社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2009－DPS－138 (6)
2009－CSEC－ 44 (6)
2009／3／5

自然言語文からのプライバシー情報検知システム
－自然言語情報の開示制御技術 DCNL の実現（１）－

渡辺夏樹* 池上未希† 片岡春乃‡ 吉浦裕†

＊ 電気通信大学大学院 電気通信学研究科 †電気通信大学 電気通信学部

*†〒182-8585 東京都調布市調布ヶ丘 1-5-1

‡日本電信電話株式会社 NTT 情報流通プラットフォーム研究所

‡〒180-8585 東京都武蔵野市緑町 3-9-11

概要 Blog や SNS からのプライバシーや機密に関わる要注意情報の漏洩が多発し，社会問題になっている．著者らは，その対策として要注意情報を検知し，適切な表現に変換する自然言語情報の開示制御技術 DCNL を開発している．本論文では，その中核となる自然言語文からのプライバシー情報検知システムについて，プライバシー情報に関する知識，自然言語処理，直接検知，想起検知，学習から成るシステム構成を提案する．まず，プライバシー情報に関する知識の表現方式，処理や知識が不完全な場合の検知方法，Web 上の知識を用いて語句の属性を推定し検知に繋げる方法を提案し，プロトタイプ実装と動作例について述べる．また，Web 上の知識を用いて文章中に記述されていないプライバシー情報を検知する方法を提案する．

キーワード 情報漏洩，プライバシー，アクセス制御，人工知能，認知科学

Detecting Private Information from Natural Language Sentences

Watanabe Natsuki* Ikegami Miki† Kataoka Haruno‡ Yoshiura Hiroshi†

* Graduate school of Electro Communications, The University of Electro-Communications

† Faculty school of Electro Communications, The University of Electro-Communications

*† 1-5-1 Chofugaoka, Chofu, Tokyo, 182-8585 Japan

‡ NTT Information Sharing Platform Laboratories, NTT Corporation

‡ 3-9-11,Midori-cho Musashino-Shi,Tokyo 180-8585 Japan

Abstract We are developing Disclosure Control of Natural Language information (DCNL) to prevent disclosure of private information from Web media. This paper describes its central component, i.e, the system for detecting private information that consists of knowledge data about privacy, direct detection, Web-based detection, and learning. The proposed method works with incomplete privacy knowledge by using Web content as its complement.

Keyword Privacy protection, Web security, Disclosure control, Access control, Natural language analysis

1. はじめに

Web 上のコミュニケーションメディアとして Blog や SNS（Social Networking Services）が注目されている．国内最大規模の SNS である mixi[1] では加入者が 1500 万人を超え，コミュニケーション手段として普及している．しかし,その一方でプライバシー情報や機密に関わる要注意情報の漏洩などの事例も発生している．

その対策として，SNS の日記の閲覧制御が挙げられる．しかし，日記ごと，ユーザごとに公開範囲を指定しなければならないためユーザの手間となり，コミュニケーションを阻害してしまう問題がある．

そこで，我々は Web 上のコミュニケーションに適した自然言語情報の開示制御 DCNL(Disclosure Control of Natural Language information)を提案し研究している[1][2][3][4]. これは，ユーザが十分配慮せずにメディアへ投稿した文章を自動的にチェックし，単語単位の開示制御を行う．

また，近年のモバイル Web の普及に伴い，未

[1] mixi は株式会社ミクシィの社名であり，同社が運営しているソーシャル・ネットワーキングサービスの名称でもある．

成年者がBlogやSNS,掲示板等のコンテンツを通し,犯罪に巻き込まれる事例が発生している.この原因は,未成年者はプライバシー情報を管理することへの認識が甘いため,不用意に自分の情報をWeb上に公開してしまうことなどが考えられている[5].

その対策として,未成年者を対象にした有害サイトアクセス制限サービス(フィルタリングサービス)の自動適用に向けた動きがある.フィルタリングサービスでは業界団体や保護者などが作成したブラックリスト,ホワイトリスト,個人設定をもとに,Webサイトへのアクセス許可・禁止を可能としている.しかし,アクセス制限はサイトに対して一律に行われ,また,健全なサイトであると認定されれば,利用の制限はされない.そのため,プライバシー情報の管理は未成年者に委ねられる.この問題にも要注意情報を単語単位で開示制御するDCNLを適用できる.

