Vol.2010-DPS-145 No.26
Vol.2010-GN-77 No.26
Vol.2010-EIP-50 No.26
2010/11/26

ブログにおける氏名の本人性判別手法の実装と評価

森　加夢偉†　　岩下　透††　　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 上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 Web リテラシの欠如が一部でみられ，特に個人のブログにおいて，記述者自身の個人情報の漏えいが起きている．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り，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，誰の情報かを判別する必要がある．本稿では，ブログに記述された氏名が，本人のものか他人のものかを判別する手法を実装し，その評価を行う．

Implementation and Evaluation of a Method to Identify the Writer's Name in a Blog

KAMUI　MORI†　　TORU　IWASHITA†　　ATSUSHI　KANAI†

Abstract　With the Personal Information Protection Law enforced and private information being digitized at a rapid pace, awareness is rising in recent years about personal information protection on the Web. On the other hand, some individuals are seen lacking in Web literacy, which led to private information leakage of blog writers themselves, particularly in personal blogs. One of the measures against such unintended leakage of personal information is to make blog writers aware how much of their private information is at risk of exposure. Doing it, however, requires the exposure risk fed back to blog writers in some way, making objective and automated evaluation essential. In automation approaches, all the personal information appearing in blog articles in free formats needs to be extracted, and then judged whom the information belongs to. This report discusses implementation and evaluation of a method to judge if each name appearing in blogs is the writer's or otherwise.

†法政大学大学院　工学研究科
Graduate School of Engineering, Hosei University
††法政大学　工学部
School of Engineering, Hosei University

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 Web 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，Web ユーザ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（Web リテラシ不足）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考えられる．

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，このシステムにおいて特に必要な機能として，本人情報を判別抽出する手法について，氏名の本人性を判別するモデルを提案した論文 12)の手法を，日本語形態素解析システムを用いて実装し，その評価と考察を行う．

本章以降では，２章で手法の概要を，３章で本人性判別モデルについて述べる．４章でモデルの実装について述べ，５章でその評価を行う．最後に６章で考察，７章で本稿のまとめを述べる．

Vol.2010-DPS-145 No.26
Vol.2010-GN-77 No.26
Vol.2010-EIP-50 No.26
2010/11/26

2. 概要

本研究は，まず日記形式で書かれたブログの本文中から個人情報を抽出し，それが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]などをすることによって，個人情報の露出状況を分り易く示すことを目的とする．

本稿では，文献 4) 5)で定義される個人情報露出量において，数値の高い「氏名」に焦点を当て，文章解析を行わず，ブログにおける氏名を本人の氏名か他人の氏名かを，判別する手法について提案した文献 12)に則り，実装と評価を行う．

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とし，ブログ作者が面識のない第三者（スポーツ選手や有名人など）も含まないこととする．氏名(Name)とは，苗字(Family Name)や名前(First Name)，あだ名など，特定の人物を指す固有名詞とする．

本手法の判別対象について以下の４つを条件とする．

- 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．
- コンテキスト依存の文章（状況を考慮しなければ意味が判断できない文章）は，判別対象としない．

3. 本人性判別モデル

あるブログ作者本人の氏名を判別抽出する，判別ブロックを用いた本人性判別モデルについて述べる．

3.1. 判別ブロックモデル

人名が含まれた一文，または文として成り立っていない場合は，氏名を含む一行を単純な文字列の塊とし，これ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．図１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，設定するパターンテーブルが含まれるかをチェックする．これをマッチングチェックと呼ぶ．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．

