一般社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
IEICE Technical Report
NS2021-129(2022-03)

DNS クエリログを用いたアクセス先ドメインの類似性に基づくユーザ属性分析

矢野安希子† 野林 大起†† 塚本 和也††† 池永 全志††

† 九州工業大学工学部 〒804–8550 福岡県北九州市戸畑区仙水町 1-1
†† 九州工業大学大学院 工学研究院 〒804–8550 福岡県北九州市戸畑区仙水町 1-1
††† 九州工業大学大学院 情報工学研究院 〒820–8502 福岡県飯塚市川津 680–4
E-mail: †yano.akiko171@mail.kyuech.jp, ††{nova,ike}@ecs.kyutech.ac.jp, †††tsukamoto@cse.kyutech.ac.jp

あらまし 観光地における訪日外国人向けの適切な案内情報の提供など，携帯端末利用者に対する効果的な情報配信の実現が期待されている．特に，多様な国や地域からの観光客が集まる場所では，国籍ごとに興味の対象が異なる場合があり，商品の購入や他地域への回遊を促すために提供すべき情報等も，その情報を受信するユーザの属性に適した内容であることが求められる．そのためには携帯端末利用者の属性を把握する必要があり，なかでも外国人観光客等の国籍に関する情報は，重要な属性情報となる．本研究では，端末の利用者が個人ユーザである場合を想定し，DNS クエリログに含まれるクライアント端末のアクセス先ドメイン情報から，クライアント端末毎の属性を分析する．ここではユーザ属性として国籍に着目し，ユーザが問い合わせたドメイン名から国籍情報を含む可能性のあるクエリログとして ccTLD を抽出し，アクセス頻度やアクセス先の分布などの特徴を分析することによって，類似した国籍の特徴を有するユーザグループを分類する手法について検討する．

キーワード ユーザ属性，クエリログ

User Attribute Analysis Based on Similarity of DNS Query Logs.

Akiko YANO†, Daiki NOBAYASHI††, Kazuya TSUKAMOTO†††, and Takeshi IKENAGA††

† Graduate School of Engineering,Kyushu Institute of Technology 1–1 Sensui-tyo, Tobata-ku, Kitakyusyu, Fukuoka, 804–8550 Japan
†† Faculty of Computer Science and Systems Engineering,Kyushu Institute of Technology 1–1 Sensui-cho, Tobata-ku, Kitakyusyu, Fukuoka, 804–8550 Japan
††† Faculty of Computer Science and Systems Engineering,Kyushu Institute of Technology 680–4 Kawazu, Iizuka-shi, Fukuoka, 820–8502, Japan
E-mail: †yano.akiko171@mail.kyuech.jp, ††{nova,ike}@ecs.kyutech.ac.jp, †††tsukamoto@cse.kyutech.ac.jp

Abstract It is expected to realize selective and effective information distribution for users of smart devices, such as providing appropriate tourism information for foreign visitors to Japan. In the places where tourists from various countries and regions tend to gather, the interest differs according to the nationality of the tourists. If the information suitable for the user’s attribute can be appropriately distributed, it is possible to urge the user to purchase goods and travel to other regions. Therefore, it is one of the most important functions to grasp the attribute of a user using a portable device, and the nationality information of foreign tourists becomes the important attribute information. In this study, we analyze the characteristics of each device using the domain information included in the DNS query log. The country code TLD (ccTLD) is extracted from the domain name queried by each user, and then the access frequency and distribution for ccTLD are analyzed. From the result, the method which classifies the user group with the features of a similar nationality is examined.

Key words User analysis，Query log

1. まえがき

新型コロナ禍収束後のインバウンド観光客の再興を見据え，その際の観光地での携帯端末を用いた観光案内やより精度の高いターゲティング広告を実現するため，携帯端末利用者の属性分析技術が求められている．特に，観光地における訪日外国人向けの情報発信は，商品の購入や回遊促進などのプロモーションのみならず，非常災害時の避難誘導などにおいて効果的な情報提供を行うためにも重要である．そのため，ユーザ属性の中でも，国籍や使用言語に関する情報の識別は重要な課題である．

これまでの取り組み事例として，スマートフォンが Wi-Fi アクセスポイントに接続した際のキャプティブポータルサイトを通して言語設定情報を読み取る手法がある．しかし，英語やフランス語，スペイン語等の多国籍で使用されている言語を利用するユーザについては，国籍を判定することは困難であった．

