

SNS 画像投稿時のプライバシー侵害予知サービスの提案

小山 貴之[†] 宋 洋[†] 町田 史門[†] 嶋田 茂[†] 越前 功[‡]

[†] 首都大学東京 産業技術大学院大学 〒140-0011 東京都品川区東大井 1-10-40

[‡] 国立情報学研究所コンテンツ科学研究系 〒101-8430 東京都千代田区一ツ橋 2-1-2

E-mail: [†] { a1119tk, a1131ys, a1151sm, shimada-shigeru }@aiit.ac.jp [‡] ieichizen@nii.ac.jp

あらまし SNS への画像投稿が増加するにつれ、それに伴うプライバシー侵害のトラブルが多く発生するようになってきた。そこで、長期間蓄積された大規模なツイートアーカイブを対象にして、プライバシー侵害の要因を分析すると共に、共通のプライバシー解釈を行うコミュニティの抽出を行った。そして、これらの結果を用いて新たに投稿される画像記事の分析から、どのコミュニティでのプライバシー侵害になるかの予測を行い、プライバシー侵害者又は被害者となり得るユーザへの事前通知を行う新たなサービスを提案する。

キーワード プライバシー侵害, SNS, 画像投稿, 自然言語処理, クラスタ分析, モバイルインタフェース, コミュニティ

Proposing of Predictive Notification Service by Privacy Invasion on Posting Image to SNS

Takayuki Koyama[†] Yang Song[†] Shimon Machida[†] Shigeru Shimada[†] Isao Echizen[‡]

[†] Tokyo Metropolitan University AIIT 1-10-40 Higashi Ohi, Shinagawa-ku, Tokyo, 140-0011 Japan

[‡] National Institute of Informatics 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo, 101-8430 Japan

E-mail: [†] { a1119tk, a1131ys, a1151sm, shimada-shigeru }@aiit.ac.jp [‡] ieichizen@nii.ac.jp

Abstract With increasing post images in SNS, invasion of privacy problems associated with it has come to be frequently occurs. Therefore, we analyze the factors on the invasion of privacy and made the extraction of communities by common privacy interpret, targeted to the large tweet archive that is stored a long period of time. We propose new services that make proactive notification to the user that may be a victim, based on the predictions that become an invasion of privacy belonged to whichever community, by an analysis of the next image article posted, using these results.

Keyword Privacy Invasion, SNS, Post Image, Natural Language Processing, Cluster Analysis, Mobile Interface, Community

1. 研究背景と目的

1.1. 研究背景

近年、高速ワイヤレスネットワーク環境の普及やモバイルデバイスの性能向上により、スマートフォンを用いて SNS (Social Networking Service) を利用するユーザが増えている。例えば、Facebook の 9 億人のユーザのうち、約 55% がモバイルデバイスからサービスを利用しており [1]、同様に Twitter ユーザのうち、約 60% がモバイルデバイスからのアクセスがあることが報告されている [2]。また、従来の SNS は掲示板やブログなどテキストデータによるコミュニケーションであったのが、現在では、画像やビデオ等のマルチメディアをその場でモバイルデバイスに記録し、SNS 記事として直接投稿する方式が主流となっている。その結果、マルチメディアの SNS 記事に占める割合は年々増加している。このようにモバイルデバイスを通じて安易にマルチメディア情報を SNS へ投稿できる

ことから、プライバシーについての配慮が欠けてプライバシー侵害が発生することが多くなっている。このような状況を裏付ける調査として、トレンドマイクロ社の Facebook 利用者に対する画像共有のプライバシー侵害の調査があり、90.6% のユーザが他人による勝手な画像投稿を不快に感じていることが報告されている [3]。

SNS への投稿記事によりプライバシー侵害が発生する場合の多くは、ユーザの不注意により投稿された後に発生しており、投稿前に事前にプライバシー侵害を防止する方式 (Opt-in 方式) を適用することが難しいといった問題がある。従って、従来 SNS 各社は事後の対応策 (Opt-out 方式) として、対応方針を示すプライバシーポリシーや投稿記事の公開範囲の推奨等を記述提示する対応を行っている。ところが、これらの Opt-out 方式の記述提示では、全てのケースを網羅するための複雑で長文の記載が行われることが多いため、

