2009－EIP－43 (2)
2009／2／13

ブログにおける個人情報漏えいレベルの定量化

安井 良介[†]，佐藤 和紀[†]，針谷 友彰[†]，金井 敦[†]，廣田 啓一[††]，谷本 茂明[††]

[†]法政大学大学院 工学研究科
[††]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

あらまし

近年，個人情報保護に対する意識が高まり，大学や企業などでは個人情報漏えい防止の対応が進められている．その一方で，ブログやソーシャルネットワークサービス（SNS）などの普及により個人が手軽に情報を発信する機会が増えたことで，無意識の内に個人情報が漏れているケースが多く見られるようになった．我々は，個人情報の漏えい状況を定量化し，客観的に評価することによる，ネットリテラシーの不足や不注意による個人情報の漏えいを防止する支援機能の検討を進めている．本稿では，個人情報の漏えい状況を定量的に評価するためのプライバシーサーチオラクル(PSO)モデルを提案するとともに，本モデルに基づき，個々の情報による個人の特定度合いを示す絞り込み量を用いて個人情報の漏えい度合いを定量化する手法を提示する．また実際のブログにおける個人情報の漏えい度合いを定量化した結果と，人間の感覚による情報の漏えい度合いとを比較することにより，提案モデルの検証を行う．

A proposal of privacy search oracle model for estimating personal information disclosure level of blog articles

Ryosuke YASUI[†], Kazuki SATO[†], Tomoaki HARIGAYA[†], Atsushi KANAI[†], Keiichi HIROTA[††], Shigeaki TANIMOTO[††]

[†]Graduate School of Engineering, Hosei University
[††]NTT Information Sharing Platform Laboratories, NTT Corporation

Abstract

In recent years, people's concern towards the protection of personal information is growing and many companies and universities are adopting disclosure prevention technologies or solutions to protect personal information they have. On the other hand, in the case of CGM such as Blog and Social Network Service (SNS), most authors (i.e. Bloggers) pay little attention to protect their information and they are disclosing personal information on their blog pages. In this paper, we propose a Privacy Search Oracle model for estimating personal information disclosure, and describe a method to quantify the level degree of information disclosure. In addition, we show the evaluation results of actual blog pages based on proposed model, and compare them with the results based on human judgments.

１．はじめに

個人情報保護法の施行に伴って，近年個人情報の保護に対する意識が急激に高まっている[1]．インターネット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漏えい対策の提案や開発が行われている [8]．これらの技術や対策の多くは，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の観点の下に，プライバシー情報（個人情報保護法における「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）の保護を中心に検討がなされている．

その一方で，最近ではブログやソーシャルネットワークサービス(SNS)等の急速な普及によって，個人がインターネット上で情報を発信できる機会が増えており，こうした CGM サービス上での情報漏えいが問題視されている．多くのブログが日記形式であり，情報発信者は個人的な日記を書く感覚で気軽に情報を発信することができる．その結果として，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るなど，個人情報が漏えいしているケースが散見されるようになっている．

我々は，こうした個人の情報発信における個人情報の保護を目的として，漏えい状況を定量化し，客観的に評価することによる，ネットリテラシーの不足や不注意による個人情報の漏えいを防止する支援機能の検討を進めている．

本稿では，個人情報の漏えい状況を定量的に評価するためのプライバシーサーチオラクル(PSO)モデルを提案するとともに，本モデルに基づき，個々の情報による個人の特定度合いを示す絞り込み量を用いて個人情報の漏えい度合いを定量化する手法を提示する．また実際のブログにおける個人情報の漏えい度合いを定量化した結果と，人間の感覚による情報の漏えい度合いとを比較することにより，提案モデルの検証を行う．

２．プライバシーサーチオラクルモデル（PSO モデル）

２．１　PSO モデルの提案

個人情報の漏えいレベルを定量化するにあたって，本研究ではまず，図１に示すような，与えられた個人情報から該当する個人を絞り込むことができるオラクルの存在を仮定する．このモデルをプライバシーサーチオラクル(PSO)モデルと呼ぶ．

