L-003

N-gram パケット解析に基づくサービス同定
Application Services Identification based on N-gram of Packets

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]
Masaki Hara[†1] Shinnosuke Nirasawa[†1] Akihiro Nakao[†2] Masato Oguchi[†3] Shu Yamamoto[†2]
Saneyasu Yamaguchi[†1]

1. はじめに

災害時には膨大なトラフィックが発生し，重要なアプリケーションのパケットを優先的に転送するような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，通信機器にてトラフィックの分類を行う必要がある．しかし，同じサイトや同一 IP アドレスのホストのサービスでも，サービス内容が多岐に渡ることがある．よって，アプリケーションや接続サイトの同定に加え，サービスの同定を行うことが重要であると考えられる．

本研究では，暗号化された通信のパケットのペイロード解析に基づくサービス同定手法を提案し，評価実験によりその有効性の評価を行う．

2. TLS セッション

2.1 TLS セッション確立手順

TLS セッションの確立は，図 1 の手順により行われる．

この手順のうち，TLS プロトコルおよび暗号スイートの決定からサーバ証明書および公開鍵の送信までは平文で通信が行われるため，DPI などでペイロード解析することによりサービスの特徴などを抽出することが可能であると考えられるが，共通鍵の送信以降は暗号化されて通信が行われるため，解析により特徴を抽出することは困難であると予想される．

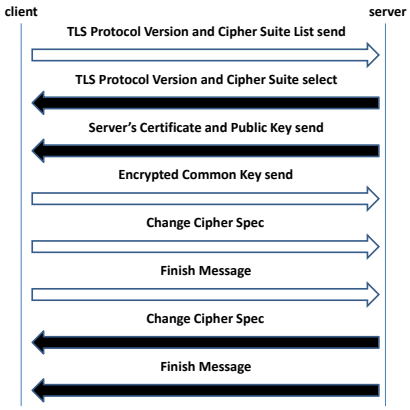

図 1 TLS セッション確立手順

2.2 TLS セッションのグループ化

近年の Web 上のサービスでは，単一の Web サイトと通信した際でも，複数のコネクションが確立される．そして，各コネクションにて TLS セッションの確立が行われる．各コネクションにて確立される TLS セッションに用いられるパケットの N-gram 出現頻度は全て同一ではなく，1 サービスの中の通信であっても異なる．我々の研究[1]により，これらは非常に類似度が高い少数のグループに分類できることが分かった．例を図 2 に示す．サービスと通信した際，各コネクションで TLS セッションが確立される．各 TLS セッションの N-gram 出現頻度の比較を行うと，相関係数が高い例と低い例に分かれる．相関係数が高い例を同一グループとしてグループ化を行うと，図内左の例ではグループ A の出現頻度は 4 回，グループ B の出現頻度は 2 回となる．このように，N-gram 出現頻度の相関係数の比較により，TLS セッションはグループ化することが可能である．

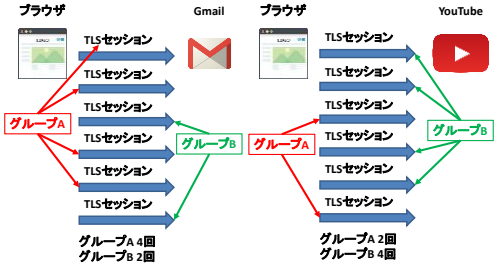

図 2 TLS セッションのグループ化

3. 提案手法

本章にて通信機器で HTTPS 通信の通信内容を解析し，サービス同定を行う手法を提案する．

本手法では TLS プロトコルおよび暗号スイートの決定からサーバ証明書および公開鍵の送信までのデータのみを解析の対象とする．また，クライアントからサーバへ送信されるデータにはサービスの特徴が現れづらいと考え，サーバからクライアントへの通信データのみを解析する．

同定は以下の手順で行う．まず，各サービスと事前に通信を行い，その通信データの解析対象部（TLS プロトコルの決定から公開鍵の送信まで）の N-gram 出現頻度を求め，TLS セッション確立群をグループに分ける．そして，グループ毎に m 件のデータを記録する（m はチューニングパラメータとする）．

次に，通信のサービスの同定は以下の様に行う．通信が行われたら，その通信のコネクションの解析対象部のデータを解析し，同様に N-gram 出現頻度を求める．そして，この N-gram 出現頻度とデータベース内の TLS セッション確立における N-gram 出現頻度との相関係数を求め， m 件の相関係数の平均が最も高いグループに分類を行う．そして，このサービスのコネクション群における各グループの出現頻度を求め，データベース内の各サービスのグループの出現頻度との修正マンハッタン距離が最も小さいサービスを同定結果とする．修正マンハッタン距離は以下の式にて計算される．

$$\sum_{k=0}^{|Gr|} d(\mathrm{freq}(\mathrm{Sv}\alpha, \mathrm{gr}k), \mathrm{freq}(\mathrm{X}, \mathrm{gr}k))$$

