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

2008－DPS－136 (12)
2008－GN－ 69 (12)
2008－EIP－ 41 (12)
2008／9／26

ブログにおける個人情報漏えいモデル

針谷 友彰†， 佐藤 和紀†， 安井 良介†， 金井 敦†
†法政大学大学院 工学研究科

あらまし
近年，個人情報保護に関する意識が高まり，情報漏えい防止技術開発や企業の努力などにより対応がなされて来ている．しかし，日記感覚で情報を発信するブログなどは，どれだけ個人情報を公開しているかを十分認識することが自分自身でもできない場合がある．このように，個人がインターネット上で情報を発信する場合，情報漏えいの意識が薄く無意識のうちに個人情報が漏れているケースがある．そこで，個人情報の漏えい状況を定量化し，客観的に把握することができれば，ネットリテラシー不足による漏えいや不注意による個人情報漏えいの防止を支援することができる．本稿では，定量化するために個人情報漏えい評価の基本モデルを提案し，個人情報漏えい度の計算法を示す．また，本モデルに従って実際のブログにおける個人情報漏えい度を計算し，計算結果と人間の感覚による漏えい度合いを比較する．

A Model of Personal Information Disclosure for Blog

Tomoaki HARIGAYA†, Kazuki SATO†, Ryosuke YASUI†, Atsushi KANAI†

†Graduate School of Engineering, Hosei University

Abstract
In recent years, many companies and parsons become paying attention to protection of personal information disclosure, and prevention technologies of information disclosure have been developed in many companies. However, in the case of CGN, many authors (e.g. blogger) don't pay attention to information of themselves on their blog pages. In this paper, we propose a model of personal information disclosure to clarify the degree of information disclosure, and describe a method to calculate the level value of information disclosure. In addition, we evaluate a concrete value of actual blog page based on this model, and compare the results of calculation with human feeling.

1．はじめに

個人情報保護法律が施行され，個人情報の保護に関する意識が高まっている[1]．インターネット上でも個人情報の保護は重要な問題となっており，情報漏えい防止技術の開発や対策の保護の提案が行われて来ている[2]-[4]．これらは，主に管理主体である企業の立場から個人情報をどのように保護すべきか，というプライバシー情報（個人情報保護法による「個人情報」や私的かつ広く開示したくない情報であるところのセンシティブな情報）の保護を主体に検討がなされている．

一方，近年，ブログや mixi 等におけるソーシャル

ネットワークサービスの普及が急速に進むにつれて，個人がインターネット上で情報を発信するようになってきている[5]．多くのブログが日記形式であるため，情報発信者が個人的な日記を気軽に書くような感覚で情報を発信している．その結果，ブログ内容から住所が推測され，顔写真や学校などの所属先がわかり，個人情報が漏えいしていると考えられるケースなどが散見されるようになってきている．個人情報の漏えい状況を客観的に評価することができれば，個人の意識が低い場合においても個人情報の漏えいの状況を認識することが容易となる．

本稿では，個人情報の漏えいを定量的に評価することを目的として，個人情報の漏えいモデルを提案し，個人情報漏えい度の計算法を示す．また，ブログにおける個人情報の漏えい状況について本モデルを適用し人間による漏えい感覚との違いを比較する．

2．個人情報の定義

2．1　個人情報とは

個人情報の保護に関する法律によると，「個人情報とは，生存する個人に関する情報であって，当該情報に含まれる氏名，生年月日その他の記述等により特定の個人を識別できるものをいう」[6]となっている．

個人情報には，氏名，住所，性別，生年月日，電話番号，職業，年収，家族構成，顔写真などが該当するが，一般には2つ以上が結合されないと個人を特定できない場合が多い．個人情報の定義として，センシティブなプライバシー情報やID等の情報を含めるものもあるが，本稿では，個人情報として最も基本的情報である個人を特定するための情報を扱う．

2．2　個人の特定

ここでは，個人を特定するということを「他人がある個人を指定し，直接的に作用できる状態であること」と定義する．他人がある個人を指定するというのは，意図する個人に目的の作用を及ぼすことができるということである．作用とは，特定の個人に対する行為である．たとえば，特定の個人にコンタクトする（会いに行く，電話する等）のは直接的作用である．個人情報が漏えいし，特定されている場合，本人に直接影響が及ぶ作用は危険性が高い．直接的作用の例を表1に示す．

