

位置情報サービスの規約・ポリシーにおける位置情報の自動抽出の検討

金杉洋・松原剛・日野智至・柴崎亮介

Study on Automated Extraction of Location Definition

from Terms and Policies of Location Based Services

Hiroshi KANASUGI, Go MATSUBARA, Tomoyuki HINO and

Ryosuke SHIBASAKI

Abstract: Personal data have been necessary to provide useful functionalities and comfortable services. However, since personal data often contain privacy concerning information, in general, notifying service detail with necessary data contents and their purpose and getting explicit agreement from a user are fundamental procedure when utilizing personal data. In terms of the agreement procedure, users do not always check terms of services and privacy policies or do not necessarily understand what they have agreed with utilization of their own personal data. This paper tries to extract spatiotemporal resolution of location information from location-based services' documents of terms of services and privacy policy.

Keywords: 位置情報サービス (Location Based Services), 利用規約 (Terms of Services), プライバシーポリシー (Privacy Policy), パーソナルデータ (Personal Data)

1. 背景と目的

位置情報をはじめとしたパーソナルデータは、日常的な行動に伴って生成され、様々なサービスの利用において不可欠となっている。同時に、パーソナルデータは個人情報やプライバシーに関わる機微情報となり得るため、利用に際しては利用者への利用項目と目的等を通知し、同意を得ることが必要とされている（個人情報保護委員会, 2016; EU General Data Protection Regulation, 2016）。しかし、多数のサービスの利用に伴って繰り返される同意取得の手続きは、多くの専門用語を含む利用規約・プライバシーポリシーの読解の困難さとその分量の多さから、利用者の負担は大きい。結果として、多くの利用者は内容を十分に理解し

金杉洋

東京大学 空間情報科学研究センター

yok@csis.u-tokyo.ac.jp

きれずに提示される規約・プライバシーポリシーに同意し、実態として利用者はサービス提供者がどんなパーソナルデータをどの程度まで収集・利用しているのか、把握・確認できていない。

こうした状況に対し、サービスやアプリの利用について、パーソナルデータの取得・利用に関わるサービス提供者と利用者との相互理解の齟齬や情報の非対称性を解決する試みが行われてきている。岩本ら(2015)はいくつかのサービスの利用規約からサービス内で利用されるパーソナルデータの項目と利用目的を独自に抽出し、それらに対応付けて可視化すると共に、利用者のパーソナルデータの利活用状況の定量化を試みている。S. Zimmeck et al, (2017) では、Play ストアで公開されているアプリを対象に、Android アプリのパーミッション要求やライブラリの利用状況

等の解析と、アプリ提供者の提示するプライバシーポリシーの自然言語解析結果を組み合わせることで、プライバシー要件に対するアプリの適合状況を推測・提示する仕組みを提案している。デバイスメーカー独自のアプリへの対応や、日本語を含む他言語への対応が改善されれば、提供者と利用者の双方に有用な仕組みとなろう。他方で、ToSBack¹⁾は利用者がサービス利用条件に同意した後、利用規約・プライバシーポリシーが更新されることに留意し、各サービスの利用規約・プライバシーポリシーを定期的にスクレイピングし、差分をGitHubでアーカイブする機能を提供している。利用者が同意した時点の条件が当該サービスで時系列に示されるとは限らないため、こうしたアーカイブの仕組みも有用となろう。

利用規約やプライバシーポリシーを通じて利用者の理解と同意を得る手続きは、サービス提供に必要なとなるパーソナルデータの項目とその取得に必要な技術的説明に加え、個人情報保護とプライバシー保護の法的な説明を含んだ複雑な文書の読解を、度々利用者に要求することが課題となっている。サービス提供者が利用規約・プライバシーポリシーを分かりやすく示すことが重要なと同時に、利用者自身の理解を支援することも喫緊の課題と言えよう。特に利用者が懸念する位置情報の利用については、利用目的や利用主体に関する詳細な説明が求められている（総務省情報通信政策研究所，2014）。

そこで本論文では、利用規約・プライバシーポリシーに記載されるパーソナルデータ項目のうち、特に利用者の位置情報に関わるデータ項目について、その時空間分解能がどのように示されているのか調査するとともに、自然言語処理技術によってどの程度その内容が自動抽出可能かを試行する。具体的な手順として、まずWebに公開された各サービスの利用規約・プライバシーポリシーをスクレイピングし蓄積する仕組みを構築する。次に、位置情報サービスを含めた50のサイ

