統計情報を用いた 個人情報露出量算出方式の検討

松永大希† 金井敦†

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．一方でブログやソーシャルネットワーキングサービス等の普及によって，パソコンや携帯電話を利用して個人が情報を気軽に発信できる機会が増え，無意識に個人情報を露出ケースが多く見られるようになってきた．このような情報露出を未然に防ぐためには，本人がどの程度情報が露出しているかを自覚しなければならない．そのために自動的に情報の露出状況を判断し，数値化して個人に示すモデルが必要である．本稿ではそのモデル中において情報の露出状況から正しく定量化評価するために，統計情報を用いて情報露出量を扱い，露出した要素と統計情報が与えられた時に個人情報露出量を算出する方式を構築した．

A study on quantifying the degree of personal information disclosure using statistical data

DAIKI MATSUNAGA† ATSUSHI KANAI†

In recent years, people's concern towards the protection of personal information is growing with the enforcement of the Personal Information Protection Law. On the other hand, in the case of CGM such as Blog and Social Networking Service, most authors pay little attention to protect their information and they are disclosing personal information on their blog pages with PCs and cell phones. Therefore model for estimating personal information disclosure and describe a method to quantify the level degree of information disclosure is necessary. In this paper, we study a method to quantify the degree of personal information disclosure using statistical data for acknowledging the more accurate situation of information disclosure in blog pages.

†法政大学大学院 工学研究科
Graduate School of Engineering, Hosei University

1. はじめに

近年，インターネットの普及により多くの人がブログやソーシャルネットワーキングサービス（SNS）等を利用して気軽に情報を発信することが可能になった．個人情報の保護に対する意識も高まってきているが[1], ブログや SNS などにおいて本人の無意識のうちに個人情報が露出してしまい問題となるケースが発生している．無意識のうちに露出した個人情報を悪意のある人が利用することでその人の住所や職場などの生活環境を特定されてしまうなど，個人の生活に危険が及んでしまう可能性がある．このような事態を防ぐためにブログや SNS においてどの程度個人情報が露出しているのかという状態を利用者や運営者に自覚させる必要があるが，どの情報がどの程度危険であるかなどの判断を個人で行うことは難しいといえる．

その対策として文章中にどのくらい個人情報が露出しているかを自動で判断し，その露出状況を誰にでも簡単な形で示すモデルが必要である．これはブログなどの文章中から個人を特定してしまう単語を抜き出し，モデルが持つデータベースと比較してその文章からどの程度個人を特定できてしまう危険性があるかを数値で表して利用者に示すためのモデル[2][3]である. このモデルが客観的な立場から露出状況を数値で表すことで利用者に個人情報の露出状況を伝え，利用者は個人情報の露出状況を自覚し個人情報の保護に対して高い意識を持つことができる．結果，個人による無防備な情報露出を避けることができ安全にブログや SNS を利用することが可能である．また，露出状況を把握することは利用者だけでなく運営者にとっても重要であり，露出状況を把握することでその対策を講じることが可能となる．よって露出状況の定量化は安全にインターネットを運用するために必要となる．

本稿ではモデルにおける露出状況の統計情報を用いた算出方式についての提案を行う. 既存の方式[4][5]では露出状況をその情報を持つ人が情報項目においてどのように分布しているかを想定し，エントロピーと同様の計算方式で平均した値を算出しているが，情報を持つ人の分布が大きく偏っている場合や想定した値と実際の値に差がある場合などには算出したその情報が持つ正確な情報量を表現することは困難になる. 正確な露出状況を表現するために, 露出状況をエントロピー的に算出するのではなく，実際に存在する統計情報からその情報を持つ人数を求め，露出状況を表現する方式について検討する．

2. 個人情報露出量の算出方式

2.1 既存の算出方式の困難さ

既存の算出方式[4][5]ではそれぞれの情報項目に関して要素の分布を仮定し, エントロピーと同様の計算を行いその値を bit に変換したものをその情報項目が持つ分解能力として，分解能力の合計を個人情報露出量としている．分解能力は露出した情報 1