また,殺人予告や自殺募集,拳銃,麻薬取引などWebが犯罪に用いられるケースもあり,自動的な犯罪検知が検討されている.この問題に対しても犯罪情報を開示制御対象にすることで,DCNLを適用できる可能性がある.

本稿では,DCNLの実現に向け,概念を整理し,システム設計とプロトタイプ実装を行う.

2. DCNL(Disclosure Control of Natural Language information)

自然言語処理により文章中の語句を認識する.語句が要注意情報を表す場合は,削除または言い換えることで要注意情報を開示制御する.

2.1. DCNLの要件

(1) 文章をチェックし,要注意情報を表す語句について言い換えるか,削除する.
(2) 閲覧者とユーザとの関係クラスを特定し,閲覧者によって開示する程度を決定する.
(3) 語句の間接的な意味を推測し,Web検索を通じて要注意情報に到達する可能性も考慮する.
(4) 言い換えにあたって,文章の意味と面白さを維持する.
(5) ユーザへの負担を最小限に抑える.

2.2. 要注意情報の検知

[3]ではDCNLの中心の課題である,自然言語文からの要注意情報の検知を,要注意情報としてプライバシー情報を取り上げ,調査・分析している.それにより,検知したプライバシー情報を2通りに場合分けしている.

- 文章に直接プライバシー情報が記載されている場合
- 文章中の語句を組合せて連想またはWeb検索を行い,その結果プライバシー情報が発見できる場合

また,日記文章に含まれるプライバシー情報を検知するために,「プライバシー情報に関する知識」として次の3種類を定義している.

(1) 【種類語】プライバシー情報の種類を表す語句
住所に関する【種類語】として「住む,暮す,在住」などが挙げられる
(2) 【NGワード】ユーザそれぞれの具体的なプライバシー情報である語句
住所に関する【NGワード】として「初台,調布」などが挙げられる
(3) 【特徴語】プライバシー情報に特徴的な接頭辞や接尾辞などの語句
住所に関する【特徴語】として「市,区,町,村」などが挙げられる

DCNLは文章中に含まれるそれらの組合せによってプライバシー情報を検知する.

3. 要注意情報の検知モデル

攻撃者がWebコンテンツから要注意情報を認識する行動をモデルとし,システム設計とプロトタイプ実装の方法を検討する.

3.1. 攻撃者の行動モデル

攻撃者はWebコンテンツの自然言語文を読んだ時に,内容を理解し,自らの知識と照らし合わせることで,要注意情報を認識する.また,自らの知識だけで足りない場合は,書籍やWeb検索など必要に応じて資料を調べる.Web検索を行う場合は,対象コンテンツの内容と自らの知識を用いてキーワードを考える.そして,検索結果のページを理解して,必要な情報を得る.認識した要注意情報を記憶し,以後Webコンテンツを閲覧した際に利用する.

3.2. システムの検知モデル

要注意情報検知モデルを図1に示す.「要注意情報に関する知識」は攻撃者が持つ事前知識に対応し,「自然言語処理」は攻撃者がWebコンテンツを理解することと対応している.「直接検知」は攻撃者がWebコンテンツから要注意情報を認識することに対応している.

