様　式　Ｃ－１９、Ｆ－１９、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成 ２８ 年　 ６ 月 ２１ 日現在

機関番号：１２６０８
研究種目：挑戦的萌芽研究
研究期間：2013～2015
課題番号：２５５４００９７

研究課題名（和文）ソーシャルメディアにおける個人情報秘匿技術に関する研究

研究課題名（英文）Privacy Preservation on Social Media

研究代表者
奥村　学（Okumura, Manabu）

東京工業大学・精密工学研究所・教授

研究者番号：６０２１４０７９

交付決定額（研究期間全体）：（直接経費）　　2,900,000円

研究成果の概要（和文）：あるソーシャルメディア上のユーザが，異なるソーシャルメディア上のどの「個人」なのかを特定する技術を開発した．また，開発した「個人」特定技術を元に，「個人」特定を困難にできる，秘匿するべき「個人」特定の手がかり情報を選択する技術を開発した．さらに，ユーザの性格を推定する技術を開発することを目的に，ユーザに性格判断テストを受けてもらった結果と，そのユーザのTwitter上でのツイート集合及びブログのエントリ集合を収集したデータセットの組を構築した．そして，このデータセットを分析することで，ツイート集合やブログのエントリ集合を入力として、ユーザの性格を自動推定する分類器を構築した．

研究成果の概要（英文）：We first presented a method for identifying which blogger is the same person as a user on Twitter, using the set of documents s/he has written. Then, based on the method that we developed, we presented a method for preventing from being identified, by indicating which clues in the documents should be hided. Furthermore, to develop a method for identifying writers' personality from their documents, we first constructed a dataset that consists of a pair of users with their personality information and a set of their tweets and blogs. We then investigated whether it is possible to identify users' personality
from a collection of their documents, and presented a method for constructing a classifier of users' personality from a collection of their documents.

研究分野：自然言語処理

キーワード：ソーシャルメディア　個人情報秘匿　著者同定　「なりすまし」検出

１．研究開始当初の背景

近年 WWW 上では，いわゆるマスメディアではない独自なメディアとして，ブログ，Twitter などに代表される，ソーシャルメディアと呼ばれるメディアが盛んになり，このメディアを介して多くの一般の人々が情報発信を行っている．一般の人々が発信している大量の情報を有効に活用しようという企業の活動も始まっている．一方で，ソーシャルメディア上で情報発信している一般の人々の個人情報の保護が声高に叫ばれるようになってきている．一般の人々が発信している情報の有効活用のため，個人情報を収集，利用しようという企業活動が過熱していること，情報リテラシーをあまり持たない情報「初心者」も数多く情報発信に加わるようになってきていることなどがその理由と考えられる．そのため，データマイニングの分野では，個人情報を保護しながら分析を行う技術として PPDM(Privacy Preserving Data Mining)[1]と呼ばれる技術が注目を集め，かつ実用化され始めている．また，このような状況から，テキスト処理技術としては，人々が発信しているテキスト情報の中の個人情報(たとえば，氏名，住所，生年月日など)を抽出し秘匿(マスキング)する技術が開発されるようになってきている(たとえば，[2])．しかし，陽に記述されている個人情報を秘匿しても，「個人」が特定できてしまう場合は多い．また，近年のように，多くの人々が複数のソーシャルメディアにおいて情報発信を行う状況下では，ソーシャルメディア間で「個人」が紐づけ(対応付け)できてしまうと，「個人」特定の手がかりとなる情報を増やしてしまうことから，それを考慮に入れた個人情報秘匿技術が必要になる．

[1] Privacy-Preserving Data Mining: Models and Algorithms, C.C.Aggrawal and P.S.Yu eds.), Springer, 2008.
[2] www.nri.co.jp/opinion/it_solution/2005/pdf/IT20050805.pdf.