全てを閲覧して理解するユーザは少なく，対応策は各ユーザの裁量にゆだねられる状況となっている．従って，記事を投稿するユーザによるプライバシー侵害の対応はほとんど行われること無く，プライバシー侵害の被害者又はその侵害が発生していることに気がついたユーザからの指摘からトラブルに発展するケースが多くなっている．しかも，このプライバシー侵害の解釈は，各ユーザによって異なるためトラブルの発生率を高めている．

1.2. 研究目的

そこで我々は，SNS への画像投稿に伴うプライバシー侵害を抑制するために，プライバシーの解釈が人によって異なることを考慮する必要があると考え，長期間蓄積された多量の SNS 投稿記事の分析を行うことにより主要な侵害の要因を見つけ，その各要因に共通するプライバシーの解釈集合をプライバシーコミュニティとして抽出することを目的とする．そして更に，これらの抽出した結果に基づき，新規に投稿された記事の分析から，どのプライバシーコミュニティに属するかかの判定を行い，そのコミュニティに関連した対応方法を事前に予知するサービスを提案する．

本稿では，上記サービスのプロトタイプを構築し，評価を行うことで，サービスの精度向上を目指す．以降，2 章で SNS 及びプライバシーについて述べ，3 章で SNS 投稿記事の分析を行い，4 章ではプライバシー情報の分析を述べ，5 章でサービスの構築を行い，6 章で評価を行う．

2. SNS におけるプライバシー

2.1. プライバシー侵害定義の検討

日本の個人情報保護法におけるプライバシー侵害の法的な成立要件は，以下の情報を本人の許可なく公開することである [4][5]．

- ・私生活の事実または，事実らしいと受け取られる情報
- ・一般的に本人の判断からして，公開して欲しくない情報
- ・一般の人には，まだ知られていない情報

しかし上記の要件には，人為的な感情でしか判別することができない表現が多く，システムによる判別を行うためには定義があいまいである．そこで，より情報処理の観点に立った表現の要件を探すと，青柳によりプライバシー侵害に関する新たな定義が提案されており，「“個人識別情報”と“不可侵私的領域”が，本人の意思に反して結びつくことである」としている [6]．ここで，“個人識別情報”とは，{氏名・性別・生年月日・住所等の個人を識別できる情報}で，“不可侵私的領域”とは，{私生活上の領域，一般人の感受性を基準

として公開を欲し不想われる領域，非公知の領域，公開によって当該詩人が現実に深いや不安の念を覚える領域}としている．しかし，これを SNS の環境にあてはめて考えると，“個人識別情報”と“不可侵私的領域”の統一的な解釈を行うのが難しく，各ユーザの裁量に任されることになり，曖昧性が残る．

2.2. プライバシーコミュニティの導入

前節での検討のように，SNS に投稿された記事に対するプライバシー侵害性の解釈は曖昧であり，同一の記事に対しても複数のユーザから異なった解釈が行われるといった，一見すると規則性のない無秩序の世界のように考えられる．これに対して，K.Strater らの研究によると，SNS ユーザ間のプライバシー解釈は，個々のユーザによって全く異なるものではなく，ある一定の特性を共通に持ったユーザ集合のコミュニティを形成するといった仮説が提唱されている [7]．本稿では，このようなコミュニティを“プライバシーコミュニティ”と命名するとともに，このようなプライバシーコミュニティに分類される投稿記事は，プライバシーを侵害していると定義する．そしてこのプライバシーコミュニティは普遍的なものではなく，SNS を構成するユーザのメンバー特性や SNS 活動が行われる時空間特性など，環境により変化するものであるとして考え，これらのシステム化には，機械学習を用いて動的な把握ができることが必要となる．

2.3. 関連研究

SNS への画像投稿時に発生するプライバシー侵害から保護を行う方式に関する研究として，Anna Squicciarini は，K.Strater らのコミュニティによる仮説に基づき，新たに SNS へ投稿される記事に対してプライバシー保護ポリシーを動的に生成し，そのポリシーを用いてプライバシー保護を行う方式を提案している [8]．しかし，この方式におけるプライバシーポリシー生成の基本は，新たに投稿される SNS 記事の分析結果からの学習方式により動的にプライバシー保護ポリシーを予測することであり，学習初期段階での予測精度は，必ずしも保証されない問題があった．又，予測されたプライバシーポリシーをユーザが実施できる状況に必ずしも無く，実施されないことが多いことが問題であった．