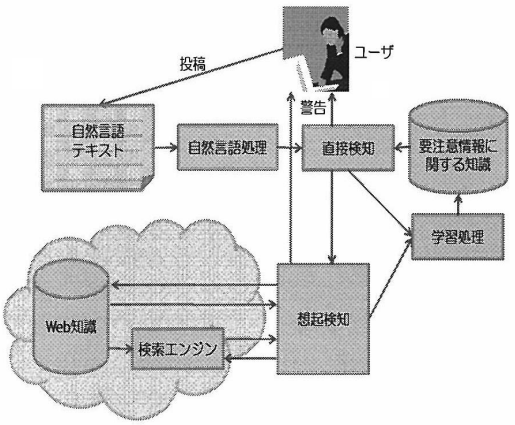

図 1 検知モデル

「想起検知」は攻撃者が自分の知識が不足している場合に外部の知識を利用する場合のうち，Web 検索を行うものをモデル化している．

「学習処理」は攻撃者が認識した要注意情報を記憶し，利用することと対応している．

4. プライバシー情報検知システムの課題

以下では要注意情報としてプライバシー情報を取り扱う．プライバシー情報を検知するためには，図 1 より以下の課題がある．

4.1. 要注意情報に関する知識について

課題 1 知識の表現形式：[3]ではプライバシー情報に関する知識として【種類語】,【NG ワード】,【特徴語】を定義している．実装に当たっては，これら知識の表現形式を検討する必要がある．

課題 2 知識の自動収集：この 3 種類の知識は膨大かつ多種多様である．ユーザの負担を減らすため，これらは自動的に収集される必要がある．

4.2. 直接検知について

課題 3 不完全性への対処：Web コンテンツの文章は口語的なため，従来の自然言語処理が失敗する可能性がある．また，自然言語処理が成功したとしても，計算機では人間のような意味理解はできないため，プライバシー情報を確定的に認識することは困難である．さらに，プライバシー情報に関する知識は膨大かつ多種多様であるため，すべてを網羅することは困難である．これらの問題を受け，システムは断片的な証拠の寄せ集めを総合的に判断する必要がある．

課題 4 基本アルゴリズム：自然言語処理ツールとプライバシー情報に関する知識のデータベースの接続をとる必要がある．その上で，自然言語文からプライバシー情報を検知するアルゴリズムを設計する．

4.3. 想起検知について

課題 5 Web 知識の利用：Web には商用のサイト，個人サイト，集合知サイトなど様々なサイトがある．これらを用いて，プライバシー情報に関する知識が網羅できない事への対処を行う．また，攻撃者と同様にシステムも Web 検索を行い関連情報について調査できる必要がある．

4.4. 学習について

課題 6 複数文章からの情報抽出：SNS や Blog は複数の文章からなる．そのため，複数の文章から抽出された情報を総合的に判断しプライバシー情報として検知しなければならない．

課題 7 検知結果の記憶と利用：SNS や Blog は時系列的な複数の記事から構成される．そのため，検知したプライバシー情報を以後の検知に利用できなければならない．

本論文では，課題 1 知識の表現形式，課題 3 不完全性への対処，課題 4 基本アルゴリズム，課題 5 Web コンテンツの利用を述べる．課題 2 知識の自動収集，課題 6 複数文章からの情報抽出，課題 7 検知結果の記憶と利用については[4]で述べる．

5. 直接検知のアルゴリズム

課題 1 知識の表現形式と課題 3 不完全性への対処を検討し，その上で，課題 4 基本アルゴリズムの設計とプロトタイプの実装を行った．

5.1. 知識の表現形式

プライバシー情報に関する知識の表現形式の要件を以下に示す．

(1) 学習可能にするために，知識の追加が容易に行える必要がある．
(2) 他の機能を追加する可能性があるため，拡張性が必要である．

上記の要件を満たすために，簡単なテーブルで知識を表すことにした．実装結果を表 1 に示す．「初台」はユーザが定義したプライバシー情報とする．

5.2. スコアによる危険度の表現

SNS や Blog の自然言語文はユーザ独自の語句で入力されるため，語句を認識するのは困難である．また，口語体によって形態素解析が失敗する可能性もある．プライバシー情報に関する知識の網羅的な準備が困難なため，文章中に含まれる【種類語】,【NG ワード】,【特徴語】を認識できない可能性もある．これら問題に対応するために，スコア方式を導入する．基本ス

表 1 プライバシーに関する知識の実装

プライバシーカテゴリ	NGワード	特徴語	種類語	基本スコア	・・・
住所			住民	2	
住所			暮す	2	
住所			住む	2	
住所		市		1	
住所		区		1	
住所	初台			10	
・ ・ ・					
職業			会社	2	
職業			勤める	2	

