小口研究室　研究紹介（2020年度）
(お茶の水女子大学理学部情報科学科)

完全準同型暗号を用いたFP-growthによるデータマイニングのアルゴリズム改良手法の提案と評価
（研究担当：種村　真由子）

研究背景

- **ビッグデータの利活用**
  IoT分野をはじめ、各種ビジネスなどにおいて大規模データの収集・活用が進んでいる
- **大規模データ処理の外部委託**
  膨大なデータ処理には高性能な計算機が必要であり、外部（クラウド等）に計算を委託するのが現実的
- **情報のセキュリティ管理の必要性**
  信用できない外部サーバにプライバシに関わるデータを平文で置くことは危険である

→外部に平文データを公開せずに委託処理を行う

研究課題・システム概要

完全準同型暗号（FHE）を用いた頻出パターンマイニングのFP-growthによる実装の改善
クライアントと委託先サーバを想定した、頻出パターンマイニングの委託処理システムを作成する。
委託先サーバはFHE暗号化されたデータを復号せず処理し、クライアントに結果を返送する。

本研究では、FP-growthアルゴリズムを使用したプログラムの実装を改善する

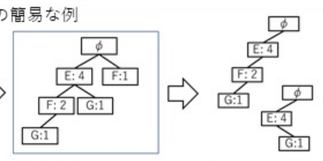
●完全準同型暗号
暗号文同士で加算が成立する**加法準同型性**と、
乗算が成立する**乗法準同型性**をもつ公開鍵暗号方式である。

●頻出パターンマイニング
トランザクションの集合から、一定以上の頻度で出現するパターンを抽出する手法。代表的なアルゴリズムとしてApriori、FP-growthがある。関連研究ではAprioriが使用されているが、本研究ではFP-growthを用いる。

- FP-growthの簡易な例

パターン	頻度
E	4
F	3
G	2

パターンを頻度順に並べる



FP-treeを構築

FP-treeの部分木を再帰的に作成
頻出アイテムセットを抽出

提案システム

クライアント

（2）データの送信準備
計算にあたり必要なパラメータ等をサーバに送信．

（3）データの送信準備
クライアント側で保持しているトランザクションデータ，最小サポート値に対応するアイテムの出現回数の閾値をFHEで暗号化，サーバに送信．

（5）FP-tree構築
受信ファイルを復号，出現回数と閾値の差の正負によって，条件を満たすアイテムを抽出し，FP-treeを構築．

（6）FP-tree走査
構築したFP-treeの走査を行う．

暗号化と復号，複雑なデータ構造を伴う処理

送信 / 結果を返送

委託先サーバ（分散処理可能）

（1）立ち上げと接続準備
通信の受付を行い，クライアントとの接続を確立．

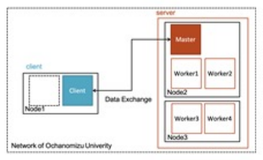
（4）委託処理
クライアントから暗号化されたデータを受信．各アイテムのサポート値を計算．
加えて各アイテムの出現回数と閾値の差を取る．クライアントに結果を返送．

暗号文で計算できる加算，乗算中心の処理

実装・実験・提案

実装・環境
- 主な使用言語：C++
- FHEライブラリとしてHelibを使用、Leveled FHEの実装を採用している

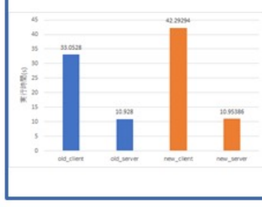
OS	CentOS 6.9
CPU	Intel(R) Xeon(R) プロセッサ E5-2643 v3 3.6GHz 6コア　12スレッド
メモリ	512GB

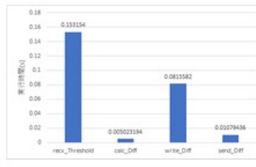


実験
入力データ
- IBM Quest Synthetic Data Generatorにより作成した人工データ
- サーバの委託処理を追加した際の実行時間
  - クライアント、サーバにおける全体の実行時間
  - 追加部分の実行時間

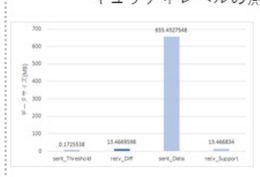
平均トランザクション長	5
最大パターンの大きさ	5
アイテム数	30
トランザクション数	9900
最小サポート値	0.1(10%)

