Vol.2014-ICS-177 No.12
2014/12/15

ステミングとN-gram共起による プライバシー関連語の抽出精度向上

佐生　明陽[1]　輪島 幸治[2]　小河　誠巳[3]　嶋田　茂[2]　古川　利博[1]

概要：近年，SNS 等が普及することで，盗撮画像や個人情報がコメント付きでインターネット上にアップロードされることによるプライバシー侵害が社会問題となりつつある．そのため，投稿されたデータがプライバシーを侵害しているかどうかを自動判定することができれば，こうした問題を素早く解決したり事前に防ぐことができるようになると期待できる．ところが，抽象的な概念であるプライバシー侵害を自動判定することは非常に困難である．先行研究では，大量の文書データからプライバシー関連データの抽出を行い，それらを解析することでプライバシー侵害の要因解析を行い，プライバシー侵害判定を行っているが，実用化には至っていない．そこで本稿では，先行研究で行われているプライバシー関連データの抽出過程における問題点を明確にし，プライバシー侵害に機微な表現の抽出精度の向上を目的として，ステミング方式の選択と N-gram 想起モデルの次数変化の組み合わせに着目した手法を提案する．また，数値実験で提案法の有効性を示す．

キーワード：セキュリティ, プライバシー，プライバシー侵害，ステミング, N-gram

1. はじめに

1.1 背景

最近，個人情報の漏洩等のプライバシーの侵害が社会問題となっている．Twiiter では，電車内等の公共の場での盗撮画像が，コメント付きでアップロードされている．今後，Google glass に代表される様なウェアラブルカメラが普及し、SNS への投稿が増加するにつれて，第 3 者からの個人情報の漏洩に関する，プライバシー侵害リスクも高まりつつあることが指摘されている [2].

しかし，SNS やインターネット上におけるプライバシー侵害の定義は，その記事が投稿される状況（位置や時間等のシチュエーション）や，ユーザの置かれる社会的位置づけ，投稿先にユーザの範囲（コミュニティー）など，多数のファクターにより影響されるため，一概に決めることはできない.

これらの状況から，先行研究ではプライバシー関連記事の抽出をして，プライバシー侵害要因の分析を行い，その分析結果を教師データとして提示するような学習方式でプライバシー侵害を検知させる方式が提案されている．しかし，未だ実用化には至っておらず，プライバシー侵害要因分析のためのプライバシー関連記事の自動抽出方法の精度向上が求められている．

[1] 東京理科大学
Tokyo University of Science
[2] 産業技術大学院大学
Advanced Institute of Industrial Technology
[3] 前橋工科大学
Maebashi Institute of Technology

1.2 目的と本稿の構成

本稿は，プライバシー侵害リスクが増加する事が危惧されている欧米諸国に向けて，英文での広範な概念でのプライバシー関連情報を習得することを目的としている．そのために，プライバシーに関連する表現，特に単語レベルの機微な表現（これを以降、PSW; Privacy Sensitive Word と略称）の抽出方式 [2] に注目し，抽出率の精度向上させる方式を検討した．また，嶋田らは，プライバシー関連記事の内容について，顕在的プライバシーと潜在的プライバシーに分けて，関連記事を得られるとしている [1] が，本稿では，プライバシーの内容については，触れない．2 章では，先行研究の概要と現状の PSW の問題点を挙げ，その問題点に対し，3 章では，文書集合の検索キーワードとなる ”privacy” にステミング処理を適用させ，語幹を用いる事で，抽象化 を行い，共起方式の次数を変化させることにより PSW の抽出率の精度向上させる方式を提案する．次に，4 章で，提案手法の実験を行い，先行研究，人間の判定した PSW との比較を用いて評価し，そして最後の 5 章