つ 1 つがどれくらいの総人数のうちから 1 人に絞ることができるかという度合いであり，それぞれの bit の合計である個人情報露出量はその文章の情報露出の度合いを表している．例えば 1bit なら 2 人に 1 人を判断できるということである．世界人口を 70 億人とすると，70 億は約 2 の 33 乗であるため個人情報露出量が 33bit になれば個人を特定することができたことになる．しかし，苗字という情報項目において「佐藤」と「伊集院」という情報要素は同じ分解能力として扱っているが，実際には「伊集院」という苗字の人は「佐藤」と比較すると少数であり「伊集院」という要素が持つ分解能力は「佐藤」が持つ分解能力よりも高くならなくてはならないため，結果として正確な個人情報露出量を表していないということになってしまう．苗字という一例を挙げたが正確な個人情報露出量を算出するにはそれぞれの要素が持つ分解能力を扱う必要がある．

また，多様な情報項目において本人を特定すること可能である物は限られてくる．例えば「住所」という情報項目は本人を特定する可能性が高いが「趣味」という情報項目は露出していたとしても本人を特定する可能性は低い．個人情報露出量を算出するために多数な情報項目から有効な情報項目を選択しなければならない．そのために本報告では「名前」や「住所」や「性別」などの住民基本台帳に記載されているような個人を特定する能力の高い情報項目に限り，個人情報露出量を算出する方式について述べる．

2.2 個人情報露出量算出までの流れ

前章での問題点を解決するような個人情報露出量を算出する方式について述べる．ブログなどの文章中に露出している個人に関する複数の情報要素に対して，統計情報を用いて総人口におけるその要素を持つ人数を知り，その人数を用いて個人情報露出量を算出することでより精度の高い結果を利用者に提供することが可能である．統計情報を用いることで，それぞれ独立した情報項目においても情報項目間の関連性を加味してその要素を持つ人数を計算し，その人数から個人情報露出量を算出することでより正確な個人情報露出量を表現することができると考えることができる．

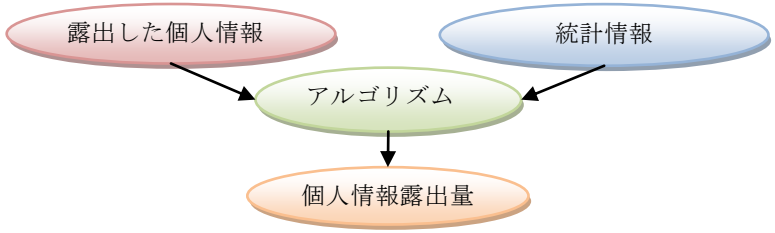

図 1　個人情報露出量算出過程

露出した要素に対してその要素が属する情報項目が統計情報に含まれていない場合や露出した要素が属する情報項目に関する統計情報が手に入らない場合は統計情報を用いて情報露出量を算出することができない．この場合は既存の方式から個人情報露出量を算出しなくてはならない．

3. 露出情報に対する個人情報露出量の求め方

3.1 情報要素と統計情報

本章では統計情報を用いて要素を持つ人数を計算し個人情報露出量を算出する方式について述べる．ここで 1 種類の情報項目に関するその要素を持つ人数の統計情報を 1 次元の統計情報，2 種類の情報項目に関するその要素を持つ人数の統計情報を 2 次元の統計情報とそれぞれ呼ぶ．

露出した要素が多数の場合，その全ての情報項目が統計情報に含まれていればその人数から容易に個人情報露出量を決定することが可能だか，実際に存在する統計情報は 2 種類の情報項目を持つものが多数であり情報項目数が増加すると入手が困難である．そのため，手に入る統計情報は 1 次元の統計情報と 2 次元の統計情報に限定する．