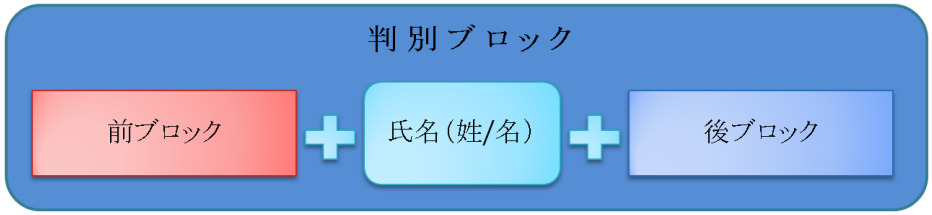

図１　判別ブロックモデルの概念図

3.2. パターンごとのスコアリング

ブロックが，パターンテーブル上のパターンが含まれれば（マッチングすれば），その対象に一定のポイント（重み）を加算する．これをスコアリングと呼ぶ．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し，それ未満の場合は他人の氏名として判別する．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．

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，本人氏名の可能性が高ければプラス側に，他人氏名の可能性が高ければマイナス側にポイントが付加することになる．

4. モデルの実装

本人性判別モデルを，形態素解析による人名の抽出から，パターンテーブルのマッチングチェック，スコアリングまでを自動的に一括で行うプログラムの実装を行った．以下，このプログラムを NI:2 と呼び，実装内容を示す．

4.1. パターンテーブル

パターンテーブル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．

以下に前ブロック，後ブロックそれぞれにおいて，マッチングすると本人氏名と判断するパターンテーブルと，他人氏名と判断するパターンテーブルを示す．

Vol.2010-DPS-145 No.26
Vol.2010-GN-77 No.26
Vol.2010-EIP-50 No.26
2010/11/26

● **前ブロックテーブル**

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

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

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．一人称であっても，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．