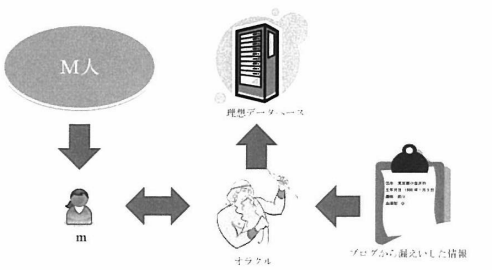

図１　プライバシーサーチオラクルモデル

仮定するオラクルは，ブログなどから漏えいした個人に関する情報を入力として受け付け，この情報を元に M 人の母集団の中からその特徴を持つm人を選び出して出力する能力を持つ．m=１の時，個人を特定できたと考えられる．本稿ではこのオラクルを，プライバシーサーチオラクルと呼ぶ．また，プライバシーサーチオラクルを用いて作成された個人データベースを理想データベースと呼ぶ．このデータベースにはあらゆる個人情報が登録され，検索条件を入力すれば条件を満たす個人を母集団から抽出できる．

PSO モデルを実際に適用するには，理想データベースを用いた図２のようになる．本稿では提案する PSO モデルに基づいて，漏えいした個人情報からの個人の特定を「理想データベースを用いて検索し，ただ１人に絞り込むことができる状態」と定義する．

図２　PSO モデルを用いた個人特定

２．２　絞り込み量の定義

個人を PSO モデルによって絞り込む時，１つの情報だけで絞り込みが大きい情報と，それほど絞り込めない情報がある．たとえば，氏名や電話番号などはある人口の母集団から一意に絞り込むことができる情報と考えられ，これらの情報が漏えいした場合，個人を特定できる情報となる．一方で，性別や身長などの情報については，ほとんどの人が同じ情報を持つので絞り込みしづらく，個人を特定しにくい情報といえる．

このように，情報には絞り込みのし易さに違いがある．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，顔写真であると考

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，個人情報の絞り込み量の概念を図 3 に示す．

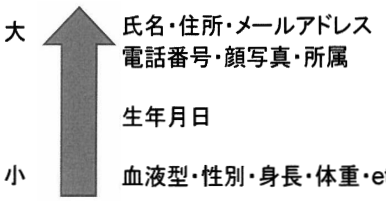

図 3　個人情報による絞り込み量の違い

本稿における絞り込み量の定義を次のように与える．まず，PSO モデルにおいて，理想データベースの登録人数を M，ある情報項目 E が漏えいした場合に M から絞り込める人数を m_E とすると，情報項目 E により個人が絞り込める割合 α_E は次の式で表される．

$$\alpha_E = \frac{m_E}{M} \qquad (1)$$

上式に対し底を 2 とする対数をとると bit として表現でき，情報項目 E に対する絞り込み量 Q_E は次のように定義できる．

$$Q_E = \log_2 \alpha_E \,[\mathrm{bit}] \qquad (2)$$

次に，絞り込み量の絶対量の定義を次のように与える．まず母数 M とし，底を 2 とする対数をとったものを M' とすると，

$$M' = \log_2 M \qquad (3)$$

となる．次に，ある個人のブログから抽出できた情報項目の絞り込み量の総和を Q とすると，

$$M' \leqq Q \qquad (4)$$

上式が等号の場合，母数 M から理想データベースから 1 人に絞り込めるといえる．また差分に関しては，M が増加しても絞り込めることを意味する．

この時，日本の人口全体を母集団として 1 人の個人を特定するための絞り込み量を，本稿における絞込み量の絶対量とする（図 4）．

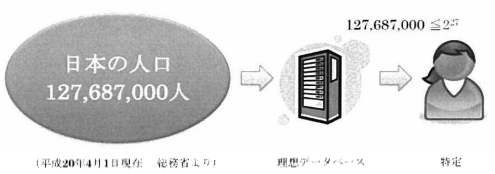

図 4　日本人 1 人を特定するための例

現在の日本の人口は 127,687,000 人であるため[6]，個人を特定するための絞り込み量は，式(3)および(4)より 2^{27}=134,217,728 以上であれば良い．つまり理想データベースでおよそ 27bit 分の絞り込み量に該当する個人情報を集めると個人を特定できたことになるといえる．したがって，本稿では各情報項目の絞り込み量の絶対値を 27 とする．