算出方式を述べるために，準備として B を露出した情報を集めた集合，また i を情報項目名，j を要素名として e_{ij} を個人が持つ要素とし，以下のように表す．

$$B = \{e_{ij}\}$$

同じ情報項目名を持つ要素 e_{ij} の集合を E_n をとして以下のように表す．

$$E_n = \{e_{i1}, e_{i2}, \cdots, e_{ij}\}$$

統計情報を T_n として 1 次元の統計情報と 2 次元の統計情報を以下のように表す．

$$T_n = T(E_x)$$

$$T_n = T(E_x, E_y)$$

統計情報を T_n から情報項目に属する要素を持つ人数を P_n として以下のように表す．

$$P_n = T(e_{xj})$$

$$P_n = T(e_{xj}, e_{yj})$$

3.2 統計情報を用いた個人情報露出量

ここでは露出した情報と手に入る統計情報の組み合わせによって発生するケースを場合分けし，それぞれの場合についての個人情報露出量算出を述べる．

説明のために露出した情報要素が属する情報項目を 3 種類とする．この場合，手に入る可能性のある統計情報は以下のものが存在する．

$$T_1 = T(E_1)$$
$$T_2 = T(E_2)$$
$$T_3 = T(E_3)$$
$$T_4 = T(E_1, E_2)$$
$$T_5 = T(E_1, E_3)$$
$$T_6 = T(E_2, E_3)$$

ここで用いる統計情報はそれぞれの情報項目に属する要素の人数を正確に記録したものであり，統計情報それぞれの人数の総計は全て等しくなる．

手に入る組み合わせの種類は露出した情報要素数が少ない方が単純であるため，1 種類の情報項目に属する要素が露出していた場合，2 種類の情報項目に属する要素が露出していた場合，3 種類の情報項目に属する要素が露出していた場合と順を追って以下に述べる．

3.2.1 露出した情報項目が 1 種類の場合の算出

露出した情報が 1 種類の情報項目に属する要素となる場合について述べる．集合は以下のようになる．

$$B=\{e_{11}|e_{11} \in E_1\}$$

統計情報で手に入ったものを用いて個人情報露出量を算出する．統計情報の組み合わせについて場合分けを考えると，露出した要素が属する情報項目を持つ統計情報が手に入らなかった場合（ケース 1），露出した要素が属する情報項目を持つ 1 次元の統計情報が 1 種類のみ手に入った場合（ケース 2），露出した要素が属する情報項目を含む 2 次元の統計情報が手に入った場合（ケース 3），露出した要素が属する情報項目を持つ統計情報が複数手に入った場合（ケース 4）が存在する．それぞれのケースの算出方式について述べる．

ケース 1

露出した要素が属する情報項目を持つ統計情報が手に入らなかった場合と統計情報が手に入らなかった場合について述べる．この場合，統計情報によって人数を求めることが可能である要素が存在しないため，露出した要素については既存の方式通り要素の出現頻度を期待値として情報項目ごとに想定した値を分解能力としてそれぞれを加算し個人情報露出量を算出する．ここでは 1 種類のみなので分解能力が個人情報露出量となる．期待値として想定した情報項目の分解能力の bit を$f(E_n)$，個人情報露出量を *IDM* と表現し，このケースでは以下のようになる．

$$IDM = f(E_1)$$

ケース 2

露出した要素が属する情報項目を持つ 1 次元の統計情報が 1 種類のみ手に入った場合について述べる．

$$B=\{e_{11}|e_{11} \in E_1\}$$
$$T_1 = T(E_1)$$

この場合, 統計情報によって人数を求めることが可能である要素が 1 種類存在する．統計情報から得た人数を分解能力として bit に変換し個人情報露出量を算出する．

$$P_1 = T(e_{11})$$
$$IDM = -\log_2\left(\frac{P_1}{U}\right)$$

ケース 3

露出した要素が属する情報項目を含む 2 次元の統計情報が手に入った場合について述べる．

$$B=\{e_{11}|e_{11} \in E_1\}$$
$$T_4 = T(E_1, E_2)$$

この場合，2 次元の統計情報に露出した要素の属する情報項目が含まれている．もう一方の情報項目については露出した要素によって決定することができないため，2 次元の統計情報において露出した要素を持つ部分の人数を全て加算することで 2 次元の統計情報を 1 次元の統計情報に縮退させ人数を求める．人数を分解能力として分解能力に変換し個人情報露出量を算出する．

$$P_1 = T(e_{11}) = T(e_{11}, e_{21}) + T(e_{11}, e_{22}) + \cdots + T(e_{11}, e_{2j})$$
$$IDM = -\log_2\left(\frac{P_1}{U}\right)$$

