様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　３０　年　　６　月　２７　日現在

機関番号：　５２６０１
研究種目：　若手研究(B)
研究期間：　2015～2017
課題番号：　１５Ｋ１６０９２

研究課題名（和文）SNS利用におけるプライバシー保護を目的とした個人プロファイル分析手法の開発

研究課題名（英文）Development of Personal Profile Analysis System for Privacy Protection in SNS Usage

研究代表者

山下　晃弘（Yamashita, Akihiro）

東京工業高等専門学校・情報工学科・准教授

研究者番号：８０５８９８３８

交付決定額（研究期間全体）：（直接経費）　　2,700,000円

研究成果の概要（和文）：本研究の目的は，日常的なSNS利用において，第三者が意図しないプライバシー情報を取得可能な状況になってしまうことを防止するシステムを開発することである．システムが自動的にSNS上の情報に基づいてプロファイル分析を行い，健全なSNS利用をサポートするシステムの実現が目標である．具体的には，投稿した内容が拡散されやすい情報かどうかを機械学習で推定するアルゴリズム，投稿内容の感情分類を推定するアルゴリズム，プライバシー情報の例として，投稿内容からユーザの職業を推定するアルゴリズムについて開発を行った．また，実システムへの応用を見据えて大規模化に向けたクラウド化や，実システムへの実装を行った．

研究成果の概要（英文）：The purpose of this research is to develop a watching system to prevent the leakage of privacy information from posted contents when using SNS. The goal is to implement a system that automatically analyzes user profiles based on information on SNS and supports healthy SNS usage. Specifically, in this study, we developed: algorithms for estimating by machine learning whether the contents posted are attracting attention or not; algorithm for estimating emotional classification of posted contents; algorithm for estimating user's occupation from posted contents. In addition, for practical use, we have used cloud environments for larger scale and experimentally implemented on real systems.

研究分野：　知能システム，機械学習，組み込みシステム

キーワード：　プライバシー保護　SNS　機械学習　自然言語処理

１．研究開始当初の背景

スマートフォンの普及を背景にTwitterやFacebookなどのSNSを利用することは，多くの人にとって生活の一部として定着している．総務省情報通信白書等によれば，2014年における日本のSNS人口普及率は40%を超えており，その後もSNS上の情報量は指数関数的に増加している．企業活動においても急速にSNSの活用が浸透しており，広告や評判分析だけではなく，就活中の学生との交流などより個人的な活用も増加している．

その一方で，SNSが抱える様々な課題も顕在化している．誹謗中傷を伴う「炎上」はその典型例であり，SNSを含むインターネット上の情報から悪意を持って個人情報の抽出を試みるネットストーカーの問題や，特に中高生の間でのネットいじめの問題など，深刻な社会問題となっている．一言に炎上と言っても様々な原因が考えられるが，たとえ犯罪告白や悪ふざけなど，社会的に好ましくない事実がネット上に書き込まれていたとしても，それを集団が必要以上に取り上げてプライバシーをインターネット上にまとめて晒し，永久に公開し続ける行為は許されるものではない．しかし，誰もが自由に情報を公開可能なインターネット上においては，炎上を本質的に解決することは難しく，SNS等を利用する者が自己責任で投稿する内容を管理する必要がある．不用意にSNSへ個人情報を投稿してしまうと，それが原因で，本人が公開を望まない情報（本研究ではこれをプライバシー情報と定義する）が第三者に取得され，何らかのきっかけで炎上した場合にその情報がまとめあげられ半永久に晒され続ける事態に発展してしまう．実際には炎上に至らないまでも，プライバシー情報がWeb上から取得可能な状態であることに本人が気付かない状況は危険である．