2．3　個人情報の形態

個人に関する情報の単位として，単体で 1 つの意味を持つ項目を情報項目として定義する[2]．情報項目は図 5 に示す通り，独立情報項目，完全従属情報項目，部分従属情報項目の 3 つに分類できると考えられる．独立情報項目とは，たとえば血液型が漏えいしてもその情報だけでは性別が判断できないなど，他の情報項目との間で関連がない，独立性を持った情報である．また完全従属情報項目とは，住所の項目で市名が漏えいすると，県名が判断できるなど，その情報項目から他の情報項目が明らかになる，従属性を持った情報である．また，住所とよく行く場所などは存在する市名などお互いの従属する情報項目を一部共有する可能性がある．このような情報を，部分従属情報項目と定義する．

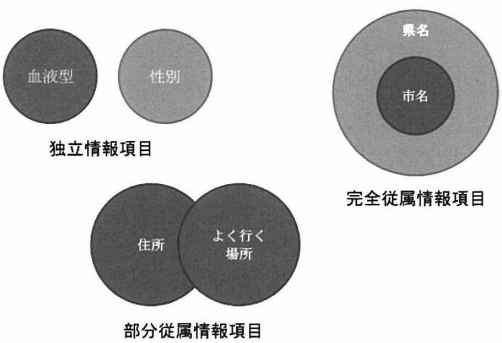

図 5　個人情報の 3 つの形態

本稿では近似的に独立情報項目と完全従属情報項目のみを扱い，部分従属情報項目による影響を考慮しないものとする．

２．４　完全従属情報項目の絞り込み量の定義

個人情報の各情報項目には，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，この場合情報項目は従属情報項目により構成される．たとえば，情報項目としての住所は図６のような階層構造になる．

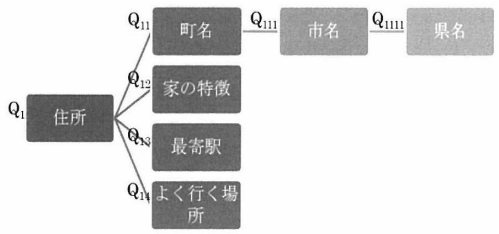

図６　階層化構造

住所は独立情報項目に該当するが，町名，市名，県名などは住所の完全従属情報項目になる．したがって，各情報項目の絞り込み量はより詳細なほど大きく，

$$Q_1 > Q_{11} > Q_{111} > Q_{1111} \qquad (5)$$

となる．また町名，家の特徴，よく行く場所，最寄駅などは同じ階層にある完全従属情報項目である．独立情報項目に対し，複数の完全従属情報項目が存在する場合には，絞り込み量の最も大きい値を採用する．

この他にも完全従属情報項目を持つ個人情報の項目は幾つか考えられ，その例を表１に示す．

表１　独立情報項目と完全従属情報項目の例

情報項目	完全従属情報項目
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

２．５　個人情報漏えい度の計算法

以上の定義に基づいて，ブログなどにおける個々の情報項目の絞り込み量からの，個人情報漏えい度の計算法について定める．

本稿では，完全従属情報項目の絞り込み量は独立情報項目の絞り込み量に組み込むものとして，独立情報項目間での計算を行うこととする．具体的には，独立情報項目が漏えいしておらず，その独立情報項目における完全従属情報項目が複数漏えいした場合に，その中で最も高い情報項目の絞り込み量を，情報項目の絞り込み量として用いる．

