

ブログにおける記述者本人の氏名判別手法の提案

森 加夢偉† 金井 敦†

あらまし 近年、個人情報保護法の施行や、個人情報の電子化が急速に進むに伴い、Webでは個人情報の保護に対する意識が高まっている。一方、個人のインターネットリテラシの欠如が一部でみられ、特にブログでは無意識な自分のブログにおいて自身の個人情報の漏えいが発生している。

こうした個人情報の不用意な露出を防止するためには、個人情報の露出状況を本人に自覚させることが、一つの解決策である。しかし、自覚させるためには、個人情報の露出状況を何らかの方法でフィードバックする必要がある、そのため客観的かつ自動的な評価が必須となる。自動化するに当たっては、自由な形式で書かれているブログ記事に出現する個人情報を抽出し本人のものとどうか判別する必要がある。

本稿では、個人情報の中でも特に重要な氏名に着目し、本人のものか他人のものかを判別するモデルを提案する。また、実際のブログ記事データをを用いパラメータを決定し評価値のしきい値を明らかにするとともにモデルを検証する。

A Methodology to extract Self-name on Blog Articles

KAMUI MORI† ATSUSHI KANAI†

Abstract In recent years, individual information has been rapidly digitized and people have had great concern about personal information protection with enactment of the Personal Information Protection act in Japan. On the other hand, incidents of personal information exposure occur frequently.

It is important to recognize the situation of information exposure to prevent the danger by information to the person. It is necessary to objective evaluation of the situation and automatic evaluation. Technology information is to distinguish self personal information on the blog article is important to extract personal information.

In this paper, we focus on "Name" which is most important information as personal information and a model to extract self name from blog articles is proposed. In addition, we estimate the parameter and threshold value and evaluate the model using actual blog articles.

1. はじめに

近年、個人情報保護法の施行や、個人情報の電子化が急速に進むに伴い、個人情報の保護に対する意識が高まっている[1][2]. 情報化社会におけるインターネット上でも個人情報保護は重要な課題となっており、様々な個人情報保護技術や情報露出対策の提案や開発が行われている。

一方、一部のユーザにおいて、インターネットユーザー一人一人のモラルの欠如や、知識不足による危険意識が薄いこと（インターネットリテラシの不足）によって、無意識のうちに自分自身の個人情報の漏えいを引き起こしている。特にブログでは、他人に見られていない個人的なノート日記を書く感覚で気軽に書き込めるため、ブログ作者（情報発信者）本人が、誰にでもわかりやすい形で自身の個人情報を発信してしまう場合がある。その結果として、ブログ内容からブログ作者本人の氏名、住所が推測されることや、顔写真、会社や学校などの所属先がわかるなど、特定されるような個人情報が露出しているケースが増えている[3].

こうした個人の情報発信において、自身の個人情報の不用意な露出を未然に防止するためには、個人情報の露出状況を本人に自覚させることが、一つの解決策である。ブログ利用者の危険意識が高まり、特定される様な危険な自身の個人情報を自ら発信しなくなり、個人情報漏えいの対策にも繋がると思われる。

しかし、その個人情報は、個人情報の保護に関する法律（2条1項）に定義されるものだけでなく、他のセンシティブ情報やライフログ等も含まれ、多種多様である。一般的な利用者は、どの情報がどの程度危険であるかを判断するのが難しいといえる。

したがって、書き込んだブログ記事の危険性を、状況に応じて書き込んだ本人に警告することや、第三者機関が状況を把握するなどの活動が重要になると考えられる。そのためには、個人情報の露出状況を把握し、客観的に評価し示さなければならない。このため、自由な形式で書かれているブログ記事から、露出度合いや露出情報などを自動的に判断する必要がある。このためには、個人ブログにおける個人情報（本人情報）を誰の情報か自動的に判別され、その情報を客観的に評価し、状況が誰にでも簡単に分かる形で示すシステムが必要である。

