第六論文

新井健介・河野和宏・馬場口登「TF-IDF 法によるユーザへの情報推薦のための匿名化処理」電子情報通信学会技術研究報告 vol. 115、no. 38、IT2015-10、EMM2015-10、2015年、51〜56 頁

一般社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
IEICE Technical Report
IT2015-10,EMM2015-10(2015-05)

TF-IDF 法によるユーザへの情報推薦のための匿名化処理

新井　健介†　　河野　和宏††　　馬場口　登†

† 大阪大学大学院工学研究科　〒 565–0871 大阪府吹田市山田丘 2–1
†† 関西大学社会安全学部　〒 569–1098 大阪府高槻市白梅町 7–1
E-mail: †arai@nanase.comm.eng.osaka-u.ac.jp, babaguchi@comm.eng.osaka-u.ac.jp ††k-kono@kansai-u.ac.jp

あらまし　収集した個人データを分析してユーザに情報推薦する場合，ユーザのプライバシーを保護するため，誰であるか特定されないよう匿名化処理を施した上で個人データを利用する必要がある．しかしながら従来の匿名化処理では，匿名化された個人データの選択基準として情報損失の度合いのみとなっているため，実際に情報推薦に活用することを想定した上での基準となっておらず，情報損失は低いが情報推薦には利用しづらい匿名化されたデータが選択される可能性がある．そこで本稿では，TF-IDF 法により，情報推薦のために必要とする個人データの属性は必要以上に汎化されないよう評価することにより，推薦対象となる属性は残しつつ匿名化処理を可能とする手法を提案する．シミュレーション結果では，推薦情報を変更すると，出力される匿名化されたデータも推薦対象の属性が残るよう変化し，推薦情報に応じた処理が実現できていることが確認できた．

キーワード　k-匿名化，TF-IDF 法，Mondrian アルゴリズム，情報推薦，プライバシー保護

Data Anonymization for Information Recommendation Based on TF-IDF Method

Kensuke ARAI†, Kazuhiro KONO††, and Noboru BABAGUCHI†

† Graduate School of Engineering, Osaka University　2-1 Yamadaoka, Suita, Osaka, 565–0871 Japan
†† Faculty of Safety Science, Kansai University　7–1 Hakubai, Takatsuki, Osaka, 569–1098 Japan
E-mail: †arai@nanase.comm.eng.osaka-u.ac.jp, babaguchi@comm.eng.osaka-u.ac.jp ††k-kono@kansai-u.ac.jp

Abstract　This paper proposes a data anonymization method considering information recommendation for users. Attaching importance to the personal data attributes required for information recommendation by TF-IDF method in the case of k-anonymization, we provide the k-anonymized personal data appropriate for recommendation. We examine through simulation that the personal data anonymized and output by our method include information needed for recommendation.

Key words　k-Anonymization, TF-IDF, Mondrian Algorithm, Information Recommendation, Privacy Protection

1. ま え が き

近年，個人データを蓄積し，個人データの分析結果を情報推薦に活用するサービスが普及しつつある．個人データとは，氏名や免許証番号，住所等の個人を一意に識別可能な個人情報だけでなく，位置情報や移動履歴，購買履歴のような単独では個人は特定できないが個人に関連する情報といった，個人の属性に関する情報を記載したデータである．個人データからは，年齢ごとの商品購入の傾向の違いなど，情報推薦に活用可能な情報を数多く得ることができる一方，個人データをそのまま活用した場合，他のデータベースとの名寄せにより個人が特定されたり，知られたくない情報が暴露されたりする可能性があるため，ユーザへのプライバシーに配慮した処理が求められる．

個人データを開示するためのプライバシー保護処理は，プライバシー保護データパブリッシング (PPDP) [1] と呼ばれ，特に，データに対してより抽象的な値に置き換える汎化処理やデータそのものを削除する抑圧処理を行うことにより，個人を一意に特定できないデータへ変換する処理を匿名化処理という．匿名化処理を行う際，処理後のデータは，k-匿名性 [2]，l-多様性 [3]，t-近傍性 [4] の観点から評価することが多いが，本稿では，最も基本的な指標である k-匿名性，および k-匿名性を実現する k-匿名化処理を対象とする．

k-匿名化処理 [5]〜[7] とは，k 人以上が同じ属性値の組み合わせを持つようにデータを汎化もしくは抑圧する処理である．k-

匿名化された結果，匿名化処理されたデータから実際の人物を特定しようとした場合，k 人以上が同じ組み合わせを持つ，つまり，$1/k$ 以下の確率でしか実際の人物を特定できない状態を実現することができる．

しかし，匿名化処理されたデータは，ユーザへの情報推薦等の利活用に用いることを想定しているにもかかわらず，匿名化処理されたデータが，必ずしも利活用しやすいデータとならない問題がある．k-匿名化処理ではその性質上，情報損失が発生することから，一般的に Prec 評価値 [7] や DM [8] 等のデータの詳しさを表す指標を用いて，匿名化されたデータの中で可能な限り情報損失が起きていないデータが選択される．例えば，表 1 の個人データを匿名化した場合，表 2 のような匿名化も考えられるが，表 3 のように一部の属性値だけ詳しくなっている状態のデータも作成可能である．利活用の環境によっては，どちらのデータが適切かは変わってくるため，必ずしも情報損失が最も少ないデータが適切とは言えない．