にて，本稿のまとめを行い，今後の展望に触れる．

2. 先行研究

2.1 先行研究の概要

プライバシー侵害要因の解析のために，SNS アーカイブから，PSW を抽出し，プライバシー関連記事を収集している研究として，町田らによる，SNS 投稿時に発生するプライバシー侵害要因に関する研究 [2] がある．ここでは，Twitter アーカイブを対象に，キーワード「privacy」を含む文書と含まない文書の出現単語頻度の適合率の比較により PSW を抽出し，プライバシー関連記事を収集して，投稿記事に含まれる画像認識と時空間情報の解析を行い，プライバシー侵害の要因を挙げ，検知するシステムを提案している．また，嶋田らによるウェアラブルカメラによるプライバシー侵害の要因分析 [3] では，YouTube アーカイブを対象にして PSW を構成し，プライバシー関連記事を収集し，感情の高揚率を用いて，プライバシー侵害要因の分析を行っている．

2.2 PSW:Privacy Sensitive Word

PSW はプライバシー関連記事を抽出するためのキーワード集である．特定のユーザーにより作成された辞書に登録された単語とその類義語の出現が確認が出来た場合に，NG ワードとして，プライバシー侵害と検知するのではなく，多量の SNS アーカイブを対象にした統計的な処理による選定を行っている．そのため，プライバシー関連情報に関する解釈が反映される可能性は高くなっている [2][3].

先行研究では，PSW を以下の様に構成している．(1) プライバシー侵害に抵触する様なキーワードを用いて，文書を検索する．(2) 検索された文書集合を構成する単語と単語出現頻度数を出力して，適合率を算出する．(3) 適合率が高い順にソートし，構成する．

2.3 PSW 精度向上に当たっての問題点

先行研究 [2][3] では，大量の SNS アーカイブを対象にした統計的な処理を基本として，キーワード「privacy」を含む文書と含まない文書の出現単語頻度の適合率の比較により PSW を抽出する方法を 提案している．しかし，いずれも人間の判断による PSW の抽出結果に比べて 50% と低い数値を示している．また，PSW を用いて得た文書集合に対するプライバシー侵害判定数も少なく，必ずしも実用的なレベルにあるとは言えない状況にある．ことなどが考えられる．

その原因として，先行研究での PSW の抽出方法は，共起抽出のための基準ワードを”privacy”といった固定表現を用いて，その類語を含めて PSW を抽出している．これは、”privacy”という単語を確実に含む文書を解析対象とすることにより，PSW の抽出精度を上げる方針をとるものである．

ところがこの方針では，英文を対象とした場合，解析対象の範囲を狭めることになり，有効な PSW の抽出には寄与しないと考えられる．また，単語間の隣接係り受けを考慮しないそれぞれ独立した単語レベルの出現頻度に基づく共起抽出を行っているので，これも抽出精度が上がらない理由となっている．

以上の点を考慮し，本稿で PSW 抽出のための検索キーワードの抽象化を行い，文書集合の選定を行う．次に，unigram，bigram，trigram の手法を用いて，提案手法の精度の評価を行う．

3. 提案手法

3.1 方式の概要

本稿では，単語のステミング処理を用いて活用形や複数系を考慮するとともに，それらの導出される類語の概念を抽象化した単語を用いて，より広範囲な概念から PSW を抽出する方式を提案する．その具体的な PSW の抽出方式の手順は、図 1 に示す処理フローで行う．

(1) 最初にキーワードのステミング処理を行い，そのキーワードに関連する類語や変形語を求める．(2) 次に Wordnet[8] により，得られた類語や変形語間の概念的な抽象化された表現を得る．(3) その抽象化されたキーワードを用いて，SNS アーカイブの各記事に対して N-gram モデルに基づく単語の抽出を行う．(4) 適合率でソートし，PSW を抽出する．

