群知能を用いたクラスタリング手法のトラヒック分類への適用と評価

須恵 匠、大下裕一、村田正幸
大阪大学 大学院情報科学研究科

背景

- ネットワークを介したサービスが多様化
  - Video on Demand
  - クラウドサービス
  - オンラインゲーム　等
- サービスにより異なる要求
  - 低遅延が必要なサービス
  - 低ジッタが必要なサービス
  - 広帯域が必要なサービス　等

↓

- ネットワークの制御を行う際にはサービスの性能要求を考慮することが必要

↓

- サービスの要求を把握することが必要

ネットワークを介したサービスの現状

- Web を介したサービスが隆盛
  - YouTube
  - Dropbox　等
    - 共通のプロトコル（HTTP）を利用

↓

- ポート番号のみではサービスの特徴の判別は不可
  - 従来は Well-known Port 番号から特定

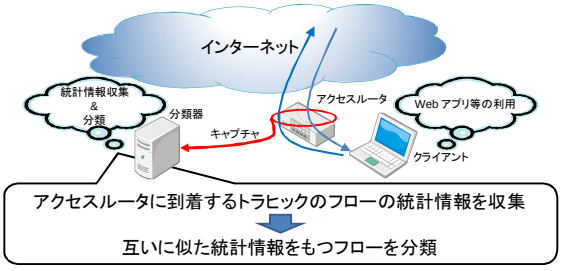
↓

- 新たなサービスの分類手法が必要

サービスを分類する手法の要件

- ネットワークを流れるトラヒックの挙動からサービスを分類できること
- サービスの性能要求ごとに分類できること
  - 個別のサービスを識別するのではなく性能要求が同様のサービスは同じカテゴリに分類
- 新たなサービスが登場した場合も適切なカテゴリに分類できること
  - 既存のものと同様の性能要求をもつ新サービスÝ 既存のカテゴリに包含
  - 性能要求が新しいサービスÝ 新カテゴリ

トラヒック統計情報を用いたサービスの分類



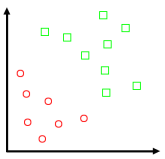
- ポート番号によらずフローの挙動で分類可能
  - 性能要求に直結するフローの挙動に関する統計情報を元に分類を行うことにより性能要求に合わせた分類が可能

統計情報によるトラヒックの分類方法

- クラスタリング：
  データの集合を類似するデータを集めた塊に分割
  - 各データがもつ N 個の属性を用いてデータ間の類似度を定義
    - 類似度の定義：ユークリッド距離
      - 最も一般的な距離の定義
    - データ x とデータ y のユークリッド距離は以下で定義

$$d(x, y) = \sqrt{\sum_{i=1}^{N} (x_i - y_i)^2}$$

    - x_i はデータ x の i 番目の値
  - 互いに類似度の高いデータが集まるようにクラスタを形成



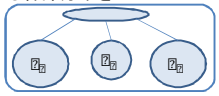
群知能を用いたクラスタリング (AntTree)

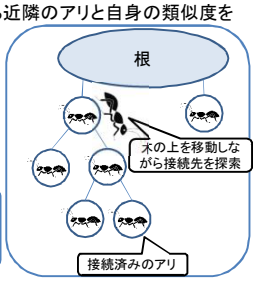
生物の自己組織化システムに学んだ計算手法

- 各アリ(＝データ)が自律的に移動しながら近隣のアリと自身の類似度を比較して互いを接続

徐々に木構造を構築

- 移動先の決定
  - 注目しているアリの子との類似度から判断
    - 類似:そのアリへ移動
    - 非類似:新たな部分木を生成
- 根に接続している各部分木を一つのクラスタと解釈





各アリが適応的に状況にあわせてつながることでクラスタを形成

新たなデータを随時投入しながら状況の変化に対応してクラスタを形成可能

研究の目的とアプローチ

- 目的
  - 群知能を用いたクラスタリング手法のトラヒックの分類への適用性の評価
- アプローチ
  - クラスタリングの入力に用いる統計情報を定義
  - Web アプリケーション利用時のトラヒックキャプチャデータを用いて統計情報を抽出
  - クラスタリングを行い分類性能を比較評価
    - 群知能を用いたクラスタリング:AntTree
      - トラヒックデータを順次読み込みながらクラスタリング
    - 比較対象:K-means 法
      - 分類対象の全トラヒックデータを取得後にクラスタリング
    - 評価のポイント
      - アプリケーションの種類が未知である / 新種のアプリケーションが出現しうる環境において適切な識別ができるか