コアはプライバシー情報に関する知識に登録されている語句の確信度を表し，結合スコアは検知されたプライバシー情報の確信度を表す．

基本スコアには以下の関係がある．

【NGワード】>【種類語】>【特徴語】

【NG ワード】はプライバシー情報そのものであるため基本スコアは最も高い．【種類語】は具体的なプライバシー情報ではないため，【NG ワード】よりスコアが低い．しかし，文章中の話題がプライバシー情報に関する可能性が高いことが推測できる．

【特徴語】は接頭辞や接尾辞など，プライバシー情報を構成する可能性がある語句の一部である．そのため，文章中にプライバシー情報となる語句が含まれる可能性を検知できるが，それらがプライバシー情報であるかは判断できず，誤検知に至る可能性が高い．そこで最もスコアが低くなると言える．

以上のような基本スコアをプライバシーカテゴリごとに計算し，結合スコアを得る．スコア方式により，自然言語処理の失敗や，プライバシー情報に関する知識の不足があっても，一定の精度で検知を実現することが可能になる．また，結合スコアは危険度を表すため，結合スコアに応じて自然言語文を色分けすると，ユーザへの警告機能が実現可能となる．

5.3. 基本アルゴリズムの設計と実装

現在使用可能な自然言語処理ツールには形態素解析ツールと構文解析ツールがある．

基本アルゴリズムを図 2 に示す．入力はプライバシー情報に関する知識と自然言語文のうちの 1 文であり，出力は検知されたプライバシー情報である．

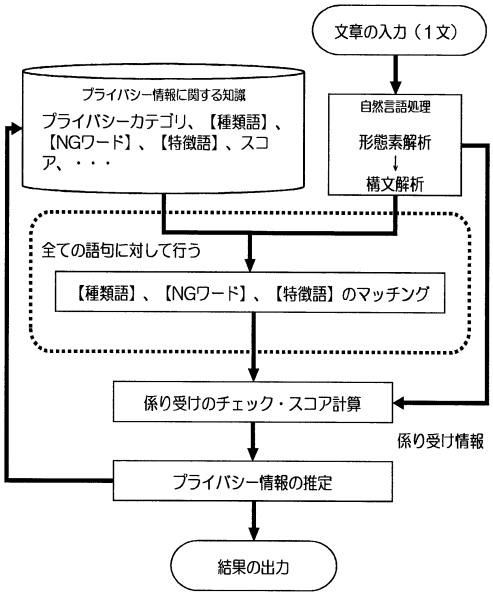

図 2 基本アルゴリズム

まず，形態素解析された文章中の語句と，表 1 で示した知識表現の【種類語】,【NG ワード】,【特徴語】をマッチングする．マッチした場合は，該当する知識表現のテーブルの行を複製し，検出されたプライバシー情報とその位置（行，文節番号）を追記して保持しておく．全てのマッチングが終了したら保持しておいた知識表現の複製をプライバシーカテゴリごとに整理する．

次に，係り受けをチェックする．これは，【特徴語】から得られる具体的なプライバシーやシステムが学習した【NG ワード】は誤っている可能性があるからである．例えば，図 3（A）において【特徴語】から得られる具体的なプライバシー「多摩市」と【種類語】である「住民」は係り受け関係にある．しかし図 3（B）では「多摩市」と【種類語】の「在住」は係り受け関係にない．この様に，係り受けをチェックすることで，検知精度を上げる事が出来る．最後に結合スコアを計算する．

以上の処理を自然言語処理ツールとして juman+knp[6]を使用し，Ruby を用いてプロトタイプ実装を行った．取り扱ったプライバシーカテゴリは住所と職業である．表 1 をプライバシー情報に関する知識として事前に与え，図 3（A）

月曜から土曜まで会社の近くの初台のマンションで暮らす単身赴任者は、実は多摩市の住民です。

(A) 文例1

今日は、京都在住の友人と多摩市内にある美味しいラーメン屋に行ってきました。

(B) 文例2

個人写真撮影の追加撮影会が決定
平成17年11月14日、15日 10時から19時 リサージュ1階
撮影対象は卒研着手者。
アルバム購入は卒業、修了が必須条件ではありません。
ちなみに、H科は昼夜合わせて52人が撮影済みらしいです。

(C) 文例3

図 3 実例日記

からプライバシー情報を検知した処理結果を表2に示す.

