RL-005

個人情報の SEM (検索エンジン広告)価格に基づいた k -匿名化手法の提案

小栗 秀暢[†1]　　曽根原 登[†2]

現代において，BigData 分析における個人の機微情報（プライバシー）の扱いは，非常に大きな関心事になっている．そのようなプライバシーを保護するために有効な手段として k-匿名化の技術がある．多くの k-匿名化の研究アプローチは，選択肢群を数学的，かつ階層的に一律に分類するため，その情報損失量を減少させることが難しい．k-匿名化を実現するための必要な計算量や，選択肢組み合わせの複雑さから，一般的な WEB サービスのような，即時応答が要求されるサービスでは利用されていない．本稿では，効率的に BigData を管理し，活用する k-匿名化サービスの実用化に向け，個人情報の SEM（検索エンジン広告）の価格に基づく選択肢のクラスタリング手法を提案する．この手法によって，個人情報の選択肢ワードという定性的な価値を，価格という定量的な価値で計測でき，マーケティング的に不必要な計算回数を減少させることが可能となる．我々は，この手法を実際のサービスデータにて実験し，価値を計測したところ，高いセキュリティレベルと高い SEM 価値の両立が可能な場合が存在することが判明した．本方式を発展させることによって，一般の WEB サービスなどでも k-匿名化を用いることが可能なレベルまで計算量を減少させ，個人情報の蓄積にかかる費用も効率化されるだろう．

A k-anonymity Method based on SEM (Search Engine Marketing) Price of Personal Information

Hidenobu Oguri[†1]　　Noboru Sonehara[†2]

Privacy is one of the main concerns in Big Data managing especially when releasing datasets involving human subjects contain sensitive information. Therefore to protect the privacy of individuals, a model that is widely used for privacy preservation in managing Big Data, is k-anonymity. Most of the approaches to achieve k-anonymity suffer from huge information loss by generalization of continuous numerical attributes and categorical attributes they depend on the attributes hierarchical structure.
It is difficult to use conventional "k-anonymity" method in the real internet services, because of the computational complexity problem and value loss problem by information loss. This paper presents a k-anonymity method defined as clustering method based on SEM (Search Engine Marketing) price of personal information for the practical use of Big Data Management services. We would evaluate the value of k-anonymised qualitative data in SEM price that is quantitative indicator. Using this method, we can calculate only the necessary data and keep a k-anonymised level. We applied the method for real data, that there is a point to be compatible at a high level both k-anonymity and the price of SEM revealed. If we develop this method, a k-anonymity will easily handle the actual service on the internet, will be able to efficiently store personal data.

1. はじめに

近年の個人情報保護意識の高まりによって，通信・サービス・その他，あらゆる業種の企業が個人情報を保持し，その維持のために多額の費用と漏えいリスクを負っている．

個人情報の価値は主に拡散した際の被害額から算出されることが多く，日本において 2011 年に発生した個人情報被害総額は 1899 億円と算出されている．[1]

企業や団体はこれらの漏えい額に対して備えるため，常にセキュリティ設備の保護に追われており，SOC, PCIDSS, ISO27001 のようなセキュリティ基準を満たすため，多額の費用を投じている．これらの基準はインターネット上のセキュリティ問題が発生するたびに増加し，減少することは無い．一方，データの保管に関しては，実際には個人情報とそれ以外の情報は分離されずに管理されていることが多い．その結果，そのデータの市場価値（market value）が解らないため，使わないデータにもコストをかけて管理している．

実際には個人情報として保持されている情報の多くは価値評価がされておらず，データ漏えいのリスクに対応するために過度なセキュリティコストをかけている．

個人情報を利用価値と利用頻度で区分し，それぞれを適切なセキュリティレベルとアクセスレベルのデータベースに保持することでコストダウンを図ることが可能となる．

そのような個人情報の流通の手段として有望視されているのが匿名化の手法である．特に k -匿名化[2]の手法は，個人情報のセキュリティレベルを規定するための基準として多く研究されている．[2]

だが，k -匿名化を実際に利用したマーケティングサービスや，インターネットサービスは非常に少ない．現状で実サービスとして k-匿名化が用いられているのは，Optimal Lattice Anonymization を用いた医療情報の提供ツールや

†1 総合研究大学院大学 複合化学研究科 情報学専攻/ニフティ株式会社
The Graduate University for Advanced Studies, School of Multidisciplinary, Informatics Department, Tokyo, Japan. NIFTY Corporation