ケース 4

露出した要素が属する情報項目を持つ統計情報が複数手に入った場合について述べる．

$$B=\{e_{11}|e_{11}\in E_1\}$$

$$T_1=T(E_1)$$
$$T_4=T(E_1,E_2)$$
$$T_5=T(E_1,E_3)$$

この場合，全ての統計情報に露出した要素の属する情報項目が含まれている．2 次元の統計情報に関してはケース 3 と同様に 1 次元の統計情報に縮退させ人数を求めることが可能であり全ての統計情報から人数を求めることができる．統計情報の部分でも触れたが同じ情報項目に関してその人数は統計情報の種類に関わらず等しくなる．よって得た人数を分解能力として分解能力に変換し個人情報露出量を算出する．

$$T(e_{11})$$
$$=T(e_{11},e_{21})+T(e_{11},e_{22})+\cdots+T(e_{11},e_{2j})$$
$$=T(e_{11},e_{31})+T(e_{11},e_{32})+\cdots+T(e_{11},e_{3j})$$
$$P_1=T(e_{11})$$
$$IDM=-\log_2\left(\frac{P_1}{U}\right)$$

3.2.2 露出した情報項目が 2 種類の場合の算出

露出した情報が 2 種類の情報項目に属する要素となる場合について述べる．集合は以下のようになる．

$$B=\{e_{11},e_{22}|e_{11}\in E_1,e_{22}\in E_2\}$$

統計情報の組み合わせについて場合分けを考えると，露出した要素が属する情報項目を持つ統計情報が手に入らなかった場合（ケース 5），統計情報によって露出した要素の一方が含まれる場合（ケース 6），2 種類の統計情報がそれぞれ 1 種類の要素を含む場合（ケース 7），露出した要素が属する情報項目を両方含む 2 次元の統計情報が手に入った場合（ケース 8）が存在する．2 次元の統計情報と 1 次元の統計情報の組み合わせはさらに存在するが，2 次元の統計情報が 1 次元に縮退可能な場合は 1 次元と見なし該当するケースに含ませた．それぞれのケースの算出方式について述べる．

ケース 5

露出した要素が属する情報項目を持つ統計情報が手に入らなかった場合と統計情報が手に入らなかった場合について述べる．この場合，ケース 1 と同様に露出した要素については既存の方式通り要素の出現頻度を期待値として情報項目ごとに想定した値を分解能力としてそれぞれを加算し個人情報露出量を算出する．露出した情報項目数が増加しその項目を持つ統計情報が手に入らなかった場合は既存の方式と同様に想定した値を分解能力とし加算し個人情報露出量を求める．

$$IDM=f(E_1)+f(E_2)$$

ケース 6

統計情報によって露出した要素の一方が含まれる場合について述べる．

$$B=\{e_{11},e_{22}|e_{11}\in E_1,e_{22}\in E_2\}$$
$$T_5=T(E_1,E_3)$$

この場合, 統計情報によって人数を求めることが可能である要素が 1 種類存在する．このケースには一方の要素のみを含む 1 次元の統計情報または 2 次元の統計情報が 1 種類のみ露出した要素を含んでいるものが手に入る場合である．2 次元の統計情報の場合はケース 3 と同様に縮退させ人数を求める．統計情報から得た人数を分解能力として bit に変換し，他方の要素については想定した分解能力として，加算し個人情報露出量を算出する．

$$P_1=T(e_{11})=T(e_{11},e_{31})+T(e_{11},e_{32})+\cdots+T(e_{11},e_{3j})$$

$$IDM=-\log_2\left(\frac{P_1}{U}\right)+f(E_2)$$

ケース 7

2 種類の統計情報がそれぞれ 1 種類の要素を含む場合について述べる．

$$B=\{e_{11},e_{22}|e_{11}\in E_1,e_{22}\in E_2\}$$
$$T_5=T(E_1,E_3)$$
$$T_6=T(E_2,E_3)$$

