情報処理学会　研究報告
Technical Report

2007－DPS－130　(55)
2007－CSEC－ 36　(55)
2007／3／2

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－

片岡春乃*　内海彰†　広瀬友紀‡　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

‡東京大学大学院　総合文化研究科

‡〒153-8902 東京都目黒区駒場 3-8-1

概要 SNS 等を対象にした自然言語情報の開示制御 DCNL(Disclosure Control of Natural Language information) を提案し、実現に向けた方針を示す。近年 Web 上のコミュニケーションメディアが注目される一方、プライバシー侵害等の問題も発生している。DCNL は、ユーザが投稿した文章からプライバシー情報を洩らす可能性のあるセンシティブフレーズを検知し、安全な表現に言い換える。センシティブフレーズは、プライバシー情報を直接表す場合、想起させる場合、単語の組み合わせによって想起させる場合がある。DCNL は、共起分析と Web 検索からセンシティブフレーズに関する知識を自動的に収集し、これらの知識を用いて閲覧者のクラスごとにユーザの文章を変換する。そのためユーザは、開示制御のルールを定義する、閲覧者のクラスごとに異なる文章を書くなどの負担を被ることなく、快適で安全なコミュニケーションを行える。

キーワード プライバシー保護，アクセス制御，自然言語処理，Web セキュリティ，認知科学

Disclosure Control of Natural Language Information to Maintaining Meaning and Interestingness

Kataoka Haruno*　Utsumi Akira†　Hirose Yuki‡　Yoshiura Hiroshi†

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

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

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

‡ Graduate School of Arts and Sciences, The University of Tokyo

‡ 3-8-1, Komaba, Meguro-ku, Tokyo, 153-8902 Japan

Abstract Disclosure control of natural language information (DCNL) is application to SNS (Social Networking Services) privacy protection. Before sentences in the communications are disclosed, they are checked by DCNL and any phrases that could reveal sensitive information are transformed or omitted so that they are no longer revealing. DCNL checks not only phrases that directly represent sensitive information but also those that indirectly suggest it. Combinations of phrases are also checked. DCNL automatically learns the knowledge of sensitive phrases and the suggestive relations between phrases by using co-occurrence analysis and Web retrieval. It transforms the sentence in different ways for different reader classes. The user's burden is therefore minimized, i.e., they do not need to define many disclosure control rules or write different sentences for different reader classes.

Keyword Privacy protection, Access control, Natural language analysis, Web security, Cognitive science

1　はじめに

近年、Web上のコミュニケーションメディアとしてBlogやSNS(Social Networking Servises)が注目されている。これらのメディアは、ヒューマンコミュニケーションを活性化する一方、プライバシー情報の漏洩や誹謗中傷などの不適切な表現が問題になっている。その解決策として、文章を投稿する際、読者を想定して内容や表現に十分配慮する、読者に応じてメディアを使い分ける、等が挙げられる。しかし、それでは手

間がかかる上、複数の相手と1つの話題を共有しコミュニケーションを図るというメディア特有の面白さが失われてしまう。そこで、ユーザが十分配慮せず投稿した文章を自動的にチェックし必要な開示制御を行う技術が必要である。

情報開示を制御する方法として、従来からアクセス制御[1]が研究されている。アクセス制御では、公共団体や事業者が持つ個人情報や機密情報が扱われていた。これらの情報はテーブル形式などで整理され、比較的種類が限定されていた。そこで、事前にアクセス制御ルールを定義し利用していた。

しかし、ヒューマンコミュニケーションでは、将来話題になる内容やそこで使われる言葉を事前に予測することが難しく、加えて、どの語句、文章からどのような情報が洩れるのか、状況や相手によってそれぞれ異なる。そのため、従来のアクセス制御を用いた場合、事前に定義されていない言葉が頻出すると予想される。それに対して一律にアクセス許可すれば、制御の意味が無くなり、一律に不許可とすると言葉が失われ、共有したい話題を相手に伝えることが出来なくなる。結果、コミュニケーションが断絶され、メディアの面白さが失われてしまう恐れがある。