そこで我々は，過去の長期間に蓄積された SNS ツイートのアーカイブを対象にしたプライバシー関連記事の抽出と分析を行い，その期間内でのプライバシー解釈のコミュニティを見つけることにより，学習の初期段階からプライバシー保護ポリシーの予測精度を向上させる方式を考える．

3. SNS 投稿記事のプライバシー侵害分析

3.1. SNS におけるプライバシー侵害のモデル化

SNS への記事の投稿によってプライバシー侵害が発生する状況をより詳細に検討するため、SNS ユーザと投稿記事との間の関係をモデル化する。まず、SNS ユーザには、記事を投稿するユーザとその記事を読覧するユーザ、及び、投稿された記事の中で注目すべき記事を指摘するユーザとの3種が考えられ、それぞれ”投稿者 T: Twitter”, “閲覧者 S: Subscriber,” 指摘者 N: Notifier”とする。そして、これらのユーザをノードとして、ノード間のリンクの位置づけで投稿記事がやり取りされるようなノード・リンクモデルを適用する。このモデルは、SNS として Twitter に当てはめて考えると、投稿者から閲覧者へのリンクはツイート、指摘者から投稿者又は閲覧者にはリプライ、閲覧者から投稿者又は指摘者にはリツイートがそれぞれ該当する。

このような SNS での記事操作に関する現実的な役割モデル（これを RWM: Real World Model と略称する）の他に、プライバシー侵害の観点からの概念的な役割モデル（これを CWM: Concept World Model と略称する）が考えられる。この CWM では、記事の投稿を行うことによりプライバシー侵害を行ったユーザを”侵害者 I: Infringer”, その投稿記事によりプライバシー侵害の被害を受けたユーザを”被害者 V: Victim”として定義する。この RWM と CWM との2つのモデルに分けたのは、投稿者が必ずしもいつもプライバシー侵害者、又は閲覧者が被害者にはならない場合があることを表現

することにあり、より正確な現象把握を行うのに有効であると考えている。更にこのプライバシー侵害からの保護サービスを行う対象は、単に被害者だけでなく、侵害者にも及ぶような Opt-in 方式の及ぶ範囲を目標にしており、それらを検討する根拠としても有効であると考えられる。

3.2. モデルに基づくプライバシー侵害の類型化

次に、SNS への画像投稿時に発生するプライバシー侵害の類型化を行い、後段の Opt-in サービスの対象を明確にする。これらの RWM と CWM の相互関係の図式から、プライバシー侵害の侵害者と被害者の関係、及びそれらの当事者としての意識があるかどうかをまとめると、表1のようになる。但し、ここでは、全体の24通りのパターンが考えられるが、その中で実際に発生し得る5パターンに注目している。この表において、白色丸記号のノード間を実線で結んでいるのが、記事の投稿の方向を中心に表現した RWM であり、灰色丸記号と点線で表現したのが CWM である。特に実線の中間の小型の長方形は、投稿記事のシンボルで、その記事をトリガーにしてプライバシー侵害が検知され、プライバシーの侵害者と被害者の関係が同定される。但し、点線の根元と先の黒丸は、プライバシー侵害に気がついている状態、白丸はそれに気がついていない状態を示す。これより、Opt-in 方式によるプライバシー侵害の予測サービスの対象は、白丸のついた#1, #3, #4, #5であることが明確になった。後段のプライバシー侵害の分析処理の結果の中で、この類型化を行ったモデルの項目が実際の SNS 投稿から抽出された記事の中でど

表1: SNS 画像投稿時におけるプライバシー侵害の侵害者と被害者の関係

	類 型	類型内容	侵害パターン	侵害気付	被害気付
#1	他者侵害型	閲覧者のプライバシー侵害となることが分かっている記事を投稿 投稿者は侵害者に、閲覧者は被害者に		有	無
#2	自己漏洩型	自分自身のプライバシー侵害となることが分かっている記事を投稿 投稿者は侵害者と被害者の両面を持つ		有	有
#3	侵害反論型	投稿された記事が閲覧者にとってプライバシー侵害となことに気が付き、投稿者への反論を行う		無	有
#4	侵害指摘型	投稿された記事にプライバシー侵害があることに気が付き、それをリツイートで投稿者と閲覧者に指摘		無	無
#5	間接侵害型	投稿された記事にプライバシー侵害があることに気が付き、それをリツイートするも、その指摘者がプライバシー侵害となる		無	無