本研究の目的は，SNS上の公開情報から第三者が推測可能なプライバシー情報を含む特徴的な情報を機械的に抽出し，利用者自身が公開情報を管理し，炎上やネットストーカーの被害を予防するシステムを開発することである．図１は典型的な炎上の経過フェーズと可能な対応策について過去の事例を元にまとめたものである．不用意な発言の防止や，早期の拡散検知，炎上の兆候が見られた場合にいち早く対策を講じることは重要であるが，最も重要な点は，SNS上に第三者に推測されたくないプライバシーを公開しないよう日常的に注意することである．本研究の着眼点は普段のSNS利用において第三者にプライバシー情報を与えないための予防的発想による対策を実現するために，それをサポートするシステムを実現することである．

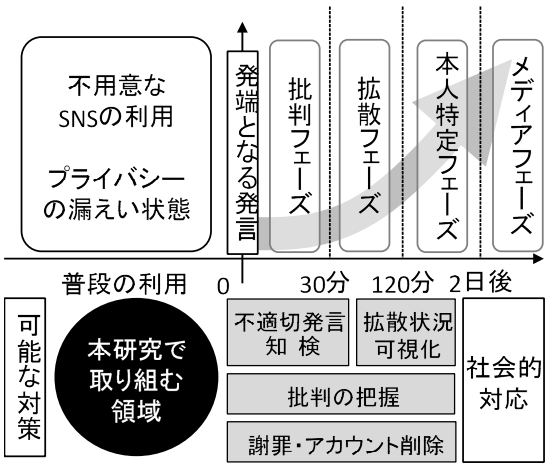

図 1 炎上の経過フェーズと可能な対策

２．研究の目的

本研究では，SNS利用におけるプライバシー情報流出を防止するシステムを実現するために，次の3点を当初の研究目的とした．

(A) SNSデータからの個人プロファイルの特徴抽出手法の開発
(B) ビッグデータに対応したクラウド型分析システムの実用化
(C) SNS見守りアプリケーションへの統合による実運用及び評価

まず，研究目的(A)を達成するために，SNS上で公開された情報源から，プライバシー情報を含む特徴的な情報を抽出する技術を開発した．また，人的なプライバシーの拡散を防止するためにも，投稿された情報が多くの注目を集めやすいかどうかを事前に推定する技術についても検討を行った．具体的に，(1)ニューラルネットワークを用いた画像付きツイートからのリツイート数の予測，(2)ランダムフォレストを用いたツイートの感情分類，(3)LIWCを用いたユーザの属性推定，の3点について研究テーマとして取り組んだ．また，研究目的(B)(C)については，本研究の開始段階で既に開発を進めていたプロトタイプシステムに新たな機能として追加することで実現する方法を検討した．以下に具体的な研究の方法と，得られた研究成果について述べる．

３．研究の方法

(1) Twitterにおけるリツイート数の予測

事前の調査によりSNSへの投稿は文章のみの投稿よりも，画像を付加した投稿のほうが周囲から注目を集めやすい傾向があることが分かった．そこで，画像と文章の両方からそのツイートがどの程度リツイートを集めやすいかを推定し，事前に注目のされやすさを評価する方法について検討を行った．具体的には，画像と文章の両方を扱うため，ニューラルネットワークを用いた機械学習によ

る手法の検討を行った．

文章については，文章を多次元ベクトルに変換する手法である Doc2Vec モデルを使用し，1,000 次元の特徴ベクトルに変換した．Doc2Vec モデルのトレーニングには Twitter からランダムに集めた 1,544,127 件のツイートを用いた．学習の際には形態素解析を行い，ベクトルが各品詞の活用形に影響されないように品詞を原形に変換した．

画像については，RGB の 3×256×256 の配列に変換した．リサイズのアルゴリズムには Bilinear 法を使用した．また，収集した全画像の画素値の平均値である平均画像を作成し，各画像から引くことで正規化を行った．

以上の処理で得られた文章の 1000 次元ベクトルと，画像の 3×256×256 次元ベクトルを入力とし，ニューラルネットワークを構成した．このニューラルネットワークにはユニット間の結合が疎らな「畳み込み層」や位置感度を低下させる「プーリング層」など，画像の特徴量抽出を担う層が含まれている[1]．ニューラルネットワークの構成には Chainer を用いた．また，画像から特徴量を抽出する畳み込み層は AlexNet[2]と GoogleNet[3]の 2 通りのモデルをベースに構成し，精度を比較した．