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

$$d(a,b)\begin{cases} LD & if\ a=0\ xor\ b=0 \\ |a-b| & else \end{cases}$$

ただし，freq (Sv α, grk) はサービス Sv α のグループ grk の出現頻度である．便宜上，同定対象もサービス X として，サービスと呼ぶ．LD は十分大きな値である．修正マンハッタン距離は，片方の値のみが 0 である場合は大きな差異があるとみなして大きな値 LD を距離に加算する．それ以外の場合は，マンハッタン距離と同一の計算を行う．

4. 性能評価

本章にて，グループ化が適切に行われているか否かと，提案手法の同定精度の評価を行う．本稿では，m(サービスごとの N-gram 出現頻度保持数)は 10，LD は 10，N-gram は 2-gram とし，グループの個数は 13(後述)とした．サービスとしては Google の 15 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 アカウント，Google ドキュメント，Google スプレッドシート）を用いた．

4.1 TLS セッション確立パケットのグループ化

本節にて，グループ化が適切に行われているか否かの評価を行う．Google の 15 サービスと複数回通信を行い，各 TLS セッション確立パケットをグループに分類した．相関係数が 0.95 を超えるもの同士は同一グループとした結果，今回 TLS セッションが分類されるグループ数は 13 となった．グループ内セッション同士の N-gram 出現頻度の相関係数の分布と，異グループセッション間の相関係数の分布を図 3 に示す．グループ内セッション間では，最低でも 0.978 の相関係数が得られており，異グループセッション間では最高でも 0.937 しか得られなかった．すなわち，異グループセッション間の相関係数が同一グループ内セッション間の相関係数を上回る事例は存在していない．よって，グループの分類は適切に行われていることが分かる．

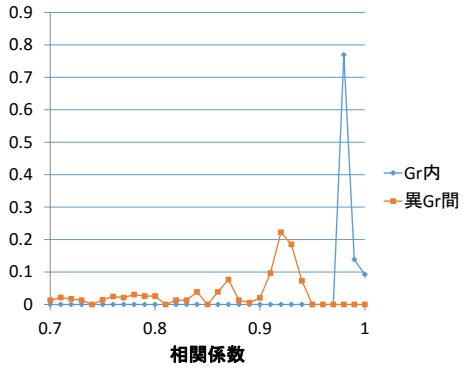

図 3　グループ内セッション間，異グループセッション間の相関係数のヒストグラム

4.2 サービス同定

本節にてサービス同定精度の評価を行う．

同定対象通信である Google の 15 サービスとデータベース内の各サービスとの修正マンハッタン距離を図 4 から図 6 に示す．まず，Google 検索に対して同定を行ったところ，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．また，YouTube に対して同定を行ったところ，同様

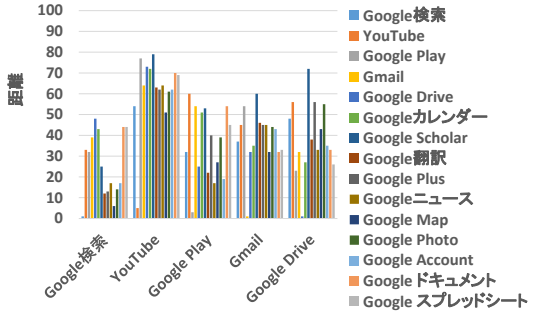

図 4　同定を行いたいサービスとデータベース内の各サービスとの距離 I

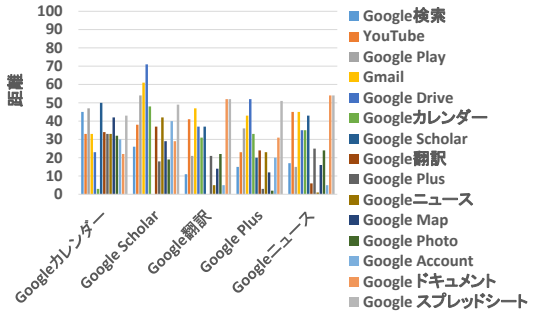

図 5　同定を行いたいサービスとデータベース内の各サービスとの距離 II

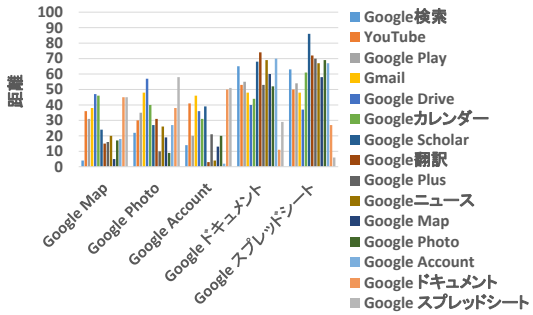

図 6　同定を行いたいサービスとデータベース内の各サービスとの距離 III

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．その他の Google サービスに対しても同定を行ったところ，本性能評価における同定成功率は 87%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．

5. おわりに

本稿では，複数のコネクションを集計してサービス同定を行う手法を提案し，その同定精度の評価を行った．サービス同定精度は，同定成功率 87%を達成し，本手法の有効性が確認された．

今後は，更に多くのサービスにて評価を行う予定である．

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。

参考文献

[1] 原雅貴・韮澤慎之介・中尾彰宏・小口正人・山本周・山口実靖, " N-gram パケット解析に基づくコンテンツ同定",信学技報, vol. 116, no. 111, NS2016-48, pp. 113-117, 2016 年 6 月.