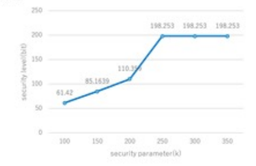
データに関するパラメータ





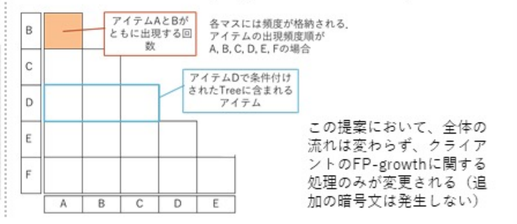
- 追加部分の通信データサイズ（左２項目が該当）
- 暗号のセキュリティパラメータkを変化させた際のセキュリティレベルの測定





新規提案手法
現システムでFP-growthを使用している部分の実装をFP-growth*に変更し、クライアントの処理を軽減する手法の提案

● FP-growth*
FP-treeのデータ構造に、追加でArrayという構造を組み込み、走査の再帰処理を軽減する手法。疎なデータセットに対して効果があるとしている。（Arrayの構造は以下の通り）



この提案において、全体の流れは変わらず、クライアントのFP-growthに関する処理のみが変更される（追加の暗号文は発生しない）

今後の課題

- FP-growth*を使用したシステムの実装、セキュリティ周りの検討
- FP-growth*の処理にFHEを適用させ、よりサーバへの委託処理を増加させる方法の検討

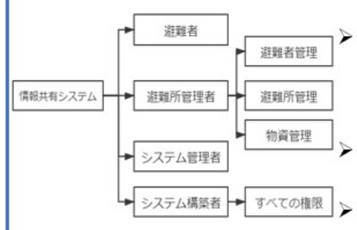
大規模災害時における情報共有システムの個人認証に関する一検討
（研究担当：周　爽）

研究背景

- 近年、日本各地で地震や台風などの災害が連続して発生している
  - 避難所で備蓄品の状況、ニーズを把握
  - 家族と知人の安否確認や避難所情報などを受け取る
    - 情報共有システムが必要
- 偽者としてシステムに登録したり、アクセスしたりすることにより、個人情報の漏えいや改ざんなどの恐れがある
  - 避難する時に身分証明書を持参していない
    - 身分証明書のみに基づく個人認証はできなくなる可能性があるため、適切な本人確認の手法を考える必要がある

提案システム

- 情報共有システム



― システム機能 ―
- 避難者は滞在している避難所や健康状態などを報告する
- 避難所管理者は所属する避難所を管理している。備蓄物資と需要物資をシステムに登録することができる
- システム管理者は各避難所の管理者の情報を管理する
- システム構築者はシステムメンテナンスのため、すべての権限を持っている

構築環境

- Node.js
- Framework：Express
- Database：MongoDB

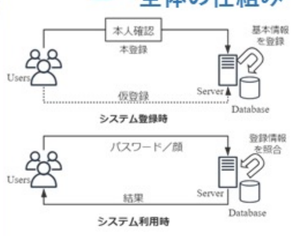
システム実装

- ホームページ

- ある避難所のユーザリスト

個人認証についての提案

― 全体の仕組み ―



➢ **基本情報**
- 氏名
- 性別
- 生年月日
- 住所
- パスワード
- 顔特徴量

➢ **システム登録時**
本登録：顔付き身分証明書をもって本人確認が完了している
仮登録：持っていない場合は本人確認が完了していない

➢ **システム利用時**
パスワードあるいは顔を入力することでログインできる
登録状態によりログイン状態が違う

― 本登録 ―

1. マイナンバーカードによる本人確認を行う際に、JPKIで本人確認ができた後、カードからデータ読み取る。パスワードを設置して基本情報を登録する。
2. 他の顔付き身分証明書を持っている場合はユーザが身分証明書を持って写真を撮影する。本人の顔特徴量と身分証明書の顔写真の特徴量との類似度により、本人の身分証明書と認められる場合、基本情報をシステムに登録する。
3. 顔付き身分証明書を持っていない場合は仮登録を行う。

― 仮登録 ―

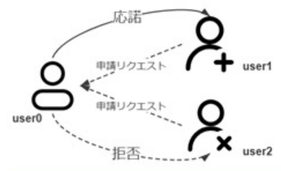
仮登録では自治体の住民票データのハッシュ値を利用することで入力したデータの真偽を判断する。

1. 他の顔付き身分証明書を持っている場合は身分証明書の写真を撮影身分証明書の基本情報を認識する。
2. 身分証明書を持っていない場合はユーザが基本情報を入力する。
   入力したデータのハッシュ値と自治体のハッシュ値を照合し、一致する場合はシステムに登録する。

