2ZF-06

Twitterの投稿画像を対象とした 個人情報流出防止のためのシステム開発に関する研究

松原心慈† 櫻井淳†

文教大学情報学部†

１．はじめに

近年，Twitter を中心とした SNS を通じたトラブルが多発しており，20 代以下の被害件数が年々増加傾向にあることが問題視されている．その種類は，詐欺被害や売春被害など多岐にわたる．また，インターネット上における不適切な投稿などによる炎上を起因とし，SNS アカウントの投稿内容から個人情報が特定され，拡散される被害も報告されている[1]．このような個人情報流出の被害は，利用者自身だけでなく，家族や所属組織など周囲にも深刻な被害を及ぼす可能性があるため，SNS 利用者にとって個人情報の適切な管理は重要である．

SNS を通じた個人情報の特定に関する既存研究として，投稿された文章から利用者の地域を推定する研究[2]や，他者の投稿に対するリアクションなどの行動データをもとに個人情報の特定を試みる研究[3]などが存在するが，投稿画像に着目した研究は調査する限り見当たらない．しかし，投稿画像においても，屋外の看板や標識などに地名や組織名が映されている場合があり，本人の意図しない形で個人情報を発信するリスクがあるため，その実態を把握することは重要と考えられる．

そこで，本研究では，個人情報流出の防止を目的とし，SNS を通じたトラブル件数が最も多いとされる中学生と高校生を対象に，Twitter アカウントの投稿画像から地名を含む画像を抽出するシステムの開発と検証を行う．これにより，学校教育におけるネットリテラシー講習などの場面において，学生が自身のアカウントの危険性を把握するためのシステム構築を目指す．

２．研究の概要

本システムの概要を図 1 に示す．本システムは，A）画像クローリング機能，B）画像内の文字認識機能，C）地名判定機能から構成される．また，本システムは，図 2 に示すように，Heroku と LineAPI を活用し，スマホで手軽に利用できるチャットボットにて開発する．Line 上のテキストボックスから TwitterID を送信（図 2 左）し，約 5 分の処理後に判定結果（図 2 右）が表示される．

まず， A）画像クローリング機能では，TwitterID を入力データとし，Tweepy とよばれる Python ライブラリを用いて，API の上限である上位 3,200 ツイート中に含まれる投稿画像をすべて保存する．次に，B）画像内の文字認識機能では，GoogleCloudVisionAPI の深層学習を用いた文字認識（OCR）機能を用いて，取得した画像内に含まれるテキストを抽出する．そして，C）地名判定機能では，そのテキストに対して，固有表現の単語が多く収録された mecab-ipadic-NEologd を用いた形態素解析を行い，名詞－固有名詞－地域－一般に該当する名詞を抽出する．この単語を地名と判定し，画像とセットで出力する．

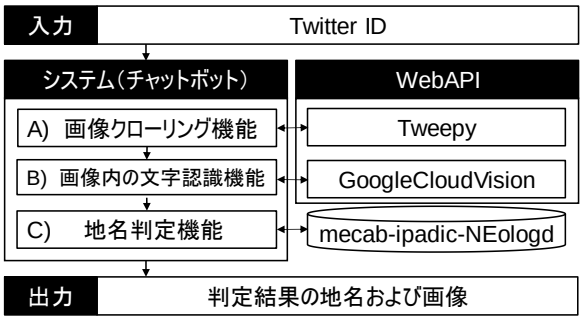

図 1 本システムの概要

図 2 画面例（左：入力画面，右：判定画面）

Development of System to Prevent Leakage of Personal Information for Images Posted on Twitter

† Shinji Matsubara,Jun Sakurai
Faculty of Information and Communications,Bunkyo University, 1100 Namegaya, Chigasaki City,Kanagawa 253-8550, Japan.

