7X-05

完全準同型暗号を用いた FP-growth による頻出パターンマイニングの分散環境への実装方式の検討

種村真由子 †　　小口正人 †

† お茶の水女子大学

1. はじめに

クラウドサービスが発展し，大量のデータをクラウド上で処理する機会が増えている．その中で．医療データ等の大規模な機密データをクラウド等に外部委託して処理する場合には，送信する情報の漏えい対策が重要となる．完全準同型暗号（FHE）は，暗号文同士の加算と乗算が成立する暗号であり [1]，これを利用することで委託先サーバに平文データを渡さずに処理を行うことが可能になる．FHE の応用例として，Apriori アルゴリズムによるトランザクションデータの頻出パターンマイニングを行った Liu らの P3CC や，その高速化を行った先行研究の [2][3][4] などがある．本研究では，FHE を使用し，同様にパターンマイニングを行う他アルゴリズムの FP-growth による頻出パターンマイニングの実装に取り組む．

2. 完全準同型暗号

完全準同型暗号は，暗号文同士の加算，乗算が成立する公開鍵暗号であり，一般に計算量が大きいことが知られている．各暗号文は，暗号の解読不可能性を高めるため，ノイズパラメータという値を付加されている．暗号文同士で計算を行うたびにこの値は増加し，特に乗算を行う際に増加が大きい．ノイズパラメータは，暗号文に対して十分に小さい必要があり，閾値を超えると復号が不可能となる．加算や乗算を行う前にブートストラップという処理を行うことで，計算後のノイズを十分に小さくすることができる [1]．復号をせずに計算を行えるという特長により活用が期待されているが，広い実用化に向けては，計算量の大きさに対しての工夫が必要となる．

3. 頻出パターンマイニング

3.1 頻出パターンマイニングと Apriori

頻出パターンマイニングは，データマイニングの一種であり，大量のトランザクションデータの中から，頻出であるアイテムの組み合わせや条件を取り出す手法である．応用例として，購買データから頻繁に購入される商品の組み合わせを抽出するマーケットバスケット分析が挙げられる．代表的なアルゴリズムでは Apriori がある．Apriori は，アイテム数 1 から小さい順に頻出アイテムの組み合わせを挙げていく方式である．

Study of Implementation Method of Frequent Pattern Mining with FP-growth in Distributed Environment using Fully Homomorphic Encryption
Mayuko Tanemura†
Masato Oguchi†
†Ochanomizu University

3.2 FP-growth

FP-growth は，本研究で用いる頻出パターンマイニングのアルゴリズムであり，先述の Apriori に対して，深さ優先探索のような形で頻出パターンの抽出を行う．FP-tree という頻出アイテムが格納されたデータ構造を再帰的に走査することで，頻出パターンを求める．FP-growth アルゴリズムの概要は以下である．

・ FP-tree の構築

FP-tree は，トライ木に類似したデータ構造である．トランザクションに含まれるアイテムで，出現頻度が閾値を超えたものを格納する．アイテムは頻出なものからルートに近い位置に格納され，各ノードはアイテムの種類と出現頻度を保持する．

・ FP-tree の走査

作成した FP-tree において，最も出現頻度の低いアイテム（A とする）を含む枝から順に走査する．A に至るまでのアイテムと頻度を格納したリストから条件付き FP-tree を作成し，走査する．tree に分岐が存在する場合には，さらにその中で最も出現頻度の低いアイテムに対し FP-tree を構築し，その tree を再帰的に走査する．分岐が存在しない場合，現在の tree に含まれるアイテムの組と条件付けに使用したアイテムを頻出パターンとして保存する．これを tree 全体の走査が完了するまで繰り返す．

4. 提案システム

本研究では，以下のようなシステムを提案する．

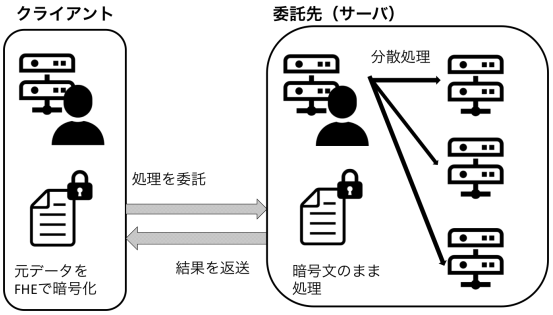

図1　システムの概要

図 1 のシステムにおいて，クライアント側で暗号化したトランザクションデータをサーバに送信し，FP-growth のを使用した頻出パターンマイニングにおける一部の計算を委託する．実装は C++ で行い，完全準同型暗号のライブラリは Helib を用いている．サーバ側はマスタ・ワーカ型の分散・並列処理を行う．FHE を扱うためのライブラリは Helib，分散・並列処理のライブラリは MPI である．

Copyright ©2019 Information Processing Society of Japan.
All Rights Reserved.