のように分布しているかを調査するが、プライバシー侵害の予測サービスの対象は、その中でも頻度の多いパターンを優先的に行う必要がある。

3.3. プライバシー侵害記事抽出のための分析処理の流れ

長期間に蓄積された SNS 投稿記事のアーカイブを対象にして、プライバシー侵害記事の抽出を行う処理の流れについて概説する。ここで抽出されるプライバシー侵害記事は、後段のプライバシーコミュニティ抽出のための入力データになるので、その抽出精度をできるだけ高めておく必要がある。この抽出処理全体の処理フローを、図 1 に示す。このフローにおいて、左側の分析の対象とするのは、投稿記事の {画像, テキスト, 属性} 成分であり、それぞれ並列に分析を行い、特徴ベクトルを生成するものとする。即ち、投稿記事は次のような 3 成分で構成されるものと仮定する。

SNS 画像記事 := 画像データ部
+ (テキストデータ部)
+ (属性データ部)
但し () 部は省略可能

これら 3 成分を用いた分析処理の概要を以下に示す。

1) PS ワードの抽出: PS (Privacy Sensitive) ワードは、テキストデータ部に含まれるプライバシー侵害に抵触するような表現を持つワードであるが、多量の SNS 投稿記事を対象にした統計的な分析により自動的な抽出方法を考える。形態素解析と頻度解析から、「プライバシー」が含まれる記事の適合率を求め、その値でソートをかけてリストの上位 20 ワードを PS ワードに選出する[9]。

2) 画像分析と顔認識: 画像分析には、顔領域の認識の他に、シーン識別や年齢判定及び個人認証が可能な商用の画像解析ソフトパッケージ[10]を用いている。これらの機能中でも、顔領域の認識精度が高いので、投稿された画像の中に顔画像が含まれているかの判定に用いるほか、後述の予測サービスで説明するように、予め登録されたユーザの顔画像と照合して個人認証を

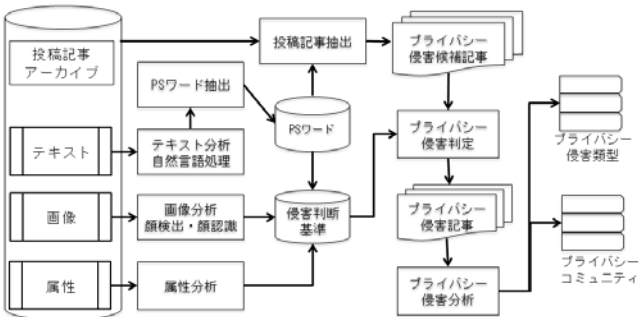


図 1: プライバシー侵害記事の抽出処理フロー

行うことにも用いる。これら認識や識別された項目の特徴ベクトルの対応位置にフラグを設定することにより、次段のプライバシー侵害判定のための知識として用いる。

3) プライバシー侵害分析: 新規に投稿された記事がプライバシー侵害となるかどうかの判定を行うために、侵害判定基準を設定する必要があるが、これを投稿記事アーカイブ内の多量の記事を用いた学習型クラスター方式で獲得する。この学習型クラスターには、最大事後確率に基づく分類関数 `classify()` で構成されるナイーブベイズ分類器を用いる。

$$\text{classify}(f_1, \dots, f_n) = \operatorname{argmax}_c p(C = c) \prod_{i=1}^n p(F_i = f_i | C = c) \dots \text{式 (1)}$$

この分類関数は、(1)式に示すように、特徴ベクトル $(f_1, \dots, f_n)$ のクラス C に依存した事後確率の積で与えられる。又特徴ベクトル f_i は、PS-ワードのフラグと、画像部の顔認識から得られるフラグ、及び属性部の時空間解析から得られるフラグで構成される。そこで、新たに投稿された記事から得られる多次元特徴ベクトルとツイートアーカイブの学習から獲得された侵害判定基準とを比較することによりプライバシー侵害判定を行う。