ヒューマンコミュニケーションに必要な開示制御は、事前にアクセス制御ルールを網羅的に定義しなくても、ある程度の信頼性を保ち、多様な自然言語文に対処できる技術である。近年、自然言語処理[2,3]、共起分析[4]、Web[5]等の関連技術が成熟してきた。本稿ではこれらの技術を利用することでWeb上のコミュニケーションを対象にした自然言語情報の開示制御DCNL(Disclosure Control of Natural Language information) を提案する。

2 例文の分析とシステム要件

2.1 データソース

ここではSNSの例としてmixi[6]を用いて分析する。mixiには、ユーザ毎にユーザページがあり、ページの閲覧者を、自分、友人、友人の友人、全体の4つのクラスに分けている。例えば、ユーザのプロフィールページにある氏名や所属のような基本的な個人情報は、それぞれどのクラスの閲覧者までアクセス可能か設定できる。しかし、投稿される日々の日記について、事前に制御ルールを定義することはできない[1]。以下、実際にSNSに投稿された日記を使用し分析する。

2.2 プライバシーに関わる表現

次の文章はmixiユーザである電気通信大学の女子学生「さやか」が投稿した日記からの抜粋である。

(文1)
「来週の就職説明会は西6号館でやるらしい。」
(文2)
「昨日、調布駅で聡子に会った。やっぱり卒業研究が大変みたい。」
(文3)
「恭輔とお台場に『鉄コン筋クリート[2]』を観に行った。」

文1について分析する。さやかは自身の所属情報{電気通信大学}、{人間コミュニケーション学科}等を公表していないので、日記文章中にも大学名を明記していない。「西6号館」は一見すると問題の無い言葉である。それは、西6号館と呼ばれる建物が多数存在し、そこから大学名が推測されると考えにくいからである。しかしこの言葉に対してGoogle検索を行うと検索結果上位10件のうち5件、電気通信大学に関わるサイトが占める。このため、所属が明らかになる可能性が高い。この問題は、制御ルールを事前に定義することで避けられる。しかし、プライバシー情報を洩らす可能性がある全ての語句を予測するのは難しい上、ユーザの手間が増えれば、SNSを使用する楽しみが損なわれる。

文2の「調布」と「卒業研究」の組み合わせは、調布で卒業研究を行う唯一の大学、電気通信大学を想起させる。つまり、それぞれの単語は比較的安全でも、組み合わせることで問題表現になる場合がある。全ての組み合わせに対して事前に制御ルールを決めることは更に困難である。

文3からは彼女の人間関係が明らかになる可能性がある。「お台場」は、デートスポットとして有名である。そのため、さやかと「恭輔」がデートに行くような親密な関係にあることを想起させる可能性がある。これは、語句が間接的

[1]日記全体に対してのアクセス制御は設定できる
[2]スタジオ4℃制作のアニメーション映画

に表す事象まで考える必要があることを示している。

これらの例文から、開示制御ルールを事前に定義することが非常に困難であるといえる。

2.3 求められる言い換え

さやか自身には 2.2 の原文をそのまま開示しても問題ない。しかし、プライバシー情報が洩れる可能性があるため、自身以外には以下のように言い換えた文章を開示するのが望ましい。

(文 1':自分、友人以外へ)
「来週の就職説明会は学科の建物でやるらしい。」
(文2':自分、友人以外へ)
「昨日、駅で聡子に会った。やっぱり卒業研究が大変みたい。」
(文3':友人へ)
「友達とお台場で『鉄コン筋クリート』を観た。」
(文3'':自分、友人以外へ)
「友達と『鉄コン筋クリート』を観た。」