この場合，露出した要素がそれぞれ 2 次元の統計情報に含まれている．それぞれの統計情報を縮退させ 1 次元の統計情報とし，人数の比率を計算しその人数を分解能力に変換し個人情報露出量を算出する．比率計算を行った人数を P_n' と表現する．

$$P_1 = T(e_{11}) = T(e_{11}, e_{21}) + T(e_{11}, e_{22}) + \cdots + T(e_{11}, e_{2j})$$

$$P_2 = T(e_{22}) = T(e_{22}, e_{31}) + T(e_{22}, e_{32}) + \cdots + T(e_{22}, e_{3j})$$

$$P'_1 = P_1 \times \frac{P_2}{U} = P_2 \times \frac{P_1}{U}$$

$$IDM = -\log_2\left(\frac{P'_1}{U}\right) = -\log_2\left(\frac{P_1}{U}\right) - \log_2\left(\frac{P_2}{U}\right)$$

ケース 8

露出した要素が属する情報項目を両方含む 2 次元の統計情報が手に入った場合について述べる．

$$B = \{e_{11}, e_{22} | e_{11} \in E_1, e_{22} \in E_2\}$$

$$T_4 = T(E_1, E_2)$$

この場合，2 種類の要素に対応した人数を統計情報から求めることが可能である．統計情報から得た人数を分解能力として個人情報露出量を算出する．

$$P_1 = T(e_{11}, e_{22})$$

$$IDM = -\log_2\left(\frac{P_1}{U}\right)$$

3.2.3 露出した情報項目が 3 種類の場合の算出

露出した情報が 3 種類の情報項目に属する要素となる場合について述べる．集合は以下のようになる．

$$B = \{e_{11}, e_{22}, e_{33} | e_{11} \in E_1, e_{22} \in E_2, e_{33} \in E_3\}$$

統計情報で手に入ったものを用いて個人情報露出量を算出する．ただし，2 種類の情報項目に関する統計情報と 1 種類の情報項目に関する統計情報の情報項目が同じ場合，2 次元の統計情報はその情報項目の条件だけに注目することで人数を得ることが可能であり，1 次元の統計情報が持つ情報を含んでいる．そのため 1 次元の統計情報を用いる必要がないので，2 次元の統計情報のみが手に入った場合と同様である．そのことを踏まえて統計情報の組み合わせについて場合分けを考えると，新しいものは 2 次元の統計情報が 1 種類のみ手に入った場合（ケース 9），2 次元の統計情報とその統計情報に含まれない情報項目に関する 1 次元の統計情報が手に入った場合（ケース 10），2 次元の統計情報が 2 種類手に入った場合（ケース 11），2 次元の統計情報が 3 種類手に入った場合（ケース 12）が存在する．統計情報が手に入らなかった場合はケース 5 の方式で算出し，1 次元の統計情報のみが手に入った場合は先に述べたケースと同様にそれぞれの人数から bit を算出し加算すれば良い．それぞれのケースの算出方式について述べる．

ケース 9

2 次元の統計情報が 1 種類のみ手に入った場合について述べる．

$$B = \{e_{11}, e_{22}, e_{33} | e_{11} \in E_1, e_{22} \in E_2, e_{33} \in E_3\}$$

$$T_4 = T(E_1, E_2)$$

この場合，2 種類の要素に対応した人数を統計情報から求めることが可能である．統計情報から得た人数を分解能力として bit に変換し，残る要素については既存の方式通り要素の出現頻度を期待値として想定した値を分解能力としてそれぞれを加算することにより個人情報露出量を算出する．

$$P_1 = T(e_{11}, e_{22})$$

$$IDM = -\log_2\left(\frac{P_1}{U}\right) + f(E_3)$$

ケース 10

2 次元の統計情報とその統計情報に含まれない情報項目に関する 1 次元の統計情報が手に入った場合について述べる．

$$B = \{e_{11}, e_{22}, e_{33} | e_{11} \in E_1, e_{22} \in E_2, e_{33} \in E_3\}$$

$$T_3 = T(E_3)$$

$$T_4 = T(E_1, E_2)$$

この場合 2 種類の要素が，2 次元の統計情報によって人数を求めることが可能であり，残る要素についてはその要素を持つ人数を 1 次元の統計情報から求めることが可能であるため，分解能力として bit に変換しそれぞれを加算することにより個人情報露出量を算出する．