3.4. 大規模な Twitter アーカイブを用いた分析処理の結果

3.2 で類型化した 5 つのプライバシー侵害パターンが実際の投稿記事でどのように分布しているかを、過去の SNS への投稿記事の解析から調査を行った。ここで、処理対象となるデータは過去 4 年間に蓄積された Twitter 記事のアーカイブ (12 億レコード) を用いた。前章で説明したプライバシー侵害記事の分析処理を適用して得られた投稿記事の中から、プライバシー侵害定義に抵触していると考えられる記事を選択した 830 件のデータである。これらのプライバシー侵害投稿記事に対して 5 つの投稿記事パターンに分類した結果を表 2 に示す。

表 2. プライバシー侵害投稿パターン

	類 型	事象数と割合
P1	他者侵害型	415 (50%)
P2	自己漏洩型	58 (7%)
P3	侵害反論型	36 (4%)
P4	侵害指摘型	210 (25%)
P5	間接侵害型	111 (14%)

これより、最も多い投稿記事は P1 の他者侵害型、即ち閲覧者のプライバシーを侵害して被害者とするような投稿記事である。P1 と P2 のパターンの比較から、本データでは自分自身の情報漏洩に比べて他者の情報を漏洩する記事が多いことが分かる。これは画像投稿

時においては自分自身を撮影するよりも他者を撮影することが多いと考えられる。更に、この中で最も割合の多い P1 の他者侵害型に分類される記事の内容を調べると、投稿者のほぼ全てが被害者のプライバシーを侵害であることは理解しているが、危険性を理解していないものであった。これは本データの抽出が PS ワードに基づいたアルゴリズムであることが原因であるとも考えられるが、前研究 [9] において顔認識による画像処理だけで抽出されたプライバシー侵害記事は非常に少なく、実際にプライバシー侵害に該当した投稿記事は PS ワードを含むものが多かった。そのため、投稿者が被害者のプライバシーを侵害していることを理解しつつも、危険性を理解せずに投稿してしまうと推定される。後段のサービスシステムでは、このようなプライバシー侵害を意識しているユーザに対してはサービス効果が少ないと判断してサービス対象から外し、むしろプライバシー侵害が起こることに気がつかないユーザへの気付きサービスの位置づけで行う方針とした。

4. プライバシーコミュニティの抽出

4.1. プライバシーコミュニティ類型抽出

プライバシーコミュニティの類型の抽出は、既に前研究で行った PS ワードの頻度に基づくデンドログラムによる方式を用いており [11]、Twitter の投稿記事アーカイブからリツイート記事を抽出し、プライバシーコミュニティの特性分析を行った。対象となる記事は、リツイートユーザが書き込みを行ったリツイート部を対象として、PS ワードを含む記事からサンプリングしたものに、ツイートの TF・IDF を用いたコサイン類似度に基づくデンドログラムによる分類を行い、プライバシーコミュニティの類型を抽出した。その抽出結果の一部を図 2 に示す。分類された投稿記事を調べると、比較的明快に類型化が行われているものは以下の 5 つであり、以降、プライバシーコミュニティの抽出には、この代表的な 5 つの類型を用いることにする。

- 1) 顔写真を晒す行為
- 2) 公開範囲を意識せず情報を晒す行為
- 3) マスメディアへユーザ情報を公開
- 4) 業務とプライベートを使い分け
- 5) 公共交通の車内における無断撮影

4.2. 特徴ベクトルの項目

SNS へ新たに投稿される画像記事のプライバシー侵害の有無の判定に加えて、プライバシー侵害をプライバシーコミュニティ類型毎に分類するためには、プライバシーコミュニティの類型に分類するための各種のシーン情報が必要となる。そのシーン情報は、{時間、

