小口研究室　研究紹介　(2016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Aprioriアルゴリズムの並列分散計算による高速化手法の検討
（研究担当：宇佐美 文梨）

Introduction | セキュアなビッグデータからの知識獲得

ビッグデータ利活用の推進により、企業や病院などで蓄積されているビッグデータについても注目が集まっている。

大量のデータを解析し新たな知識を得ることで、今までにない手法や着眼点を得るためには、年々膨れ上がるデータを捌き、要求により複雑化していく計算量の大きなアルゴリズムを扱えるマシンパワーが必要となる。

データマイニングを行う際、クラウドサービス等を提供する第三者の計算資源が一般的に用いられるが、高い機密性を有するデータを外部のストレージに預けることで、委託先に勤務する従業員等による流出等、新たなセキュリティ上の問題が生じる。

頻出パターンマイニングの例 | Aprioriアルゴリズム

POSレジなどに蓄積された購買情報から、お客さんがよく買う商品パタンを知りたい！

①探索対象となる商品の集合(パタン)を決める 初期値は パタン長 = 1
②その商品パタンの購入者数(サポート)を数える
③閾値と比べ、頻出パタンを決定する
パタン長++;
該当なしなら終了

例 トランザクション(取引履歴)

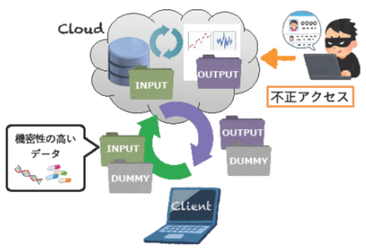
	アイテム（商品）			
	D	B	I	A
T1	○	○	○	○
T2	○	○	○	×
T3	○	×	×	×
T4	○	○	×	○
T5	×	○	○	×
T6	×	○	○	○
T7	○	○	×	×

例 最低サポート数（閾値）= 4

パタン長	1	2	3
①	{D},{B},{I},{A}	{D,B},{B,I},{I,D}	{D,B,I}
②	5,6,4,3	4,4,2	2
③	{D},{B},{I}	{D,B},{B,I}	{}

プライバシ保護データマイニング

高い機密性を有するデータから、適切なセキュリティ管理のもと知識を獲得するための技術の総称。ここでは①匿名化、②秘匿計算を紹介する。

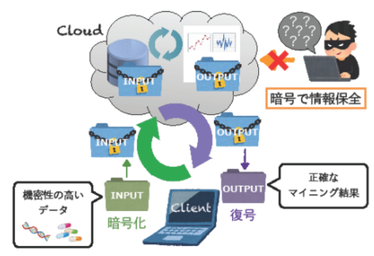


①匿名化

真の値にノイズを混ぜ、統計的な手法で規則性を知る

Merit
- 計算量が小さい

Demerit
- プライバシを保護すればするほど誤差が大きくなる
- 真の値を含むデータが外部に知られる



②秘匿計算(準同型暗号方式)

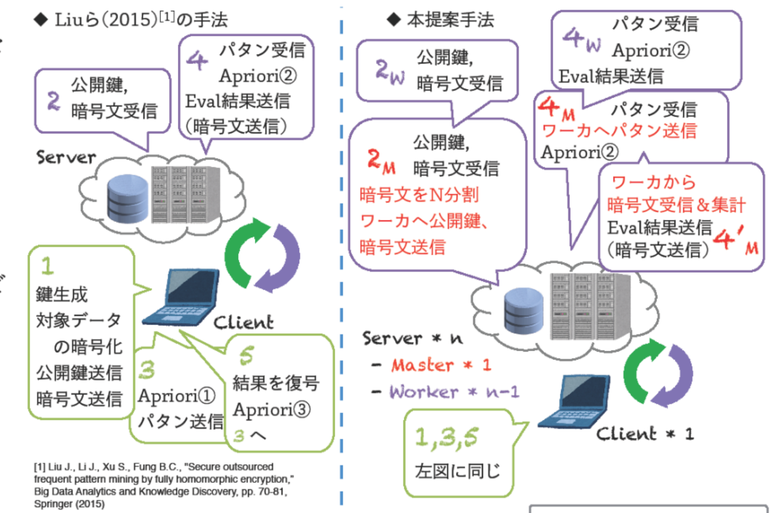
暗号化したままAND・OR演算ができる完全準同型暗号によって暗号したデータを使用し委託計算

Merit
- 正確な結果が誤差なく得られる
- 暗号文のまま計算委託できるので保全性が高い

Demerit
- 計算量が大きい

統計的な手法を用いにくいセキュアなデータからマイニング結果を得るには、準同型暗号を用いた秘匿計算が有効だが、一方で計算量が膨大になってしまう。

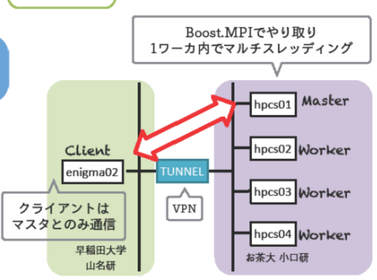
Method | 完全準同型暗号を用いた委託データマイニング



[1] Liu J., Li J., Xu S., Fung B.C., "Secure outsourced frequent pattern mining by fully homomorphic encryption," Big Data Analytics and Knowledge Discovery, pp. 70-81, Springer (2015)