従属関係がある場合の例を図７に示す．各情報項目を E_n，その情報項目に対する絞り込み量を Q_n とする．情報項目 E_{11}，E_{12}，E_{13}，E_{111} が漏えいしたとした時，まず E_{11}，E_{111} に対する絞り込み量 Q_{11}，Q_{111} を比較すると次の関係が成り立つ．

$$Q_{11} > Q_{111} \qquad (6)$$

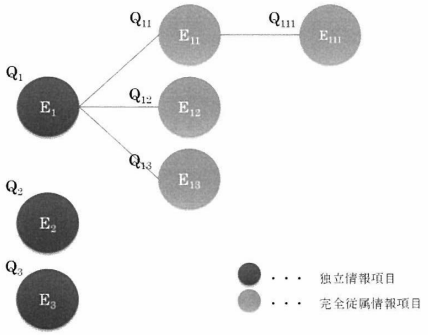

図７　独立情報項目と完全従属情報項目の例

この時，E_{11}，E_{111} の漏えいによる絞り込み量として Q_{11} を採用する．次に E_{11}，E_{12}，E_{13}，の絞り込み量の比較を行い，この中で最も大きい絞り込み量を，E_1 の絞り込み量 Q_1 とする．

一方，独立情報項目が漏えいし，その独立情報項目における完全従属情報項目も漏えいしている場合には，独立情報項目の絞り込み量のみを用いる．

以上示した独立情報項目の絞り込み量の与え方による情報項目全体の絞り込み量の計算式を示す．あるブログから抽出できた各独立情報項目の絞り込み量を Q_E とすると，情報項目の絞り込み量の総和 Q は以下で表すことができる．本稿ではこの値を各情報項目の漏えいによる個人情報漏えい度とする．

$$Q = \sum Q_E \qquad (7)$$

3．適用例

前節に述べた個人情報の情報項目の絞り込み量と，個人情報漏えい度の計算式に従って，適用例として，日本国内における絞り込み量の付与について述べる．

3．1　情報項目と絞り込み量の設定

ブログにおける個人情報漏えい度の計算と評価を行うにあたって，評価用の仮パラメータとして情報項目と絞り込み量の設定を表 2 のように与えた．パラメータの内，絞り込みに大きく影響する情報項目として氏名，電話番号，メールアドレス，所属，顔写真の 5 つの項目はどれか 1 つ漏えいした段階で個人を絞り込めたものとして，絞り込み量を 27bit とする．ここで所属とは，ある個人の勤務先，部署，役職などが漏えいした場合であり，一意であるものと考える．その他の情報項目に関しては，絞り込みの程度に関する詳細な調査が困難であったため，今回は絞り込みの程度を仮定した上で，仮の数値を便宜的に設定した．幾つかの情報項目について絞り込みの程度の仮定を示す．

まず氏名については同姓同名が存在する場合があるが，今回はその確率は考慮せず一意であるものとした．氏名を構成する完全従属情報項目の内，苗字については，日本の苗字は約 25,000 種以上存在し，図 8 に示すように約 7000 種の苗字によって，日本の人口の 96,27%をカバーできるとされている[4]．ただし，それぞれの苗字についての出現頻度は異なる．個々の出現頻度の仮定により絞り込み量の設定を行うことも可能だが，今回は便宜的に出現頻度の期待値をとり，その値から絞り込み量を算出した．名前についても今回は同様の絞り込み量とした．

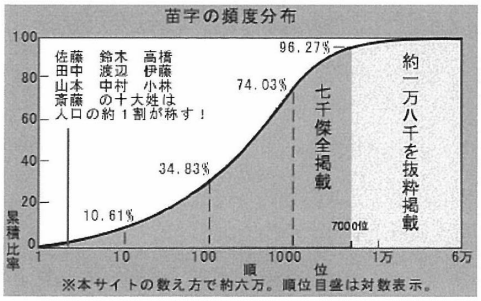

図 8　苗字の頻度分布

次に住所については，ほぼ個人を絞り込むのに役立つ情報といえる．ただし，母数を日本国内で考えた時に 27bit であれば個人を特定できたことになるが，世帯住居など 1 つの住所に必ずしも 1 人が住んでいるとは限らないので 1bit 分減らし，絞り込み量は 26bit とした．また県名は 47 の都道府県から 6bit とし，市名は各都道府県の市の数の平均をとり，町名は調査できなかったため暫定的な設定として，絞り込み量を与えた．これらの情報項目についても理想的には各県の人口分布などを調査した上で，詳細な絞り込み量を設定することが望ましい．また家の特徴として，一軒家・アパート・マンションの 3 種とし，2bit とした．なお最寄駅については，全国に 9716 駅[7]存在しているため 13bit とし，よく行く場所に関しては最寄駅と同程度の絞り込みが可能と考え，同様の値とした．