†2 国立情報学研究所
National Institute of Informatics

[5][9]，μ-argus を用いた，公共情報の提供ツールなどがある．[5]　これらのサービスは通常のインターネットサービスのように自由に使うことは難しい．扱える情報の種類や利用法が決められており，即時的な提供を行っていない．これは，各選択肢の組み合わせパターン数の問題に起因している．

k-匿名化を実現するためには，膨大な種類の選択肢同士の組み合わせ計算が発生する．(識別子，準識別子，センシティブ情報，非センシティブ情報等)

重要な個人情報は重要な属性が多くなっていく傾向があるため，必要なデータを組み合わせるほど，匿名化の複雑度が増していく．また，個人情報は常に変化することから，定期的に再計算可能な匿名化の方法が求められる．

インターネット上のリアルタイムサービスを提供するためには，通常 3 秒以内にページを表示しなくてはいけないとされている．2012 年の研究では，e-Commerce サイト等でのサイト表示が 2 秒以上になった場合，40%の顧客が別のサイトで買うことを考えるという実験結果がある．[3]

k-匿名性を用いたサービスが現実的に利用可能になるためには，２秒程度のレスポンスが求められる．図 1 はデータベースのデータ数とのレスポンス時間の推移について調査した図となる．横軸は 1 テーブル内のデータ行数を示し，縦軸はクエリに対するレスポンス時間を示す．調査対象は like クエリと count クエリになる．通常のアプリケーションにおいて，前処理として全体テーブル確認クエリを走らせる場合を想定した．

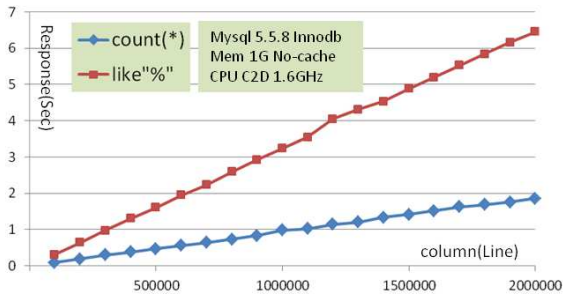

図 1： クエリ数とレスポンス時間の関係性

このグラフ[図 1]によると，like クエリのレスポンス時間は，column 数×9.4e-09 秒程度．count クエリで column 数×3.24e-06 秒程度が必要になる．この環境下で，2 秒以下でのレスポンスを実現できるのは，like クエリを引いた場合 60 万行．Count クエリの場合 200 万行程度となる．今後のサーバシステムの進歩に合わせ，これらの数字が改善されるとしても，現実的に k-匿名化がされたデータを快適にサービスで利用できるには 10^6～10^9 行レベルの計算量/選択肢量以下に抑える必要がある．

k-匿名化を実施した場合の論理的な限界回数は，2 以上の選択肢を持つ項目出現数の相乗値の合計になる．[図２]

K_n ： 計算回数

A_n ： 属性

$A_{n(c)}$: その属性内の選択肢バリエーション数

$$K_{(1)} = A_{1(c)} + A_{2(c)} + A_{3(c)} + \ldots A_{m(c)}$$

$$K_{(2)} = (A_{1(c)} * A_{2(c)}) + (A_{2(c)} * A_{3(c)}) \ldots$$

$$K_{(3)} = (A_{1(c)} * A_{2(c)} * A_{3(c)}) + (A_{2(c)} * A_{3(c)} * A_{4(c)}) \ldots$$

...

$$K_{(n)} = \Pi(A_{1(c)}, A_{2(c)}, A_{3(c)} \ldots A_{n(c)})$$

図２：k-匿名化の選択肢組み合わせ数

属性の組み合わせ数が多くなると相乗的に計算量が増加するため，属性が多い個人情報であるほど計算量は指数的に増加する．例えば，ニフティ社の持つ ISP 会員ユーザデータの属性種類は 47 種類あり，それぞれにユーザの属性（男性/女性，料金プランなど）が 2～49 種類のデータに分類されて保持されている．その計算量を試算すると，最大で 1.51e+38 個の組み合わせとなる．

通常のデータベース処理能力 10^9 と比較すると 10^29 程度の乖離がある．現在のスーパーコンピュータの計算速度が 10^18/秒程度であることから考えても，このレベルの能力差は数年程度で改善できるとは考えづらい．

Incognito[12]は，上記の全パターンを計算するのではなく，k 値＜２が検証されたパターンを含む計算を排除することで計算量を減少させている．だが，最終的に計算するべきパターン数が，全体の何%になるのかはデータ種類毎に異なるため，計算量の事前試算が難しい．

通常のサービス事業者の観点で考えると，このような最大値の試算結果は，バッチ処理スケジュールと，結果保存用 HDD の容量に影響を与える．サービス運営側にとって，計算結果が事前した想定容量よりも大きくなる場合，サービス停止の恐れがあるために推奨できない．膨大なパターンの中から，想定する時間内にバッチ処理と HDD 容量が収まるように，技術者が主体的に決定できる仕組みが必要である．