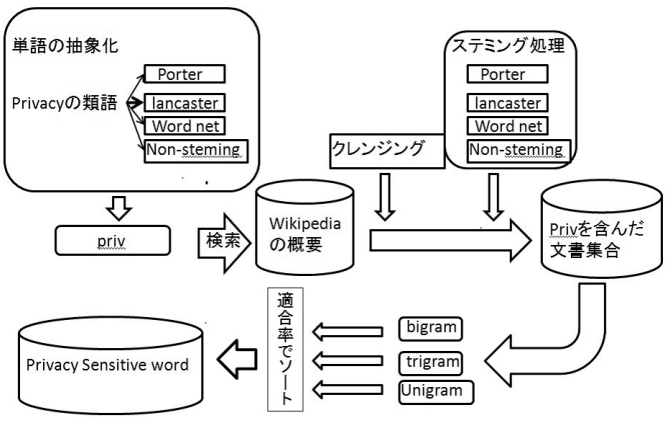

図 1 提案手法

3.2 ステミング処理

語形が変化する単語の語幹でマッチングを行う処理を指す．英語等の言語は，意味的には同じ単語で語形が変化する．そのため，複数の語形を統一する処理が自然言語処理では，求められる．代表的なステミング手法には，Porter algorithm[5]，The Lancaster Stemming Algorithm[6] がある．

3.3 単語の抽象化

単語の複数形や名詞や副詞等の派性語を考慮するために，単語に対して，Porter algorithm，The Lancaster Stemming Algorithm のステミング処理を用いて，単語を短く纏める事を，本稿では，抽象化と表現する．

例えば，「privacy」には，「private」等の意味合いが近い単語がある．これらの単語もプライバシーに関する文書を含んでいる事が想定される．そのため，包括的に文書を検索するために，検索ワードの抽象化を行う．

3.4 提案手法手順

PSW の抽出方法は以下の様に細分化される．

Step1:文書集合構成のための検索ワードの抽象化

抽象化した語を「privacy」という単語から作る．

Step2:文書集合構成のための検索ワードの選定

抽象化した単語を用いて，対象データに含まれる PSW を抽出する．

Step3:文書集合のクレンジングとステミング処理

テキストデータのクレンジングを行う．クレンジングとは，テキストデータの部の解析を進めると記号や顔文字等のノイズが発生し，適切なワードを検出する事が出来ない．その結果，データ解析が失敗する原因にもなりうる．そこで，ノイズの除去を行うために，不要な文字列を除外した．除外対象となる不要な文字列としては，冠詞や記号、顔文字をストップワードとした．また，ステミング処理として，Porter algorithm，The Lancaster Stemming，Algorithm と WordNetLemmatizer とを適用させた．

Step4:PSW の摘出

bigram，trigram を用いて，PSW の抽出を行う．

4. 実験

本稿の提案手法に則って，PSW を構成する．実験の目的は，提案手法と先行研究を評価し，抽出方法の違いによる精度を明らかにする事である．

評価する基準としては，正解データとして，人間が判定した PSW を抽出し，正解データとして，自動抽出手法がどれだけ正解データに近いかをコサイン類似度を用いて判定する．

4.1 実験環境

本研究の実験環境は，OS は，CentOS，メモリは，16GB，プログラミング言語は，python2.6，NLTK，Gensim を用いた．

4.2 対象データ

本研究では，文書集合の wikipedia の abstract の中から，より広範な文書を収集するために，「privacy] に The Lancaster Stemming Algorithm を適用させて，「priv」という単語を検索ワードとした．そして，この単語を含んでいる文章は，「privacy」や「private」，「privately」等の単語を集める事が出来る事から，プライバシーに関連する表現を持つ文章集合と仮定した．

4.3 Wikipedia abstract を用いた PSW 抽出

Step1: 文書集合構成のための検索ワードの抽象化

従来の単一のキーワードを用いた関連検索方式では，”privacy” の単語自体を含む文書だけが対象となり，その類語や同意語を含む文書は対象にならなかった．

そこで，本稿では，広範な文章集合を扱うために、検索キーワードの類語に代表的なステミング処理である Porter algorithm，The Lancaster Stemming，Algorithm，を適用して正規化するとともに，更にそれらの類語に WordNetLemmatizer を用いてより上位の抽象的な概念に相当するワードも対象となるようにした．

表 1 に検索ワード選定のための前処理を行った結果を示す．これらの処理を行わない場合に対して，Porter algorithm，The Lancaster Stemming，Algorithm，WordNetLemmatizer をそれぞれのワードに対して適用させた結果を示している．