分類対象のトラヒックデータ

- Web アプリケーション利用時に発生したパケットをキャプチャしたデータを使用
  - 5 種類のサービスタイプを定義

Download	クラウドストレージ(Bitcasa[3]、Box[4]、Dropbox[5])からファイルをダウンロード
Upload	クラウドストレージ(Bitcasa[3]、Box[4]、Dropbox[5])へファイルをアップロード
Video Live Streaming	動画をリアルタイムで視聴(Ustream[6] の生中継)
Video on Demand	動画ファイルをダウンロードしながら再生(Ustream[6] の録画配信、YouTube[7]、ニコニコ動画[8])
Interactive	ユーザの操作に応じた返答(Google Map[9]、Yahoo!地図[10])

  - 各種類 20 個ずつのトラヒックデータを使用

[1] https://www.bitcasa.com [2] https://www.box.com [3] https://www.dropbox.com
[4] http://www.ustream.tv [5] http://www.youtube.com/ [6] http://www.nicovideo.jp/
[7] https://maps.google.co.jp [8] http://map.yahoo.co.jp

トラヒックデータのタイプ別特性

- サービスタイプによって異なるフロー挙動

Download	多量のレスポンス・一定間隔のレスポンス・一定間隔のリクエスト
Upload	特に少量のレスポンス・特に断続的なレスポンス・断続的レスポンス
Video Live Streaming	多量のレスポンス・一定間隔のレスポンス・一定間隔のリクエスト
Video on Demand	多量のレスポンス・断続的なレスポンス・断続的リクエスト
Interactive	比較的少量のレスポンス・断続的なレスポンス・断続的リクエスト

⇒ 挙動の違いを捉え分類できる統計情報が必要

分類に用いる統計情報

- Web アプリケーションのフロー挙動特性が反映される統計情報を検討
  - アプリケーションによってリクエストやレスポンスのタイミング・大きさに違い

用いた統計情報

- ペイロード付パケット到着レートの下り方向と上り方向の比の対数
  - $R = \log \frac{\text{下り方向のペイロード付パケット到着レート}}{\text{上り方向のペイロード付パケット到着レート}}$
- 下り方向のペイロード付パケット到着レートの変動係数
  - $CV_{\text{下り方向のペイロード付パケット到着レート}} = \frac{\text{標準偏差}(\text{下り方向のペイロード付パケット到着レート})}{\text{平均}(\text{下り方向のペイロード付パケット到着レート})}$
- 上り方向のペイロード付パケット到着レートの変動係数
  - $CV_{\text{上り方向のペイロード付パケット到着レート}} = \frac{\text{標準偏差}(\text{上り方向のペイロード付パケット到着レート})}{\text{平均}(\text{上り方向のペイロード付パケット到着レート})}$

$\text{下り方向のペイロード付パケット到着レート}$: 下りパケット到着レート
$\text{上り方向のペイロード付パケット到着レート}$: 上りパケット到着レート
標準偏差: 標準偏差
平均: 平均

評価指標

- 正確度:クラスタリングの分類性能を測る指標

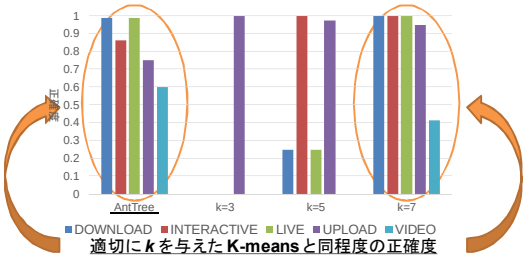
$$\text{正確度} = \frac{\text{ラベル } i \text{ のクラスタに分類されたトラヒックデータ数}}{\text{アプリケーションタイプ } i \text{ のトラヒックデータ総数}}$$

- 各アプリケーションのトラヒックが適切なクラスタに分類された割合
  - 分類先クラスタのラベルと一致したトラヒックデータの割合

クラスタへのラベル付与条件

- クラスタ内最多のアプリケーションタイプのトラヒックデータが 3 個以上含まれる場合
  - そのアプリケーションタイプを当該クラスタのラベルとして付与
- それ以外の場合
  - クラスタはラベルなし

評価結果



適切に ***k*** を与えた **K-means** と同程度の正確度

- AntTree は順次到着するデータを正確に分類可能
  - 各アリが自律的に類似データを探索し所属するクラスタを判断
    - 類似するデータのクラスタへ所属 or 新たなクラスタを構成

まとめと今後の課題

- まとめ
  - 群知能を用いたクラスタリング手法の評価
    - 次々に到着するトラヒックデータをサービスタイプごとに分類が可能
    - 全データを読み込んでから分類する K-means と同様の正確度
- 今後の課題
  - 分類の正確度の向上
  - ネットワーク制御との連携についての検討