本稿では、このシステムに特に必要な機能として、本人情報を判別抽出する方式について、本人の氏名を判別する手法を提案する。

本章以降では、2章で手法の概要を説明し、3章で本人氏名の判別のアプローチ方法とそのアルゴリズムについて述べる。4章でアルゴリズムに則ってパラメータを決定し、5章でその考察を行う。最後に6章で本稿のまとめを述べる。

†法政大学大学院 工学研究科

Graduate School of Engineering, Hosei University

2. 概要

本研究は、日記形式で書かれたブログ記事を検索し、その中の本文中から個人情報を抽出し、本人のものか他人のものかを判別する。最終的に個人情報ごとに定量化し、危険度計算[3]をすることによって、個人情報の露出状況を分かり易く示すことを目的とする。以下の図1に方式の全体の流れを示す。

まずブログ検索において、ブログサイトから日記形式で書かれたブログを特定し、URLを取得する。その後ブログの構造を解析し、ブログ記事本文のみをテキストで抽出する。

次に抽出したテキストの形態素解析を行い、個人情報を抽出する。抽出した個人情報を、誰の情報か判別し、ブログ作者本人の個人情報のみを抽出する。

最後に、本人にとってどれ程の情報なのか定量化計算することで、ブログ作者に露出状況を分かり易く表示する。

本稿では、文献[4][5]で定義される個人情報露出量（個人情報項目の各要素を、その要素を持つ個人を区別する分解能力で、数値化したもの）において、数値の高い「氏名」に焦点を当て議論する。本人の氏名か、作者本人以外の氏名かを、判別する手法を検討する。

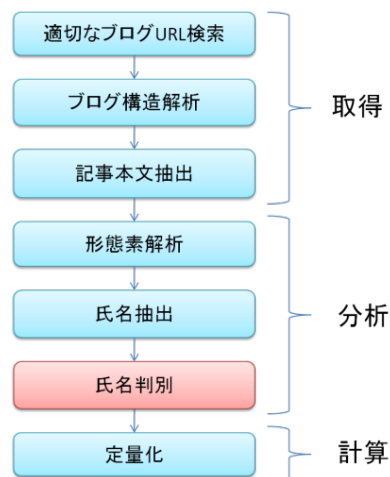


図1 方式の流れ

ここで作者本人以外の氏名を「他人の氏名」と呼び、作者の友人や知人だけでなく、面識のない第三者（スポーツ選手や有名人など）も含むことにする。

なお本稿において氏名(Name)とは、苗字(Family Name)や名前(First Name)、IDなど、人を指す固有名詞とする。

3. 本人名判別モデル

3.1. ブログの特徴とアプローチ方法

本稿で判別の対象としているブログ記事本文は、文章が体系的に（正しい文法に則って）書かれておらず、自由な表現がなされている。サイトにより異なるが、一般的にブログは以下のような特徴を持っている。

- ・時系列（記述した時間）に並ぶ
- ・字数制限・形式制限なし
- ・絵文字・顔文字・記号など、特殊文字が利用される
- ・頻繁な改行
- ・文の語り（記述）形式が一人称であることが多い
- ・主観的形容詞（感覚・感情を表す形容詞）が多用される

新聞記事や論文などの「正しい日本語」で「規則的に書かれた文章」を、従来の自然言語処理を利用して分析し、氏名を抽出する手法は存在しており、その研究・実装は既に行われている。これら手法は、主に文法解析を利用している。日本語の文法解析は2種類あり、言語表現に使用された文法的約束に関する解析で、品詞、活用形など単語の品詞レベルの属性を決定するための形態素解析と、それぞれの単語が文中で果たす役割を決定するための構文解析である。

以上より、従来の自然言語処理では、本研究の目的である「ブログ作者本人の情報」を抽出することが困難であることがわかる。そこで、前述したブログの特徴を生かし、多様な記述に対応する文章パターンによる抽出方式を提案する。但し、以下を前提条件とする。

- ・日本語で書かれた、日記形式のブログを判別対象とする
- ・引用文（括弧などを利用した文など）は、判別対象としない
- ・一文中（或いは一行中）に2つ以上の名前がある文章は、判別対象としない
- ・コンテキスト依存は考慮しないものとする

