機械学習を用いた複数のネットワーク間の時系列データの補間に関する研究

長岡技術科学大学 電気電子情報工学専攻 修士 2 年
ネットワーク特性評価研究室
會澤一輝

研究背景

ネットワークの研究において，実験的にネットワークのパフォーマンスを測定する際に，試験用のフローのデータが必要

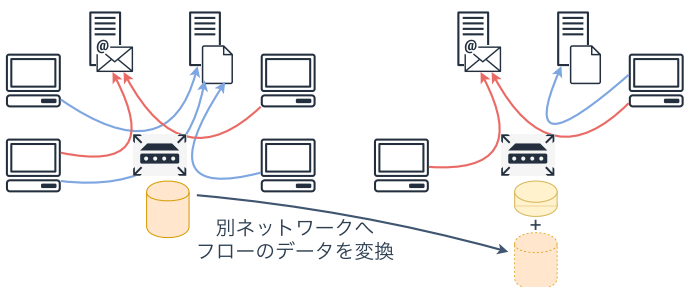
- 実際のネットワークのデータはあまり公開されていない

既存の実データの組み合わせから，異なるネットワークのフローを他のネットワークのフローに変換する方法を検討

フロー 2 つのノード間のセッション開始から終了までの一連の通信

目的

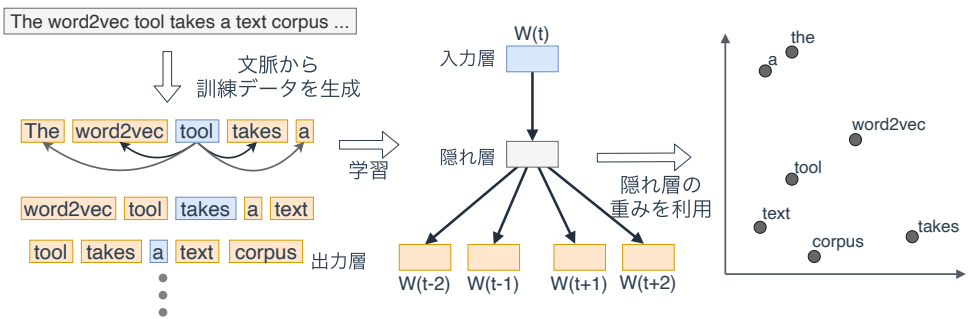
既存のフローの情報を用いて，他のネットワークのフローを生成



Word2Vec[1]

文章の文脈の情報から，ベクトル空間への単語の埋め込みを学習する手法

- 文章中の隣接する単語群を訓練データとし，入力された単語に隣接する単語が出力されるようにニューラルネットワークを学習する
- 隠れ層の重みをベクトルとして利用する

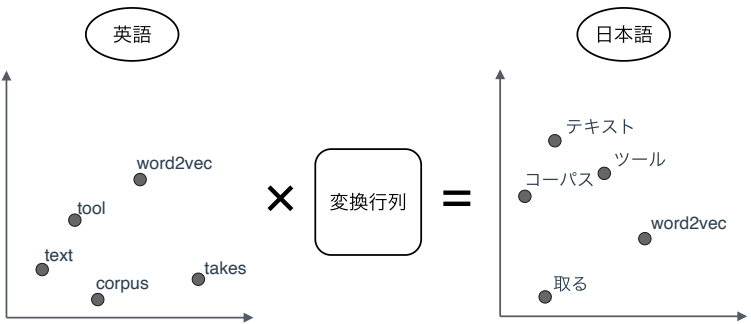


生成された単語ベクトルから，類似する意味合いの単語の発見や，単語間の関係性の分析が可能

Word2Vec を用いた単語の翻訳 [2]

生成された単語ベクトルから，2 つの言語間の翻訳を学習

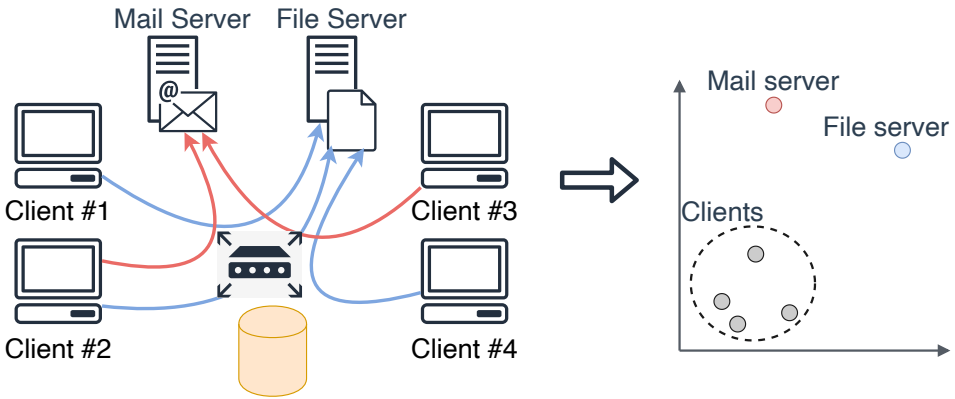
1. 対訳が存在する単語ペア群から，異なる言語間の翻訳を行う変換行列を計算
   - 計算には確率的勾配降下法を使用
2. 変換行列を未知の単語の翻訳に適用