【特徴語】である「市」をもとに，プライバシー情報の可能性が高い「多摩市」を取り出すことに成功している.また,【種類語】である「会社」から文章中の話題が職歴に関するプライバシー情報の可能性があることを検知している.

6. Web 知識の利用アルゴリズム

課題5で述べた通り，Web上の知識はプライバシー情報に関する知識が網羅できない事への対処に利用できる．以下に属性推定と関連情報の発見を攻撃者の行動モデルを例に定義し，アルゴリズムを検討する.

6.1. 属性推定

図3（A）に示した文章がある場合，攻撃者は「暮す」という表現や係り受け関係などから「初台」がプライバシー情報だと認識する.しかし，システムでは【NG ワード】として「初台」が登録されていないと「初台」をプライバシー情報として検知できない．そこで，Web上の知識を用いて「初台」の意味的な属性が地名であると推定することで,【NG ワード】として検知することに繋げる.

システムは，図3（A）の【種類語】の「暮す」から文章中の話題がプライバシー情報に関する可能性が高いことを検知できる.しかし,「初台」が【NG ワード】として事前に登録されていない場合は「初台」を具体的なプライバシーとして検知することができない．つまり，文章中の話題がプライバシー情報に関する可能性があるが，具体的なプライバシー情報（【NG ワード】または【特徴語】）を見つけられない場合にWeb

表 2 処理結果

#dialy1_090128

プライバシーカテゴリ	NGワード	特徴語	種類語	プライバシー	結合スコア	日付、行
住所		市	住民	多摩市	6	090128-0
住所	初台		住民	初台	15	090128-0
住所		市	暮す	多摩市	6	090128-0
住所	初台		暮す	初台	15	090128-0
職歴			会社		2	090128-0

上の知識を利用する必要があると分かる.

これは5章の直接検知のアルゴリズムにおいて，文章中に【種類語】のみが含まれている場合にあたる．この場合に，文章中の名詞の意味的な属性が【種類語】と対応するかどうかを判定する.例えば,【種類語】が住所の場合の対応する属性は地名である.

以上のアルゴリズムを，Wikipedia を Web 上の知識として用い，設計とプロトタイプ実装を行った（図4）.Wikipedia は Wiki をベースにした大規模 Web 百科事典であり，幅広い分野の記事をカバーしている．また，記事の冒頭部分が定型的であるという特徴を持つ[7].

Wikipedia 利用の条件は上記の通り，文章中に【種類語】のみが含まれる場合である．その際の検索キーワードは文章に含まれる種類語を除いた全ての名詞である．これらのキーワードの1つずつを DCNL から WikipediaAPI を起動して検索する．検索結果から属性情報を取り出すためには，Wikipedia の定型性を利用する.例えば，地名に関する記事の一番最初の一文は「～地名。」「～地域。」などで終わることが多く，最後の語句から地名であることが判断できる．検索結果にこのようなパターンが現れた場合に，検

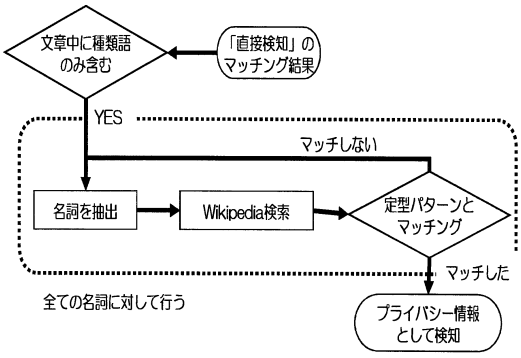

図 4 属性推定のアルゴリズム

索キーワードの属性が地名であると推測することができる.