そこで本研究ではこの属性判別に Domain Name System(DNS) クエリログを用いることを検討する．このクエリログには様々な情報が含まれており，ネットワークの運用やセキュリティの観点で多くの情報を得ることが可能である．例えば，観光地に自治体等が管理するフリー Wi-Fi アクセスポイントを設置することで，適切な利用許諾を得たうえで，この Wi-Fi に接続するクライアント端末の DNS クエリログを自治体等が収集することが可能となる．そこで，入手した DNS クエリログに含まれるクライアント端末のアクセス先ドメイン情報からクライアント端末毎の属性を分析する．携帯端末の利用者が個人ユーザである場合を想定し，DNS 問い合わせ先ドメインの類似性から，ユーザの属性情報を抽出する手法を検討する．

本稿では，2 節において DNS ログ情報の利活用に関する関連研究について紹介する．3 節では，DNS ログ情報からの国籍情報に着目したユーザ属性分析に関する提案手法について延べ，4 節では提案手法を実際の DNS ログデータに適用した場合の結果を示す．5 節で結果に関する議論と考察を述べ，最後に 6 節でまとめる．

2. 関連研究

DNS クエリログを利用した検出の例として，悪性ドメイン検出手法がある．文献 [3] では，DNS クエリログの送信挙動を表す分散ベクトルを構築し，アノマリ型の検出手法より悪性端末検出を行った．さらに文献 [1] では，悪性 DNS クエリを Word2Vec [4] により 3 つのクラスタに分類し，それぞれが共通の原因のクエリのみで構成できることを示した．このように DNS クエリはユーザが通信する際に必ず発生するものであるため，クエリログを分析することによってネットワーク内の感染端末を検出することに活用することが期待されている．

他の目的にも DNS クエリログは用いられており，また文献 [2] では，ユーザ通信行動抽出を行なっている．文字列として類似する FQDN をグルーピングし，データを次元圧縮した後，時刻，ユーザ，問い合わせドメインの観点から，これらを 3 階のテンソルと捉え，非負値テンソル分解アプローチより同時にアクセスが発生しやすいドメインと通信時間帯のパターンを抽出している．そこで本研究では，DNS クエリログを用いた新たな検出対象である，ユーザ属性の分析を目的とする．

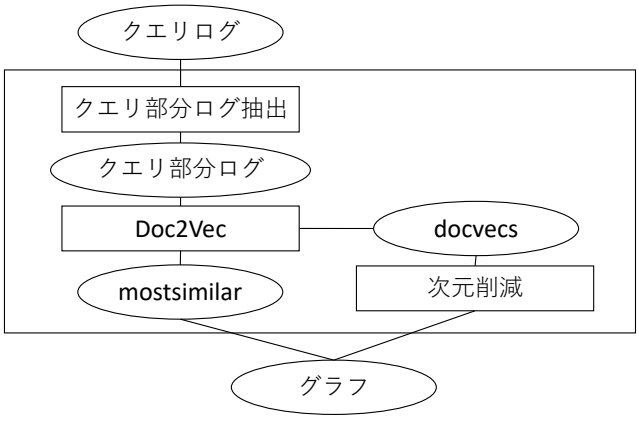

図 1　提 案 手 法

3. 提案手法

本稿では，ユーザ同士のアクセス先ドメインの類似性に基づく属性分析を提案する. その独自性は，(1) 各ユーザのクエリログ情報をひとつの文章と見立て，文章をベクトル化することで機械学習に使えるようにするアルゴリズムである Doc2Vec [5] を利用し類似性を出したこと，(2) ユーザ個人を判別するのではなく，DNS 問い合わせドメインの類似性から同一国籍と思われるユーザ同士をグループに分類し分析したこと，の 2 点である．クエリログを文章と見立てることで，ユーザのアクセスの順番に依存する特徴の出力を狙うことが可能となる．またユーザをグループに分けることで，ユーザ個人を見るだけではわかり得ない特徴を獲得できると考えた．これにより，単一言語国家だけでなく複数の公用語が存在する多言語国家についての判別も可能になる．

DNS クエリログには，問い合わせ IP アドレス，問い合わせ先ドメイン名の情報が含まれている為，ネットワーク利用時に発生する DNS クエリログからユーザの国籍に関する属性情報を抽出することができると考える．DNS クエリログの問い合わせ先トップレベルドメイン (TLD) のうち，country code TLD (ccTLD) に着目し，ccTLD の傾向からユーザの国籍を推定する手法を提案する．提案手法では ccTLD の国別問い合わせ回数を降順に並び替えた情報 (提案手法 1) と，ccTLD の問い合わせ履歴を順に並べた情報 (提案手法 2) の 2 つを，ユーザの特徴として利用する. 次に抽出したユーザのアクセス先情報から，Doc2Vec を用いて 100 次元の特徴ベクトルを算出する. つまり，DNS クエリから抽出された各ユーザの問い合わせ先 ccTLD の情報をひとつの文章と見立て，Doc2Vec を用いて得られる特徴量から類似性を推定する. ユーザの特徴量に基づく他のユーザとの問い合わせに対する類似度を出力することで分析が可能となる. また，獲得した 100 次元のユーザ特徴量を，主成分分析により次元削減を行い，2 次元情報として可視化することで類似性を確認する．