文 1 について、さやかの友人は既に彼女の所属を知っているため、原文を開示しても問題ない。文 1'は、自分、友人以外への文章で、問題のある語句の表現を換えている。これは、問題のある語句を一旦特定すれば困難ではない。文 2'は、問題となる 2 つの語句の組み合わせを特定し、どちらを削除すべきか決めるため複雑である。文 3 では、「恭輔」との親密な関係を明かす可能性があるので友人にも公開できない(文 3')。さらに、自分と友人以外には、特別な関係の存在自体も隠したいので、「お台場」も伏せる必要がある(文 3'')。しかし、単純にこれらの語句を削除すると、文章の意味や面白さが損なわれる可能性がある。そのため、「恭輔」を「友達」に言い換えることで、意味と面白さの維持を目指す。

以上より、閲覧者のクラスによって原文をそのまま開示するか、どのように言い換えるかを判定する必要があるといえる。この解決策として、閲覧者のクラス毎に異なる文章を書く方法がある。しかしそれでは、ユーザの手間がかかる上、複数の相手と 1 つの話題を共有するという SNS の面白さを削ぐため不適切である。

2.4 DCNLのシステム要件

以上を踏まえ、DCNL のシステム要件を見出した。

(1) コミュニケーションにおける文章をチェックし、プライバシー情報を洩らす恐れのある全ての語句について洩れないように言い換えるか、削除する。
(2) 言葉の意味によって、どんな情報が洩れる可能性があるかを推測して言い換える。また、間接的な意味や、言葉の組み合わせから生じる意味も推測する必要がある。
(3) 言い換えにあたって、文章の意味と面白さを維持する必要がある。
(4) ユーザへの負担を最小限に抑える。例えば、それぞれのクラスに対して言い換えルールを詳細に定める、異なる文章を書くといった負担をかけない。そのため、言い換えルールは出来るだけ自動的に学習する必要がある。

3 DCNL の実現に向けて

3.1 既存の技術

DCNLの実現に向け以下の技術を使用する。

(1) 自然言語解析

自然言語解析は形態素解析、構文解析、意味解析、文脈処理[2]から成る。形態素解析は文の単語を特定し、構文解析は単語間の文法的な関係を特定する。これらの技術は確立されており、様々なソフトウェアツールが利用できる[3]。大規模な電子辞書も提供されている。また、固有名詞を検出するツールなども使用できる。意味解析は文およびその語句の意味を解析する。文脈処理は、文脈によって意味がどのように変化するかを解析し、テキスト全体としての意味を推定する。意味解析と文脈処理は技術的に確立されておらず、アプリケーションに依存した知識も必要であるため、実現が難しい[7]。そこで、意味解析や文脈処理を直接実行せずに、共起分析等の、より実行しやすい技術を用いて同等の効果を得ることが望ましい。

(2) 共起分析とWeb検索

共起分析は2つの単語が同時に出現する頻度を統計的に測定[4]し、どの程度密接に関連付けられるかを分析する。Google等のWeb検索はよく知られた技術[5]である。これらの技術は現在も研究されているが、基本技術が確立しており