２．研究の目的

そこで本研究課題では，ブログと Twitter などのように，複数のソーシャルメディアにおいて情報発信を行っているユーザを対象とした個人情報秘匿技術の研究を行う．ソーシャルメディアにおいては，自身の属性，行動，人間関係に関して記述する人も多い．それらの情報に関する記述を元に「個人」を特定する技術を開発するとともに，「個人」を特定する際に手がかりとなった情報のうち，どの情報を秘匿することにより，「個人」の特定を困難にできるかを明らかにする技術を開発する．より具体的には，a. あるソーシャルメディア(たとえば，Twitter)上のユーザが，異なるソーシャルメディア上のどの「個人」なのか(たとえば，どのブログを記述しているのか)を特定する技術を開発する．b. ソーシャルメディア上の記述から，ユーザの属性，行動，人間関係を抽出，推定し，それらを元に「個人」を特定する技術を開発する．c. a, b.で開発した「個人」特定技術を元に，「個人」特定を困難にできる，秘匿するべき「個人」特定の手がかり情報を選択する技術を開発する．

３．研究の方法

本研究課題では，

a. あるソーシャルメディア上のユーザが，異なるソーシャルメディア上のどの「個人」なのかを特定する技術を開発する．

b. ソーシャルメディア上の記述から，ユーザの属性，行動，人間関係を抽出，推定し，それを元に「個人」を特定する技術を開発する．

c. a, b.で開発した「個人」特定技術を元に，「個人」特定を困難にできる，秘匿するべき「個人」特定の手がかり情報を選択する技術を開発する．

４．研究成果

a.では，あるソーシャルメディア(たとえば，Twitter)上のユーザが，異なるソーシャルメディア上のどの「個人」なのか(たとえば，どのブログを記述しているのか)を特定する技術を開発した．b.では，a.で開発した「個人」特定技術を元に，「個人」特定を困難にできる，秘匿するべき「個人」特定の手がかり情報を選択する技術を開発した．

Twitter とブログの同一性判定システム

同一性判定システムの概要

１つの Twitter アカウントを入力とし，同一のユーザが作成したブログを含むブログ集合を，同一のユーザが作成したと考えられる順にランキングするタスクを考える．教師データとしてはユーザ自身により関連付けられた n 組の Twitter とブログのアカウントを使用する．

Twitter アカウントとブログアカウントの関連付いているペアを正例，Twitter アカウントと関連のない各ブログアカウントとのペアを負例とし，正例が負例より上位にくるように，ランキング SVM[3] に基づくランキング学習を行う．

素性

SVM の学習に用いる素性には，大きく分けて，類似度に基づく素性，ユーザ固有な表現に基づく素性，投稿時間に基づく素性の３つの素性を用いる．以下では各素性の詳細を述べる．

類似度に基づく素性

ペアとなる Twitter とブログ，それぞれを１つの文書とみなし，それらの文書間で定義された類似度を素性として用いる．本

研究では，3 つの類似度を素性として使用する．

（1）IPA 辞書で定義されている 69 個の品詞細分類ごとに，Twitter およびブログそれぞれで出現した形態素の頻度を要素とする形態素ベクトルを作成し，それらの Jaccard 係数を類似度として使用．

（2）Twitter およびブログそれぞれで出現した内容語の tf-idf 値を要素とする形態素ベクトルを作成し，それらの余弦類似度を使用．

（3）Twitter およびブログそれぞれで出現した任意の記号列の tf-idf 値を要素とする形態素ベクトルを作成し，それらの余弦類似度を使用．

ユーザに特有な表現を考慮した素性

Schwartz ら[4] は Twitter を対象としたユーザの同一性推定タスクにおいて，ある特定のユーザ 1 人だけが使用する文字列が同一性推定の有力な手掛りとなったと報告している．そこで本研究でもあるユーザに特有な表現を考慮した素性を用いる．

ただし，Schwartz らの研究における実験設定では事前に対象となるユーザの集合が既知であることから，ある表現がそのユーザに特有な表現であるかどうか判別可能であるのに対し，本研究では判別対象とするユーザの母集団が既知であることを仮定していないことから，ある表現がそのユーザに特有な表現であるかどうかは判別できず Schwartz らが使用した素性をそのまま使用することはできない．