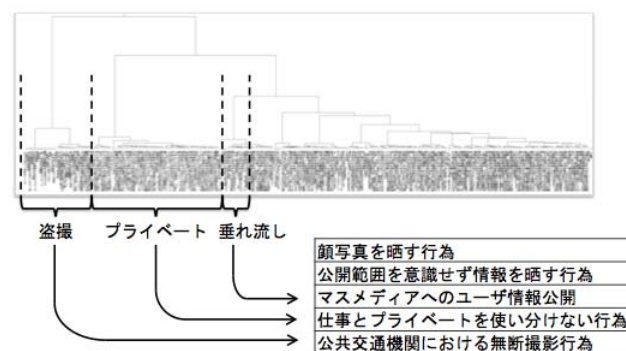


図 2：デンドログラムによる類型抽出の結果

場所、人、行動内容、公開範囲、…} 等が考えられ、投稿記事の各データ部に含まれる項目を総合的にまとめると表 3 のようになる。このうち、時刻・場所・行動内容については各データ部から得られる情報を挙列している。なお、投稿記事に含まれる情報に表 2 の類型の項目が該当する場合には、これらの項目を用いてプライバシーコミュニティの分類を行う。そして、これらの項目のリストは、多次元の特徴ベクトル形式 $FV[b1, b2, \dots, b19]$ で構成し、ナイーブベイズの学習過程により各コミュニティのクラスターパラメータを得る。

表 3:FV;特徴ベクトルの項目

項目	シーン情報		
b1	風景	場所	画像 データ部
b2	顔認識	人	
b3	人数		
b4	顔認証		
b5	性別		
b6	年齢		
b7	食事風景	行動内容	Exif 情報
b8	撮影時間	時間	
b9	撮影場所	場所	
b10	時刻ワード	時間	テキスト データ部
b11	地名ワード	場所	
b12	固有名詞	人	
b13	アカウント名		
b14	行動ワード	行動内容	属性 データ部
b15	投稿時間	時間	
b16	位置情報	場所	
b17	役職ワード	人	
b18	プロフィール文	行動内容	
b19	閲覧者数	公開範囲	

5. プライバシー侵害の予測サービス

5.1. プライバシー保護対象の事前定義による

Opt-in サービスの提案

プライバシー侵害となる記事の漏洩を防ぐためには被害者となりうるユーザが公開されたくない情報を前もって定義し、その定義に該当した投稿記事を投稿者に事前に通知することが有効である。ただし、投稿記事に被害者情報が含まれていても被害者がプライバシー侵害を事前に定義していない場合、その被害者がプライバシー侵害についてどのように考えているか不明である。そこで、投稿記事から被害者の情報が含まれているかを判定し、プライバシー侵害コミュニティに該当した場合には、投稿者に対して投稿が問題となることを提示する Opt-in サービスを提案する。

このサービスでは、投稿記事をシーン項目表の各項目で得られる情報とプライバシー知識に基づいてナレッジベースによるプライバシー侵害判定とコミュニティ分類を行う。プライバシー知識は前節の投稿記事を我々が5つのコミュニティに分類した結果を使用する。プライバシー侵害記事であると該当した場合、適切な対応方法を取るよう通知を行う。通知内容には投稿記事に含まれる被害者の顔画像を含まないように取り直しもしくはフィルターを掛け、該当したコミュニティで検出されているシーン項目を修正するように促す。

これにより、被害者はプライバシー侵害となる状況を投稿者に SNS 投稿する前に通知することができる。さらに、投稿者が修正を無視した場合、既に流出した情報は抑えられないが、無視したことが被害者へ通知が届くので速やかに削除依頼等の対策を取ることができる。一方、投稿者は通知を受けることにより、対応方法の支援から、どの情報が問題であることを認識することができ、投稿操作を一旦休止するので、プライバシー侵害のトラブルを避けやすくなる。また、サービスによる対応方法を適切に取った場合、その行為がサービスログに記録されるため、投稿記事のプライバシー侵害に対する責任が投稿者からサービスに変わり、SNS 投稿への責任が軽くなる効果も期待できる。

5.2. サービスシステムのサーバ部実装検討

サービスシステムのサーバ部の実装ブロック図を図3に示す。そのデータ処理の流れを説明すると、まず被害者は、自身の顔画像と公開されたくないシーンをサービス上に登録する。一方、投稿者は SNS に投稿しようとしている投稿記事をサービスに送信する。サービス内では投稿記事をシーン情報に応じて解析を行い、コミュニティ分類を行う。得られた結果から、被害者のプライバシー侵害に該当する場合はコミュニティに応じた対応方法の処置を行わなければ被害者とト

ラブルになる可能性があることを投稿者に通知する。通知後、対応方法に対して投稿者が受理/却下/破棄をサービスに送信する。

5.3. クライアント部

本サービスのクライアントは Android アプリケーションとして開発を行い、投稿者と被害者は Android 端末から SNS を利用しているものとする。クライアントでは被害者が顔画像とシーンの登録を行い、投稿者が Opt-in 方式による投稿記事のプライバシー侵害予測通知を行える GUI (Graphical User Interface) を提供する。