３．実証実験

実証実験では，本システムが投稿画像から地名を特定できることの可能性を検証する．

入力データは，Twitterプロフィール文や投稿内容から中学生または高校生と思われるアカウントを人手の作業により抽出した81件（以下，実験用アカウント）とする．

３．１　実験内容

本実験では，実験用アカウントの投稿画像を対象に，目視でテキストが確認できる 2,281 枚（以下，実験用画像）から地名が含まれる画像の手動による判別結果と，本システムの出力結果をもとに精度評価を行う．なお，精度評価にはF値を用いる．

３．２　実験結果

目視による実験用画像の結果と本システムによる結果から作成した混同行列を表1に，そこから求めた適合率，再現率，F値の結果を表2に示す．実験結果から，再現率が 0.82 と高い結果となっており，目視で地名が確認できる画像の多くは本システムで判別が可能であるといえる．正解結果の画像を詳細に確認すると，図3に示すように，天気予報や災害情報などのスクリーンショット（図3左）が半数程度を占めていた．また，少数ではあるが，運動会時の撮影写真の背景に映った学校用テントや，卒業写真の証書（図3右）や学校看板など，個人情報の特定につながる画像も検出された．しかし，図4のように，画像内の撮影角度やテキストのフォントの影響により，B）画像内の文字認識機能において地名を正確に抽出できていないケースが確認できた．

一方，適合率は 0.72 と再現率と比べ低い結果となった．これは，人名を地名と誤判定したものや，画像上に表示される文章の改行が影響し，単語の分かち書きの処理の際に「成」や「向」などの一文字に誤って分割され，地名と判定されたものが主な原因であった．

上記の実験結果より，学生への啓発活動に有益な情報を抽出できたため，システムの一定の有用性は示せた．しかし，一文字の単語を地名と多く誤判定するなど，その検出精度には課題が残る結果となった．これに対しては，地名判定機能の辞書データを拡張し，抽出単語を限定する方法が考えられる．具体的に，形態素解析に用いる辞書データを本システムの利用地域に合わせ拡張し，抽出単語を利用地域の周辺やその略称，組織名や施設名などに限定することで，B）画像内の文字認識機能で誤判定があった際，単漢字や人名などによる地名の誤判定を防ぐことが期待できる．加えて，辞書データの拡張によって，地名以外の個人情報流出のリスクを含む画像の抽出が可能な利点がある．

表1　混同行列

		システムによる判定	
		地名あり	地名なし
目視による確認	地名あり	TP: 277 枚	FN: 60 枚
	地名なし	FP: 107 枚	TN: 1,837 枚

表2　実験結果

適合率	再現率	F 値
0.72	0.82	0.76

図3　正解結果の画像例
（左：スクリーンショット，右：卒業証書）

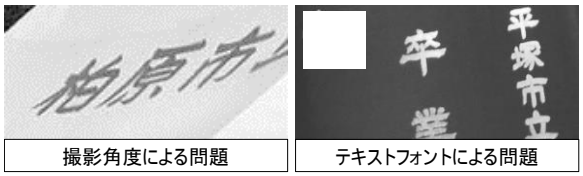

図4　テキストの抽出に失敗した画像の例

４．おわりに

本研究では，Twitter の投稿画像に着目し，地名を含む画像を出力するシステムを開発した．そして，実証実験により，本システムが投稿画像から地名を特定できることの可能性を示した．

今後は，システムの精度向上に加え，組織名や施設名などの抽出可能な名詞を増やすことで，投稿画像中に含まれる多くの個人情報流出リスクを検出し，それらを可視化することで，本システムの有用性を証明したいと考えている．

参考文献

[1] 山下晃弘，上村卓史，川村秀憲，鈴木恵二：SNS プライバシー保護とリスク管理の検討ーソーシャルモニタリングツールの開発に向けてー，デジタルプラクティス，情報処理学会，Vol.6，No.2，pp.150-158，2015.

[2] 奥村賢俊，彌冨仁：Twitter における個人情報推定のための基礎検討，ファジィシステムシンポジウム講演論文集，日本知能情報ファジィ学会，Vol.32，pp.701-704，2016.

[3] 畑田裕二，矢谷浩司：ソーシャルメディア上の行動データから流出する個人情報の定量的分析，第 80 回全国大会講演論文集，情報処理学会，Vol.80，pp.195-196，2018.