そこで本研究では以下の手順で作成した素性をユーザに特有な表現に基づく素性として使用する．

（1）ある Twitter とブログのアカウントのペアが与えられた場合，その Twitter データ中に 2 回以上出現した形態素のうち学習に使用する他の n - 1 個の Twitter アカウントで一度も使用されていない形態素をその Twitter アカウントに特有な表現とみなす．

（2）それらの表現のうちランキング対象のブログ中で 2 回以上出現した表現の数を，その Twitter とブログのアカウントのペアに対する素性として使用する．

たとえば，1,000 人の Twitter ユーザの中で 1 人のユーザのみが，複数回使用する形態素が 5 つある時，それらの形態素はそのユーザに特有な表現であると考える．その上で，それら 5 つの形態素のうち，ランキング対象とするアカウントのブログに出現した形態素の数を，ユーザに特有な表現を考慮した素性として使用する．

投稿時間を考慮した素性

Twitter やブログにはその日体験した出来事が投稿される場合が多いと考えられることから，同一のユーザにより，近い時間に投稿された Twitter とブログには，その出来事に関連する語が出現している可能性が高いと考えられる．

そこで投稿時間を考慮した素性として，近い時間に投稿された Twitter とブログで共起する固有名詞の細分類の割合に関する素性を導入する．

具体的には，ある Twitter とブログのアカウントのペアが与えられた場合，そのブログに含まれる各記事ごとに，ブログの投稿時間から前 24 時間の間に投稿された Twitter の全投稿をまとめたものを 1 文書とみなし，対象のブログとの類似度を算出し，全ブログ記事で平均を取った値を素性とする．Twitter のテキストとブログの類似度は 7 つの固有名詞の細分類ごとに算出する．ただし，本素性は対象の語が共起したかどうかを重視するため，類似度の計算に使用する形態素ベクトルの各要素の値は頻度ではなく，出現した場合に 1，出現しなかった場合に 0 となるような 2 値とし，類似度としては余弦類似度を用いる．

Twitter とブログの著者同一性判定実験

実験に用いるデータ

本研究の目的の 1 つは，複数のソーシャルメディアのアカウントを関連付けられたくない場合に，どのような特徴からそれらのアカウントを関連付けられてしまうかを分析することである．このため，実験に使用するアカウントとしては，ユーザ自身が関連付けを行っていない 2 つのソーシャルメディアのアカウントであることが望ましい．しかし，そのようなデータを収集することは困難であることから，ユーザ自身が関連付けている Twitter とブログのアカウントの組を疑似的なデータとして使用する．具体的には，Twitter のプロフィール欄にアメーバブログへのリンクが張られている場合に，その Twitter とブログのアカウントの組を同一のユーザにより作成されたものと考え，実験データとして使用する．

ここで，Twitter やブログにはプロフィール欄やアカウント名などのユーザの同一性の推定の手掛りとなりうる様々な情報が含まれていると考えられるが，本研究では Twitter とブログの投稿から得られる情報のみをユーザの同一性推定に使用する．これは，実際に複数のアカウントを関連付けられないようにしているユーザは，複数のアカウントを関連付ける要因となるような情報をプロフィール欄等に載せていないと考えられるためである．

実験に使用するアカウントの具体的な収集の手順は次の通りである．

手順 1 日本語の Twitter アカウントから，そのプロフィール欄にアメーバブログの URL が記載されているアカウントを収集する．

手順 2 収集されたアカウントから下記の条件 1，2 を満たすアカウント 3,000 組を抜き出し正例として使用する．また，負例を生成するため条件 2 を満たすブログ

アカウントも 10,000 アカウント収集.

条件 1 2013 年の 3 月 1 日から 10 月 31 日の期間中に Twitter の投稿数が 11 以上

条件 2 2013 年の 3 月 1 日から 10 月 31 日の期間中にブログの投稿数が 6 以上

収集された Twitter アカウントの 8 ヵ月間の平均投稿数は 1,635, ブログの平均投稿記事数は 60.2 であった.