表1　直接的作用の例

カテゴリ	具体例
会う	直接会う、訪問販売、誘拐、ストーカー
電話する	いたずら電話、勧誘、アポイントメントセールス
メールする	ダイレクトメール、スパムメール、ウイルスメール、フィッシング詐欺
送る	ダイレクトメール等の手紙

個人を特定するには，複数の個人情報が必要となる場合が多い．たとえば，氏名のみ電話番号のみでは，個人を特定することができない．また，年収と職業などの複数の項目が判ったとしても，その情報から個人を特定することはできない．「会う」の作用を及ぼすには，本人の住所や氏名，または，所属先の所在地や頻繁に訪れる場所などの情報が必要である．

このように，個人情報の漏えい度を定量的に評価するには，多くの情報とその組み合わせを評価しなければならない．そこで，個人情報漏えい状況との関係を論理的にモデル化し表現することにより，情報漏えい状況の程度を定量化する必要がある．

3．個人情報漏えいモデル

3．1　漏えいレベル

個人情報の漏えい状況を定量的に評価するに際し，漏えいレベルを定義する．表2に漏えいレベルの定義を示す．

表2　漏えいレベルの定義

レベル	漏えいレベルの定義
1	個人を特定できる情報がほとんど漏えいしておらず、 個人を特定するのが非常に困難である状態
2	このままでは個人を特定できる状態とは言えないが、 他の情報がかなり漏えいしている状態
3	探偵などの専門家を通すことで 個人が特定できる可能性がある状態
4	地図や電話帳などの他のデーターベースなどを 使用することで個人を特定することが可能な状態
5	氏名・住所・電話番号など個人を 完全に特定できる情報が漏れている状態

漏えいレベルは，個人情報の漏えい状況をレベル分けしたものである．レベル5では，個人情報が完全に漏えいしている状態で直接的作用を及ぼすことができる．この場合，氏名，住所等の個人情報が漏えいし，本人が特定されている状態である．レベル1では，個人を特定できる情報が漏えいしていない．レベル2, 3, 4では個人情報が漏えいしているが特定までには至っていなく，漏えいの程度が違う．個人情報の中でも特定の度合いに違いがある．

3．2　個人情報の特定度の違い

複数項目の組み合わせではなくて単体で一つの意味を持つ項目を情報項目として定義する．個人を特定するという観点で情報項目を見てみると，個人を特定するのに大きな効果とそれほど大きな効果がない項目がある．たとえば，住所や氏名が漏えいしてしまうと，「会う」という作用が可能となり，個人を特定できたこととなる．また，結婚暦の情報は他の情報と組み合わせたとしても直接的な作用を及ぼすことができず，特定できたとは言えない．

このように，情報項目には特定のされやすさに違いがある．直接的作用に深く関係し，最も個人の特定に結びつく項目は，氏名，住所，電話番号，所属，メールアドレスであると考えられる．また，血液型，身長，結婚暦などの情報はそれだけでは個人を特定する情報としてほとんど価値がないといえる．

上記を考慮し，個人情報の特定度の違いを図1に示す．

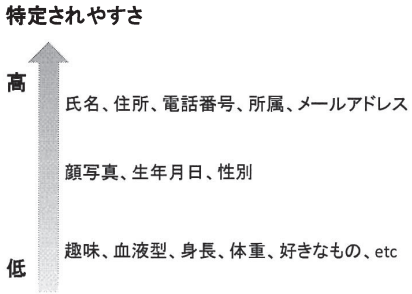


図1　個人情報の特定度の違い

3．3　漏えい値の基本的考え方

個人情報の各項目には，個人を特定する度合があると3．2で述べた．特定度の違いによる漏えい状況を表現するために，個人情報の各項目に漏えい値を設定する．氏名や住所といった特定度の高い項目の漏えい値を高く設定し，その他雑多な情報は漏えい値を低めに設定する．漏えいしている情報の漏えい値を加算していくことで漏えい状況を定量化する．

3．4　個人情報の階層化

個人情報の各項目には，その情報を構成するサブ情報が部分的に漏えいする場合があり，この場合，一つの情報項目がサブ情報項目を構成する場合がある．たとえば，住所は図2のような階層構造になる．