トの規約・ポリシーから、基本的な形態素解析を通じて得られる単語を用いて、利用される位置情報の時空間分解能を抽出・考察するとともに、自動抽出の可能性を検討する。

2. 利用規約・プライバシーポリシー収集と分析

2.1 スクレイピングによるデータ収集

各サービスのWebサイトで公開されている利用規約とプライバシーポリシーを、定期的に収集・蓄積し、更新箇所として差分を可視化する仕組みとして図1のシステムを試作した。ToSBackの仕組みを踏襲し、各サービスのWebに公開された規約とポリシーのスクレイピングし、GitHubのリポジトリに蓄積する。更新履歴の管理と差分の可視化はGitHubの機能をそのまま利用している。主な機能追加・変更点として、日本語対応やサポート外となっていたライブラリの更新などに加え、利用規約・プライバシーポリシーをPDFで公開しているサービスも少数ながら存在することを考慮し、HTMLだけでなくPDFについてもテキスト変換し、アーカイブに追加できるようにしている。スクレイピングにあたり、サービスごとに異なるWebサイトのデータ構造に対応するため、サービスごとに設定ファイルを設けている。

2.2 位置情報サービスと時空間分解能の抽出

スクレイピングによって得られた利用規約・プライバシーポリシーのテキストデータをもとに、位置情報を取扱うサービスを抽出する。本論文で

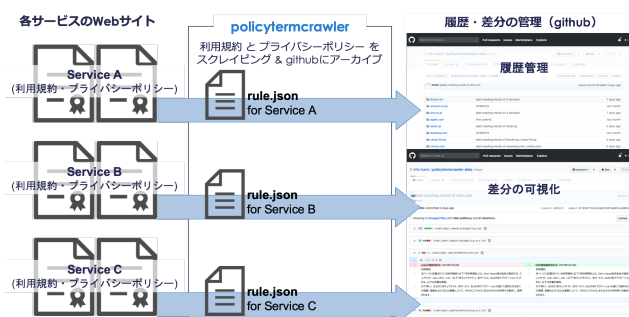
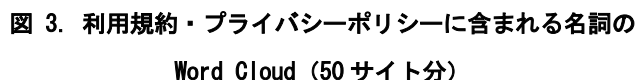


図 1. 利用規約・プライバシーポリシーのスクレイピングとGitHubでのアーカイブ

位置情報の空間分解能として、具体的な数値が利用規約やプライバシーポリシーに記載されることは稀であり、通常は「GPS」「Wi-Fi アクセスポイント」「携帯電話基地局」「Bluetooth」「IP アドレス」などの測位手法が記載される．そこで、上記で抽出した位置情報サービスにこれら測位手法に関わる単語（名詞）が共起する場合に、空間分解能が記載されているものとする．一方で、時間分解能には測位手法のような固有名詞がないため、「頻度」「間隔」「タイミング」のような時間に係る単語の有無を調査する．なお、形態素解析には、kuromoji²⁾を、日本語辞書には mecab-ipadic-NEologd³⁾を使用した．

本論文執筆時点では、位置情報を利用するサービスに限らず、恣意的に選択した 50 件のサービスから利用規約・プライバシーポリシーを取得している。本論文では 2019 年 8 月 16 日時点で取得されているデータを用いる。

図 2 にスクレイピングにより得られた 50 サービスの利用規約・プライバシーポリシーの文字数をまとめた。利用規約とプライバシーポリシーの文字数の合計の最小値と最大値は、それぞれ 4,147 文字と 87,353 文字、平均値は 25,118 文字でサービス間の差異が顕著に表れている。サービス提供者ごとの特徴も影響しているだろうが、特に、決済に関わるサービスの分量が大きいことなどから、取得するパーソナルデータの種類とその取扱に加え、サービス内容に関わる法令等への対応の有無により、分量に差異があることが窺える。図 3 には利用規約・プライバシーポリシーから得られた名詞 9,832 種類の出現頻度に基づく Word Cloud を示した。「位置情報」の出現頻度は全体を