実験設定

本研究における実験では, 収集された関連付きアカウント 3,000 組を 1,000 組ずつ 3 つに分割し, それぞれ学習データ, 開発データ, テストデータの正例として使用する. 負例は正例に含まれる Twitter アカウントと別途収集した 10,000 ブログアカウントから生成する. 正例に対する負例の数は学習時には 9, 99, 999 の 3 つの値で, テスト時には 99, 999, 9,999 の 3 つの値で実験を行った. 学習時とテスト時で正例と負例の割合を一致させていないのは, 実際にこのシステムを用いる場合を考えると, テスト時にどのくらいの負例が存在するか事前に分からないことから, 学習時とテスト時で正例と負例の割合を一致させるという設定が現実的ではないと考えられるためである.

実験結果の評価には, Twitter アカウント 1,000 個中, 候補ブログのランキング上位 1 つ, 5 つ, 10 つ中に正解のブログが含まれているアカウントの数(以下, Top1, Top5, Top10), および, 正解のブログの順位の逆数の平均(Mean Reciprocal Rank: MRR) を使用する. ランキング SVM のツールとしては, SVMrank を使用し, パラメータ C は開発データにおいて MRR が最大となる値を使用し, それ以外のパラメータはデフォルトの値を使用した.

実験結果

1 つの正例に対する負例の数と精度の関係

まず, 1 つの正例に対する負例の数と精度の関係を明らかにするための実験を行った. 結果を表 1 に示す.

表 1 1 つの正例に対する負例の数と精度の関係

Top1/MRR		テスト時に 1 つの Twitter アカウントに対して候補とするブログ数		
		100	1,000	10,000
学習時における	10	787/0.835	646/0.711	323/0.386
正例に 1 つに対	100	804/0.849	681/0.739	356/0.426
する全事例の数	1,000	807/0.851	679/0.738	369/0.434

学習時の正例の数はいずれも 1,000 事例であり, 負例の数をそれぞれ 1 つの正例に対して 9 事例, 99 事例, 999 事例に変化させて実験を行った結果, 負例の数が 9 事例の場合と 99 事例の場合を比較すると 99 事例の場合の方がより推定精度が良いことが確認できる. しかし, 負例の数が 99 事例の場合と 999 事例の場合を比較すると推定精度にあまり違いがないことが確認できる.

また, この傾向はテスト時における 1 つの Twitter アカウントに対して候補とするブログの数に依らないことも確認された. 対象の Twitter とブログが, 同一著者が作成したものかどうかの 2 値分類を行うような実験設定では, 学習時とテスト時の正例と負例の数を一致させた方が良い精度が得られる場合が多いことが一般的に知られているが, 本研究のようにランキング問題として捉えた場合は, 正例と負例の割合はあまり重要ではないと考えられる.

素性ごとの効果

各素性の有効性を確認するため, 素性を 1 種類ごとに除いた場合の精度を調べた. 結果を表 2 に示す.

表 2 各素性を除くことによる精度の変化

除く素性	Top1	Top5	Top10	MRR
類似度に基づく素性	537	637	661	0.585
特有な表現を考慮した素性	618	755	815	0.687
投稿時間を考慮した素性	642	768	819	0.707
すべての素性を使用	679	795	834	0.738

類似度に基づく素性, ユーザに特有な表現を考慮した素性, 投稿時間を考慮した素性はいずれも, これらの素性を除くことによりシステムの精度が大きく低下しており, これらの素性の有効性が確認できる. 特に, 類似性に基づく素性が有効であることがわかる.

関連付け防止システム

関連付け防止システムの概要

本節では, 関連付けを行っていない Twitter とブログのアカウントを持つユーザを想定し, 新たに投稿されたブログ記事に同一性推定において重要となる手掛かり語が含まれていた場合に, その語に関する情報とともに警告メッセージを提示するようなシステムの構築について検討する.