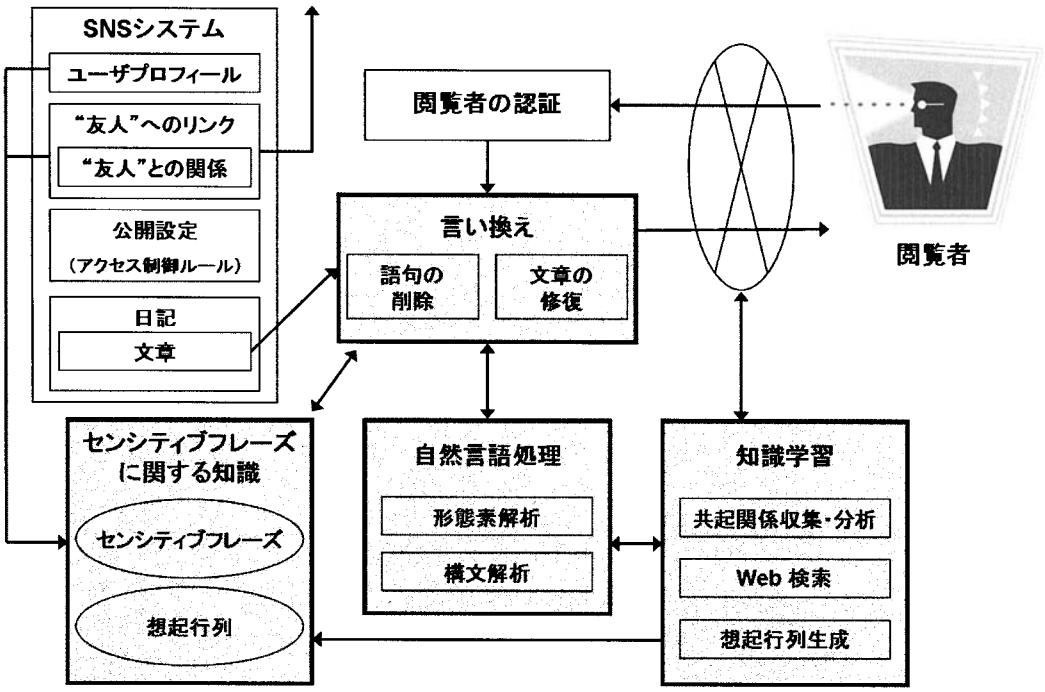

図1 DCNLのシステム構成案

ソフトウェアツールの利用も可能である。

3.2 DCNL のシステム構成

DCNLのシステム構成案を図1に示す。DCNLは4つのコンポーネントから構成される。SNSの日記にアクセスするとき、閲覧者の認証が行われクラスが特定される。DCNLはクラスに基づき原文を安全に開示できるように言い換えて、閲覧者に送信する。言い換え処理を以下の通りに行う。

始めに、「自然言語処理」が、原文の語句とフレーズを認識する。次に、「センシティブフレーズに関する知識」を用いて、フレーズがセンシティブ情報(プライバシー情報などの要注意情報)を直接表すか、間接的に想起させるか、あるいは無関係かの判断を下す。その結果、センシティブ情報を洩らす可能性のあるフレーズを削除する。最後に、削除によって生じた文法的な誤りを修復し、意味や面白さの維持も出来るように最終的な文章を生成する。

センシティブフレーズに関する知識は、センシティブフレーズと想起行列から構成される。想起行列は、共起分析とWeb検索の結果に基づいて生成される。想起行列が自動的に学習されることが望ましい。

3.3 センシティブフレーズと想起行列

はじめに、「フレーズ」の拡張した概念を定義する。ここでフレーズとは、個数1以上の単語の集合である。言語学の定義において、フレーズは文法的な構造をもつ単語の続きとされているが、本稿の定義はそれとは異なる。本稿の定義では、例えば文2の「調布」「駅」「卒業」「研究」の単語の組み合わせもフレーズとする。したがって、フレーズはDCNLの全ての言い換えの単位になる。

センシティブフレーズとは、センシティブ情報を洩らす恐れがあるためユーザが伏せようとするフレーズである。図1のセンシティブフレーズは投稿前に用意され常に更新されるものである。センシティブフレーズの初期集合は、ユーザのプロフィールに含まれるフレーズとする。SNSでは参加の際、基本的なプロフィールの入力が必要である。また、”友人” 関係のあるユーザ同士のページは相互にリンクされ、”友人”ユーザとの関係の説明（例えば、同級生、男友達等)をすることもできる。さらに、共通の所属や趣味をもつユーザ同士は”コミュニティ”と呼ばれるグループに参加している。これらを利用することで、ユーザ自身のプロフィールからは得