1) サービス登録者クライアント：被害者は顔画像の登録と公開されたくないシーンの登録を行い、該当する投稿記事がサービスに送られてきた際に、その投稿記事情報を被害者に通知する。図4に顔・シーン登録と投稿記事の通知が行われた場合の端末画面を示す。被害者が登録した顔画像とシーンに該当した投稿記事をサービスが検知すると、被害者へ通知が行われる。

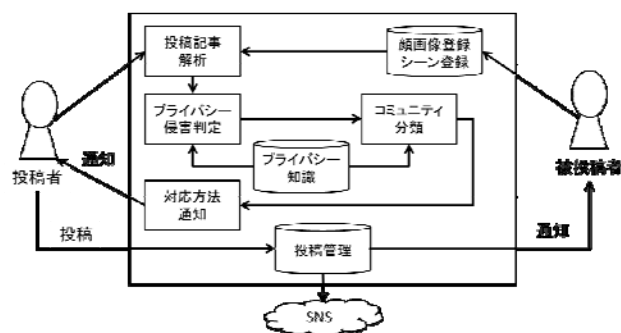


図3:サーバブロック図



図4:顔・シーン登録画面と通知画面

2) 投稿者クライアント：投稿者は Android アプリケーションによって作成した投稿記事をサービスに投稿することで、投稿記事内容に被害者のプライバシー侵害情報が含まれているかを調べ、通知する。通知結果に対して、投稿者は修正を行うかどうかを判断して投稿を行う。図5にサービス投稿画面と分析結果画面の端末画面を示す。投稿者が投稿記事データをサービ

スに投稿することで、該当する被害者の情報修正通知が送られる。



図 5：サービス投稿画面と分析結果画面

5.4. プライバシー侵害予測の実装

新規投稿記事からのプライバシー侵害を予測するための方式として、3.3 節で示したナイーブベイズ分類器を基にした学習型クラスター方式を用いる。この場合、プライバシーの侵害の有無を判定するための 2 値分類を、プライバシーコミュニティの区分を行う多値分類に含めて行うような実装を行った。即ち、その分類目標は、表 5 のように 6 値のクラスター問題として実装した。学習対象とする特徴ベクトルは、表 3 に示したシーン項目を用いるが、テキストデータ部の地名ワードと固有名詞には ENAMDICT[12]を活用し、地名ワードは p;place-name, st;stations ラベル、固有名詞は s;surname, u;person-name, g;given name, f;female given name, m;male given name ラベルの形態素で識別している。

6. 評価

6.1. 評価概要

開発したプライバシー侵害の予知サービスの精度を測定した。具体的には、プライバシー侵害判定の精度を測定することにより、サービス側が投稿記事をプライバシー侵害記事であると正しく分類できているかを評価した。評価データには 3.4 節のプライバシー侵害記事の分析処理結果によってプライバシー侵害定義に抵触していると考えられるプライバシー侵害記事 830 件とプライバシー侵害ではない記事 830 件の記事のうち学習データとして用いなかった 330 件を実験データとして使用した。

6.2. プライバシー侵害判定精度

実験データのうち、分類器が評価記事データを正しくプライバシー侵害記事として判定できているかを評価する。評価方法として、プライバシーコミュニティのクラスター:C1～C5 をまとめてプライバシー侵害が有

ると分類した記事がどれだけ正解しているかを適合率 (Precision)によって、正解記事のうち検出した記事がどれだけ含んでいるかを再現率 (Recall)によってそれぞれ算出した。また、適合率と再現率の調和平均から F 値 (F-measure)を求め、検出精度も評価した。評価記事データによる評価結果を表 4 に示す。

表 4：適合率と再現率、F 値による精度評価

適合率	再現率	F 値
88.7%	43.7%	58.5%

適合率は高い値を得ることができたが、再現率は低い値であることが分かった。このことから、特徴ベクトルを用いたナイーブベイズクラスターによる分類方法では、プライバシー侵害記事を正しく検出し誤検出は発生しにくい、プライバシー侵害記事の検出率が低いと言える。そのため、プライバシー侵害記事として検出できない記事を分類器が検出できるようにシーン項目を検討する必要がある。