4. 実　　験

4.1 クエリログ部分ログ抽出

提案手法である DNS クエリログによるアクセス先 TLD に基づく分析を実施するため，2020 年 10 月 1 日から 2020 年 10 月 31 日までの九州工業大学の DNS サーバのクエリログを用いた．学内 Wi-Fi ルータが割り振る 131.206.224.0/22，150.69.224.0/22 に当てはまる IP アドレスを抽出し，分析には，HPE ProLiant DL360 Gen10 (Intel Xeon 1.8GHz, 4 core, Mem32GB) を用いた．1 つの IP アドレスに対し 1 人のユーザが利用していると仮定し，また，ccTLD は IANA の Root Zone Database より抽出し，外国人留学生在学状況より本学に在籍していないと思われる国籍の ccTLD であり顕著なドメインハックに利用可能 ccTLD は排除した．

4.2 特徴ベクトルの抽出

利用する DNS クエリログから提案手法における 2 パターンの特徴を用いて分析を実施した．Doc2Vec は gensim ライブラリより利用した．1 人のユーザのアクセス先 TLD をカウントし回数の多いものから降順に ccTLD を並べ，一行のリストにまとめた．これを一つの文章と見立て，gensim ライブラリにて model を作成した．model 作成時のパラメータとして単語の並びを考慮しより優れた精度がでることが報告されていることから，前後の単語から対象単語を推測するニューラルネットワークである Distributed Memory(DMPV) を用いた．また，単語の予測に何単語を用いるかを指定する window ではサイズを 1 に指定した．これは，ユーザのアクセス先 ccTLD の要素数が 3 以下であるユーザが半数ほどであること，また離れた要素に比べ隣り合う要素により関係性があると考察したことからである．作成次元は 100 次元に指定した．それぞれのユーザの特徴量ベクトルを算出する docvecs メソッドと cos 類似度を求める most_similar メソッドを使用した．docvecs メソッドでは，各ユーザの特徴量が 100 次元で出力され most_similar メソッドでは各ユーザと cos 類似度の高いユーザを 10 選び出力した．

```
[('kr', 5), ('cn', 2), ('in', 2), ('fr', 2)]
[('fr', 7), ('in', 7), ('cn', 5), ('id', 2), ('tw', 2),
('th', 2), ('kr', 1)]
[('th', 1), ('fr', 1), ('cn', 1), ('kr', 1), ('tw', 1)]
[('th', 8), ('kr', 6), ('vn', 2), ('id', 1)]
[('cn', 5), ('kr', 3), ('vn', 1)]
[('id', 1), ('th', 1), ('kr', 1)]
[('in', 13), ('kr', 2), ('fr', 1)]
[('in', 1), ('id', 1), ('tw', 1)]
[('cn', 4), ('fr', 4), ('kr', 3), ('th', 2), ('tw', 2),
('id', 2), ('ph', 2), ('in', 1)]
[('kr', 4), ('vn', 1), ('fr', 1), ('cn', 1)]
```

```
['kr', 'cn', 'in', 'fr']
['fr', 'in', 'cn', 'id', 'tw', 'th', 'kr']
['th', 'fr', 'cn', 'kr', 'tw']
['th', 'kr', 'vn', 'id']
['cn', 'kr', 'vn']
['id', 'th', 'kr']
['in', 'kr', 'fr']
['in', 'id', 'tw']
```