● **後ブロックテーブル**

本人氏名と判断するパターン

後パターン１）断定表現

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

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くり

がお送りしました**／が送る**

※ ** …[～／．／（／　／っ／ッ／！／＊／☆／♪]のいずれかの記号

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

さん／さま／様／くん／君／ちゃん／先輩／先生／教授／嬢／氏

後パターン６）パターン伝聞の表現

*そうだ／*そうだね／*そうだよ／*らしい／*らしいね／*らしいよ／*のよう／*のようね／*のようだよ　　※ * …ワイルドカード

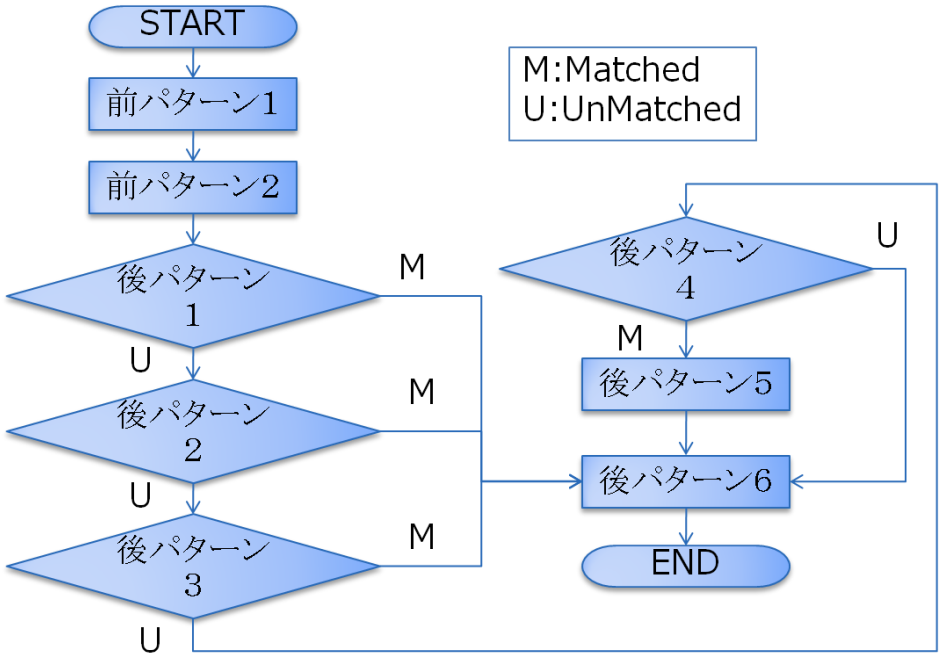

図２　マッチングチェックのフローチャート

4.2.　マッチングチェック

それぞれのブロック内に，パターンテーブルが含まれるかをチェックする，マッチングチェックするパターンテーブルの順番を示したフローチャートを，図２に示す．マッチングチェックは単純なフロー設計になっていない．例えば，後ブロック１）～３）のマッチングチェックを行い，マッチングした場合は，後ブロック４），５）のマッチングチェックは行わないなどである．変則的なフローにすることで，効率的に行うように設計されている．

4.3.　NI:2 の動作フロー

形態素解析から結果出力までの NI:2 の動作フローを図３に示す．フローの右側には，それぞれのフローフェーズにおいて利用するライブラリやデータベースなどを示す．形態素解析や人名探索には Sen を利用し，形態素の保存に利用するデータベースは PostgreSQL，マッチングチェックには前述したパターンテーブルである．

Vol.2010-DPS-145 No.26
Vol.2010-GN-77 No.26
Vol.2010-EIP-50 No.26
2010/11/26

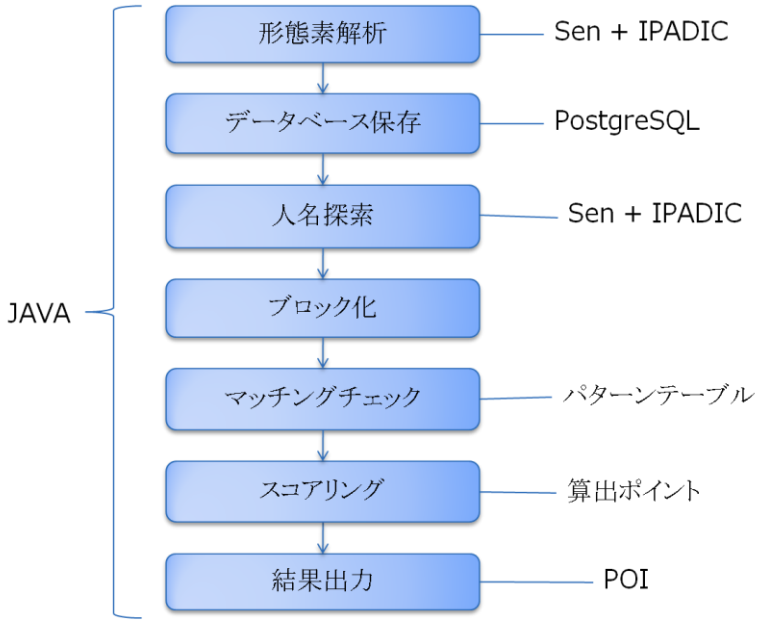

図３　NI:2 の動作フロー

4.4.　日本語形態素解析システム Sen

本手法の分析フェーズの自動化には，日本語形態素解析システム Sen(vr. 1.2.2.1)を利用している．Sen は C++で開発されている MeCab を Java に移植したものである．対象言語の文法の知識（文法のルール）や辞書（品詞等の情報付き単語リスト）を情報源として用い，自然言語で書かれた文を形態素の列に分割し，品詞を判別する．

Sen の辞書には IPADIC(vr.2.6.0)を利用している．品詞や単語の情報が格納されており，その情報を参照することで，「品詞」，「活用形」，「読み」等の情報を得ることができる．

5.　NI:2 の評価

5.1.　評価用サンプル

実際のブログを調査して収集した，評価用のサンプルを用いて，NI:2 を評価する．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種類の判別ブロックの集合である．

表１　マッチング件数とポイント

パターンテーブル	本人氏名　群	他人氏名　群	ポイント
前パターン1	15	1	71
前パターン2	0	1	−100
後パターン1	302	3	95
後パターン2	41	0	100
後パターン3	1	0	100
後パターン4	44	37	−36
後パターン5	31	118	−81
後パターン6	0	0	0
マッチング合計	434	160	
サンプル合計	760	301	