3.2. 判別ブロックモデルと判別アルゴリズム

あるブログ作者本人の氏名を抽出するためには、非体系的な文章や多様な記述方法に対応する必要があることに加え、誰の情報か判断するために意味を理解する必要がある。そこで、日記形式で書かれたブログから作者本人の氏名を抽出する手法を提案する。

3.2.1. モデルと適用手順

氏名が含まれた文を抽出する。また、文となっていない場合は、氏名を含む、一行を抽出するものとする。取り出した文字列を判別ブロックと呼ぶ。判別ブロック内の氏名を中心に、前方にある文字列を前ブロック、後方にある文字列を後ブロックと定義する。以下の図に判別ブロックモデルの図を示す。

それぞれのブロック内に、あるパターンが当てはまるか（マッチングするか）をチェックする。パターンとは、ブロック内の文字列の一部分であり、文字列全てをマッチングさせるものではない。パターンの内容については後述する。

パターンごとにポイント（パターンの重み）を設定し、そのパターンが当てはまるごとに、ポイントを加算していく。一判別ブロックにおける合計ポイントが、一定数（しきい値）以上であれば本人の氏名とし、それ未満の場合は他人の氏名として判別する。具体的なポイントとしきい値の設定については後述する。

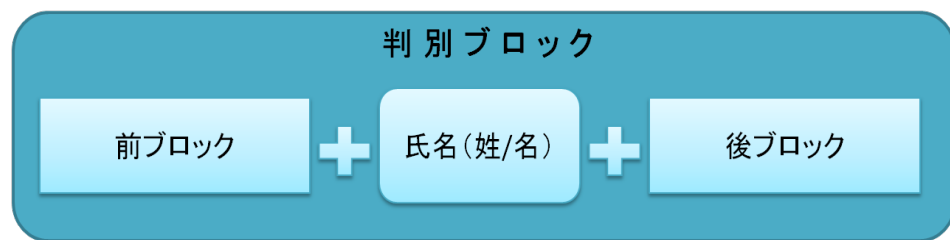


図2 判別ブロックモデルの図

判別の流れを説明する。手順は、以下のとおりである。

- ① 抽出した氏名が含まれた文字列が、一文として成り立っているか評価する
- ② 氏名が含まれた一文または一行を、判別ブロックと設定する
- ③ 判別ブロック内の氏名より前方の文字列を、前ブロックと設定する
- ④ 判別ブロック内の氏名より後方の文字列を、後ブロックと設定する

- ⑤ 前ブロックにおいて前パターン1)、前パターン2)があるかを判定する
- ⑥ 該当するパターンがあった場合、それぞれ設定されたポイントを加算する
- ⑦ 後ブロックにおいて後パターン1)～後パターン6)があるかを判定する
- ⑧ 該当するパターンがあった場合、それぞれ設定されたポイントを加算する
- ⑨ 合計ポイントがしきい値を超えるものを本人氏名と判定して抽出する

3.2.2. ブロック内のパターン

パターンは前ブロックと後ブロックのそれぞれ別々にあり、その中に、本人氏名と判断するパターンと、他人氏名と判断するパターンがある。前者はプラスのポイント、後者はマイナスのポイントであり、その合計ポイントを評価値とする。

以下に前ブロック、後ブロックそれぞれの中のパターンについて示す。

3.2.2.1. 前ブロック

ここでは前ブロックにおいて、本人氏名と判断するパターンと、他人氏名と判断するパターンの定義を示す。

本人氏名と判断するパターン

前パターン1) 一人称を含む

- 例)
- | | | |
|------|------|------|
| ・私 | ・わたし | ・ワタシ |
| ・僕 | ・ぼく | ・ボク |
| ・俺 | ・おれ | ・オレ |
| ・自分 | ・じぶん | ・ジブン |
| ・おいら | ・あっし | ・ミー |

他人氏名と判断するパターン

前パターン2) 自分との関係性を示す修飾語を含む

- 例)
- | | |
|--------------|--------|
| ・隣の、隣に | ・左の、左に |
| ・近くの、近くに | ・右の、右に |
| ・友人の、友人に | |
| ・知り合いの、知り合いに | |