(2)ツイートの感情分類

SNS 上への投稿は主観を伴った感情的な投稿も散見される．また，何らかの情報が拡散する場合も，その投稿に対する主観的な感想が付与されて拡散されることが多い．マスメディアと異なり SNS が基本的に個人間を繋ぐメディアである特性上，そこに流れる情報は感情を伴った情報が比較的多いことが特徴である．各投稿に内包される感情を機械的に推定することが可能になれば，情報がどのような受け取られ方で拡散しているかの可視化や，例えばネガティブな印象で拡散している場合にアラートを送ることなどが可能になる．そこで本研究では SNS 上の投稿からその投稿の感情を推定する手法について検討を行った．

まず，感情の種類を定義する必要がある．感情表現辞典[4]では，感情を表現する単語を,喜,好,昂,怒,哀,厭,驚,怖,安,の 10 種類に分類している．しかし，細分化された感情推定は直感的ではなく，本研究の目的であるツイートの感情分類には適さないと判断し，今回はより簡略化して喜,怒,哀,楽,と「無感情」の 5 つに分類することにした．喜怒哀楽はそれぞれ感情表現辞書から喜={喜,好,昂}，怒={怒}，哀={哀,厭,驚,怖}，楽={安}として，感情表現辞書の単語を，それぞれの感情を表現する語として割り当てた．また，本研究では，感情表現辞書に記載された単語の類義語を検索し，その類義語も喜怒哀楽を表現する単語(感情語)とした．

一般的に，喜怒哀楽の各分類の感情語が Twitter 上にどの程度出現しているか事前に調査したところ，喜:31%，怒:2.4%，哀:25%，楽:0.6%，無感情:41%という結果になった．

ツイートを特徴ベクトルとして表現する方法には様々な手法があるが，単に文章中の感情語出現頻度や,TF/IDF 法では，否定文など文の構造を考慮した表現に対応できないうえ,新出単語に対応できない問題がある．そこで本研究では，Word2Vec と呼ばれる単語をベクトル化する手法を用いた．Word2Vec で生成されるベクトルは，その距離が近ければ，単語の意味も近いため，予め感情語を登録すれば意味的に似た単語を包括的に推定可能である．また,否定文や疑問文になる場合も特徴ベクトルの変化によって分類が可能である．

生成した特徴ベクトルを感情で分類するための分類器を構築するために，今回は高速で精度が良く,大量データに適したランダムフォレストを用いて検証を行った．

(3) LIWC を用いたユーザの属性推定

SNS 上には年齢や性別，職業などのプライバシーにかかわる情報が数多く存在する．また，直接それらが記載されていなくても，投稿された内容から推定可能な場合もある．Schwartz らは Facebook の投稿に含まれる単語とトピックの使用頻度から辞書を作成し，性格，性別，年齢を判別できることを示している [5]．また IBM の那須川らは Twitter のツイートからユーザの性格を推定できることを示している [6]．那須川らのこの研究では，語彙を抽象化してカテゴリ化するためのツール Linguistic Inquiry and Word Count(以後 LIWC) [7]が使われている．LIWC は，この研究以外にも，選挙運動の分析や著者の年齢推定、性格推定などの様々なタスクの素性として使用されている．LIWC はそれぞれのカテゴリごとの単語を内包した辞書を持ち，カテゴリ内の単語の現れる回数から特徴量を算出する．例えば，“Sad”のカテゴリであれば”cry”や”alone”，“lost”などが定義されている．

しかしながら LIWC の原本は英語であり，複数の言語に翻訳されているものの，公式サイトには日本語に翻訳されたデータは公開されていない．そこで本研究では，ユーザの属性推定に有効な LIWC を日本語に翻訳し，ユーザの属性推定に有効か検証を行った．翻訳の対象は LIWC2015 の全 73 カテゴリのうち，日本語には存在しないカテゴリを除く 66 カテゴリとした．具体的な翻訳手法としては，Word2vec を用いて各単語のベクトル表現を取得し，翻訳結果がカテゴリに属しているかをベクトルとカテゴリの cos 類似度の値から判別する手法を開発し，翻訳を行った．