表 1 検索ワード選定のためのステミング処理

Non-stemming	Porter	Wordnet	Lancaster
privately	privat	private	priv
private	privat	privately	priv
privacy	privaci	privacy	priv
privateer	privat	privateer	priv
privilege	privileg	privileg	priv
privileged	privileg	privileged	priv
privata	privata	privata	privat
privative	priv	privative	priv
privity	priviti	privity	priv
privatbank	privatbank	privatbank	privatbank

Step2:文書集合構成のための検索ワードの選定

表 1 の結果から，The Lancaster Stemming は，「privacy」を「priv」と 1 番短く抽象化した．

従って，より広範なプライバシーに関連するワードを含む文書集合を対象にするために，The Lancaster Stemming を用いて，単語を抽象化するのが適当と判断できる．抽象化した単語「priv」を含む文書を検索した結果，約 2 万文書が抽出された．

Step3 文書集合のクレンジングとステミング処理

抽出した 2 万文書に対して，クレンジング処理を行う．その結果に対して，Porter algorithm，The Lancaster Stemming，Algorithm，そして WordNetLemmatizer を適用させ，各単語の名詞や副詞，形容詞を原型にした．

Step4: PSW の抽出

文書集合を構成する単語に対して ”priv” の共起語を，N-gram モデルを用いて抽出した．抽出した PSW の詳細は付録に示す．

4.4 人間が判定した PSW

本稿における正解データとして，文章の意味合いから，プライバシーに関連するキーワードを人間の主観を用いて，抽出する．手順としては，wikipedia に対して，”privacy” をキーワードとして検索をして，プライバシーに関連する語が含まれる文書を抽出した．それらの abstract を対象にして，プライバシー侵害に関連すると考えられるワードを，5 名の学生が，文章の意味合いを考えて，全員の主観が一致した単語として抽出した．

その結果を表 2 に示す．表 2 には，人間が判定した PSW を表記している．

表 2 人間が判定した PSW

privacy	commissioner	law	human
right	individual	security	anonymous
information	company	protection	consumer
personal	control	freedom	cryptography
data	invasion	internet	lawyer
act	legal	online	private
privacy	commissioner	law	human
right	individual	security	anonymous
information	company	protection	consumer
personal	control	freedom	cryptography

4.5 提案手法を用いてのプライバシー関連記事の抽出

本節では，提案手法を用いて抽出した PSW を用いて，プライバシー関連記事を検出し，プライバシー関連記事の内容を確認した．

step1:提案手法の手順で抽出した PSW(表 A·1，A·2，A·3，A·4) を用いての SNS 投稿記事検索

google ブログの 2014 年 8 月 20 日以前のブログ，検索エンジン「Google」「Baidu]，SNS 投稿記事を検索する事が出来る「Social Searcher」上で，PSW の適合率の高いトップ 3 のキーワードを用いて，プライバシー関連記事検出を行う．

step2:抽出したプライバシー関連記事の内容判定

検出されたトップ 100 の文書に対して，プライバシー関連のある内容であるかを学生 5 人で，目視にて判定する．判定する基準としては，全員の主観が一致した記事とした．

目視で確認したところ，プライバシー関連記事は，(1) プライバシーの概念についての記事，(2) 国，個人・企業のブログ及び，(3) プライバシーポリシー記事，(4) プライバシー，もしくは情報セキュリティに関連する技術紹介記事，(5) プライバシー侵害記事を抽出する事が出来た．表 3 にプライバシー侵害記事を抽出の結果を示す．表記されている数字は，各ステミング手法と頻度手法を用いた場合の件数である．表 3 から，今回提案した手法はプライバシー侵害記事も抽出する事が出来る事を確認した．

ここで，プライバシー侵害記事があると判定する基準は，記事の画像かコメントのいずれかに対して，第 3 者が本人の了承を得ずに投稿していると判定出来る場合を指す．

表 3 プライバシー関連記事の該当件数

ステミング手法＋ N-gram	Baidu	Google blog
Porter-Unigram	0	3
Porter-bigram	3	11
Porter-trigram	3	5
Lancaster-Unigram	0	0
Lancaster-bigram	0	0
Lancaster-trigram	0	0
Wordnet-Unigram	1	15
Wordnet-bigram	3	14
Wordnet-trigram	4	8

