7S-06

複数コネクションのクラスタリングに基づくサービス同定

原雅貴[†1]　韮澤慎之介[†1]　中尾彰宏[†2]　小口正人[†3]　山本周[†2]　山口実靖[†1]

工学院大学 工学研究科 電気・電子工学専攻[†1]　東京大学 大学院 情報学環[†2]
お茶の水女子大学 理学部 情報科学科[†3]

1 はじめに

東日本大震災で我が国が経験したように，大規模災害時には通信網に輻輳が発生するため，被災者の救済や減災に重要なサービスを優先するトラフィック制御が求められると考えられる．そのためには，通信機器にてサービス同定などのトラフィックの分類を行う必要がある．簡易なサービス同定手法として，送信先 IP アドレスを用いる手法があるが，同一のアドレスで動画共有サービス，クラウドサービス，メールサービスなどの複数のサービスを提供している場合もあり，その様な場合は同定が困難である．

そのため，DPI などのパケット解析に基づく同定精度のさらなる向上も重要であると考えられる．ただし，近年の通信の多くが暗号化されておりペイロードの解析は困難となっており，解析可能な部分のみを用いての同定が重要となる．また，DPI と機械学習に基づくアプリケーション同定により高い精度を実現する手法[1]や DPI に基づく高速な接続サイト同定手法[2]が提案されているが，どちらの手法においても同定対象が異なり，サービスを同定することはできない．加えて，機械学習に基づく手法は学習に長い日数を要し，同定の時間の短縮も重要な課題となっている．

本稿では，我々が過去に提案した手法である複数コネクション解析に基づくサービス同定手法[3]の同定精度に関する考察を行う．

2 既存手法

2.1 アプリケーション同定, 接続先サイト同定

暗号化された通信(HTTPS 通信)の解析に基づくアプリケーション同定手法として，機械学習を用いてヘッダ情報，パケット長，パケット到着間隔などの統計情報や N-gram を解析し，アプリケーションを同定する手法[1]が提案されている．N-gram の解析を行っていない手法[4]と比較すると同定精度の向上が見られ，N-gram の解析はアプリケーション同定において有効であることが分かる．しかし，当該研究はアプリケーション同定でありサービス同定とは同定対象の粒度が異なっているほか，学習に長い日数を要することが課題となっている．

筆者らが過去に提案した手法として TLS セッション確立に用いられるフローの N-gram を解析し，接続先サイトを同定する手法[2]がある．この手法も，サービス同定と比較すると同定対象の粒度が異なっている．これらの手法における N-gram の解析はいずれも 1 サービスの単一コネクションのみを解析対象としており，複数コネクションに対する N-gram の解析は行われていない．

2.2 サービス同定

筆者らが過去に提案したサービス同定手法として複数コネクション解析に基づくサービス同定手法[3][5]がある．当該手法では，サービスにアクセスした際に確立される全てのコネクションに対して TLS セッション確立に用いられるフローの N-gram 出現頻度を調査し，各コネクションとの相関係数を比較することによりクラスタリングを行っている．そして，各グループの出現回数に対して我々が修正したマンハッタン距離を計算し，最も距離が小さいサービスを同定結果として出力している．通常のマンハッタン距離と修正マンハッタン距離は次の様に求められる．通常マンハッタン距離 (A, B) = $\sum_{i=1}^{G} d_{org}(a_i, b_i)$，$d_{org}(a_i, b_i)=|a_i-b_i|$．修正マンハッタン距離 (A, B) =$\sum_{i=1}^{G} d(a_i, b_i)$，*If ai=0 xor bi=0, d(ai, bi)=LD*．それ以外の場合，$d(a_i, b_i)=|a_i-b_i|$．ただし，LD は十分大きな値とする．