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，本人の氏名と人が判断した判別ブロックを 760 件と，他人の氏名と判断した判別ブロックを 301 件用意した．

5.2.　ポイントの値の決定

表１に，評価サンプルにおけるパターンテーブル別のマッチング件数と，ポイントの定義に沿って算出したポイントの一覧を示す．

5.3.　しきい値の有効範囲

本人の氏名か他人の氏名かを判別する評価値(ポイント)のしきい値範囲を求める．図４は，横軸を評価値，縦軸を本人の氏名であると認識した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．この認識率は，文献 12)と同様の方法で求めている．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．

図４のグラフにおいて，本人氏名群の認識率の高さを優先する範囲 A か，他人氏名群の誤認識率の低さを優先する範囲 B のどちらかしか選択できず，トレードオフを考えなければならない．よって，本人氏名群の認識率が高く，他人氏名群の誤認識率が低い，理想的なしきい値範囲を求められないことがわかる．

Vol.2010-DPS-145 No.26
Vol.2010-GN-77 No.26
Vol.2010-EIP-50 No.26
2010/11/26

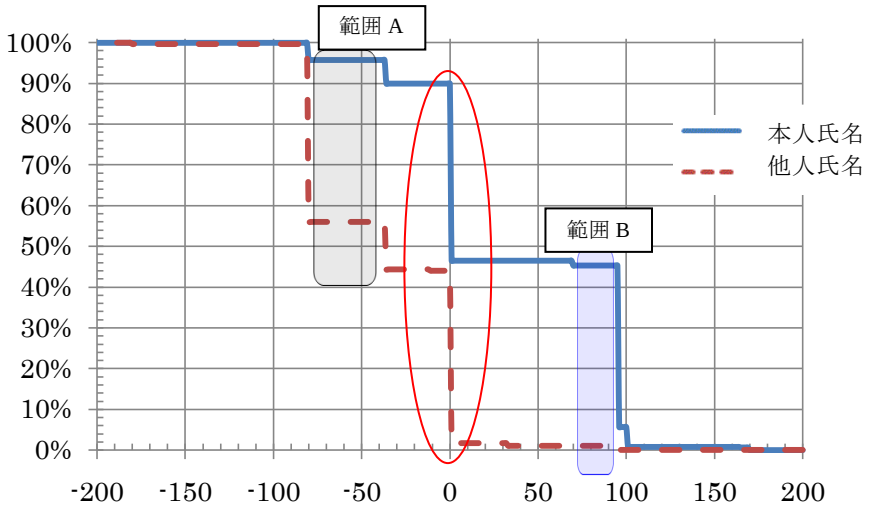

図４　ポイントごとの認識率を示すグラフ

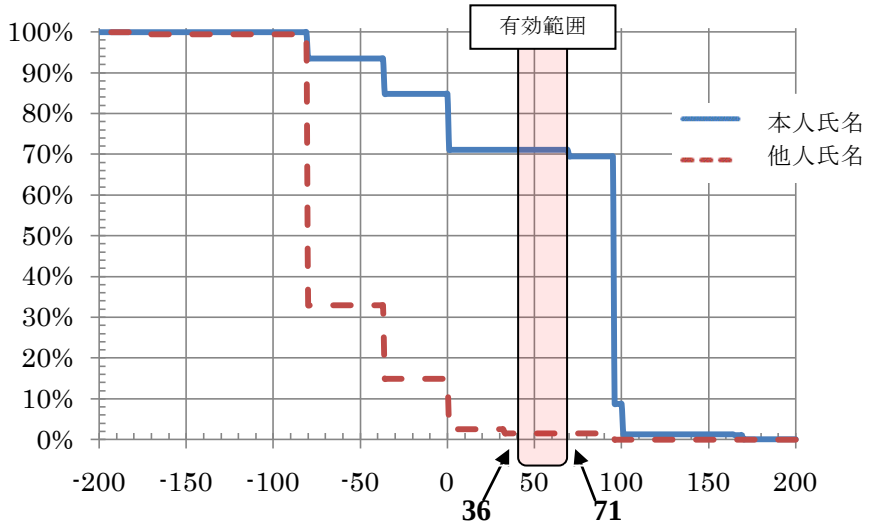

図５　Sen による誤検出を補正した後の図

5.4.　Sen による誤検出の補正

図４からは理想的なしきい値範囲を求められなかった．そこで，赤い丸で囲まれた部分に注目する．ここは Sen が人名を検出できなかった，または誤った部分を人名であると認識した，誤検出による 0 ポイントのサンプルが多く含まれる．

そこで，誤検出になるサンプル群を除外することで，グラフを補正する(ポイントも全て再計算する)．その結果を図５に示す．図５のグラフより，本人氏名の認識率が高く，かつ他人氏名の誤認識率が低い部分である，36 ポイント以上，71 ポイント以下の範囲が，判別手法の有効性がある範囲と考えられる．

6.　考察

6.1.　NI:2 の有効範囲

前章で求めた，しきい値の有効範囲について考察する．5.4.で求めた Sen による誤検出を補正した上での有効範囲は，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み，かつ他人氏名を含む判別ブロックが，誤って本人氏名であると誤って認識する場合を少なく含む範囲である．この範囲において，本人氏名の認識率は 71.1%，他人氏名の誤認識率は 1.5%である．この有効範囲において，本人性を判別できるといえる．