4.6 先行研究の方式との比較

本節では，プライバシーに関連する PSW の自動抽出処理の結果の妥当性を評価するために，人間により抽出された PSW を正解データとして比較する．

そのため，本節の実験では，提案手法を用いて自動抽出した PSW と人間が評価した PSW との評価の一致度をコサイン類似度を用いて定量的にみて，評価する．また，先行研究と提案手法の比較もコサイン類似度を用いて同時に行う．その結果を表 4 に示す．

表 4 の左側はステミング手法と頻度手法（unigram, bigram, trigram）を示している．表の右側は，各ステミング手法のコサイン類似度を示している．

表 4 から，正解データに最も類似している手法は，④の wordnet と bigram を併用した手法だという事が，コサイン類似度の値が先行研究の手順である⑫に比べて，9 倍高い数値である事から分かった．

5. まとめ

本稿では，Wikipedia を対象として，検索キーワードのステミングと抽象化の前処理を行い，N-gram モデルによる関連検索を行う方式により，従来の統計的な関連検索方式に比べて PSW の抽出精度が向上することを示した．

表 2 から，文書集合に対する検索ワードのステミングと抽象化に関しては，The Lancaster Stemming Algorithm を用いると最も抽象化される事が提案手法によって分かった．また抽象化する事により，Wordnet と bigram を用いた PSW は，先行研究よりも人間が判定して抽出した PSW

表 4　先行研究のコサイン類似度

ステミング手法＋ N-gram	コサイン類似度
①Porter-Unigram	0.04
②Porter-bigram	0.42
③Porter-trigram	0.23
④Lancaster-Unigram	0.00
⑤Lancaster-bigram	0.00
⑥Lancaster-trigram	0.00
⑦Wordnet-Unigram	0.05
⑧Wordnet-bigram	0.45
⑨Wordnet-trigram	0.15
⑩bigram	0.4
⑪trigram	0.15
⑫先行研究の手順で抽出した PSW	0.05

とのコサイン類似度が，約 9 倍高い数値を示し，プライバシーに関連する表現を含む記事をより多く抽出することが出来た．この結果は，検索キーワードを抽象化する事で，プライバシー関連語を広く収集することが出来た事と，bigram や trigram を用いた事から，係り受けを考慮し頻度解析を行った結果，人間が判定した PSW により近い PSW を構成出来た事が起因すると考える．

今後の課題としては，twitter 等の他の文書集合を用いての比較実験を行う必要があると考える．また，本研究の正解データとする人間が判定した PSW は、5 名の研究員により評価したものであるが，更に多くの評価者による PSW を採取して，より精度の高い正解データを構成し，提案手法による PSW の抽出結果との類似度を測定する予定である．

6. 謝辞

本研究を進めるにあたり，産業技術大学院大学の嶋田研究室所属の諸氏によるデータの一部を提供して頂いた事に心から感謝申し上げます．

付　　録

A.1 Proof of the First Zonklar Equation

実験 II の結果を以下の表 A·1，A·2，A·3，A·5 に示す．上から，Porter-Unigram，Porter-bigram，Porter-trigram，Lancaster-Unigram，Lancaster-bigram，Lancaster-trigram，Wordnet-Unigram，Wordnet-bigram，Wordnet-trigram，Unigram(先行研究の手順 [1])，bigram，trigram である．それぞれ適合率が高い順にソートをした結果である．

表 **A·1**　提案手法の PSW 集

1	2	3	4	5
privileg	privaci	depriv	privatis	koprivnica
quot	invas	surveil	protect	commission
wep	geoloc	quot	injunct	privaci
privy	kopriv	capriv	privat	depr
millhon	governing_body	snafu	nonsect	quot
lsbf	elektrim	snafu	nonsect	seedbox
privately	privacy	privatization	privileged	deprivation
preserving	protection	surveillance	invasion	protect
wep	geolocation	injunction	privacy	defamation
privately	privacy	privileges	privatization	privileged
preserving	invasion	protection	surveillance	commissioner
wep	geolocation	privacy	defamation	expectation