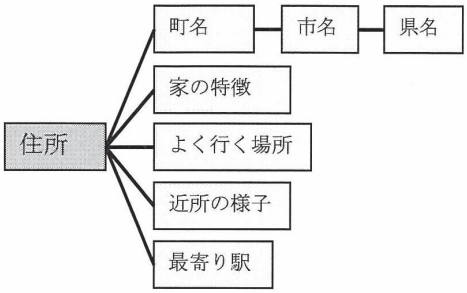


図2　住所の階層化構造

住所は，県，市，町名からなっているのでそれぞれの項目がサブ情報項目になる．また，よく行く場所，最寄り駅，家の特徴などは住所を特定するのに役立つ項目になる．この他にもサブ情報項目を持つ個人情報の項目があり，その例を表3で示す．

サブ情報項目の一部が漏洩している場合は，親の情報項目は完全に漏えいしていないが，一部が漏えいしていると考え，その親情報項目の値を利用せずサブ情報項目の値とする（親情報の値よりも低くなる）．

表3　サブ情報項目を持つ個人情報

情報項目	サブ情報項目
氏名	苗字
	名前
	あだ名
	ハンドルネーム
所属	学校名・勤務先
	学科・部署・役職
	学校所在地・勤務地
生年月日	誕生日
	年齢

3．5　複数項目の漏えいによる特定度の上昇（相乗効果項目）

複数の項目が漏えいすると，個人特定に決定的に結びつく場合と，どんなに多量の情報が漏えいしても個人を特定されない場合もある．このような情報項目を相乗効果項目と呼ぶ．

たとえば，趣味，血液型，身長，体重，好きなものなどの情報がいくら漏えいしても個人を特定することは難しい．それらの情報が組み合わさっても特定度合いはあまり変化がない（図3）．

一方で，複数の項目の結びつきで特定が可能になる例として，氏名と住所がある．氏名と住所が漏えいす

ると，個人が特定されているといえる．実際に会いに行くことも可能であるし，ダイレクトメールを送りつけることもできる．このように，一項目では特定されないが，複数項目が漏えいした場合，個人特定を特定できる場合がある（図4）．また，重要な情報項目が複数組み合わさる場合において，前節で説明したように，サブ情報項目しか漏えいしていない場合もあり，この場合，完全に個人を特定できるとは言えない．しかし，かなり個人情報特定に近づいている事は間違いない．このため，個人情報特定レベルは，漏えいした，しないの二値ではなく，中間の値を持つとして考える．

複数項目の漏えいによる特定度の上昇する場合は，漏えい値同士の積として計算する．組み合わさることにより特定可能とならない項目は漏えい値の加算であるのに対し，組み合わせることにより特定可能となる項目同士は漏えい値の積を取るので，特定度が大きく上昇する．氏名と住所が漏えいしている場合，氏名＋住所という組み合わせは個人を特定できるので，積とする．特に，作用に影響を与える氏名，住所，電話番号，メールアドレス，所属の項目の組み合わせは積として計算する．

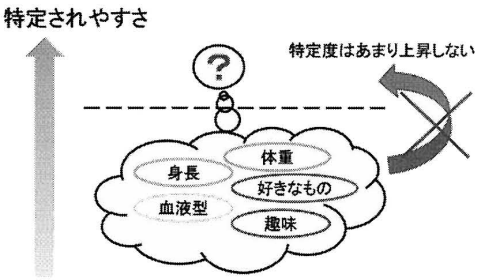


図3　項目同士が結びつかない例

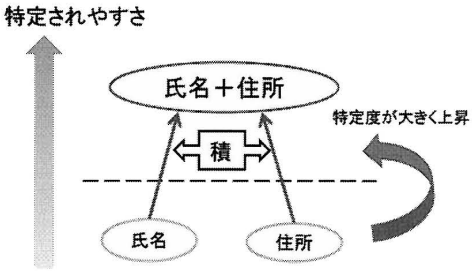


図4　項目の組み合わせによる漏えい値の上昇

3．6　漏えい値の重み

個人情報の項目にはそれぞれ漏えい値を設定したが，サブ情報項目を持つ個人情報の項目の場合は，漏えい値が中間値となる．住所を例に取ると，住所が完全に漏えいしている場合，住所そのものの漏えい値を用いるが，県名と市名までしか漏えいしていない場合はサブ情報項目である県名と市名それぞれの漏えい値の総和を住所の漏えい値として扱う（この場合，住所が完全に漏えいしている場合よりも値が小さくなる）．また，複数項目による特定度の上昇の際の積は，それぞれの情報項目の中間値を用いて計算を行う．