ここで重要なことは、品詞は関係ないことである。一人称であるからといって、主語であるかどうかや、修飾語が連体修飾か連用修飾かは考慮しない。

3.2.2.2. 後ブロック

同様に、後ブロックにおいて、本人氏名と判断するパターンと、他人氏名と判断するパターンの定義を示す。

本人氏名と判断するパターン

後パターン1) 断定表現

- | | | |
|----|------|----------|
| 例) | ・だ. | ・でした. |
| | ・だよ. | ・でございます. |
| | ・です. | ・だよ～ん. |

後パターン2) 氏名を引用する語尾

- | | | |
|----|-----------|-----------|
| 例) | ・という. | ・という者です. |
| | ・といいます. | ・と申すものです. |
| | ・という名前です. | |
| | ・と申します. | |

後パターン3) 氏名から続く特定の口語的な文の締めくり

- | | |
|----|------------|
| 例) | ・がお送りしました. |
| | ・が送る. |

他人氏名と判断するパターン

後パターン4) 格助詞・係助詞からはじまる

- ・が ・の ・を ・に ・へ ・と ・から ・より ・で ・や
 ・は ・も ・こそ ・でも ・しか ・さえ ・だに

後パターン5) 敬称からはじまる

- | | | |
|----|------|----|
| 例) | ・さん | ・様 |
| | ・くん | ・君 |
| | ・ちゃん | |

後パターン6) パターン伝聞の表現

- | | |
|----|-------|
| 例) | ・～そうだ |
| | ・～らしい |
| | ・～のよう |

3.2.3. ポイントの定義

前述したパターンにマッチングすれば、その対象に一定のポイントを付加する。その合計ポイントを評価値とし、一定の値（しきい値）以上のものを本人の氏名と判別する。

設定するポイントは、A：本人氏名と判断するパターンが当てはまる確率、B：他人氏名と判断するパターンが当てはまる確率、P：評価値として

$$P = \frac{A - B}{A + B} \times 100$$

の式で算出されたものとする。

本人氏名と判断するパターンが当てはまる確率とは、本人氏名を含んだ全ての判別ブロック群に、ポイント設定をする対象のパターンが当てはまるか走査して、当てはまった割合のことである。また、他人氏名と判断するパターンが当てはまる確率とは、その逆である。

この計算方法により、本人氏名と判断するパターンの場合は、計算結果のポイントがプラスになり、加算する。反対に他人氏名と判断するパターンの場合は、必然的にポイントがマイナスとなり、減算する。したがって、本人氏名であればプラス側に、他人氏名であればマイナス側にポイントが付加することになる。

4. パラメータの決定

4.1. 決定に用いたサンプルについて

ここでは、実際のブログを調査し、実例サンプルに基づいてパラメータを決定する。パラメータ決定に用いるサンプルは、日本語で書かれた日記形式のブログ記事であり、人間が読み、本人の氏名か他人の氏名かを判断した、2つの判別ブロックの集合である。調査したブログサイトは Yahoo!ブログ、goo ブログ、FC2、livedoor ブログ、Ameba ブログである。

今回はサンプルとして、本人の氏名と人が判断した判別ブロックを 766 件と、他人の氏名と判断した判別ブロックを 215 件、用意した。

4.2. ポイントの値（パターンの重み）の決定

ここで、前述したポイントの定義に沿って、ポイントの値を算出する。

最もマッチング件数が多い後パターン1を例にあげる。後パターン1の、本人の判別ブロック群におけるマッチング件数÷本人の氏名と判断した判別ブロック群=Aであるから、 $A = 561 / 766 \approx 0.73$ となる。また、後パターン1の、他人の判別ブロック群におけるマッチング件数÷本人の氏名と判断した判別ブロック群数=Bであるから、 $B = 3 / 215 \approx 0.014$ となる。

よって、

$$P = \frac{A - B}{A + B} \times 100 = \frac{0.73 - 0.014}{0.73 + 0.014} \times 100 = \frac{0.716}{0.744} \times 100 = 0.9623 \dots \times 100 \approx 96$$