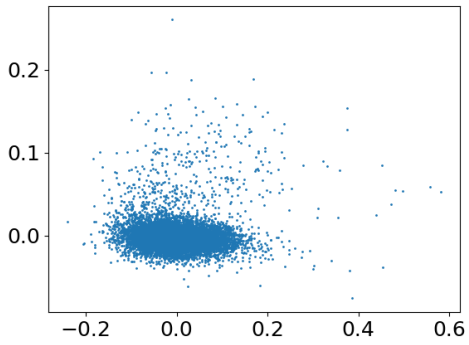

図 2　提案手法 1

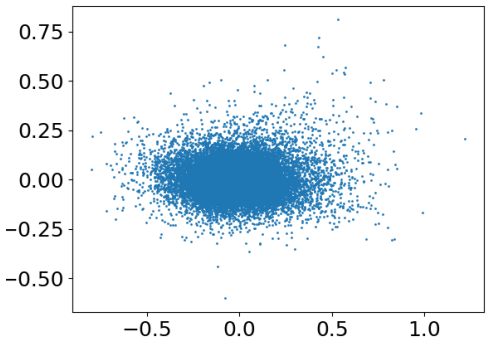

図 3　提案手法 2

```
['cn', 'fr', 'kr', 'th', 'tw', 'id', 'ph', 'in']
['kr', 'vn', 'fr', 'cn']
```

```
['kr', 'cn', 'in', 'fr']  =>
[ 2.1167449e-03 -4.0363502e-03 -1.4642774e-03  4.1042641e-04
  3.9511416e-03 -1.7104612e-03 -1.5750190e-03  8.6015780e-03
  8.3607296e-04  3.3028624e-03  2.4335459e-04 -7.8532472e-04
 -6.8357838e-03  5.0253700e-03 -6.1476217e-03 -2.0336439e-03
 -1.2912489e-03 -2.0030420e-06  1.3294478e-04 -9.9176099e-04
 -5.0236494e-03 -3.9914693e-03 -5.6786686e-03 -1.6462982e-03
  4.7129109e-03 -4.8882170e-03 -2.3795024e-03 -1.4369822e-03
 -4.8972811e-03  8.2532726e-03 -6.4627705e-03  1.9778375e-04
  5.1550725e-03 -4.0071947e-03 -3.6245764e-03  8.3869509e-03
  4.5749871e-04 -2.5843654e-03  2.9854155e-03  6.3599064e-04
  5.6164023e-03  3.4087547e-03 -4.3514613e-03  1.6332314e-03
 -6.3278824e-03 -5.8565149e-04  3.8855928e-03 -4.1519823e-03
 -8.1507489e-05  1.3658085e-03 -5.5339495e-03 -6.3954396e-03
 -1.0426870e-02 -3.0225546e-03  3.8592170e-03  8.0396369e-04
  2.9643532e-03 -2.1992463e-03  2.5233061e-03  2.3231050e-03
 -6.9547328e-04 -4.1057896e-03 -3.2563529e-03 -4.4590095e-05
 -3.0312999e-03 -1.2774356e-03  5.5692825e-03 -3.4475490e-03
 -1.0902004e-05 -2.0758156e-04  5.8352244e-03  5.6334361e-03
  1.7820842e-03 -4.7522681e-03  2.1246811e-03 -4.4088764e-03
 -3.9728708e-04 -2.6131507e-03 -1.1593234e-03  2.4969338e-03
  9.5973387e-03 -6.9093257e-03 -1.4230614e-03 -1.3327978e-03
 -2.3684232e-03 -3.1108023e-03  9.5134019e-04  5.5767000e-03
 -1.8760277e-03  3.8015300e-03 -2.8606909e-03  1.2216936e-03
  6.4429189e-03 -6.4056227e-04  7.5399620e-03 -7.4662231e-03
 -5.2923635e-03 -2.1148506e-03 -2.9477468e-03 -5.3431783e-03]
```

4.3 次元削減

次元削減には主成分分析である PCA を用いた．Scit-learn を使い，100 次元から 2 次元のデータへ次元削減した．

結果のグラフを図 4. 3, 4. 3 に示す．

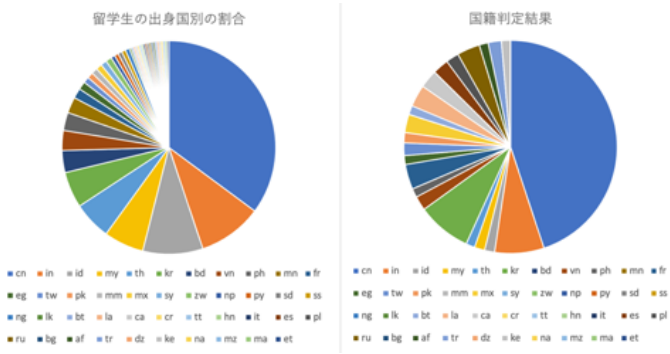

図 4　九工大の留学生数及び国籍判定に基づく国別の割合

5.　ユーザ属性分析と考察

5.1　グラフから見たユーザ属性分析

図 4 に ccTLD の国別問い合わせ回数を降順に並び替えた特徴を用いた分析結果，ccTLD の問い合わせ履歴を順に並べた特徴を用いた分析結果を示す．どちらのグラフにおいても，座標 (0, 0) を中心に大きなクラスタを形成していることが確認できる．このことから，このクラスタが今回のユーザの大半を占める日本人学生であると推測できる．このクラスタの中心から離れるにつれ，ユーザ数が少なくまばらになっている．また，今回利用した DNS のクエリログは学内のデータを用いていることから，ウイルス等に感染して正常でない挙動をする通信は大きな割合を占めないと仮定すると，クラスタから離れたユーザが留学生の通信であると推測することができると考えた．また、図の結果より，アクセス履歴の順番より問い合わせ回数に基づく特徴抽出の方が明確に国籍に関する属性の違いを抽出できると考えられる．