3．7　漏えい度

情報漏えいを数値化したものを漏えい度とし，以下のように定義する．

漏えい度＝各項目の漏えい値の和＋組み合わせによる値の総和

上記個人情報漏えいレベルの概念を図5に示す．

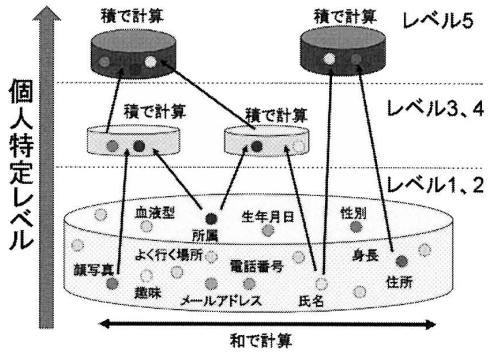


図5　個人情報漏えいレベルの概念

4．計算法

前節の個人情報漏えいモデルに従って，漏えい度を計算する手法を述べる．

4．1　パラメータ設定の基本的な考え方

作用に大きく影響する項目として氏名，住所，電話番号，メールアドレス，所属を基本5項目として定義する．基本5項目は漏えい値を高く設定し，その他の情報は，相対的にパラメータを低く設定する．

サブ情報項目については，親情報項目の値よりも小さな値を設定し，総和が親情報項目の値を超えないよ

うにする．

議論の分かれるところであるが，ここでは，基本 5 項目と同レベルの情報項目として，顔写真と生年月日を重要な項目と考え，基本 5 項目と同じ値とする．ただし，直接作用には寄与しないと考えられるため相乗効果項目とはしない．

4．2　パラメータ設定

個人情報の各項目に漏えい値のパラメータを設定する．パラメータは表 4 のとおりである．パラメータの値はここでは評価用の仮段階のものであり，氏名，住所，電話番号，メールアドレス，所属を 100 の基準とした．前節で示したように，顔写真や生年月日は相乗効果項目とはならないが特定度が高い項目なので 100 とした．

複数項目の組み合わせによる漏えい値の積の一覧を表 5 に示す．なお，この組み合わせも評価用の仮組み合わせであり，今後，拡張変更して行く予定である．

表 4　漏えい値のパラメータ

氏名	
氏名	100
苗字	10
名前	10
あだ名	2
ハンドルネーム	1
住所	
住所	100
町名	10
市名	5
県名	2
家	2
最寄駅	7
よく行く場所	2
実家	6
町名	3
市名	2
県名	1
出身地	1
電話	
電話	100
生年月日	
生年月日	100
年齢	10
月日	10
顔写真	
顔写真	100

所属	
所属	100
学校名・勤務先	10
学科・部署・役職	10
学校所在地・勤務地	10
メールアドレス	
メールアドレス	100
その他	
性別	1
趣味	1
車	1
職業	1
アルバイト	1
サークル	1
家族構成	1
結婚歴	1
病歴	1
好きなもの	1
血液型	1
体格	1
身長	1
体重	1
資格	1

表 5　複数項目の組み合わせによる積の一覧

項目数	組み合わせ
2	氏名＋住所
	氏名＋所属
	氏名＋電話
	氏名＋メール
	住所＋電話
	所属＋電話
	所属＋メール
	電話＋メール
3	氏名＋住所＋所属
	氏名＋住所＋電話
	氏名＋住所＋メール
	氏名＋所属＋電話
	氏名＋所属＋メール
	氏名＋電話＋メール
4	氏名＋住所＋所属＋電話
	氏名＋住所＋所属＋メール
	氏名＋住所＋電話＋メール
	氏名＋所属＋電話＋メール
	住所＋所属＋電話＋メール
5	氏名＋住所＋所属＋電話＋メール