我々が修正したマンハッタン距離では，どちらか片方のみのグループ出現回数が 0 回であることを大きな差異としている．

当該手法の Google15 サービス（Google 検索，YouTube, Google Play，Gmail，Google Drive，Google カレンダー Google Scholar，Google 翻訳，Google Plus，Google ニュース，Google Map，Google Photo，Google Account，Google ドキュメント，Google スプレッドシート）における同定結果[3]は図 1，図 2，図 3 の通りである．Google 検索に対する同定に着目すると，データベース内の Google 検索との修正マンハッタン距離が最も小さく，正しく同定が行われていることが分かる．YouTube に対する同定についても，同様にデータベース内の YouTube との修正マンハッタン距離が最も小さく，正しく同定が行われていることが分かる．全 15 の Google サービスの同定では，同定成功率が 93%となり高い精度でサービス同定が実現できていることが分かる．しかし，例外的に Google Plus のみが Google Photo と誤答しており，この原因の解明と対策の考察が必要であると言える．

3 考察

本章にて複数コネクション解析に基づくサービス同定手法における同定精度に関する考察を行う．

当該手法における Google15 サービスに対しての同定成功率は 93%である．誤答したのは Google Plus であり，誤答先は Google Photo である．Google Plus と Google Photo

---

Service Identification Based on Multiple Connection Clustering
†1 Masaki Hara, Shinnosuke Nirasawa, Saneyasu Yamaguchi
†2 Akihiro Nakao, Shu Yamamoto
†3 Masato Oguchi

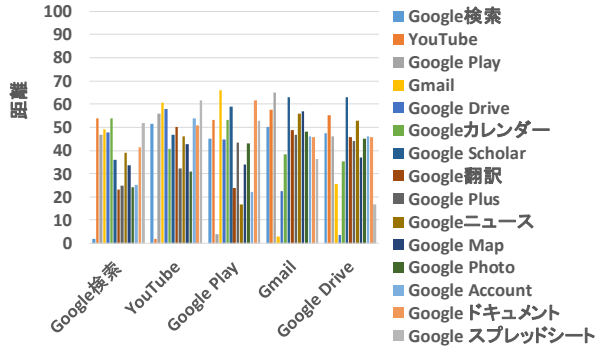

図 1 同定を行いたいサービスとデータベース内の各サービスとの平均距離 I [3]

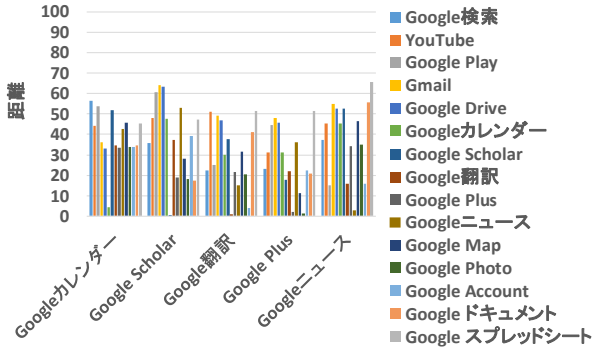

図 2 同定を行いたいサービスとデータベース内の各サービスとの平均距離 II [3]

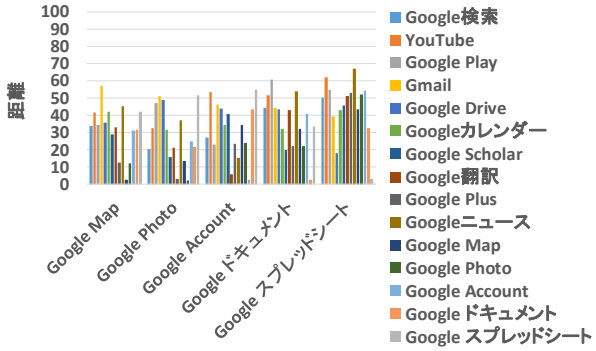

図 3 同定を行いたいサービスとデータベース内の各サービスとの平均距離 III [3]