翻訳した日本語の LIWC を用いて，Twitter 利用者の職業推定を行った．推定手法は，SVM，ランダムフォレスト，ロジスティック回帰のそれぞれの手法で検証した．

(4) 実システムへの適用とクラウド化
研究代表者らは，本研究の申請段階において，過去の炎上事例を詳細に分析し，危険ワードフィルタや情報拡散の可視化による SNS 見守りシステムの開発を進めていた．図 2 は開発した見守りシステムの画面である．

図 2 開発中の SNS 見守りシステム

本研究で得られた成果はこれらのシステム上に適用し，評価を行うこととした．また，より多くのユーザ数に利用して頂くことを想定し，大規模化を前提としたクラウド化を進めることも本研究の目標とした．これらの実システムとしての評価とフィードバックについては，株式会社調和技研などの外部企業と密に連携して取り組んだ．

<引用文献>

[1] 谷岡貴之，深層学習（機械学習プロフェッショナルシリーズ）(2015)
[2] Alex K., et al: ImageNetClassification with Deep CNN
[3] Christian Szegedy, et al. Google Inc: Going deeper with convolutions
[4] 中村 明:感情表現辞書，東京堂出版
[5] H. A. Schwartz, J. C. Eichstaedt , M. L. Kern, “Personality, gen- der, and age in the language of social media,” The open- vocabulary approach, 2013.
[6] 眞．山本 ，哲．那須川，“LIWC2001 手作業翻訳の方針と 半自動翻訳手法の提案，” 言語処理学会 第 22 回年次大会，2016.
[7] “LIWC,”https://liwc.wpengine.com/

４．研究成果
(1) Twitter におけるリツイート数の予測
Twitter を対象に，そのツイートの注目度の指標として RT 数に着目し，投稿の内容から RT 数をを予測するモデルを構築して実験を行った．一般的に画像付きツイートの方が注目を集めやすいことから，文章と画像の両方を入力とするマルチモーダルなニューラルネットワークを構築し，深層学習のアーキテクチャによって予測を行った．図 3 は，本研究で利用したニューラルネットワークのモデルである．

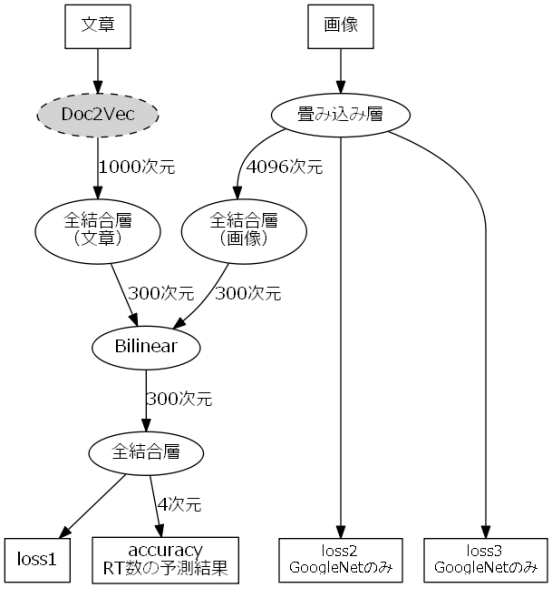

図 3　使用したニューラルネットワーク

本研究では，問題設定として，RT 数を予測する回帰問題と，RT 数をカテゴリ分割し，あるツイートが属するカテゴリを予測する分類問題の両面で実験を行った．その結果，回帰問題では予測値と実際の値の相関係数が 0.39，分類問題では，予測精度が 43.75%となった．