6.3. プライバシーコミュニティ分類精度

実験データのうち、プライバシーコミュニティ分類によって判定した結果を表 5 に示す。

表 5. プライバシーコミュニティ分類精度

	クラスター	検知率
C1	侵害有；顔写真を晒す行為	50%
C2	侵害有；公開範囲を意識せず情報を晒す行為	35%
C3	侵害有；マスメディアへユーザ情報を公開	63%
C4	侵害有；業務とプライベートを使い分け	50%
C5	侵害有；公共交通の車内における無断撮影	4%
C6	プライバシー侵害無し	98%

C1～C5 において検知率は 4%～63%となり、平均は約 40.4%となり十分な精度とは言えない。正しいコミュニティが出力できない原因は学習データが少ないことだけでなく、あるコミュニティの各学習データにおけるシーン項目表の特徴ベクトルにばらつきがあり、標準偏差が大きくなっている問題が考えられる。つまり、学習データを増やした場合とコミュニティを細分化した場合の評価を比較し、どちらが影響を及ぼしているか分析する必要がある。

7. まとめ

SNS の画像投稿記事のプライバシー侵害を事前に抑制するため、SNS ユーザのプライバシーに対する解釈の違いに着目し、プライバシー解釈の共通項をプライバシーコミュニティと定義し、過去の SNS 投稿記事アーカイブを対象にした分析により投稿記事をコミュニ

ティに分類してプライバシー侵害を予測する方式を開発した．さらに，そのシーン情報に基づいてプライバシー侵害予測通知システムのプロトタイプを構築し，コミュニティ分類の精度を測定した．

今後は構築したシステムの実用性を評価するため，多量の被験者によるサービスの利用評価や分析を行う予定である．

謝 辞

本研究の自然言語解析にご協力頂いた東京大学知の構造化センター荒牧英治講師と馬場一貴氏に感謝申し上げますとともに，画像解析における顔認識ソフトウェアパッケージ OKAOVision を提供頂いたオムロン株式会社に謝意を表す．

参 考 文 献

- [1] 「Facebook の全ユーザは 9 億人突破，モバイルは 5 億人—上場申請書改訂版を提出」，<http://jp.techcrunch.com/archives/20120423facebook-s-amended-s-1-500-million-mobile-users-paid-300m-cash-23-million-shares-for-instagram/>, 2012/11/14
- [2] 「Twitter, 1 日の投稿数が 4 億を達成- 60%がモバイルを利用」，マイナビニュース，<http://news.mynavi.jp/news/2012/06/08/022/index.html>, 2013/03
- [3] 「SNS と携帯電話の利用におけるプライバシー意識調査」，トレンドマイクロ: <http://jp.trendmicro.com/jp/about/news/pr/article/20120410035738.html>, 2013/03
- [4] 『宴のあと』事件，東京地判昭和 39 年 9 月 28 日判時 385 号 12 頁
- [5] 『石に泳ぐ魚』事件，東京地判平成 11 年 6 月 22 日判時 1691 号 91 頁
- [6] 青柳武彦，日経新聞 個人情報保護に行き過ぎ 公益との調和必要 プライバシー論議深めよ，2005.09.19
- [7] K. Strater and H. Lipford. Strategies and struggles with privacy in an online social networking community. In British Computer Society Conference on Human-Computer Interaction, 2008.
- [8] Anna Squicciarini, Smitha Sundareswaran, Dan Lin, A3P: adaptive policy prediction for shared images over popular content sharing sites, ACM New York, NY,
- [9] 高田さとみ，小山貴之，町田史門，宋洋，嶋田茂，"SNS 画像投稿時に発生するプライバシー侵害の要因分析"，電子情報通信学会 EMM 研究会技術報告，2012
- [10] OMRON: <http://www.omron.co.jp/ecb/products/mobile/>, 2013/03.
- [11] 宋洋，小山貴之，町田史門，嶋田茂，"プライバシーコミュニティ検知に向けた SNS 投稿記事の時系列分析"，電子情報通信学会 DE 研究会技術報告，2012
- [12] ENAMDICT: http://www.csse.monash.edu.au/~jwb/enamdict_doc.html, 2013/03