6.2. 関連情報の発見

図3(C)に示した文章がある場合,攻撃者は「リサージュ1階」をキーワードにWeb検索を行う.そして,リサージュが電気通信大学内にある建物であることを発見し,筆者が電気通信大学の学生であると推測する.このように,Web検索を用いることで文章中に直接書かれていないプライバシー情報を発見できる.

本論文では,Web検索を検知システムが行うためのアルゴリズムを検討する.そのために,攻撃者の行動モデルをより詳しく分析する.

最初に,攻撃者は「卒研」から筆者が大学生であると推測する.さらに,「リサージュ1階」が特徴的な建物を表すことを認識する.次に,「リサージュ1階」をキーワードにWeb検索をする.最後に,Web検索結果のページから筆者が電気通信大学の学生であると推測する.

つまり,攻撃者はまず,(1)検索しようと考え,(2)検索のキーワードを決定する.そして(3)検索結果から情報抽出を行うことで関連情報を発見する.

以上の3点に着目してアルゴリズムを検討した結果を以下に示す.

(1) 検索のタイミング.「卒研」は【種類語】として検知できる.しかし,「リサージュ1階」が特徴的な語句であることを判断する事は計算機では困難である.そのため,「卒研」などの【種類語】がある場合とする.

(2) 検索のキーワード.「リサージュ1階」が特徴的な語句であることを判断することが困難なため,検索キーワードとして特定できない.そのため,文章中にある名詞を組み合わせることで検索キーワードを作成する.

(3) 検索結果からの抽出方法.計算機でWebサイトの内容を理解するのは困難なため,ドメインに着目する.「リサージュ1階」をWeb検索した結果には,電気通信大学のドメインがトップ3を占めている.

7. まとめと今後の課題

自然言語の開示制御DCNLの実現に向け,攻撃者の行動モデルに基いた,プライバシー情報検知システムのシステム構成を提案した.まず,攻撃者の行動モデルを確立し,そのモデルに基づいてプライバシー情報検知の課題を以下の通り見出した.課題1プライバシー情報に関する知識の表現形式,課題2知識の効率的な準備,課題3不完全な処理や知識からの情報抽出,課題4基本アルゴリズム,課題5Webコンテンツの利用,課題6複数文章からの情報抽出,課題7検知結果の記憶と利用.

それらのうち課題1知識の表現形式,課題3不完全性への対処について,拡張や学習に適した知識表現および断片的な検知結果の組合せからプライバシー情報の漏洩度合いを数値化する方法を提案した.また,それらに基づいて課題4基本アルゴリズムを設計し,プロトタイプ実装を行った.

さらに,課題5Web知識の利用について,文章中の語句の属性推定の方法,および文章中に表れないプライバシー情報の検知方法を提案した.今後の課題は以下の通りである.

(1) 提案システムの評価

(2) 他のアプリケーションへの適用

参考文献

[1] 片岡春乃, 他; 意味と面白さを維持する自然言語情報の開示制御技術の提案—SNSのプライバシー保護への試適用—, 情報処理学会 第38回コンピュータセキュリティ研究会 (2007)

[2] H. Kataoka, et al.; Disclosure Control of Natural Language Information to Enable Secure and Enjoyable Communication over the Internet, 15th International Workshop on Security Protocols (2007)

[3] 片岡春乃, 他;自然言語情報の開示制御技術DCNLの実現に向けて—プライバシー情報検知手法—,情報処理学会 第40回コンピュータセキュリティ研究会 (2008)

[4] 水谷桂子 他; 自然言語文からのプライバシー情報検知システム—自然言語情報の開示制御技術 DCNL の実現(1)—,情報処理学会 第44回コンピュータセキュリティ研究会 (2009)

[5] 子供たちのインターネット利用について考える研究会, "第一期 報告書", 2008

[6] 黒橋禎夫,長尾真,日本語形態素解析システム "JUMANver5.1"

[7] 中山浩太郎 他; 自然言語処理とリンク構造解析を利用したWikipediaからのWebオントロジ自動構築,日本データベース学会誌 Vol.7,No.1