られないセンシティブフレーズを収集することもできる。

想起行列は、フレーズがどのくらいの強さでセンシティブフレーズを想起させるかを表す。フレーズの概念を拡張したため、複数の単語の組み合わせによる想起もフレーズとセンシティブフレーズの想起関係として表現される。図 2 に示すように、想起行列の各行はフレーズ、各列はセンシティブフレーズである。行数はシステムで扱うフレーズの数 N、列数はユーザに関する全てのセンシティブフレーズの数 M である。想起行列の要素 S(i,j)は i 行のフレーズが j 列のセンシティブフレーズを想起させる強さを表す。

想起行列は、共起分析とWeb検索の結果に基づいて自動的に生成される。文2では、「調布」と「電気通信大学（電通大）」に強い共起関係があるため、「調布」はセンシティブフレーズである「電通大」を想起させる。想起の強さは共起の度合いである。共起分析には過去のSNSに投稿された日記と既存のWebテキストを使用する。文1では、「西6号館」のWeb検索結果上位10位のうち電通大のサイトが5つ含まれるので電通大を想起すると認識される。この場合、想起の強さは検索結果中のセンシティブフレーズを含むサイトの数とランクを基に計算される。

3.4 アルゴリズム

(1) 想起行列の生成

共起分析とWeb検索を使用する。共起分析から、i行のフレーズとj列のセンシティブフレーズとの正規化された共起度C(i,j)を得る(1<=i<=Nかつ1<=j<=M)。Web検索からはi行のフレーズからj列のセンシティブフレーズへの正規化された到達度R(i,j)を得る。想起行列の要素S(i,j)は、C(i,j)とR(i,j)の大きい方の値とする。想起行列は、例えばユーザが日記を投稿した直後に生成される。文1の想起行列を図3に示す。

センシティブフレーズ / フレーズ	sensitive phrase 1	sensitive phrase 2	....	sensitive phrase j	....	sensitive phrase M
phrase 1	S(1, 1)	S(1, 2)		S(1, j)		S(1, M)
phrase 2	S(2, 1)	S(2, 2)		S(2, j)		S(2, M)
⋮						
phrase i	S(i, 1)	S(i, 2)		S(i, j)		S(i, M)
⋮						
phrase N	S(N, 1)	S(N, 2)		S(N, j)		S(N, M)

図2 想起行列の構造

(2) 語句の削除

1<=i<=Nかつ1<=j<=Mの全てのS(i,j)が閾値T以下の場合、文章中の全てのフレーズがセンシティブフレーズを想起させないので単語を削除する必要はない。Tよりも大きいS(i,j)が存在する場合、削除後のSにおいて全ての要素がTよりも小さくなるように単語を削除する。なお、もし単語 W が削除されるならば、W を含む全てのフレーズに対応する行が想起行列から削除される。削除の方針としては、例えば、削除する単語の個数の最小化が考えられる。

(3) 文章の修復

削除後の文章は、文法的な不備があったり、文章としての面白さを失っている可能性がある。文法的な不備は構文解析技術で特定し、自然言語生成技術を用いて修復できる。しかし、DCNLではそれに加えて文章の面白さを修復する技術が必要である。

4 シミュレーション

提案方式の手順を、例文を使用して説明する。

(文1)
「来週の就職説明会は西6号館でやるらしい。」
(文 1':自分、友人以外へ)
「来週の就職説明会は学科の建物でやるらしい。」

まず{電気通信大学}や{恭輔}を含むセンシティブフレーズが洗い出され、想起行列が生成される。語句を削除するアルゴリズムにより、閾値 T を越えた｛西 6 号館｝から彼女の所属である{電気通信大学}が想起されることが分かる。そのため、{西 6 号館}を削除し、文章の修復処理が上位概念の「建物」を代入する。しかしこ

センシティブフレーズ / フレーズ	電気通信大学	恭輔	...
来週	1.17E-01	2.56E-04	
就職	9.84E-02	5.69E-05	
説明会	3.10E-01	3.88E-04	
西6号館	5.30E-01	0	
...			
来週 就職	1.02E-01	6.64E-05	
来週 説明会	9.03E-02	6.17E-05	
...			
来週 就職 説明会	1.60E-01	1.06E-04	
来週 就職 西6号館	5.71E-01	0	
...			
来週 就職 説明会 西6号館 やる らしい	0	0	