第三者からの本人確認

仮登録から本登録の状態への移行

➢ 友達になる



➢ 仮登録のユーザの写真を利用することで本登録の友達から本人確認を受ける

まとめと今後の課題

➢ **まとめ**
情報共有システムを提案
個人認証の仕組みを提案

➢ **今後の課題**
全体的に個人認証についての仕組みを更に考えていく
提案した仕組みをウェブアプリケーションとして実装する

室内動作解析のためのドメイン適応による合成データ活用の検討
（研究担当：礒井　葉那）

研究背景

- Deep Learning による動画像解析-
◆ 高齢者や子どもの見守りへ
◆ 解析精度は主に学習用データの量と質に依存
◆ 室内動作解析のための十分なデータが存在しない

- 合成データセット-
◆ 現実の動画像の収集・ラベル付けは非常に高コスト
◆ CGを使って画像データを自動的に生成

- ドメイン適応-
◆ ソースデータとターゲットデータを一緒に学習して、ラベルなしでもターゲットデータを解析
◆ 合成動画像のドメイン適応はまだ高精度な方法が確立していない

研究課題

◆室内の行動解析に利用可能な合成動画像データセットを構築

◆ドメイン適応を用いた学習方法によって、合成動画像を実データ解析に活用することを目指す

→実際に人が部屋の中で動くような合成動画像を作成し、実写の動画像を用いて評価

作成した合成動画像

● **写実的な出来栄え**
実動画像と、それに写実的で人間の目で見てそっくりな合成動画像を作成
部屋の様子やカメラの位置・角度なども同様

● **部屋内で人が動く**
Ochahouse を模した部屋の中で、歩く・立ち止まる・座る・座っている・立ち上がる・横になる・寝ている・起き上がるの7つの動作を行う

● **各動画像は約3~7秒、データ数は以下の通り**

	歩く	座る	座っている	立ち上がる	寝る	寝ている	起き上がる
合成データ	997	747	1118	780	250	250	250
実データ	96	44	56	51	32	39	32

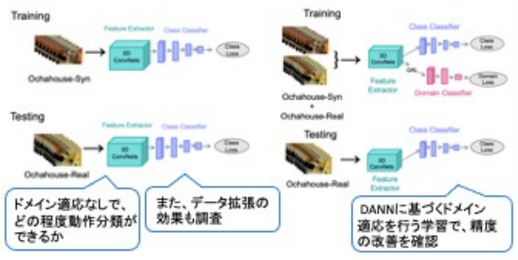
←実データ

←合成データ

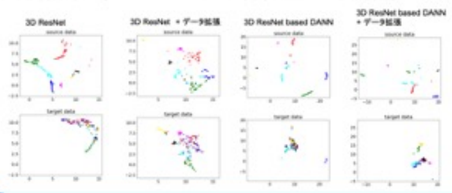
unity

評価実験

実験



実験結果（UMAPによる特徴量の可視化）



実験結果（精度）

実データの教師あり学習（比較用）　ベースライン

学習手法	ドメイン適応	精度
3D ResNet (target only)	-	81.14
3D ResNet (source only)	-	**28.08**
3D ResNet (source only) + データ拡張	-	**48.57**
TemPooling + データ拡張 ※	o	17.14
TA³N + データ拡張 ※	o	28.57
3D ResNet based DANN	o	**45.55**
3D ResNet based DANN + データ拡張	o	**38.57**

→ドメイン適応なし・データ拡張なし（ベースライン）では、十分な精度で解析できないが、
ドメイン適応・データ拡張により、精度が向上（先行研究のTemPooling, TA³N よりも高精度に）

※ [関連研究] M.H.Chen et al.(Temporal Attentive Alignment for Large-Scale Video Domain Adaptation. ICCV2018)

まとめと今後の展望

- まとめ-
◆ 室内の動作解析のための合成動画像データを作成
◆ そのままでは、合成動画像を用いても実動画像の解析はできない
◆ データ拡張と、DANNに基づくドメイン適応により、実データ解析精度が向上

- 今後の展望-
◆ 解析精度を上げるさらなる工夫
・manifold mixup などの中間層でのデータ拡張
・VDB などの手法による敵対的学習の安定化
・さまざまな学習手法・ドメイン適応手法による比較

http://ogl.is.ocha.ac.jp/

oguchi@is.ocha.ac.jp