一方，所属については，学校名・勤務先は企業総数が約 1,500,000[5]あることから 21bit とし，学科・部署・役職などは学校や企業により異なり厳密に調査できないため，今回は暫定的に 10 とした．学校所在地・勤務地なども調査は困難だが，複数の住所に存在する企業なども考慮した上で，企業総数×3 として与えた．

また，その他の情報について，体格について 5 段階に分類し，身長・体重についても 10 段階に分類を行った．こうした身体的特徴に関する数値は，それぞれの個人に異なる値だが，時間的な変動や区間内での誤差を踏まえた上で類型化を行うことが望ましいと考えられる．その他の情報項目については漏えいした場合の絞り込みの程度の判断が難しいため，基本的に漏えいしたかしていないかでの判断とし，1bit とした．

これらのパラメータに関しては，今後の検討を踏まえて拡張，変更していく予定である．また，各情報項目に関する母集団の違いによる絞り込み量の違いも生じるため，詳細なパラメータ設定は今後の検討課題とする．

表 2　パラメータ設定

情報項目	絞り込み量 [bit]	備考
本名		
本名	27	同姓同名は考慮しないと一意に決定できる
苗字	18	期待値より
名前	18	苗字と同じに
あだ名	1	有・無の 2 種
HN	1	有・無の 2 種
住所		
住所	26	ほぼ一意に決定できる
町名	14	47×17×30

市名	9	47×17
県名	6	47 の都道府県
家	2	一軒家・アパート・マンション
最寄駅	13	全国の駅数は 9716 個
よく行く場所	13	最寄駅と同等とする
実家住所	27	一意に決定できる
実家町名	14	47×13×30
実家市名	9	47×13
実家県名	6	47 の都道府県
出身地	6	47 の都道府県
連絡先		
電話	27	一意に決定できる
生年月日		
生年月日	15	年齢×月日
年齢	6	平均年齢が約 80 歳より
月日	9	365 日より
顔写真		
顔写真	27	一意に決定できる
所属		
所属	27	200,000 個
学校名・勤務先	21	1,500,000 個
学科・部署・役職	3	10 個
学校所在地勤務地	22	1,500,000×3 個
メールアドレス		
メールアドレス	27	一意に決定できる
その他の情報		
性別	1	男・女の 2 種
趣味	1	有・無の 2 種
車	1	有・無の 2 種
職業	4	16 種の職業だけと考える
アルバイト	1	有・無の 2 種
サークル	1	有・無の 2 種
家族構成	1	有・無の 2 種
結婚歴	1	有・無の 2 種
病歴	1	有・無の 2 種
好きなもの	1	有・無の 2 種
血液型	2	A,B,O,AB の 4 種
体格	2	5 種
身長	3	10 種
体重	3	10 種
資格	1	有・無の 2 種

4．PSO モデルの適用と比較調査

4．1　調査法

PSO モデルの適用による各情報項目への絞り込み量の付与と個人情報漏えい度の計算について，実際のブログを対象として調査を行い，人の主観による個人の特定のしやすさとの比較を行った．

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．調査件数は 203 件である．また調査したブログの書き込み内容について，個人情報として公開されている情報項目をチェックし，PSO モデルに従って個人情報漏えい度の計算を行った．

4．2　主観的モデルとの比較

主観的に個人情報の定量評価を行うモデル[2]（以下，主観的モデル）において，個人特定の定義は「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」である．また直接的な作用とは，表 3 に示すような個人情報の漏えいにより起こるインシデントである．

表 3　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・スパム・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

次に主観的モデルでは，個人情報の危険度レベルを表 4 のように定義している．

表 4　危険度レベルの定義

レベル	危険度レベルの定義
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

このような判断基準によりブログ内で公開されている個人情報の情報項目をもとに定量評価を行なった主観的モデルと，今回行った各情報項目に対する絞り込み量の付与と個人情報漏えい度の計算式に基づく「絞り込み量を用いて客観的に定量化を行ったモデル」(以下，客観的モデル）との比較を図 9 に示す．また図 9 は調査件数 203 件に対してそれぞれ危険度レベルの判定を行い，横軸に絞り込み量によるプロットをしたものである．