表 **A·2**　提案手法の PSW 集

6	7	8	9	10
underprivileg	upriv	privada	independentindep	priviti
internet	freedom	differenti	secur	consum
defam	usabl	tort	surveil	invas
kaminen	koprivshtits	graphpad	utsub	privatbrauere
postsecond	paroch	coeduc	unaccredit	fundingtyp
chonkin	telegraaf	karibib	mouawad	quattleba
deprived	privatisation	koprivnica	underprivileged	upriver
commissioner	differential	internet	concern	freedom
wired	tort	authentication	invasive	transparency
deprived	deprivation	privateers	privatisation	koprivnica
protect	concerns	differential	consumer	internet
wired	tort	authentication	transparency	invasive

表 **A·3**　提案手法の PSW 集

11	12	13	14	15
privatbank	privolzhski	kopriva	elektroprivreda	privilegium
onlin	concern	data	equival	advoc
gnu	piraci	violat	protect	commission
privet	grafenwald	millhon	cunimondo	jcpc
schooltype	motorbo	amp	equ	own
laatst	sja	smic	poin	quot
privatised	privatized	privateering	independentindep	privity
consumer	security	advocate	equivalent	information
preserving	surveillance	protection	gnu	invasion
underprivileged	upriver	privatised	privatized	privateering
laws	users	freedom	security	equivalent
preserving	protection	surveillance	invasion	liberties

A.2 Proof of the First Zonklar Equation

表 A-5 で，提案手法と比較をした，先行研究 [2] の手順で抽出した PSW を記す．

表 **A·4**　提案手法の PSW 集

koprivnik	priverno	koprivna	privado	koprivica
inform	user	guard	preserv	financi
authent	transpar	freedom	supermarket	internet
penab	isesak	psps	inokashir	privileg
healthc	corsair	offlin	montessor	coed
koprivnik	blueb	depo	vpns	dhafr
privada	privatbank	privolzhsky	privilegium	kopriva
online	electronic	data	user	financial
analytics	piracy	expectation	violation	protect
privates	privity	independentindep	privada	privatbank
information	online	electronic	issues	data
gnu	analytics	piracy	commissioner	protect
preserving	protection	surveillance	invasion	liberties

表 **A·5**　先行研究 [2] の手順で抽出した PSW

privacy	privately	privileged
koprivnica	underprivileged	upriver
privatized	privateering	independentindependent
privata	privatbank	privolzhsky
elektroprivreda	privilegium	koprivnik
deprived	privatised	privity
kopriva	deprivin	

参考文献

[1] 嶋田茂, ユーザ参加型景観サービスに含まれる潜在的プライバシーの保護策, 信学技報, vol. 111, no. 287, EMM, pp. 77-82, 2011.15.

[2] 町田史門, 小山貴之, 宋洋, 高田さとみ, 嶋田茂, SNS 写真投稿に起因するプライバシー侵害の類型化とその保護策, 信学技報 EMM, 2012.09.27.

[3] 奈良育英, 高田さとみ, 高田美樹, 他, SNS の感情分析をトリガーにしたウェアラブルカメラによるプライバシー侵害の要因分析, 信学技法 HCS, 2013.08.16.

[4] 片岡春乃，渡辺夏樹，水谷桂子，吉浦裕, "自然言語情報の開示制御技術 DCNL の実現に向けて" , 情報処理学会研究報告. CSEC, [コンピュータセキュリティ] 2009.

[5] H. Kopka and P. W. Daly, "Porter Stemming Algorithm ",Daniel Waegel CISC889 2011.

[6] H. Kopka and P. W. Daly, "The (un)official Lovins stemmer page" ,published in "Mechanical translation and computational linguistics", 11:22-31, 1968.

[7] Alan F.Westin, "Social and Political Dimensions of Privacy" ,Journal of Social Issues，Vol.59,no.2,2003.

[8] "WordNet A lexical database for English" ,http://wordnet.princeton.edu/，2014.