後パターン1におけるポイントは96ポイントとなる。

表1に実際のマッチング件数と、算出したポイントの一覧を示す。

表1 実際のマッチング件数とポイント

パターン	本人氏名 群	他人氏名 群	付加ポイント
前パターン1	18	0	100
前パターン2	0	5	-100
後パターン1	561	3	96
後パターン2	83	0	100
後パターン3	3	0	100
後パターン4	7	25	-85
後パターン5	1	152	-100
後パターン6	0	3	-100

4.3. しきい値の有効範囲

ここでは、本人の氏名か他人の氏名かを判別する評価値のしきい値範囲を検討する。

図3のように、実例サンプルをそれぞれ分けた判別ブロック群に、本人名判別手法を適用し、認識率を求めることで決定する。

図4は、横軸を評価値、縦軸を氏名の認識率とするグラフである。横軸に対して垂直に、しきい値を引いた場合、その右側が認識率である。

実線は、本人氏名を含む判別ブロックを、本人氏名が含まれると正しく判別した認識率を表している。点線は、他人氏名を含む判別ブロックを、本人氏名が含まれると誤って判別した認識率を表している。

グラフより、本人氏名の認識率（実線）が高い部分である、0ポイントより大きく96ポイント未満の範囲が、判別手法の有効性がある範囲と考えられる。

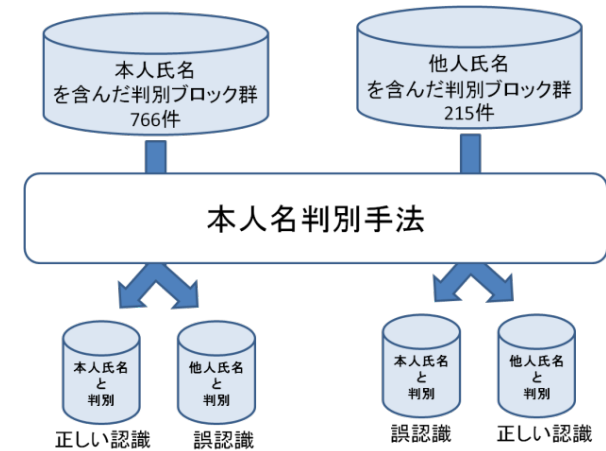


図3 本人名判別手法の精度検証

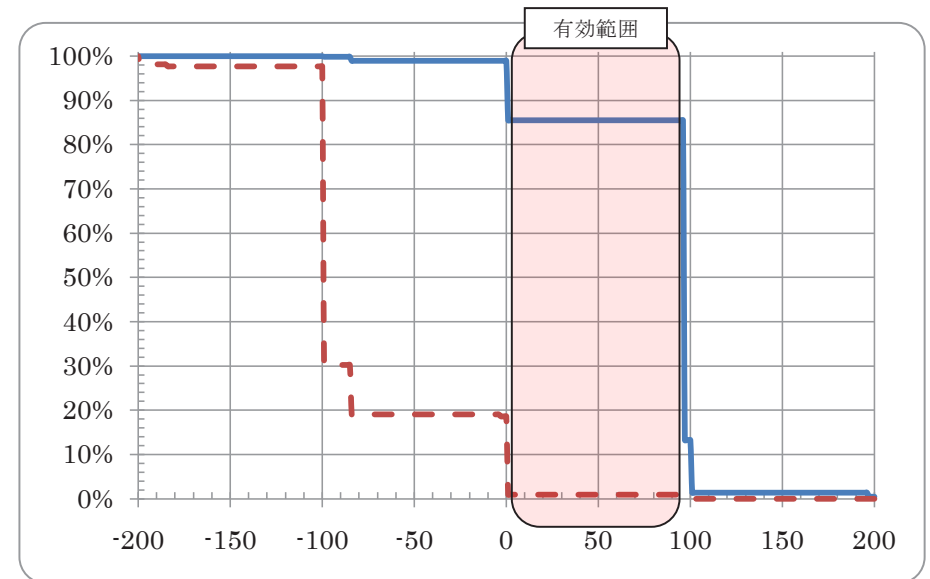


図4 実際の判別ブロック群における認識率

5. 考察

前章で決定した、しきい値の有効範囲について検討する。

有効範囲は本人氏名を含む判別ブロックが、正しく認識する場合を多く含む（本人氏名が含まれると正しく判別した認識率が 85.5%の）範囲である。その中で、認識できなかったものは、評価値が 0 ポイントであるものと、マイナスポイントのものであり、判別ブロック群全体の 14.5%存在する。実際に調査して抽出した判別ブロックの例を挙げる。