もう一つの問題点としては，このような膨大な計算量の存在に加え，その選択肢における最良の匿名化パターンの基準がないことである．特に定性的な言葉で作られた属性は抽象化パターンが無限に存在するため，どれを利用して良いかの基準が存在しない．

例えば，「年齢」という，比較的定量的なデータを用いた場合であっても，大きく分類して３種類のデータの抽象化パターンが存在する．

1. 数学的な階級化：パレート分析，スタージェスの公式やデシル分類など，計算で階級を作成．
2. マーケティング的な階級化：年代や学校など，一般的な分析結果と対照するための区分で階級を作成．
3. 限定的な用途の階級化：特定のマーケティング会社が利用するためだけの階級区分．例えば飲酒できる年齢以上にのみ広告を送りたい場合などに，20 歳以上/20 歳以下

の分類が必要になる．

この上記の３パターンの中でも，分類の種類は無限に存在しており，どのパターンが目的に合致しているのかが判定できず，また，事前に社会ではどのような分類にニーズがあるのかを知ることは出来ない．

実際の業務で k -匿名化を行う場合, データの抽象化を行った際には，データ内のノイズ的な分類を「その他」などの意味のない分類に入れてしまうことが多い．

処理するデータが大きくなる程，データの一次処理と実際に利用するマーケターは分業化されることが多くなる．安全性を重視する k-匿名化と利用価値を重視するマーケターにはギャップがあり，分析に必要な項目が削除されてしまう可能性がある．

これらの問題は一般的な個人情報データに対して，どのような抽象化を行うことが適当であるかの基準が存在しないことに起因している．

この k-匿名性の計算回数を，通常のデータベースで利用可能な回数まで減少させ，かつ，何らかの基準となるデータと対照して，利用性をなるべく損なわないことが，実サービスへの匿名化適用のために必要である．

本文は，計算回数を減少させ，マーケティング的な価値減損を最小限に抑える k -匿名化について検討し, 実データに適用し，その効果を検証する．

2. 過去研究

k-匿名化についての研究は多くされている．まず，匿名化とは，ユーザを特定できないようにパーソナル情報を加工することである．

ここでパーソナル情報とは「属性」と「属性値」として表現されるユーザに関する情報であり，あるユーザのパーソナル情報をテーブルのレコードとして表現する.そして，単一の属性ではユーザを特定できないが，複数組み合わせるとユーザを特定できる可能性のある属性の組合せを準識別子(quasi-identifier,QID) と呼ぶ．

また，ユーザを特定された状態で開示されることが望ましくない属性をセンシティブ属性(sensitive attribute : SA) と呼ぶ．

この時，もし攻撃者があるユーザの QID の属性値を知っていたとすると，そのユーザのレコードを特定できてしまい, SA の属性値を知られてしまう．これを防ぐために，QID の属性値を一般化して，より抽象的な値にする方法が知られている．そして，QID の属性値によって識別されるレコードが少なくとも k 個以上ある場合，そのテーブルは k-匿名性を満たすという[2].

k-匿名化を実現するための手法として， Datafly 方式[4] [7]や μ -Argus 方式[4] [8]などのアルゴリズムが主に使われており，公共データや医療データの匿名化アプリケーションとして提供されている．

それらの k-匿名化手法は，データの出現数に合わせて切り落としや抽象化を行い，データの出現数を k 値以下にする．

多くの匿名化アルゴリズムは，上記のような情報の変更の組み合わせにデータを Re-coding し，抽象化を行うことで利用者を特定するデータの組み合わせ出現数を k 値以下にすることで成立する．

Re-coding は大きく分けて局所的な変更である Local Suppression と，選択肢全体の組み合わせから変更を行う Global Recoding の二種類が存在する．主にデータの組み合わせを考える場合は Global Recoding によって抽象化レベルを測りながら複数の抽象化判定を試す形になることが多い．Global Recoding は，非常に多くの手法や変更するべき内容が存在する．そのため，正しい手法を効率よく探し出すのは難しい．

これらの k-匿名化の手法の評価基準は，主に情報損失のレベルを基準としている．k-匿名化におけるセキュリティ（安全性）のレベルを k 値のレベルと規定すると，現在の評価軸である情報損失レベルとは常に相反する状態になる．