文章の特徴ベクトルのみから RT 数を予測する実験や，画像の特徴ベクトルのみから RT 数を予測する実験も実施したところ，文章と画像の両方の特徴ベクトルから RT を予測したほうがより精度が高い結果になった．つまり，画像と文章の両方のデータを合わせて学習することで，個別に学習させた時よりも，より高い精度で RT 数を予測させることが可能であることが分かった．

(2)ツイートの感情分類
実証実験を行うために実データを収集する必要がある．本研究では，「怒」や，「悲」などの典型的な感情語でツイートを検索し，得られた約 4000 のツイートの感情を改めて目視で確認し，それぞれに喜,怒,哀,楽,無感情のラベル付けを行ったものを使用する．データ数はそれぞれ,喜が 659,怒が 1605,哀が 480,楽が 344,無感情が 700,用意した．テストデータのうち,90%を教師用,10%を評価用として実験を行った．

テストデータと評価データを入れ替えて 10 回実験を行った際の正解率の平均値は 74.6%であった．また,感情語で検索したツイートではなく，Twitter 全体からランダムに抽出したツイートに対しても感情推定を実施した．無作為に抽出した 355 個のツイートに対して提案手法で感情推定した結果について，20 代男性 4 名に目視でその結果が，妥当，不適切，どちらともいえない，の 3 段階

で評価してもらった．その結果，妥当 35.5%，不適切 40.8%，どちらともいえない 23.7%という結果となった．不適切が多くなった原因は，無感情のツイートに感情ありのラベルを付与してしまった誤りが多いためであり，今後は無感情のツイートの検出精度向上が課題である．また，同じ文章でも，前の文を受けて感情が変わる文章は正しく推定できないため，前後のツイートの感情の推移を考慮した推定方法の検討などが必要である．

(3) LIWC を用いたユーザの属性推定

英語版しか存在しない LIWC を日本語に翻訳するために，図 4 に示す翻訳の仕組みを新たに構築し実装した．日本語に翻訳した LIWC を用いて，ユーザのツイート内容からそのユーザの職業推定を行った．ユーザ 1 人の過去のツイートに対して，LIWC 辞書の 66 カテゴリに含まれる単語の出現頻度を計算し，66 次元の特徴ベクトルを構成してそれを入力として機械学習を行った．使用したアルゴリズムは RBF カーネルを用いた Support Vector Machine(SVM)，ランダムフォレスト，ロジスティック回帰の 3 つである．3-fold cross-validation の検証方法で各アルゴリズムのパラメータを最適化した．

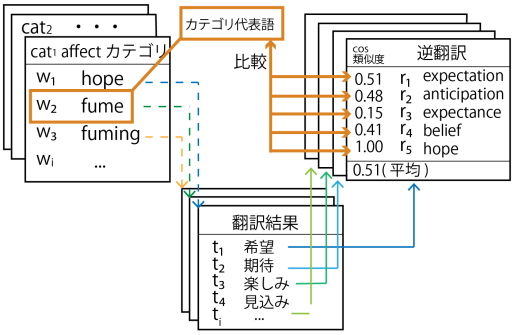

図 4 LIWC を日本語に翻訳する仕組み

8 種類の職業を 8 カテゴリとして各ユーザの多クラス分類を職業別に行なった (図 5)．

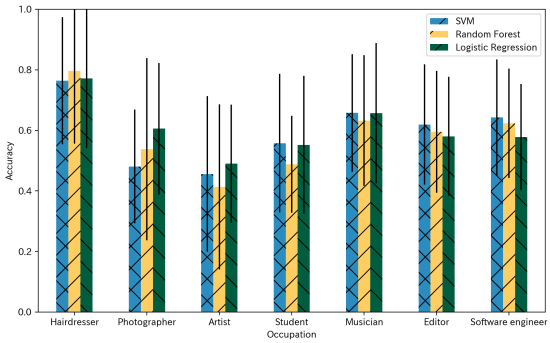

図 5 LIWC を用いて職業分類した際の精度

職業別に分類精度の偏りはあるが，どのアルゴリズムにおいても平均して 62%程度の分類精度となった．また美容師は精度の開きが大きいが，ソフトウェアエンジニアは開きが小さいのはユーザ数の違いによる．一定の分類精度が得られたことから，日本語に翻訳された LIWC の辞書は多少ノイズを残しつつも，属性推定に有効であることが確認できた．