前節の結果から Twitter とブログの類似度, ユーザに特有な表現, および, 近い時間に投稿された Twitter とブログで共起する固有名詞は, 2 つのアカウントの同一性推定の手掛りになることを確認した. このうち, Twitter とブログの類似度は Twitter およびブログ全体から計算される値であることから, この値から同一性推定の手掛かりとなった特定の語を検出するのは難しいと考えられる. また, Kacmarcik らの研究[5] は高頻度の機能語の置換を行うことで, 著者推定の推定率を低下させる防止方法を報告していた. しかし, ブログの投稿ごとに著者が高頻度で使用する機能語のマスキングを著者に求めることはユーザへの負担が大きいと考えられる. 一方, ユ

ーザに特有な表現や，近い時間に投稿された Twitter とブログで共起する固有名詞を検出することは比較的容易であり，修正が必要となる語の数も少ないと考えられることから，これらの表現を検出しマスキングすることで，同一性推定の防止を行うことを考える．

関連付け防止システムの評価実験

実験設定

本実験では，Twitter を使用していたユーザが新しくブログを使い始める場合に，Twitter のアカウントが知られている第三者にブログの存在を知られるのを防ぎたいという状況を想定し，システムの評価を行う．このため，本評価では以下の条件を満たすユーザを実験に使用する．

条件 1 2013 年 7 月 1 日以前に Twitter での投稿がある

条件 2 2013 年 7 月 1 日以降にブログの投稿がある

上記の 2 つの条件に当てはまるユーザは，上で説明したテストデータの正例 1,000 組のうちちょうど 800 組存在した．これら 800 組を正例とし，正例 1 組に対して負例数は 9,999 事例として評価実験を行う．正例のブログの中には 2013 年 7 月 1 日以前の投稿が存在するものもあるが，7 月 1 日以降にブログを始めたとの想定から，それ以前の投稿は使用しない．

評価方法

本評価では，正例のブログにおいて 7 月 1 日以降に新しく投稿される記事につき，新しく投稿される記事が第三者に関連付けられる危険性を含んでいるか判定する．判定を行う指標として，2 つの値 rank と rr-diff を求める．

rank は Twitter とブログの同一性判定システムにより出力されるランキングの値であり，rr-diff は投稿後と投稿前の rank の逆順位の差である．

rank は Twitter のユーザと同一のユーザが作成したと考えられる度合いであり，rr-diff の値は，新たなブログ記事の投稿によりどの程度ランキングが上昇するかの危険性を表している．これら 2 つの値のどちらかが，それぞれ定めた閾値 rankth，rr-diffth 以上となる場合，第三者に関連付けられる危険性を含んでいると判定する．すわなち，あるユーザの Twitter が与えられた場合に，同一ユーザにより作成されたブログの候補 10,000 ブログの中で，同一ユーザにより作成されたブログのランキングが rankth 以上となるか，ブログの投稿前と投稿後で順位が大きく変動し逆順位の差が rr-diffth を超えた場合に，対象の投稿を著者同定の危険性の高い投稿であるとみなし，著者同一性推定防止システムの処理対象とする．

具体的には，危険性を含んでいると判明した新しいブログ記事からユーザに特有な表現と固有名詞を除き，再度 rank と rr-diff を求め，各閾値との比較を行う．再計算した rank と rr-diff の値がそれぞれの閾値未満となる場合，そのブログ記事に対してはユーザに特有な表現と固有名詞のマスキングは有効であったと考え，その割合によりシステムの有効性を評価する．本評価では rankth の値を 20 と，rr-diffth の値を 0.01 として評価した．

7 月 1 日以降のブログの中で，古い記事から順に 1 記事ずつ追加し危険性の有無を判定する．ここで，Twitter と負例のブログの記事データは，着目している正例記事が投稿された時間までのデータを使用する．危険性がある場合，上記のマスキングが有効であるか調べる．危険性がない場合，次の記事に危険性が有るかを判定する．危険性を含む記事が見つかるまで危険性の有無の判定を行う．したがって，ユーザ 1 人に対してマスキングが有効か否かの判定は 1 回のみとなる．