これらの評価基準は社会の状況によっても変化する．例えば, 2012 年の日本で発生した T ポイントツールバーの例などが挙げられる．個人情報の収集に関する規定を目立つ場所に記載していなかったということで，多くのユーザがクレームを送り, 事業方針を変更した事例がある. [11] このような問題が発生した場合，事件の影響は問題を起こした企業だけでなく，同様の事業を行っている同業他社に対してもユーザからの質問や疑いが多くなり，事業方針を変更せざるを得なくなる．

今後，政府が安全性に関する法的なガイドラインを定め, 仮に k の値が法律の定める基準を満たしていたとしても，大きなセキュリティ事故などが発生するたびに基準は変化する．常に変化するユーザの要望に対して，個人情報の安全性レベルを柔軟に対応させることが求められる．

だが，情報損失のレベルについて，ユーザが使っても良い，と考えるレベルの抽象化と，実際にマーケティング担当者が使いたい, と考えるレベルにはギャップが存在する．その指標は情報損失量では評価することができない．

そのため，k-匿名性における評価の指標として，情報損失量ではなく，実際に使われるマーケティング的な価値や顧客の状況に応じて変化する指標を提案する．

k-匿名化の安全性と有益性を評価する指標を，実際の事業で使用できる概念によって行うことによって，k-匿名化の事業利用が可能となる．

3. 個人情報のＳＥＭ価格をベースとした k-匿名化手法 の提案

個人情報に対して，匿名化処理を実施した後の情報価値の変化について，検索エンジン広告（SEM）の価格価値によって評価する方法を提案する．

現在，定性的なマーケティングの世界は，検索エンジンマーケティング（SEM）が標準として確立されている．例えば，選択肢を抽象化して k -匿名化を行った場合，元のデータと比べてマーケティング的な価値にどれだけの変化が発生したのか，具体的な減損レベルを金額として計測できる．

また，ＳＥＭの金額的価値は現在のユーザの嗜好とも合致しているため，常に変化している．例えばクリスマスシーズンに関する情報は夏休み期間に必要とされないだろう．ならば，その期間に不要な情報を排除することによって，計算量や無駄なデータのアーカイブ量を減らすことが可能になる．

将来的に k -匿名化に関するデータの売買が SEM のように一般的になれば，それを利用することの方が効率的である．だが現在，そのような市場が存在しないことから，SEM を代替指標として用いて換算するアプローチを実施する．

SEM の価格決定プロセスはオークション形式である．入札会社はその語に対して投じることのできる広告費の上限値を入札し，他社と金額を競う．

他社広告費:$C_1, C_2 \ldots C_n$　($C_1 > C_2 \ldots > C_n$)
自社最大広告費：M
$C_n > M$ の場合，広告費は
[M*広告クリック数]
$M > C_n$ の場合，広告費は
[(C_n+ 最低金額単位)*広告クリック数]
広告クリック数＝0 の場合，金額は最低値

図３： SEM の金額決定プロセスサンプル

SEM のオークションは，１つの商品を争うのではなく，広告表示位置の順番を争う仕組みである．そのため，SEM の価格は他の会社の価格との関係で決まり，自社の入札価格が低い場合，表示位置が下がる．[図 4]

これらの価格は，全て買い手（広告をクリックするユーザ）と売り手（広告出稿会社）が存在しているときにのみ成立するものであり，過去において一度もユーザがクリックしたことのない広告は０円となる．

図４：ＳＥＭ結果画面のサンプル

現在では，非常に多くの事業者がＳＥＭを利用している．そのため，ＳＥＭの価格を参照すれば，どのような業種でどのような概念に対して売買が発生しているかが解るため，経済状況を表現する指標となっている．

実際に選択肢を全てＳＥＭの価格に変換してみると，k -匿名化を実施した場合にオリジナルのデータよりも価値が高い状況が存在することが明らかになった．

以下の表は実験的に行ったものだが[表 1]，SEM の価格を比較すると，抽象化後のデータの方が，安全性（k 値）も高く，広告的価値が高い場合があることが解る．

k=1

年齢	人数	SEM価格	合計価格
12才	2	@¥18	¥36
13才	7	@¥33	¥231
14才	1	@¥0	¥0
15才	3	@¥29	¥87
合計	13		¥354

抽象化・匿名化

k=13

年齢	人数	SEM価格	合計価格
10代	13	¥37	¥481
合計	13		¥481

表 1. SEM 価格と k 値の変化サンプル

必ずしも，詳細なデータ＝価値あるデータではないため，一番高い経済効果のあるレベルまで抽象化する．多くは抽象化によって各選択肢の種類数は減少し，選択肢あたりの顧客存在数は多くなるため，必然的に安全性は向上する．

このような安全性と利用価値の両立する選択肢群を辞書として用意しておき，各種の匿名化サービスと組み合わせて利用することにより，k -匿名化を実サービスに利用することが可能になる．