のグループ出現回数データベース内の各グループの出現回数の分布を図 4，図 5 に示す．Google15 サービスにおいて出現するグループ数は 11 であるが，Google Plus と Google Photo に出現するグループは Group *a* と Group *c* の 2 グループのみであり，Google Plus における出現グループの種類と Google Photo における出現グループの種類は同じであることが分かる．このことから，出現グループが同じである場合の同定は困難であるということが分かる．また，各グループの出現回数に着目すると，Google Plus と Google Photo では出現回数の分散に差異があることが分かる．修正マンハッタン距離を用いた手法では出現回数が 0 か否かのみを大きな差異としており，出現回数の差に関する考慮が不十分であると考えられる．そのため，出現回数の差に着目し出現回数の分布考慮した分類を行うことにより，同定精度の改善が可能であると期待できる．

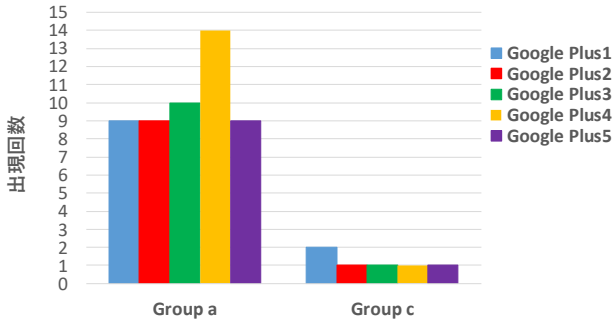

図 4 Google Plus のグループ出現回数分布

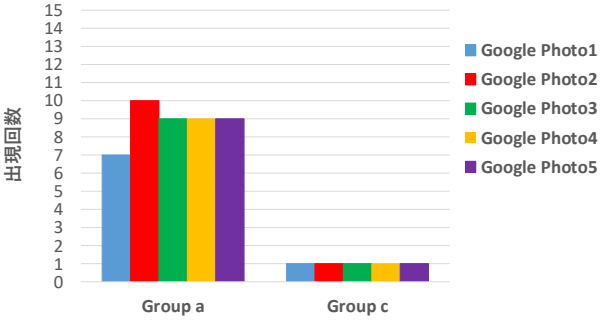

図 5 Google Photo のグループ出現回数分布

4 おわりに

本稿では複数コネクション解析に基づくサービス同定手法の同定精度改善に関する考察を行った．

今後は，同考察を用いた手法の開発を行っていく予定である．

謝辞

本研究は JSPS 科研費 25280022, 26730040,15H02696 の助成を受けたものである。

本研究は、JST、CREST の支援を受けたものである。

文　献

[1] 岩井貴充・中尾彰宏，“アプリケーション毎のトラフィック制御を目的とする N-gram を用いた網内機械学習によるモバイルアプリケーション同定手法”，信学技報, vol.

[2] Masaki Hara Shinnosuke Nirasawa Akihiro Nakao Masato Oguchi Shu Yamamoto Saneyasu Yamaguchi,"Fast Application Identification Based on DPI N-gram", 2016 IEEE 17th International Conference on High Performance Switching and Routing Workshop Program,June,2016.

[3] 原雅貴・韮澤慎之介・中尾彰宏・小口正人・山本周・山口実靖，” 複数コネクション解析に基づくサービス同定”，信学技報, vol. 116, no. 214, DE2016-14, pp. 13-18, 2016 年 9 月

[4] 岩井貴充・中尾彰宏，“アプリケーション特化型 QoS 制御のための網内機械学習によるモバイルアプリケーション同定”，信学技報, vol. 114, no. 477, NS2014-260, pp. 487-492, 2015 年 3 月

[5] M. Hara, S. Nirasawa, A. Nakao, M. Oguchi, S. Yamamoto, and S. Yamaguchi, “Service Identification by Packet Inspection based on N-grams in Multiple Connections,” 7th International Workshop on Advances in Networking and Computing, 2016