評価実験結果と考察

本評価実験に使用した閾値では，正例 800 組のうち，rank または rr-diff の値が閾値以上となる記事の投稿があり，評価の対象となる正例は 370 組であった． 370 組の中で固有名詞とユーザに特有な表現をマスキングする関連付け防止システムが有効に働いたのは 61 組，有効に働かなかったのは 309 組となった．すなわち，第三者に関連付けられる危険性を含むブログ記事に対し，関連付け防止を行うことができた割合は 16.5%であった．

さらに，ユーザの性格を推定する技術を開発することを目的に，ユーザに性格判断テストを受けてもらった結果と，そのユーザの Twitter 上でのツイート集合及びブログのエントリ集合を収集したデータセットの組を構築した．そして，このデータセットを分析することで，ユーザの性格を，ツイート集合やブログのエントリ集合を用いて，推定可能であるかどうか検討を行い，ツイート集合やブログのエントリ集合を入力として、ユーザの性格を自動推定する分類器を構築した．

[3] Joachims, T.: Optimizing Search Engines Using Click-through Data, Proc. of KDD'02, pp. 133-142 (2002).

[4] Schwartz, R., Tsur, O., Rappoport, A. and Koppel,M.: Authorship Attribution of Micro-Messages, Proc. of EMNLP'13, pp. 1880-1891 (2013).

[5] Kacmarcik, G. and Gamon, M.: Obfuscating Document Stylometry to Preserve Author Anonymity, Proc. of COLING-ACL'06, pp. 444-451 (2006).

５．主な発表論文等

（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計　2 件）

1. Noriyuki Okumura, Manabu Okumura, A Construction of Knouledge Base for Personality Estimation based on Submitted Text Data in Twitter or Blogs, KEOD2015, pp. 418-423 (2015)，査読有

2. Miho Matsunagi, Ryohei Sasano, Hiroya Takamura, Manabu Okumura, cquiring Activities of People Engaged in Certain Occupations, the 14th Pacific Rim International Conference on Artificial Intelligence (PRICAI 2016) 2016 年 8 月 24-26 日，査読有（掲載は確定しているが印刷前の為、巻及び頁は未定）

〔学会発表〕（計　4 件）

1. 木原 裕二，笹野 遼平，高村 大也，奥村 学，複数のソーシャルメディアアカウントの関連付け防止システムの構築，情報処理学会第 216 回自然言語処理研究会，2014 年 5 月 23 日，東京工業大学

2. 奥村紀之，金丸裕亮，奥村学，感情判断と Big Five を用いたブログ著者の性格推定に関する調査，2015 年度人工知能学会全国大会，2015 年 6 月 2 日，公立はこだて未来大学

3. 奥村紀之，奥村学，オンラインでの振る舞いから想定される人物像の特徴，電子情報通信学会「言語理解とコミュニケーション」研究会，2015 年 6 月 5 日，徳島大学

4. 馬緤美穂，笹野遼平，高村大也，奥村学，職業ごとの行動に関する知識の収集，情報処理学会第 222 回自然言語処理研究発表会，2015 年 7 月 15,16 日，首都大学東京秋葉原サテライトキャンパス

〔図書〕（計　0 件）

〔産業財産権〕
○出願状況（計　0 件）

名称：
発明者：
権利者：
種類：
番号：
出願年月日：
国内外の別：

○取得状況（計　0 件）

名称：
発明者：
権利者：
種類：
番号：
取得年月日：
国内外の別：

〔その他〕
ホームページ等

６．研究組織
(1)研究代表者
奥村　学（OKUMURA MANABU）
東京工業大学・精密工学研究所・教授
研究者番号：60214079

(2)研究分担者
（　　）

研究者番号：

(3)連携研究者
（　　）

研究者番号：