判別ブロックの例

- 1) 大山真志, 参上. ⇒ 0 ポイント
- 2) 紗季なんか全然だけど疲れた ⇒ 0 ポイント

いずれの例も、パターンが一切マッチングしない。特に例 2 は完全に日本語として成り立っていない。この様なパターンの「マッチングなし」の場合が、「本人氏名である」と正しく認識できなかった原因である。

今後の課題として、主に 3 点ある。1 点目は、調査件数を増やすことである。しきい値の信頼性向上を図る。2 点目は、パターンを更に増やすことである。本人氏名の集合に対し、プラスの合計ポイントのときに、100%の認識率がある範囲を作り、しきい値設定の最適化を図る。3 点目は、プログラム実装をして、自動的に判別することができるか検証する必要があることである。

6. おわりに

本稿では、ブログ記事から抽出した氏名を、判別ブロックモデルを利用し、ブログに特化したパターンをマッチングさせることで、本人のものか、他人のものか判別を行う手法について提案した。氏名を中心に前後で分割しパターン化することで、自由な文形式に対応でき、パターン辞書を増やすことで精度を上げられる様な拡張性がある。

考察においては、パターンごとにポイントを設定し、判別ブロックにマッチングさせ、評価値（合計ポイント）を求め、判別に有効なしきい値の範囲で、判別することができた。0 ポイントより大きく 96 ポイント未満の範囲において、85.5%の判別精度があり、妥当性があると考ええる。

参考文献

- 1) NRI セキュアテクノロジー株式会社：情報セキュリティに関するインターネット利用者意識 2006, <http://www.nri-secure.co.jp/news/2007/pdf/vol3-1.pdf>
- 2) 総務省 情報通信政策研究所：ブログの実態に関する調査研究の結果, <http://www.soumu.go.jp/iicp/chousakenkyu/data/research/survey/telecom/2008/2008-1-02-2.pdf>
- 3) 針谷友彰, 佐藤和紀, 安井良介, 金井敦：ブログにおける個人情報漏えいモデル, 情報処理学会研究報告, 2008-EIP-041, Vol.2008, No.91, pp.65-70 (2008)
- 4) Ryosuke Yasui, Atsushi Kanai, Takashi Hatashima, Keiichi Hirota：The Metric Model for Personal Information Disclosure, ICDS2010, vol.68, pp.112-117, 2010
- 5) 安井良介, 金井敦, 廣田啓一, 畑島隆：個人情報記述レベルの定量化手法の検討, DPSWS2009, Vol.2009, (2009)
- 6) 安井良介, 佐藤和紀, 針谷友彰, 金井敦, 廣田啓一, 谷本茂明：ブログにおける個人情報漏えいレベルの定量化, 情報処理学会研究報告, 2008-EIP-043, Vol.2009, No.11, pp.9-16(2009)
- 7) 森加夢偉, 金井敦：ブログにおける個人情報抽出方式, 法政大学工学部卒業論文(2009)
- 8) 長尾真, 黒橋禎夫, 佐藤理史, 池原悟, 中野洋：言語の科学 9 言語情報処理, 岩波書店 (1998)
- 9) 天野 真家, 宇津呂 武仁, 成田 真澄, 福本 淳一, 石崎 俊, 自然言語処理 (IT Text), オーム社(2007)
- 10) 徳永 健伸, 情報検索と言語処理 (言語と計算), 東京大学出版会(1999)
- 11) 北 研二, 津田 和彦, 獅々堀 正幹, 情報検索アルゴリズム, 共立出版(2002)