6.2.　パターンテーブルの取りこぼし

しかしその中でも，本人氏名を含む判別ブロックであるにもかかわらず，本人氏名であると判別できなかった誤認識の多くは，本人氏名と認識するいずれのパターンにもマッチングしなかったサンプルである．これは Sen の誤検出分を除いた全体(496 サンプル)の約 28%にあたる 137 サンプルである．これらの主な原因は，パターン定義に存在しないパターンを含むサンプルであるということだけでなく，氏名の前後にスペースや括弧などのイレギュラーな文字列（記号）が含まれていたことである．

6.3.　Sen における形態素解析精度の限界

5.3.の Sen による誤検出を補正する前において，図４の赤い丸で囲まれた部分は，Sen の誤検出による 0 ポイントのサンプルが多く含まれた部分である．これは Sen の形態素解析の誤りであり，形態素解析の精度の限界によるものと考えられる．この誤検出は，本人氏名群において 264 件，他人氏名群において 106 件である．

人間の作業による検証において，この誤検出はないため，自動化における特有の問題である．

Vol.2010-DPS-145 No.26
Vol.2010-GN-77 No.26
Vol.2010-EIP-50 No.26
2010/11/26

6.4. 改善策と課題

以上を改善することを含む今後の課題は，主に３点ある．１点目は，調査件数を増やすことである．２点目は，パターンを更に増やすことである．３点目は基本のパターンに類似しているにもかかわらず，検出されないものに対応する別の新しいアルゴリズムの検討である．

Sen における形態素解析精度の限界の問題は，現段階で根本的には解決できないため，違う形態素解析システムを用いることも検討するべきと考える．

7. 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文献 12)に則り，その実装と評価を行った．

自動化実装に際し，誤検出という形態素解析ツールにおける精度の限界があり，判別に理想的なしきい値範囲を求められなかった．しかし，この誤検出を取り除くことでしきい値範囲を求められ，判別できた．

36 ポイント以上，71 ポイント以下の範囲において，本人氏名の認識率は 71.1%，他人氏名の認識率は 98.5%の判別精度で判別が可能であり，本手法は妥当性があると考える．しかし，十分な判別精度ではないため，まだ実用上改善の余地があると言える．

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)
12) 森加夢偉，金井敦 : ブログにおける記述者本人の氏名判別手法の提案，DPSWS2010，Vol.2010，pp.149-154 (2010)