4．3　レベル定義

漏えい度のレベル分け基準値を表 6 に示す．

表 6　基本漏えい度の数値によるレベル分け

レベル定義				
レベル1	レベル2	レベル3	レベル4	レベル5
0～99	100～199	200～999	1000～9999	1000～

レベル 5 では個人が特定されている状態である．氏名，住所，電話番号，メールアドレス，所属の基本的な 5 項目うち相乗効果のある最低二項目が漏えいした場合，個人がほぼ特定されると考えられる．2 項目の組み合わせによる漏えい値は 10000 となるので，レベル 5 は 10000 以上と定義した．以下，レベル定義の目安を示す．

レベル 5

・基本 5 項目のうち 2 項目が完全に漏えいしている．

例：　氏名（100）× 住所（100）= 10000

レベル 4

・相乗効果項目のうち，完全にひとつが漏えいし，もうひとつが不完全に漏えいしている場合．

例：　氏名（100）× 県市町（17）= 1700

レベル 3

・基本 5 項目のうち 1 項目が完全に漏えいしていて残りの項目がわずかに漏えいしている

・相乗効果項目のうち複数項目が部分的に漏えいしている

例：・氏名（100）× 県，市名（7）= 700

・苗字（10）× 県，市名（7）×学校名（10）= 700

レベル 2

・基本 5 項目が一つのみ漏えいしている.

・個人を特定できない情報が多く漏えいしている.

例：・氏名（100）

・性別，趣味，好きなもの，etc ＜ 200

レベル 1

・特定に至る情報がほとんど漏えいしていない

5．モデルの適応例

5．1　調査法

インターネット上で一般に公開されているブログを対象にし，書き込み内容を調査した．個人情報漏えいモデルに従って，個人情報が公開させている情報項目をチェックし，漏えい度の計算を行った．

5．2　結果

基本漏えい度がレベル 5，レベル 4，レベル 3 のケースを表 7，表 8，表 9 に示す．

表 7　レベル 5 の適応例

漏えい度	100206		
氏名	住所	所属	生年月日
苗字	市名	学校名・勤務先	顔写真
名前	県名	学科・部署・役職	性別
あだ名	よく行く場所	学校所在地・勤務地	趣味
ハンドルネーム	出身地		職業
			好きなもの
			血液型
			身長

表 8　レベル 4 の適応例

漏えい度	2206		
氏名	住所	所属	生年月日
あだ名	市名	学校名・勤務先	性別
ハンドルネーム	県名	学科・部署・役職	趣味
		学校所在地・勤務地	職業
			家族構成
			好きなもの
			血液型

表 9　レベル 3 の適応例

漏えい度	295	
氏名	住所	生年月日
苗字	町名	性別
	市名	趣味
	県名	アルバイト
	よく行く場所	好きなもの
		血液型

レベル定義に従ってあらかじめ人間がレベル分けしたものが漏えい度と一致することを確認した．ただし，人間の主観的な評価にとどまり，厳密な評価をする必要がある．

6．おわりに

本稿では，個人情報漏えいモデルを提案し，個人情報の漏えい状況の定量的評価を試みた．また，本モデルに従って実際のブログについて漏えい度を計算し，人間による情報漏えい感覚との比較を行った．その結果，モデルの基本的考え方には妥当性があり，個人情報漏えい度合いを定量評価できるものであることが分かった．

今後の課題としは，人間の主観評価等による情報項目漏えい値のチューニング，精度の向上，情報項目 DB および相乗効果項目 DB の充実，レベル 5 以上の作用別漏えい度の算出モデルの構築を行う予定である．

なお，現在はブログの文章を読み人間が情報項目を抽出しているが，これを自動化しないと実用的ではないため，情報項目自動抽出手法についても研究を進める予定である．

参考文献

[1] NRI セキュアテクノロジーズ株式会社，“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] 岡本栄司：プライバシ保護のための要素技術の動向，情報処理学会誌，Vol.48.No.7（2007 年 7 月）

[3] 小松文子：プライバシ保護のためのアーキテクチャ，情報処理学会誌，Vol48.No.7（2007 年 7 月）

[4] 谷本，廣田，山本，千田，畑島，高橋，金井：次世代プライバシ保護サービスのコンセプト提案，情報処理学会論文誌，Vol.49.No7（2008 年 7 月）

[5] 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

[6] 内閣府，“個人情報の保護に関する法律”，
http://www5.cao.go.jp/seikatsu/kojin/index.html