$$P_1 = T(e_{33})$$

$$P_2 = T(e_{11}, e_{22})$$

$$IDM = -\log_2\left(\frac{P_1}{U}\right) - \log_2\left(\frac{P_2}{U}\right)$$

ケース 11

2 次元の統計情報が 2 種類手に入った場合について述べる．

$$B=\{e_{11},e_{22},e_{33}|e_{11}\in E_1,e_{22}\in E_2,e_{33}\in E_3\}$$

$$T_4=T(E_1,E_2)$$
$$T_5=T(E_1,E_3)$$

この場合 2 次元の統計情報は独立ではなく情報項目に重なりが存在するため，統計情報の関連性についても考慮して個人情報露出量を算出する必要がある．3 種類の要素を持つ人数は 2 種類の要素を持つ人数に残る要素を持つ比率を乗算することで算出することができるが，このケースの場合は正確な比率を求めることが不可能であり，統計情報を用いて比率を表現しなければならない．統計情報を用いて正確な比率を表現できるのは統計情報の重なり部分の情報項目に属する要素の人数における統計情報から得ることができる人数と，全体における統計情報から得ることができる人数の比が等しい場合である．本方式では正確な比率を得ることが不可能であるためこの比率を正確であると仮定し，統計情報から求めることができる人数に乗算し個人情報露出量を算出する．1 種類の要素を持つ人数は場合分けの時に述べたように 2 次元の統計情報から得ることができる．

$$P_1=T(e_{11})$$
$$P_2=T(e_{11},e_{22})$$
$$P_3=T(e_{11},e_{33})$$
$$P'_1=P_2\times\frac{P_3}{P_1}=P_3\times\frac{P_2}{P_1}$$
$$IDM=-\log_2\left(\frac{P'_1}{U}\right)$$

ケース 12

2 次元の統計情報が 3 種類手に入った場合について述べる．

$$B=\{e_{11},e_{22},e_{33}|e_{11}\in E_1,e_{22}\in E_2,e_{33}\in E_3\}$$

$$T_4=T(E_1,E_2)$$
$$T_5=T(E_1,E_3)$$
$$T_6=T(E_2,E_3)$$

この場合統計情報が持つ情報項目はそれぞれ重なりがある．ケース 11 と同様にそれぞれの 2 次元の統計情報は独立ではなく重なりが存在しケース 11 では 2 種類の統計情報によってその比率を表現することを行っている．ケース 12 の場合 3 種類の統計情報が存在するため全てを用いて人数を表現しようとして情報項目の重なりによって比率を計算すると，その人数は前提となる 2 種類の要素を持つ人数と異なってしまう．よってこのケースでは全ての統計情報は用いらず 3 種類の統計情報の中から 2 種類を選び比率計算から人数を求め個人情報露出量を算出する．用いる統計情報の組み合わせは 3 種類存在し，ここでの統計情報の組み合わせは露出したどの要素によって統計情報が重なっているかということである．本来正確な値を算出していれば比率計算を行った人数は全て等しくなるはずであるが，統計情報が持つ要素の分布は一様ではないため統計情報から算出した比率が実際の比率と異なる可能性もある．値が異なった場合は統計情報を用いて最も人数が少ないものは最も個人を特定する可能性があるということであり，算出した人数についての評価が困難である場合最も個人を特定することが可能なものを適用することが適当であると考えられる．よってこのケースにおいて，比率計算を行った人数が異なった場合最も人数が少なくなった値を分解能力として個人情報露出量を算出する．

$$P_1=T(e_{11})\quad P_2=T(e_{22})\quad P_3=T(e_{33})$$
$$P_4=T(e_{11},e_{22})\quad P_5=T(e_{11},e_{33})\quad P_6=T(e_{22},e_{33})$$

$$P'_1=P_4\times\frac{P_5}{P_1}=P_5\times\frac{P_4}{P_1}$$
$$P'_2=P_4\times\frac{P_6}{P_2}=P_6\times\frac{P_4}{P_2}$$
$$P'_3=P_5\times\frac{P_6}{P_3}=P_6\times\frac{P_5}{P_3}$$
$$P'_{min}=min(P'_1,P'_2,P'_3)$$
$$IDM=-\log_2\left(\frac{P'_{min}}{U}\right)$$