まず，各選択肢についての価値評価を行ってみる．

属性Ｓに所属するユーザ全員に広告を打つ場合の価値は，[各選択肢の人数×各選択肢の広告単価]となる．広告単価を SEM ツールから取得し，各選択肢に人数と単価を掛け，合計したものを選択項目全体の広告価値と考える．

$S_{(e)}$:Attribute S value
$S_{(k)}$:k-anonymised S value
$s_1, s_2, \ldots s_n$: QID
$c_1, c_2, \ldots c_n$: number of QID
$e_1, e_2, \ldots e_n$: SEM price of QID

$s_{1\,(e)} = c_1 * e_1$
$s_{2\,(e)} = c_2 * e_2$
...
$s_{n(e)} = c_n * e_n$

} Multiply the SEM Price and number of people

○Amount of Attribute Value

$S_{(e)} = \Sigma\{ c_i * e_i\}$

○Reduction rate of k-anonymized data Value

$M_{(k)} = S_{(k)} / S_{(e)}$　　$M(k) = S(k) / S(e)$

図５： 匿名化後の価格変化の評価方法

これにより，例えば自社のユーザ群の中で特徴的な出現率を記録するユーザ群が存在していたとしても，出現率が低いためにノイズとして排除される可能性を排除できる．匿名化によって自動的に削除されたデータの中に，マーケティング的な価値が高いものが存在していた場合，その価格の減損率を確認し，一定以上の価値減損であった場合に，その変更を取りやめることができる．

例えば，実験によって職業のアンケート調査を行ったところ「教員」の価値は非常に高いことが判明した．だが，世間的に教員に従事している人の数は非常に少なく，少ないが故に広告的価値がある場合もある．

もし，自社のデータを調査した際，教員の数が少なく，使いにくいと考えた場合，他のデータと比べて相対的に多いならば，そのデータの特徴として価値化される．

注意が必要な点として，この手法は語の価値のみを判定しているため，抽象化後のデータの方が元データよりも価値が高くなる場合が存在する．ここでの価格はあくまでも価値評価のための基準としての価格である．本来ならば，データを売る場合は詳細で使いづらいデータであっても，オリジナルデータの方が価値は高いだろう．

SEMPB k-anonymity Method::={

Input: テーブル T を匿名化する，P_1 P_2 は選択肢変更のバリエーションに SEM 価格を算出して加えたもの．

Output: テーブル T_2 ：k-匿名化され，かつ SEM 価格が最大のもの．

for i_1 = 1 to l **do**
// l : 抽象化パターンの存在数 $(P_1,P_{2\ldots\ldots})$
for i_2 = 1 to m **do**
// m : それぞれの抽象化パターンの選択肢数 P_i
int $P[i_2]$
// n : 各パターンのデータ行数 P_i
for i_3 = 1 to n **do**
int r_{i3} // Pattern_i_2 の SEM 価格.
int c_{i3} // Pattern_i_2 のデータ出現数.
int $p_{i2} = p_{i2} + \{ r_{i3} * c_{i3} \}$
// Pattern_i_2 の SEM 価格の総額
$P[i_2] = p_{i2}$ // SEM 価格を配列化
// $P[i_2] = \{ p_1, p_2, \ldots, p_{i2},\}$
$i_3 = i_3 + 1$
end for
$i_2 = i_2 + 1$
end for
// Temporary table R にデータを入れ込む
Connect Database
query = Update Table R set price = $\{ p_1, p_2, \ldots, p_{i2},\}$;
$i_1 = i_1 + 1$
end for

// SEM 価格から計算プライオリティを設定
Int k = 1 // k-anonymised check
Int *num* = 1
while k <2 **do** // 2-anonymised level
Connect Database
query = Select *word pattern* from R order by *price* desc limit *num*, 1 ;
Table copy T into T_2
query = Update T_2 set $\{At_1,At_2.. At_{i1}\}$=*Word pattern* of Max price
k = Select count(min($At_1,At_2.. At_{i1}$)) from T_2 group by ($At_1,At_2.. At_{i1}$) ;
num = *num* +1
end while

Response: T_2 : k-匿名化が検証され，かつ SEM 価値が最も高い組み合わせとなる **}**

T：匿名化するべきデータ

	At1	At2
ID	年齢	出身地
1	12	東京
2	14	NewYork
3	16	京都
4	12	大阪
:	:	:

T2：匿名化後データ

At1	At2
年齢	出身地
10代	日本
10代	アメリカ
10代	日本
10代	日本
:	:

表２： Input テーブルと Output テーブル

P1： 年齢データの抽象化パターン

Original	Price	Pattern1	Price1	Pattern2	Price2	・・	count
12才	¥367	小学生	¥414	10代	¥467	・・	33
13才	¥563	中学生	¥446	10代	¥467	・・	25
14才	¥304	中学生	¥446	10代	¥467	・・	48
15才	¥508	中学生	¥446	10代	¥467	・・	60
16才	¥360	高校生	¥514	10代	¥467	・・	47
：	：	：	：	：	：	：	：

P2： 出身地データの抽象化パターン

Original	Price	Pattern1	Price1	Pattern2	Price2	・・	count
東京	¥261	日本	¥441	アジア	¥322	・・	25
京都	¥404	日本	¥441	アジア	¥322	・・	10
大阪	¥365	日本	¥441	アジア	¥322	・・	8
New York	¥384	アメリカ	¥234	北米	¥130	・・	48
San Francisco	¥342	アメリカ	¥234	北米	¥130	・・	36
：	：	：	：	：	：	：	：

R：価格による優先度決定テーブル

Base	ID	Word Pattern	Price
At1	Original	{12,13,14…}	¥148,838
At1	Pattern1	{小学,中学,高校…}	¥218,361
At1	Pattern2	{10代,20代,…}	¥241,586
：	：	：	：
At2	Original	{東京,大阪, …}	¥93,574
At2	Pattern1	{日本,アメリカ…}	¥107,167
At2	Pattern2	{アジア,北米…}	¥60,008
：	：	：	：

表３：アルゴリズムに必要なテーブル群

〇事前準備

現状の選択肢群，及び匿名化候補となる選択肢群を SEM ツールに入れ込み，広告価値を算出する．(make table: P1,P2 …)

〇匿名化プロセス

1. 各データの出現数と価格を用いて，属性全体の価格を求める．
2. 各属性の価格をデータベースに格納し，価格が高い順番に抽象化する語の候補を取得する．
3. テンポラリーテーブル上にユーザデータをコピーして，価格が高いワードに書き換える．
4. その状態で匿名化の検定を行い，2-匿名状態以上である時にそのパターンを採用する．

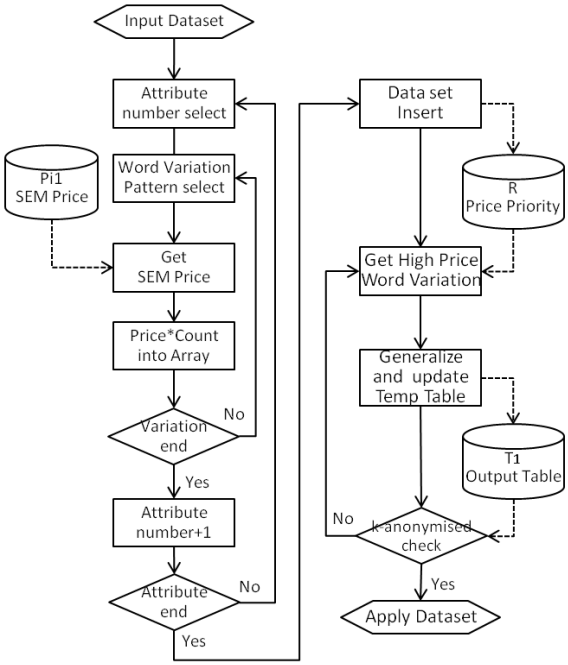

図６： 提案アルゴリズムのフロー

4. 本手法の実データへの適用例

上記の手法について，実際に業務で利用した個人情報データを k 匿名化して個人特定できない状態に変換し，価値変化を確認する実験を行った．実データのプロパティは図 9 の通り．

属性	選択肢種類	k値	SEM価格	計算時優先度
職業	10	86	¥1,307,878	3
年代	8	17	¥749,360	5
住所	47	39	¥690,763	4
年齢	73	1	¥550,713	6
性別	2	4280	¥549,109	1
婚姻	2	5464	¥524,724	2
年収	11	29	¥0	7
総額			¥4,372,547	

表４： オリジナルデータのプロパティ

検索広告の単価調査には google の SEM ツール (https://adwords.google.com/ko/TrafficEstimator/Home) を利用した．（2012 年 10 月 19 日実施）

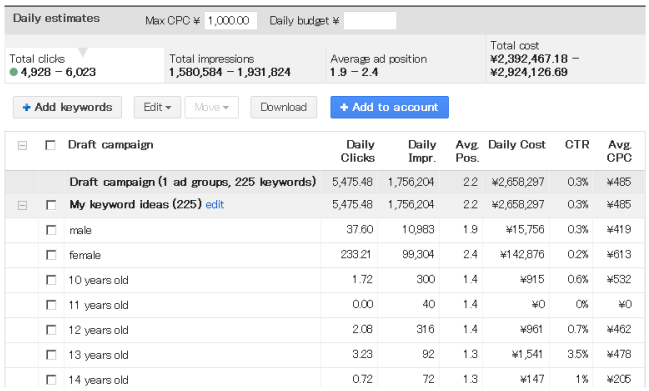

図７： Google SEM ツール

(https://adwords.google.com/ko/TrafficEstimator/Home)

