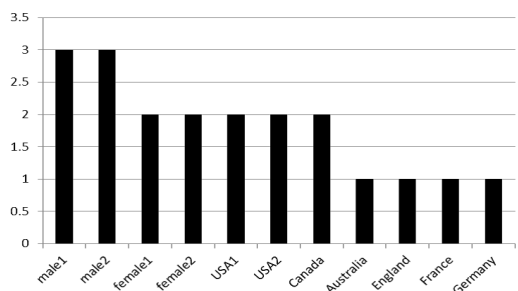


図 3: 戦略 4 適用後

この一連の流れをシステム化したものが図 4 である。まず利用者は、保存したいテーブルをシステムにアップロードする。システムは、与えられたテーブルから頻度情報を作成する。そして、テーブルに含まれている属性を提示し、利用者に索引を作るべき属性を選択させる。利用者は、頻度情報が明らかになると予想されるカテゴリ型の属性や偏った頻度分布を持つ属性を選ぶ。システムは、選ばれた属性のヒストグラムを生成する。さらに、利用者は誤検出率を入力する。すると、システム側では戦略 1 の適用を開始し、さらに属性の頻度情報を見てどの戦略を適用するかを自動的に判断し、戦略 2 から戦略 3、もしくは戦略 4 へと適用していく。そして戦略適用後のヒストグラムを利用者に提示する。利

用者は、頻度情報が隠れるのに十分なヒストグラムであると判断すると、ブルームフィルタ索引生成を承認する。するとシステムは、入力された誤検出率を用いてブルームフィルタの長さを決定し、ブルームフィルタ索引を生成する。生成されたブルームフィルタ索引が利用者に返されて、本プログラムは終了する。

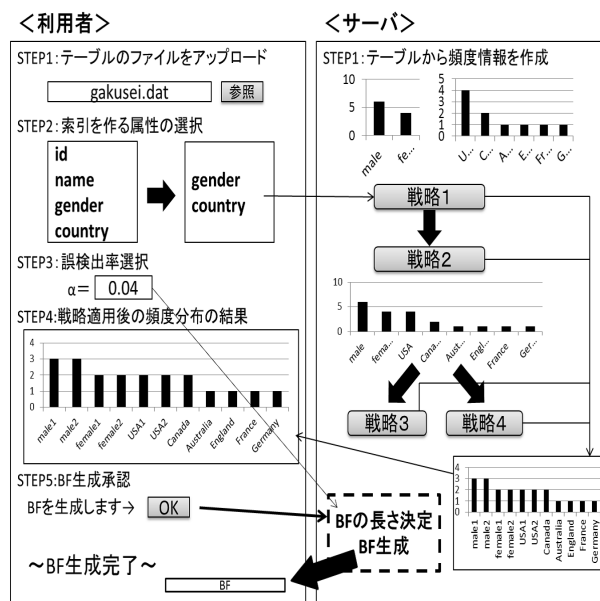


図 4: システム概略図

## 4 まとめと今後の課題

クラウド上の DBaaS において、利用者はデータの管理者からも機密情報を隠すことを要求する。そこで、データを暗号化してサーバに保存し、暗号化されたまま問い合わせを行う EDBMS が提案されてきた。本研究では、先行研究で定義された、第三者による統計情報を用いた攻撃からデータを守る安全性の指標を用いてノイズ付きのブルームフィルタ索引を生成するプログラムを提案し、実装した。今後、本システムで生成されたブルームフィルタ索引を実際の EDBMS に適用できるように、システムを改良していく予定である。

## 参考文献

- [1] H.Hacigumus, B.Iyer, C.Li, and S.Mehrotra: "Executing SQL over Encrypted Data in the Database-Service-Provider Model", Proceeding of the ACM SIGMOD International Conference on Management of Data, pp.216-227,(2002).
- [2] C.Watanabe and Y.Arai: "Privacy-Preserving Queries for a DAS model using Two-Phase Encrypted Bloomfilter", Proceeding of International Conference on Database Systems for Advanced Applications (2009).
- [3] 金子 静花: "Encrypted DBMS のための安全かつ高速な多属性索引の諸検討", お茶の水女子大学大学院人間文化創成科学研究科理学専攻修士論文,(2013)