図3 文1の想起行列

のままでは文章の面白さが欠けるので、{西 6 号館}と所属の{人間コミュニケーション学科}の関係から「学科の」を加える。

(文 2)
「昨日、調布駅で聡子に会った。やっぱり卒業研究が大変みたい。」
(文 2':自分、友人以外へ)
「昨日、駅で聡子に会った。やっぱり卒業研究が大変みたい。」

最初の例と同様に処理を行う。フレーズの定義により、語句の全ての組み合わせが想起行列の行にある。その結果、「調布」「駅」「卒業」「研究」の組み合わせが問題表現として特定される。このうち、「調布」を削除したときに想起行列の全ての要素が閾値を下回るので、「調布」を削除する。文章の修復アルゴリズムにより、文法的な不備が無いことを確認する。

(文3)
「恭輔とお台場で『鉄コン筋クリート』を観た。」
(文3':友人へ)
「友達とお台場で『鉄コン筋クリート』を観た。」
(文3":自分、友人以外へ)
「友達と『鉄コン筋クリート』を観た。」

ここでは、異なる2つの閾値を使用する。文3"により大きな閾値を使うことで、"友人"により多くの情報を公開する。文1と同様に処理を行うが、単純に「恭輔」を削除しては文章の意味と面白さが欠けるので、彼との関係である{友達}に言い換える。しかし、ユーザや友人のプロフィールから「お台場」をセンシティブフレーズと判定するのは難しい。そのため、この種のセンシティブフレーズを検知する新たな方法が必要である。

5 まとめと今後の課題

自然言語情報の開示制御DCNLを提案し、実現に向けた方針を示した。DCNLの実現性はいかにユーザへ負担をかけずに文章の問題表現を検知できるかにかかる。提案方式は、意味解析、文脈処理の代わりに共起分析とWeb検索を使用することによって、問題表現を検知する。つまり、本方式はコンピュータ処理が困難な意味を扱う代わりに統計処理を用いるので自動処理が可能である。今後の課題は以下の通りである。

(1) 理論的基礎の確立

例えば情報理論に基づいて「共起度」と「到達度」の意味を明確にし、理論的に根拠のある想起行列生成方法を確立する。

(2) 機能、ユーザビリティ、保守性の評価

提案したシステムを実装し、例文から見出した要件を満たすか評価する。

(3) 面白さの維持

文章の面白さを維持するための方法を考案する。本稿では情報漏洩の防止を中心に検討したが、今後文章の面白さの維持に関して具体的な方針を検討する。

(4) 他のアプリケーションへの適用

BlogやSNSだけではなく、異なったクラスの人々に自然言語文を開示する他のアプリケーションへの適用を検討する。

参考文献

[1] Ross Anderson, "Security Engineering", John Wesley & Sons, 2001.

[2] 長尾真(編),『自然言語処理』,岩波書店, 1996

[3] 松本裕治, 高岡一馬, 浅原正幸, 工藤拓, "茶筌と南瓜による日本語解析 --構文情報を用いた文の役割分類", 人工知能学会誌, Vol.19, No.3, pp.334-339 ,2004

[4] Julie Weeds and David Weir, "Co-occurrence retrieval: A Flexible Framework for Lexical Distributional Similarity," Computational Linguistics, Vol.31, No.4, pp.439-475, 2005

[5] Pierre Baldi, Paolo Frasconi and Padhraic Smyth, "Modeling the Internet and the Web: Probabilistic Methods and Algorithms," John Wiley & Sons, 2003

[6] ソーシャルネットワーキングサービス "mixi," , http://mixi.jp/

[7] 吉浦裕,『不完全表現からの意味の復元モデルを用いた理解方式の研究』, 東京大学博士論文, 1997