IP2Vec[3]

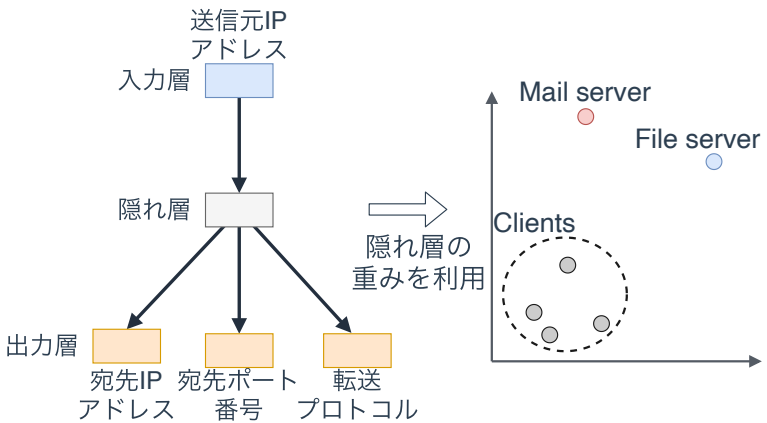
ニューラルネットワークを用いて IP アドレスやポート番号などのカテゴリ値をベクトル空間上の埋め込みに変換する手法

- フローに含まれる情報から，ネットワーク内に出現するクライアント PC やプリンタ，サーバなどの IP アドレスをベクトル表現に変換



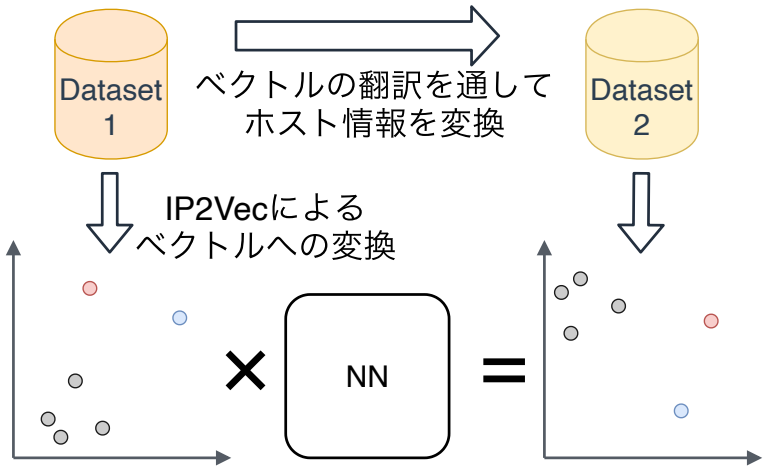
IP2Vec

1. フローデータから訓練データを生成
2. IP アドレス，送信先ポート番号，転送プロトコルの組み合わせが訓練データになる
3. 入力に対し，特定の出力を返すように隠れ層を学習させる
4. 隠れ層の重みをベクトルとして利用



提案法

1. ネットワークのフロー情報をデータセットとし，2 つのデータセットに対し IP2Vec による学習を行う
2. 2 つのネットワーク間で，共通するポート番号，プロトコルを元にベクトルの翻訳を学習する



実験

異常検出システム評価用のフローベースのデータセット CIDDS-001 [4] を利用

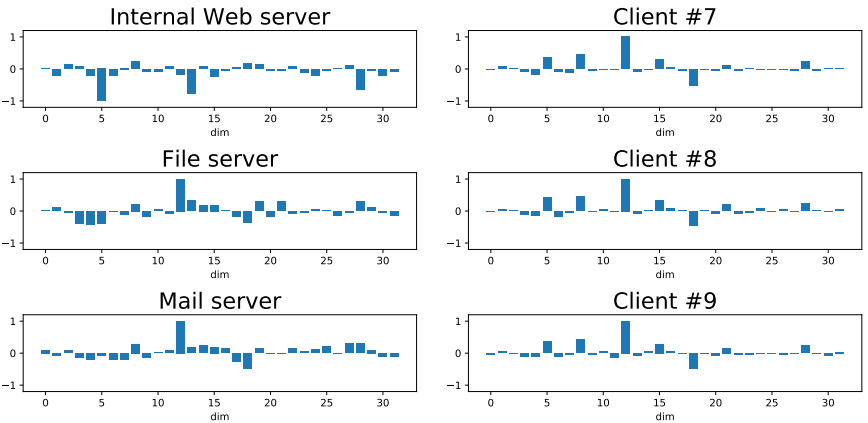
- 複数のクライアント PC やプリンタ，サーバ群から構成されるネットワークのフロー情報のデータセット