Liuらの手法を広域分散環境で実験（VPN環境 + 並列分散処理）

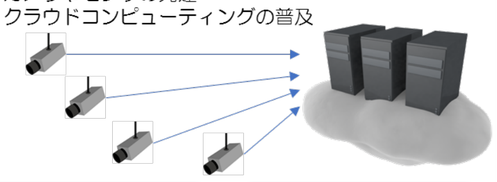
目的：Apriorの完全準同型暗号を用いた委託計算の高速化
方法：広域分散環境を構築＆Boost.MPIを用いた分散並列計算を実現



大規模データ分散プラットフォームApache Sparkを用いた分散並列機械学習
（研究担当：加藤 香澄）

研究背景

ライフログの取得・蓄積と利用の拡大
- カメラやセンサの発達
- クラウドコンピューティングの普及



一般家庭にサーバやストレージを設置し、解析処理を行うのは困難

動画像データが集約し、解析処理によってクラウドに大きな負荷

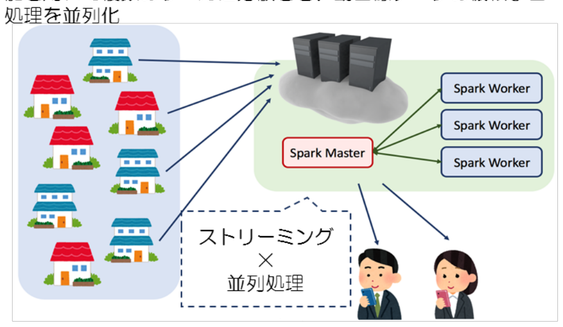
Apache Sparkを用いてChainerによる解析処理を並列化し、クラウドにかかる負荷を軽減

Spark
- 分散コンピューティングプラットフォーム
- 高速かつ汎用的

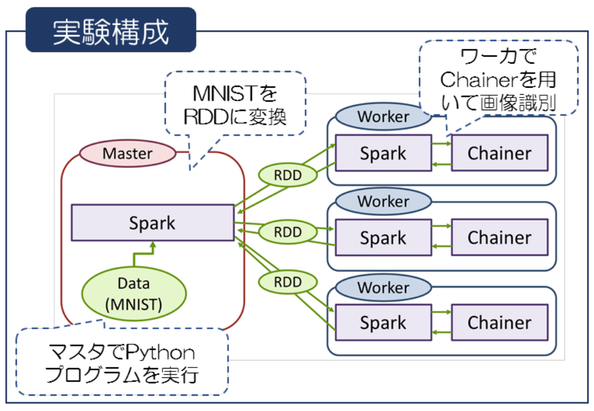
Chainer
- ディープラーニングフレームワーク
- 柔軟性、直感的、高機能

研究概要

各家庭から収集した動画像データをSparkのストリーミング機能を用いて複数のワーカに分散させ、動画像データの機械学習処理を並列化



実験構成

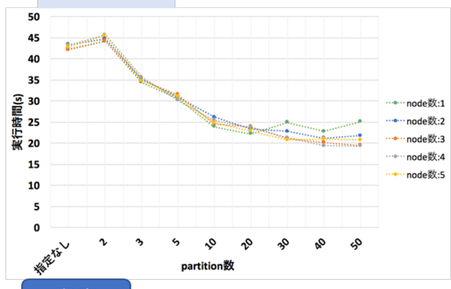


実験結果

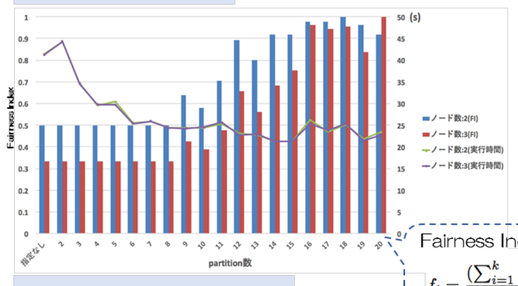
Sparkによる並列機械学習処理部分の実験を実施

変化させるパラメータ
- 読み込ませるデータのpartition数
- ワーカのノード数

実行時間



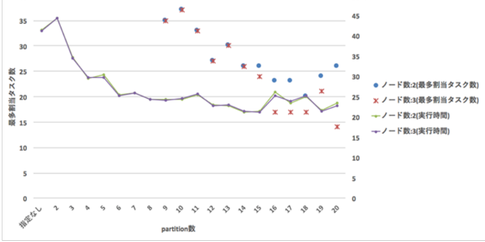
Fairness Index



Fairness Index

$$f_i = \frac{(\sum_{i=1}^{k} x_i)^2}{k \sum_{i=1}^{k} x_i^2}$$

最多割り当てタスク数



考察
- ノード数変化による実行時間の変化ほぼ無し
- 公平性の高さと実行時間が反比例しない
- タスクを最も多く割り当てられたノードの処理時間による律速の影響ほぼ無し