3.3 4 種類以上の情報項目における個人情報露出量算出

4 種類の情報項目に属する要素が露出していた場合，手に入る 2 次元の統計情報は 6 種類，1 次元の統計情報は 4 種類手に入る．露出している情報数が増加するほど手に入る統計情報数も増加し，露出した情報とその中から統計情報を選ぶ場合分けの種類も増加する．モデルにおいて統計情報を用いた個人情報露出量を算出するにはその全ての場合について算出方式を検討し一般化を行う必要がある．以下に 4 種類の要素が露出していた場合の統計情報を用いた関連付けの一例を示す．要素間の直線はその要素が属する情報項目に関する統計情報を表している．

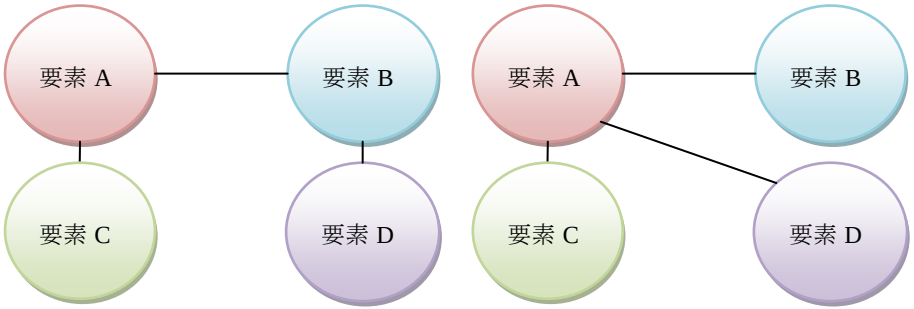

図 2　4 種類の要素における統計情報を用いた関連付け例

4.　算出方式と個人情報露出量の考察

4.1　実際の統計情報を用いた算出例

ここでは提案した統計情報を用いた個人情報露出量について実際の統計情報を用いての算出を試みる．算出を行うために以下の情報が露出していたとする．

住所：東京都
年齢：20 歳
性別：男性

今回手に入った統計情報は，「都道府県別人口」と「人口増減率と年齢各歳別人口」の 2 種類[6]である．

(1)　従来の方式による個人情報露出量の算出

従来の方式[4]では露出した情報に対して以下のように分解能力を設定している．この分解能力はそれぞれの情報項目に関して要素の分布を想定しそのエントロピーを分解能力して設定している．分解能力を加算した結果個人情報露出量は 18bit となる．

住所：鳥取県　　→11bit
年齢：80 歳　　→6bit
性別：男性　　→1bit

(2)　統計情報を用いた個人情報露出量の算出

露出した情報に対して住所と性別に関しては「都道府県別人口」，年齢と性別に対しては「人口増減率と年齢各歳別人口」の統計情報を用いることができる．よってこの場合はケース 11 の算出方式で個人情報露出量を算出する．またここでは統計情報と同じ年代の世界人口を得ることができなかったため，従来の方式と同様に世界人口は 70 億として算出を行う．

鳥取県で男性の人口は人数であるため小数点第一位を四捨五入し計算すると 309424 人と求められる．また 80 歳で男性の人口は 402000 人である．2 種類の統計情報は性別において情報項目の重なりが存在するため比率計算は要素を男性として，男性の中での統計情報から得た人数と世界人口における人数の比が等しいとして計算を行う．男性の人数は 62130000 人であるため，比率計算は

$$P'_1 = 309424 \times \frac{402000}{62130000} = 2002$$

となり，人数は 2002 人となる．この人数から個人情報露出量を算出すると約 22bit となる．

4.2　算出した個人情報露出量の考察

実際に存在する統計情報を用いて個人情報露出量を算出した．従来の方式で算出した値は 18bit であり割合は約 26 万人に 1 人，統計情報を用いて算出した値は 22bit であり割合は約 419 万人に 1 人と結果は大きく異なった．エントロピーでは情報項目が持つ 1 つ 1 つの要素の分解能力を表現できないため，エントロピーを用いて個人情報露出量を算出した場合，算出例のように分布が少ない要素を持つ時の情報の露出状況が実際のよりも少ない値となってしまう可能性がある．統計情報を用いて実際に要素を持つ人数を扱うことでこのような事態を防ぐことが可能であると考えられる．