処理の流れの概要は以下である．

1. クライアントでのデータの準備
   (a) 元データを FHE で暗号化する．
   (b) アイテムの候補を作成し，サーバに送信する．
2. サーバでの委託処理
   (a) クライアントからアイテムのサポート値を計算し，クライアントに結果を返送．このときマスタ・ワーカ型の分散処理を行う．
3. クライアントでの FP-tree 構築
   (a) サーバからファイルを受信，サポート値の条件を満たしたアイテムを取り出す．
   (b) アイテムを頻出順に並び替え，非頻出アイテムを除いたトランザクションを作り直す．
   (c) FP-tree を構築する．
4. クライアント側での FP-tree 走査
   (a) 出現回数の低いアイテムを含むアイテムを含むアイテムの組から走査をする．
   (b) 走査中の木の分岐がある場合は条件付 FP-tree の作成をし，再帰的に tree を走査．

5. 実験

クライアントプログラムとサーバプログラムを，同一のマシン上で動作させ，クライアントプログラムにおいて実行時間を測定した．

5.1 実験概要

アイテム数 10, 20, 30 の場合に対して，それぞれトランザクション数を 1000 から 5000 まで 1000 ごとに変化させた場合の実行時間を，クライアント側プログラムにおいて測定した．

実験で用いた計算機は，OS は CentOS6.9，CPU は Intel® Xeon® プロセッサー E5-2643 v3，メモリが 512GB である．

実験のためのデータセットの生成は，IBM Quest Synthetic Data Generator で行った．作成データは，アイテム数とトランザクション数を変化させ，平均アイテム長 5，最大パターンの大きさを 5，生成されうるパターン数を 100 に固定した．また，この実験において，サーバ側のワーカ数はすべて 1 としている．

5.2 実験結果

サポート値計算，FP-tree 構築，FP-tree の走査，全体の実行時間をそれぞれ測定したものを図 2に示す．「その他」の時間には，データの暗号化と復号の処理時間，サーバとの通信時間が含まれる．

サポート値計算時間はアイテム数，FP-tree の構築時間はトランザクション数にそれぞれ比例している．その他の処理はアイテム数やトランザクション数ではほぼ変化しない．一方，FP-tree の走査時間は，アイテム数とトランザクション数の増加に伴い大きく増加している．特にアイテム数の増加により，走査時間が長くなっており，最大ではトランザクション数 3000 の場合, アイテム数が 10 から 30 に増加すると，走査時間は約 5.69 倍になっている．これは再帰的に条件付き FP-tree を生成する部分においての処理時間が長くなっているためであると考えられる．
また，アイテム数をさらに増加させた場合，処理時間全体は指数関数的に増加していくと考えられる．

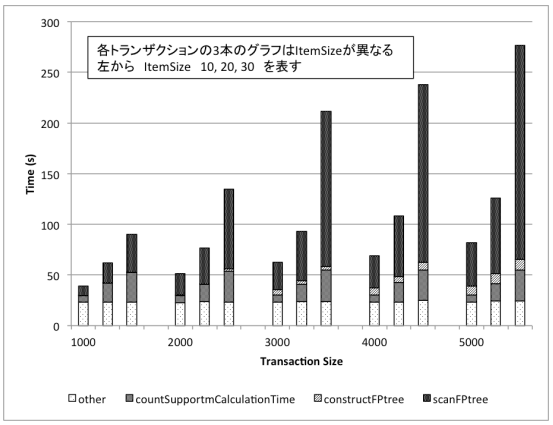

図2 実験結果

6. まとめと今後の予定

完全準同型暗号を使用した FP-growth アルゴリズムによる頻出パターンマイニングのシステムを実装し，実行時間の測定を行った．その結果，FP-growth の処理時間がアイテム数とトランザクション数の増加に伴い非常に大きくなってしまったことから，より大きなサイズの入力データに耐えうる，効率的な処理方法の検討が必要である．また，FP-growth の tree 走査部分の処理を部分的にサーバへ委託することで，実行時間とクライアント側の使用リソースを削減に取り組む予定である．また，今回の実験は同一ネットワーク内のみで行ったが，外部のネットワークからサーバに接続する実験の実施を考えている．

謝辞

本研究は一部，JST CREST JPMJCR1503 の支援を受けたものである．
本研究を進めるにあたり，早稲田大学の山名早人先生をはじめ，多くの先生方のご協力をいただいております．深く感謝申し上げます．

参考文献

[1] C. Gentry: A Fully Homomorphic Encryption Scheme. Doctoral dissertation. September 2009.
[2] 高橋卓巳, 石巻優, 山名早人: SV パッキングによる完全準同型暗号を用いた安全な委託 Apriori 高速化. DEIM Forum 2016 F8-6. 2016.
[3] 今林広樹, 石巻優, 馬屋原昂, 佐藤宏樹, 山名早人: 完全準同型暗号による安全頻出パターンマイニング計算量効率化. 情報処理学会論文誌データベース (TOD), Vol. 10, No. 1. 2017.
[4] 山本百合, 小口正人: 完全準同型暗号を用いた秘匿データマイニング計算のデータベース更新時の分散処理による高速化. dicomo2018. 2018.

Copyright ©2019 Information Processing Society of Japan.
All Rights Reserved.