(4) 実システムへの適用とクラウド化

本研究では，ユーザの属性推定アルゴリズムを研究開発するためデータ収集の時点から大規模化を見据え，クラウド環境である Amazon AWS 上にシステムを構築し取り組んできた．また，株式会社調和技研が運営する SNS 見守りシステム上に新しい機能として研究成果を実装するなど，実用化に向けた取り組みを実施してきた．

図 6 は，上記(2)で示した感情分析のアルゴリズムを実際にシステム化した例である．

しかし，本研究期間中において，アルゴリズムとしての評価実験は実施できたものの，実システムとして組み込んだ際の有用性に関しては研究期間内での十分な評価を実施できていない．この理由としては，アルゴリズムをシステムとして組み込む開発に時間が必要であることと，利用ユーザ数についても本研究期間中には，大規模化を実証するほどの規模で検証することができなかったことが原因である．しかしながら，蓄積したデータに基づくアルゴリズムの評価実験については一定の成果が得られたため，実用化と大規模化の検討については，今後も継続して研究協力者と協力しながら進めていく予定である．また，実用化に向けた検証の成果については得られた時点で学会等に報告していく予定である．

図 6 SNS データから感情推定するシステム

５．主な発表論文等

〔学会発表〕（計 8 件）

① Toshiki Tomihira, Akihiro Yamashita, Katsushi Matsubayashi: Estimation of user attributes using LIWC and application to SNS, Proceedings of the First International Symposium on AI for ASEAN Development, Phuket(Thailand), 2018

② 浅妻 佑弥, 山下 晃弘, 松林 勝志: SNS 上への発言の特徴分析に基づくユーザの属性推定, 1T-03, 情報処理学会第 80 回全国大会, 東京, Mar. 2018

③ 冨平 準喜, 山下 晃弘, 松林 勝志: LIWC を用いたユーザー属性推定手法の検討と SNS データへの応用, 1T-02, 情報処理学会第 80 回全国大会, 東京, Mar. 2018

④ 松林 圭, 松原良和, 今野陽子, 小野良太, 井上祐寛, 山下晃弘: 特定の話題に関する対話エージェントの実現に向けた特徴語及び状態抽出法, 5Q-04, 2-pp565-566, 情報処理学会第 79 回全国大会, 名古屋, Mar. 2017

⑤ 藤原裕樹, 山下晃弘, 松林勝志: Twitter から獲得した会話データに基づく雑談対話システムの開発, 3M-01,情報処理学会第 78 回全国大会, 横浜, Mar. 2016

⑥ 新田大悟, 山下晃弘: Twitter 上に投稿された画像への深層学習の適応および内容の分析, 5W-03,情報処理学会第 78 回全国大会, 横浜, Mar. 2016

⑦ 松林圭, 五味京祐, 古川和祈, 松尾祐佳, 松原良和, 日諸マルセロ優次, 中村拓哉, 山下晃弘, 松林勝志: Twitter 上に投稿された文章に基づく感情推定法とその応用に関する検討, 5W-06,情報処理学会第 78 回全国大会, 横浜, Mar. 2016

⑧ 山下 晃弘, 松林 勝志: 学生の SNS 利用実態の調査とプライバシー漏洩リスク評価に関する検討, 第 14 回情報科学技術フォーラム(FIT2015), F-037, 愛媛, 2015

６．研究組織

(1)研究代表者

山下　晃弘（YAMASHITA, Akihiro）

独立行政法人国立高等専門学校機構　東京工業高等専門学校・情報工学科・准教授

研究者番号：80589838

(2)研究協力者

松林　勝志（MATSUBAYASHI, Katsushi）

独立行政法人国立高等専門学校機構　東京工業高等専門学校・情報工学科・教授

中村　拓哉（NAKAMURA, Takuya）

株式会社調和技研・代表取締役