具体的には，以下のような手順によってデータの抽象化処理を行ない，一番価値が高い形で匿名化済データとして保持する実験を行った．この手法によってデータの価値を定めた結果は以下の通り．

この手法によって，本データをｋ-匿名状態にした上で，データの利用価値を最も損失しない形で抽象化した場合，価値が 60.6%に変化したことが判明した．

属性	変更前価格	k値	匿名化実施	変更後価格	変化量
職業	¥1,307,878	86	抽象化	¥951,813	72.78%
年代	¥749,360	17	抽象化	¥622,910	83.13%
住所	¥690,763	39	データ削除	¥0	0.00%
誕生日	¥550,713	1	データ削除	¥0	0.00%
性別	¥549,109	4280	そのまま利用	¥549,109	100.00%
婚姻	¥524,724	5464	そのまま利用	¥524,724	100.00%
年収	¥0	29	データ削除	¥0	0.00%
合計	¥4,372,547	2		¥2,648,556	60.57%

表５：匿名化実施後の価格変化量

最終的にＩＤを削除し，連結不可能匿名化したことで本匿名化作業は完了された．

5. マーケティング価値とｋ-匿名性の関連性調査

本データを用いて，ｋの値が異なるデータ抽象化パターンをいくつか作成し，それぞれの価値変化を調査する実験を行った．

実験に使用したデータは，ユーザの年齢データで，全部で５種類の抽象化案を作成した．

1. 年齢（元データ）
2. 年代区分
3. ５年区切り
4. 所属区分（小学生/中学生等）
5. 個別調整（15 才以下/60 才以上を包含）

上記のデータに対して，データの価値がどのように変化するか調査を行った．

	年齢	年代	5年区切り	所属	カスタマイズ
サンプルデータ種類	12才	10代	10-14	小学生	15才以下
	13才	20代	15-19	中学生	16才
	14才	30代	20-24	高校生	17才
	15才	40代	25-29	大学生	18才
	…	…	…	大人	…
	83才	70代	65-69	退職	59才
	88才	80代	70 over	null	60才以上
データ種類数	73	8	13	6	46
最少出現数(k)	1	17	70	4	79
最大出現数	317	2746	1469	8440	866
出現数の平均	150	1367	841	1822	238
出現数の標準偏差	106	1024	475	2980	111

表６： 各パターンの匿名化後の価値

それぞれに対して価値評価を行ったところ，上記のような結果が出た．セキュリティレベル（ｋ値）とＳＥＭの価格/価値に対しては有意な関係性は無い．「年代区分」は，広告価値も安全性の両方が高い状態の高い状態になる．「５歳区切り」は情報損失が少ないにも関わらず，広告価値は０であるという結果が出た．

これらのデータをｋ値の標準偏差や平均値等と比して並べたものが上の図である．マーケティング価値とセキュリティレベル（ｋ値）の有意な結果は見えなかった．これにより，安全性の基準とマーケティング的な価値には特に関連性は無く，その抽象化後の価格とその安全性を確認することで，データの価値を定めることが可能となる．