5.2　国 籍 推 定

次に，対象とした DNS クエリログのうち，特徴的な ccTLD へのアクセス状況から留学生の国籍推定を行った．本推定では，ccTLD 毎の問い合わせ先を分類し，明らかに企業または団体が利用している，ccTLD の国に関連しないクエリを抽出し，それらの問い合わせ先を除外することで，国籍判定に活用する．問い合わせ元 IP 毎に ccTLD へのアクセスをカウントすることで，問い合わせをした端末の国籍判定を実施する．一般的に，Web 利用時またはメール内の広告画像への問い合わせにより，一つの端末から複数の ccTLD への問い合わせが発生することが多い．そこで，本研究では，一つの端末から複数の ccTLD へのクエリが有る場合には，クエリが最も多い国をその端末の国籍であると判定してカウントを行った．

今回利用したクエリログについては，該当の期間は新型コロナウイルス感染拡大が落ち着いており，後期日程開始と共に，学内 Wi-Fi の利用者がある程度確保できると想定できた．さらに，学外からの来訪者が少なく，特に外国からの来訪者はほぼ無いと考えられ，学内 Wi-Fi を利用する外国人は，九工大の留学生のみであると想定できた．図 4 に九工大の留学生数に基づく国別の割合（左）と今回の国籍判定に基づく国別の割合の結果を示す．この結果より，国籍判定の結果が，九工大の留学生の出身と同じ傾向になることが確認できた．一部留学生が多いにも関わらずクエリ数が少ないケースが存在するが，これは留学生が大学に来ていない（オンラインによる講義），PC やスマホを持ち込んでも大学内の Wi-Fi を利用していない等が想定できる．また，国籍判定結果と留学生の国籍情報の類似性を示すユークリッド距離は 0.13 という結果となり，このことからも非常に類似性が高いことが確認できた．以上の結果より，今回の国籍判定方法により，ある程度の国籍判定が可能であるという実現可能性を示すことができた．一方で，今回の分析では，対象とした IP アドレスが学内 Wi-Fi 経由の通信のみであったことから，国籍を抽出できない国も存在した．そのため，今後は Wi-Fi だけではなく，他の問い合わせ元 IP アドレスについても調査する必要がある．また，今回は ccTLD のみを用いて国籍判定を実施しているが，gTLD（general top level domain，例えば com 等）の問い合わせ先を分析することで国籍判定が可能になるよう改良が必要である．

6.　ま　と　め

端末の利用者が個人ユーザである場合を想定し，DNS 問い合わせ先ドメインの特徴量の類似性から日本人ユーザと留学生ユーザを大まかに判定することができた．

今後の課題として，ドメインハックに用いられる ccTLD を排除する等，精度向上に向けて検討を行う必要がある．

謝辞　本研究の一部は，情報通信研究機構の委託研究による成果を含む．ここに記して謝意を表す．

文　　献

[1] 佐藤彰洋, 中村豊, 小倉光貴, 野林大起, 池永全志, "ブラックリストに基づく検出の効率化に向けた悪性 DNS クエリ分類手法", インターネットと運用技術シンポジウム論文集,Vol.2018,pp.100–105,2018.

[2] 畑中耕太朗, 木村達明, 滝根哲哉, "非負値テンソル分解による DNS クエリログからのユーザ通信行動抽出", 電子情報通信学会技術研究報告 NVolume121,umber102,pp.57-62.

[3] 仲宗根一成, 北口 善明, 山岡 克式, "機械学習を用いた DNS クエリ/応答のログ解析による悪性端末検出手法の提案" 電子情報通信学会信学技報 IN2018-129,2019.

[4] T. Mikolov et al., "Distributed Representations of Words and Phrases and their Compositionality," Advances in Neural Information Processing Systems, pp. 3111–3119, 2013.

[5] Quoc Le, Tomas Mikolov, "Distributed Representations of Sentences and Documents," Proc. of the 31st International Conference on Machine Learning, PMLR 32(2), pp.1188–1196, 2014.