通じて 154 件で、上位 3%内に含まれており、今回取得しているサービス内では比較的頻出する単語といえよう。そのため、対象とする「位置情報サービス」に限ると、29 件 (58%) のサービスが該当する結果となった。

得られた位置情報サービス 29 件について、2.2 で示した空間分解能に相当する測位手法（位置推定手法）に関する単語の出現頻度を図 4 に示した。29 件のうち 25 件 (86%) のサービスにおいて何らかの位置推定手法に言及しており、空間分解能が抽出できる可能性を示している。また、GPS や Wi-Fi などの直接的な測位手法よりも、IP アドレス

利用を含むサービスが多いが、これはアクセス元（地域）の特定などでも IP アドレスが利用されることが影響していると推測できる。一方で、位置情報サービスであっても、空間分解能に触れていない4件については、今回抽出対象とした単語以外が使用されている可能性など、精査が必要であろう。

同様に時間分解能についても同様に 2.2 で示した単語の出現有無を調査したが、29 件のうち 5 件のみで極少数が確認されるに留まった。恣意的に選択した単語であるため、改めて名詞以外も含めた検討が必要であることに加え、利用規約・プライバシーポリシーに時間分解能の記載が無いことも踏まえた検討が今後必要であろう。

4. まとめと今後の課題

本論文では、利用規約・プライバシーポリシーに記載されるパーソナルデータ項目のうち、特に利用者の位置情報に関わるデータ項目について、その時空間分解能がどのように示されているのか調査するとともに、自然言語処理技術によってどの程度その内容が自動抽出可能かを試行した。

利用規約・プライバシーポリシーをスクレイピングし蓄積する仕組みを構築し、その上で基本的な形態素解析を通じて得られる単語を用いて、時空間分解能の抽出可能性を検討した。29 件の位置情報サービスを対象に試行したところ、空間分解能は「GPS」などの測位手法に関わる単語が 86% のサービスに含まれていることが分かった。一方で、時間分解能については、抽出する単語の選択に加え、時間分解能に関する記述の有無からの精査が必要であろう。

謝辞

本研究は、国立研究開発法人科学技術振興機構戦略的創造研究推進事業（社会技術研究開発）「人と情報のエコシステム」研究開発領域における研究成果の一部として実施した。

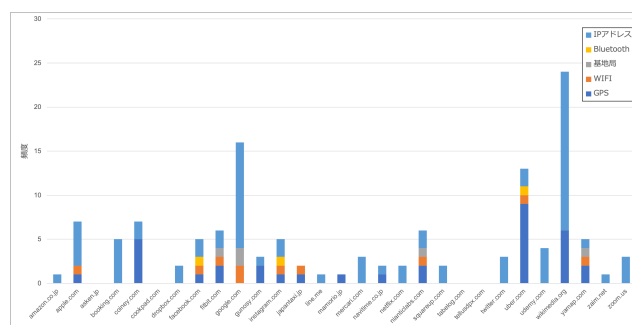


図 4. 位置情報サービスの規約・ポリシーにおける空間分解能（位置推定手法）の出現頻度

補注

- 1) ToSBack, <https://tosback.org/>
- 2) Kuromoji, <https://GitHub.com/atilika/kuromoji>
- 3) mecab-ipadic-neologd, <https://GitHub.com/neologd/mecab-ipadic-neologd>

参考文献

- 個人情報保護委員会(2016), 個人情報の保護に関する法律についてのガイドライン（通則編）, <<https://www.ppc.go.jp/files/pdf/guidelines01.pdf>>.
- EU General Data Protection Regulation (2016), <<https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679>>.
- 岩本佑太, 山内正人, 砂原秀樹(2015) パーソナル情報可視化システムの提案, マルチメディア・分散・協調とモバイルシンポジウム(DICOMO 2015), pp.1870-1871, 2015
- S. Zimmeck, Z. Wang, L. Zou, R. Iyengar, B. Liu, F. Schaub, S. Wilson, N. Sadeh, S.M. Bellovin, J.R. Reidenberg (2017), Automated Analysis of Privacy Requirements for Mobile Apps, NDSS'17: Network and Distributed System Security Symposium, 2017
- 総務省情報通信政策研究所（2014）位置情報の利用に対する意識調査, <<http://www.soumu.go.jp/iicp/chousakenkyu/data/research/survey/telecom/2014/location-info.pdf>>.