→ 効率的な分散が行えていない

まとめと今後の課題

まとめ
- Sparkを用いてChainerの機械学習処理を並列化させた
- 実験から、効率的な分散が行えていないことがわかった

今後の課題
- 分散環境のさらなる調査
- MNIST以外のデータを用いて実験

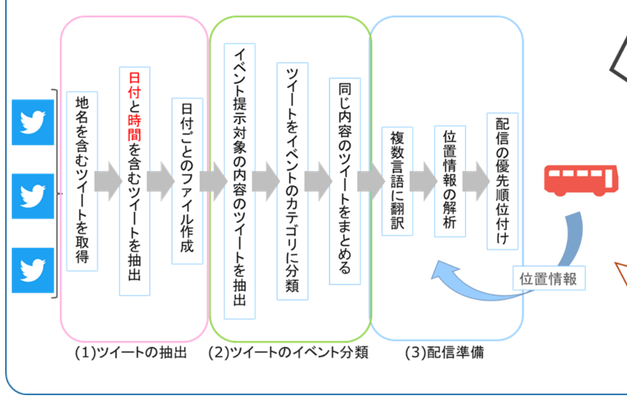
場所と時間の制約条件に対応するSNSデータを用いたタイムリーなイベント提示システム
（研究担当：工藤 瑠璃子）

研究背景

- ◆ 2020年の東京オリンピック開催を受けインバウンドが増加しているが、観光情報の発信は不足
- ◆ 恒常的に人気のあるスポットはガイドブックなどから取得可能だが、今まさに開催されているイベントは情報を取得しにくい
- ◆ 地理的・時間的な制約がある旅行者などが必要とする「その時」「その場」で役立つ情報配信は少ない

時間とともに移動していく旅行者などにリアルタイムに有益な情報を配信するシステム構築

提案システム



- ツイートの抽出
  Twitter APIのキーワード検索で地名をキーワードに設定し,地名に紐付いたツイートを取得
  ツイートの本文に日時を含むツイートを抽出
- ツイートのイベント分類
  イベント提示対象の内容のツイートを抽出
  抽出されたツイートをイベントのカテゴリごとに分類
- 配信準備
  提供する情報を複数言語に翻訳
  ユーザと提供するイベントの位置情報を取得

BeaconCast
ネットワーク仮想化に対応したアクセスポイント
インターネットにアクセスできない場合のメッセージ配信を実現
→ 通信契約をしていない層へ情報配信可能

ツイートの分類評価

◆ 有用な情報をカテゴリ分類

音楽イベント　寄席・お笑い　舞台　映画
展覧会　ポケモンGO　その他

◆ 2つの分類器で比較

分類の流れ
1. ツイートをMeCabを用いて形態素解析
2. 名詞・動詞・形容詞・副詞
3. Bag of Wordsでツイートを数値化
4. 学習データの作成 (正解・不正解のラベルを与える)
5. 学習データよりモデルを作成
6. 未知要素の分類結果で精度検証

◆ SVM (support vector machine)

① 有用/有用でないの判別結果

Precision	Recall	F値
0.9175	0.8396	0.8768

線形入力素子を利用して2クラスの識別器を構成

② カテゴリ分類の判別結果

カテゴリ	Precision	Recall	F値
音楽イベント	0.8484	0.9032	0.8750
寄席・お笑い	0.5357	0.9375	0.6818
舞台	0.3500	0.8750	0.5000
映画	0.5454	0.8571	0.6666
展覧会	0.3703	1.0000	0.5405
ポケモンGO	1.0000	1.0000	1.0000

◆ ランダムフォレスト

弱学習器を決定木とするバギング

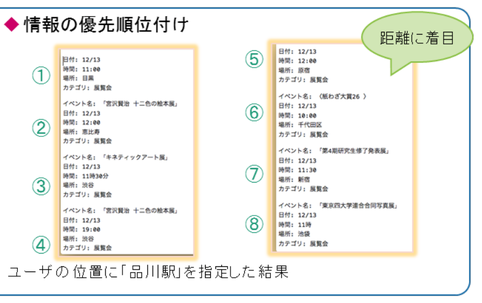
カテゴリ	Precision	Recall	F値
音楽イベント	0.8750	0.7583	0.8125
寄席・お笑い	0.7948	0.7561	0.7750
舞台	0.5000	0.4000	0.4444
映画	1.0000	0.7272	0.8421
展覧会	0.8667	0.8125	0.8387
ポケモンGO	1.0000	1.0000	1.0000

SVM, Random Forestの両分類器で全体として良い結果 → Random Forest を採用
分類精度：Random Forest の方がF値が高い
処理速度：多カテゴリに分類可能な Random Forest が速い

配信準備

◆ インバウンド対応のため多言語翻訳

◆ 情報の優先順位付け



ユーザの 位置に「品川駅」を指定した結果

今後の課題

●ツイートの分類精度、テキスト情報整理の精度の向上
●地名・日付・時間を含むツイートのみが対象になっているため解析対象のツイートの種類が少ない
→この３点の情報が欠けている場合に補う手法を提案し、解析するツイートの種類を増やす

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

oguchi@is.ocha.ac.jp