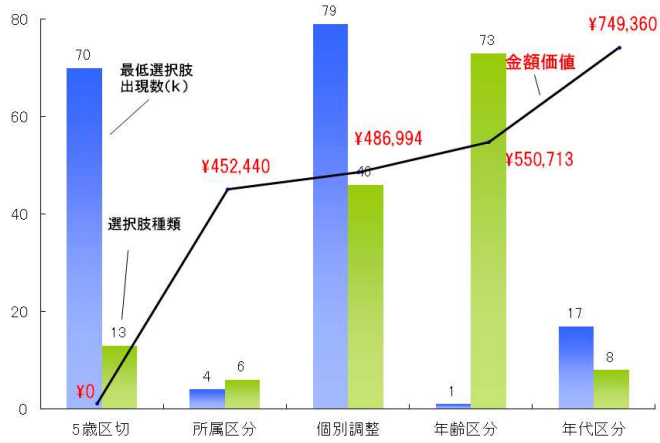

図８：金額価値と選択肢出現数との関係

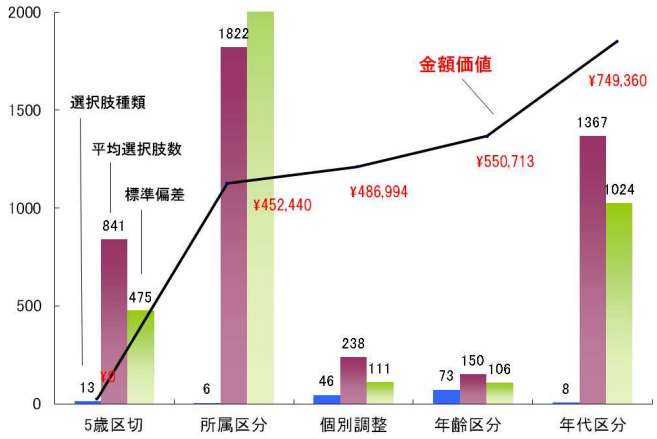

図９：金額価値と選択肢数との関係性

6. 結論

本実験の結果により，k-匿名化処理済のデータをＳＥＭの価格によって価値化し，評価の指標とすることによって，実データに対する抽象化を実施する際の優先度をつけることが可能となった．

また，データの利用価値とｋ-匿名レベルとの間には特に関係性は存在せず，定性的な価値基準が存在すれば，安全と価値の両立ができる点が存在することが判明した．

このようにデータ分析の価値によるプライオリティを明確化することにより，今後，匿名化データの価格データが整備されることによって，その時々の最適な匿名化を自動的に実施する仕組みが可能になると考える．

我々は，Lattice structure[10]の構造を参考にして，データの分析プライオリティを定め，一定価格以下のデータ分析を省くアルゴリズムを提案する．

また，検索エンジン広告を用いることによって，新しい語の出現に対して抽象化のパターンを組むことも可能となる可能性があり，自動的な語の意味解析と合わせて実現するという期待が持てる．

価格と安全性の最適解を求めることによって，そのデータの最大価値を測ることも可能となる．かつ，データ流通の観点からも，価格の高い匿名化済データのみを通常利用するデータベースに保持しておき，元データとの独立した運用を実現することが出来ることでコストの削減効果も期待できる．

だが，反面，ＳＥＭの価格を利用することから，問題点も多く存在する．

1. 価格の決定が毎日のように変化するため，日々の対応が求められる．

2. 価値の定義が検索エンジンに入力したユーザであることから，「will=検索して知りたい事項」と「be＝現在の自分の姿」の違いによるデータの扱いが難しい．

3. 検索エンジンで価格が付いていない概念や，同意義語への対応

4. この方式はあらゆる言語に適用できるのか

上記のような問題について，今後も検討していく必要がある．

参考文献

1) 2011 年情報セキュリティインシデントに関する調査報告書,NPO Japan Network Security Association(JNSA)日本ネットワークセキュリティ協会 セキュリティ被害調査ワーキンググループ

2) Sweeney, L, k-anonymity: a model for protecting privacy, Int. J. Uncertainty, Fuzziness and Knowledge-Based Systems, Vol. 10, pp. 557.570 (2002)

3) Akamai Technologies and Helen Yang and Noelle Faris, Akamai Reveals 2 Seconds as the New Threshold of Acceptability for eCommerce Web Page Response Times, September 14, 2009

4) Mitsubishi Research Institute, Inc. 情報技術研究センター 松崎和賢,データ匿名化の現状に関する一考察. 医療・統計分野を中心とした国内外の動向.2011-7-8

5) 日本情報経済社会推進協会(JIPDEC),パーソナル情報の利用のための調査研究報告書,2011-3

6) Josep Domingo-Ferrer, Francesc Sebe and Agusti Solanas, A polynomial-time approximation to optimal multivariate microaggregation. Comput. Math. Appl.,55(4):714-732, 2008.

7) Latanya Sweeney, k-anonymity: a model for protecting privacy. International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, March 2002.

8) Marek P. Zielinski and Martin S. Olivier, How appropriate is k-anonymity for address-ing the conflict between privacy and information utility in microdata anonymisation,

9) El Emam K and Dankar FK and Issa R and Jonker E and Amyot D and Cogo E and Corriveau JP and Walker M and Chowdhury S and Vaillancourt R and Roffey T and Bottomley J, A globally optimal k-anonymity method for the de-identification of health data, September--October 2009

10) Daniel C. Barth-Jones, How should we understand re-identification risks under HIPAA?, 2011

11) 日経コンピュータ, 2012/8/30, p.10

12) Kristen LeFevre David J. DeWitt Raghu Ramakrishnan, Incognito: Efficient Full-Domain K-Anonymity,SIGMOD '05 Proceedings of the 2005 ACM SIGMOD international conference on Management of data, Pages 49-60, 2005