4.3　算出方式の考察

露出した要素が属する情報項目数が 3 種類以下の場合についてそれぞれ手に入る統計情報を場合分けし，個人情報露出量を算出する方式について提案した．露出した情報数が増加すると手に入る可能性のある統計情報や統計情報の組み合わせの場合分けの数がさらに増加するため一般化することが困難になってしまう．よって実際に存在する統計情報の種類やその統計情報の年度などを調査し，統計情報を用いることが可能な情報項目について検討を行う必要がある．

3 種類の要素が露出している場合において 2 次元の統計情報を 2 種類以上用い，個人情報露出量を算出する際にその人数を求めるために統計情報から人数の比率を求めた．統計情報は 2 次元が持つ情報項目 2 種類に関してそれぞれの 1 次元の統計情報に

縮退させることは可能であるが, 2 種類の情報項目それぞれの 1 次元の統計情報から 2 次元の正確な統計情報を得ることは困難である．それは 1 次元の統計情報それぞれの比率と 2 次元以上の統計情報の各要素の人数の比率が等しいとは限らないからである．本稿では正確な人数の比率を入手することが困難であるため，人数の比率を手に入る統計情報から表現するために 2 種類の 2 次元の統計情報において，全体における 2 種類の要素を持つ人数と，ある要素を持つ人における種類の要素を持つ人数の比率が等しいとし, 統計情報が持つ情報項目の重なり部分の要素を用いることで算出を行った．この仮定が正しいならばケース 12 において用いる統計情報が異なっていた場合も，3 種類の要素を持つ人数を計算しているため結果は全て等しくなるはずであるが，実際の比率は複雑であるため算出した結果も異なっている．よってこの仮定から求めた比率と実際の比率を比較しどの程度誤差があるのかを検討しなければモデルの個人情報露出量を算出する方式に組み込むのは難しい．また, 3 種類の 2 次元の統計情報から 2 種類を用いて要素を持つ人数を求め，それぞれの人数が異なる場合は人数の最も少ない値を適用し個人情報露出量として算出した．これは仮定の下で起こりうる結果であり，複数の値が算出された場合その結果は全て同じ割合で起こりうるということである．これも仮定の比率と実際の比率を比較して誤差などを検討し，複数の結果が出た場合にも最適なものを選択し個人情報露出量として算出する必要があるが，本稿では検討に至らなかったため同じ割合で起こりうるならば最も露出している結果を選択することでモデルが知らせた露出状況が実際の露出状況よりも低いという事態を避けるためである．

既存の方式では情報項目それぞれから分解能力を bit として求め，それらを加算することで個人情報露出量を算出しているが，提案した方式では露出した統計情報から人数を求めその人数を bit 変換し個人情報露出量としているため，それぞれ統計情報を用いてそれぞれ分解能力として bit を求め，bit の演算から個人情報露出量を算出するための手段を提案する必要がある．

5. おわりに

本稿では，モデルにおける露出状況の統計情報を用いた算出方式についての提案を行った．結果として露出した情報数が少ない場合にはより正確な個人情報露出量を算出することが可能となった．

今後の課題として露出した情報数が増加した場合について条件を場合分けしそれぞれの個人情報露出量算出方式を定義し，一般化を行う必要がある．また，算出した個人情報露出量の評価と計算精度の向上を目指す．

参考文献

1) NRI セキュアテクノロジー株式会社: 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) Ryosuke Yasui, Atsushi Kanai,Takashi Hatashima, Keiichi Hirota : The Metric Model for Personal Information Disclosure, ICDS2010, vol.68, pp.112-117, 2010

4) 安井良介, 金井敦, 廣田啓一, 畑島隆: 個人情報記述レベルの定量化手法の検討, DPSWS2009, Vol.2009, (2009)

5) 安井良介, 佐藤和紀, 針谷友彰, 金井敦, 廣田啓一, 谷本茂明: ブログにおける個人情報漏えいレベルの定量化, 情報処理学会研究報告, 2008-EIP-043, Vol.2009, No.11, pp.9-16(2009)

6) 総務省統計局刊行, 総務省統計研修所編集: 日本の統計 2011