クライアント PC やサーバのホストに関する情報を変換できるか確認

- 2 日分のデータを分割し，別ネットワークとして利用
- 1 日目のホストのベクトルを 2 日目の同じホストのベクトルに変換できるか？

実験結果

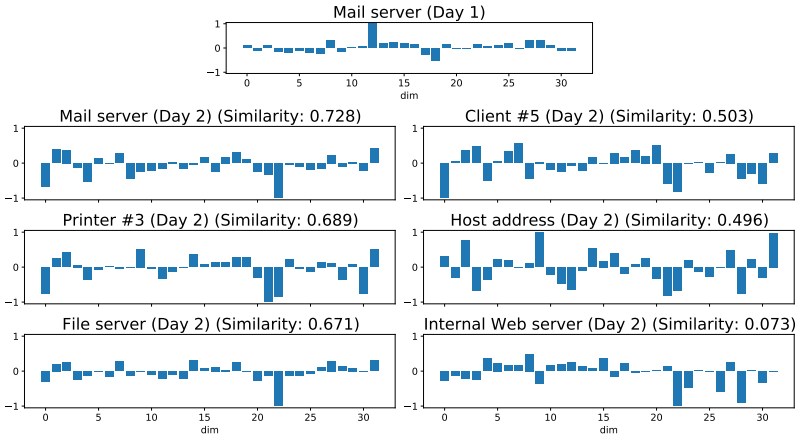
- IP2Vec を用いて，1 日目のデータからホストの情報を 32 次元ベクトル空間上に埋め込み



- 同じ用途で用いられるホスト間で，類似するベクトルが得られた

実験結果

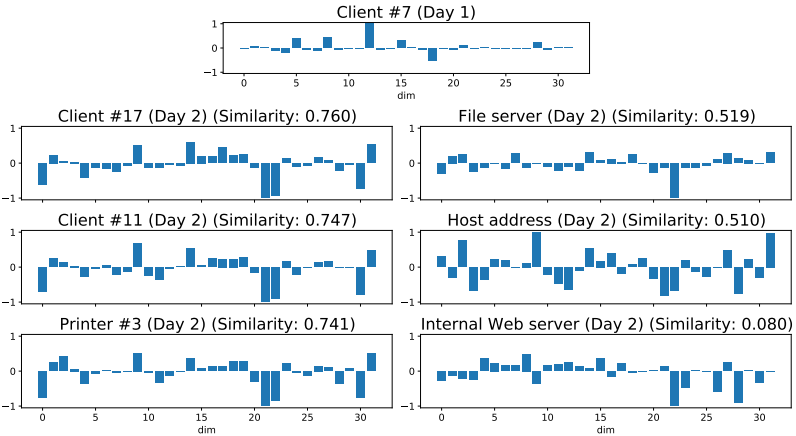
- 1 日目のデータのベクトルを，2 日目のデータの空間へ変換して比較
- 変換後のベクトルと，類似度の高い上位 3 件と下位 3 件を示す



- 異なるデータセット間で埋め込みのベクトルを変換

実験結果

- 1 日目のデータのベクトルを，2 日目のデータの空間へ変換して比較
- 変換後のベクトルと，類似度の高い上位 3 件と下位 3 件を示す



- 同じ用途のホスト同士で類似度が高くなる

まとめ

まとめ

- IP2Vec によりクライアント・サーバ群のベクトル空間への埋め込みを学習
- 埋め込みを他のネットワークのベクトル空間上に変換
  - ベクトルの変換により，同じ振る舞いをするホストを特定

今後の予定

- データが十分に存在しないネットワークのフローを，他のネットワークのフローを元に生成
- 既知のネットワークのホストに関する情報をもとに，他のネットワーク内に存在する似た振る舞いのホストを特定

Reference

[1] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean.
Distributed representations of words and phrases and their compositionality.
In *Advances in neural information processing systems*, pp. 3111–3119, 2013.

[2] Tomas Mikolov, Quoc V Le, and Ilya Sutskever.
Exploiting similarities among languages for machine translation.
arXiv preprint arXiv:1309.4168, 2013.

[3] Markus Ring, Alexander Dallmann, Dieter Landes, and Andreas Hotho.
Ip2vec: Learning similarities between ip addresses.
In *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*, pp. 657–666. IEEE, 2017.

[4] Markus Ring, Sarah Wunderlich, Dominik Grüdl, Dieter Landes, and Andreas Hotho.
Flow-based benchmark data sets for intrusion detection.
In *Proceedings of the 16th European Conference on Cyber Warfare and Security (ECCWS)*, pp. 361–369.
ACPI, 2017.