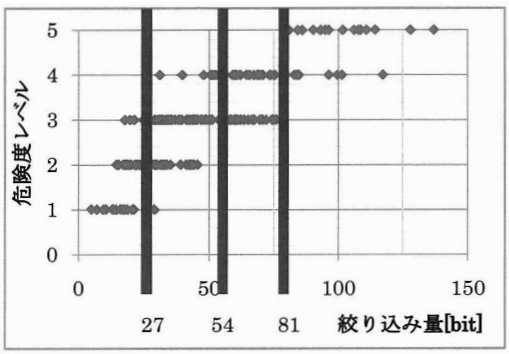

図 9　主観的モデルと客観的モデルとの比較

図 9 より，主観的モデルにおける危険度レベルと，客観的モデルによる絞り込み量の総和の間には，概ね正の相関が得られたともと考えられる．また本稿で規定した絞り込み量の絶対量 27bit の倍数である 54bit,81bit の地点に縦線を引くと若干のブレはあるものの，軸を中心として危険度レベルとがだいたい一致することが分かった．主観的モデルでは実際に作用を及ぼす危険があるのはレベル 5 である．しかし客観的モデルにおける絞り込み量でレベル 5 に相当する箇所はおよそ 81bit から分布が始まる．またレベル 4 は 50~81bit に分布が集中し，レベル 3 は 27～81bit に分布，レベル 2 は 27～50bit に分布し，レベル 1 は 27bit までに収束していることがわかる．

客観的モデルでは，日本における絞り込み量の絶対量として，27bit を個人特定が可能な絞り込み量と定義した．しかし絞り込み量の総和により 27bit で理論的に個人を特定できたとしても，仮定した理想データベースは実際には存在しないため，実際の個人の特定は非常に困難であることが多い．また主観的モデルと客観モデルの比較評価から，個人に対する何らかの直接的作用を及ぼすには，絶対量の約 3 倍の絞り込み量が必要であると考えられる．情報項目と絞り込み量の値などパラメータとしての設定をより厳密に評価する必要があるが，PSO モデルと現実とのギャップはおよそ 54 ビット程度と考えられ，この 3 倍という値が主観的モデルと客観的モデルとの誤差ということになるといえる．

5．おわりに

本稿では，プライバシーサーチオラクルを仮定した上で，個人情報漏えいレベルの定量化を行う PSO 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，提案モデルに基づいて，個人情報の情報項目と絞り込み量をパラメータとして与え，実際のブログについて絞り込み量を用いて漏えい度を計算した客観的モデルと，人間による主観的な判断を行う主観的モデルとの比較を行った．その結果，客観的モデルの基本的な考え方には十分な妥当性があり，誤差はあるものの，個人情報の漏えい度合いを定量的に評価可能であることが分かった．

今後の課題としては，情報項目と絞り込み量の決定のための調査と精度の向上，時間の経過による情報項目の劣化に対する絞り込み量の定義，および部分従属情報項目に関する個人情報漏えい度の計算方法の検討と，個人情報漏えいによる直接的な作用による危険度の分析などを行っていく予定である．

なお，現在はブログの文章を人間が読み，規定された情報項目を人手で抽出しているが，こうした作業を自動的に行う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] 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)

[3] 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

[4] 日本の苗字 7 千傑
http://www.myj7000.jp-biz.net/

[5] 総務省，“企業数に関するデータ”

http://www.stat.go.jp/data/jigyou/2004/kakuhou/gaiyou/08.htm
[6] 総務省，“統計データ”
http://www.stat.go.jp/data/index.htm
[7] 愛知県公式Webサイト，“統計データ 鉄道駅数（都道府県別）”
http://www.pref.aichi.jp/cmsfiles/contents/0000008/8567/2-(2).xls
[8] 谷本茂明，廣田啓一，山本太郎，千田浩司，畑島隆，高橋克巳，金井敦“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,No.7, pp.2440-2455(July 2008)