ここで，馬場口らが提案する HIFI [9] が想定するような，駅やテーマパーク，商業施設といった限定された空間を考えると，サービス提供者はどのような情報をユーザに推薦したいかをあらかじめわかっていることが多い．例えば，商業施設では，衣服類を販売する店舗がユーザに対して衣服の情報を推薦したい場合，性別や身長に応じて適切な衣服情報を推薦したいことが想定されるため，表 2 のような両者の性別が残っているデータを用いて推薦することが望ましい．反対に，表 3 のような生年月日の属性を詳しく出力し性別や身長の属性を詳しく出力していないデータが選択された場合，ユーザ毎に応じた適切な情報を推薦することができない．言い換えると，推薦したい情報がわかっているのであれば，その推薦情報の対象グループに合わせた匿名化処理を行うことが求められる．

そこで本稿では，推薦情報が対象とする人物の属性値が決まっており，推薦情報と対象の属性値に関する規則が存在すると仮定出来る環境において，規則から推薦情報が要求する属性を TF-IDF 法 [10] によって評価することにより，評価の高い属性を必要以上に汎化せず，詳しく出力する k-匿名化アルゴリズムを提案する．TF-IDF 法とは，文書検索において，ある文書中における単語の出現頻度を表す TF と，他の文書中におけるその単語の出現頻度の逆数である IDF を用いて，その単語がその文書中に固有であり，かつ重要であるかを示す指標である．TF-IDF 法によって推薦情報が個人データ中の人物にとって固有であり，重要であるかが評価できるため，推薦に必要とされる属性を基準にして匿名化処理を行うことができる．さらに，疑似的に生成した個人データと推薦情報を用いてシミュレーションすることにより，提案手法が推薦情報の分類に用いる属性を活かしたまま匿名化処理することが可能であることを検証する．

2. k-匿名化処理の概要

匿名化処理とは，個人データ内の属性値に対して，属性値を削除する抑圧処理や抽象的な属性値に置き換える汎化処理などのデータ変換を実施することにより，個人を一意に特定できないようにする処理である．本節では，匿名化の際に満たすべき性質と，本稿で利用する匿名化アルゴリズム Mondrian の概要，さらに匿名化後のデータの選択の際に利用される，匿名化処理によるデータの情報損失の割合を示す指標を述べる．

表 1　匿名化前の個人データの例.

ID	氏名	性別	生年月日	職業	身長
1	佐藤太郎	男性	1973 年 1 月 1 日	自衛官	176cm
2	高橋健	男性	1997 年 2 月 2 日	研究者	172cm
3	鈴木一郎	男性	1993 年 5 月 5 日	清掃従事者	174cm
4	小谷零	女性	1977 年 8 月 8 日	管理的公務員	159cm
5	山田花子	女性	1993 年 5 月 1 日	一般事務従事者	153cm
6	谷ひろみ	女性	1997 年 2 月 9 日	教員	155cm

表 2　2-匿名性を満たす匿名化後の個人データの例.

ID	氏名	性別	生年月日	職業	身長
1	*	*	1970 年代	*	*
2	*	男性	1990 年代	*	170cm 代
3	*	男性	1990 年代	*	170cm 代
4	*	*	1970 年代	*	*
5	*	女性	1990 年代	*	150cm 代
6	*	女性	1990 年代	*	150cm 代

表 3　2-匿名性を満たす別の匿名化後の個人データの例.

ID	氏名	性別	生年月日	職業	身長
1	*	*	1970 年代	*	*
2	*	*	1997 年 2 月	*	*
3	*	*	1993 年 5 月	*	*
4	*	*	1970 年代	*	*
5	*	*	1993 年 5 月	*	*
6	*	*	1997 年 2 月	*	*

2.1 k-匿 名 性

k-匿名性とは，匿名化されたデータから実際の人物を特定しようとした場合，データ中の k 人以上が同じ属性値の組み合わせを持つために区別できず，$1/k$ 以下の確率でしか実際の人物を特定できない性質であり，k-匿名性を満たすようにデータを変換することを k-匿名化と呼ぶ．データ変換の処理としては，より抽象的な値に変換する汎化処理，属性値を消去する抑圧処理，雑音を付加する摂動処理等があるが，k-匿名化処理においては，氏名や免許証番号のような個人を一意に特定可能な属性は抑圧，年齢や性別のようないくつかの項目を組み合わせることにより個人を一意に特定可能な属性は汎化またはこれ以上汎化できない場合は抑圧することにより，個人を一意に特定できないデータに変換する．

表 1 に示す個人データに対して，2-匿名性を満たすよう変換した個人データの例を表 2，表 3 に示す．表 1 の個人データでは，氏名が個人を一意に特定可能な属性，性別，生年月日，職業，身長が組み合わせにより個人を一意に特定可能な属性であるため，k-匿名化処理を施した場合，氏名は抑圧，性別，生年月日，職業，身長は汎化を試した後，汎化しても k-匿名性を満たさない場合，抑圧される．表 2 では，ID1 と ID4，ID2 と ID3，ID5 と ID6 の属性値が一致し，一人だけのデータは存在

しないため 2-匿名性を満たし，表 3 では，ID1 と ID4，ID2 と ID6，ID3 と ID5 の属性値が一致しているため，2-匿名性を満たしているデータとなっている．

2.2 k-匿名化アルゴリズム

k-匿名化処理では，個人データ内の属性を次元とした多次元空間において，各グループには最低 k 人いるよう各レコードを分類した後，各グループ内のレコードの属性値を，グループを代表する属性値（例：数値の範囲やより大きな区分の値）に変換する．グループの作成には，空間分割による手法 [5], [6] とクラスタリングによる手法 [11] が存在する．本稿では，k-匿名化手法の代表的なアルゴリズムであり，空間分割による手法である Mondrian アルゴリズムを用いる．

Mondrian アルゴリズム [5] は, 全ての人物（レコード）を一つのグループとした初期状態から分割を繰り返すことにより，グループに含まれる人物の数が k 人以上という k-匿名性の条件を満たすグループを出力するトップダウン方式のアルゴリズムである．Mondrian アルゴリズムにおける空間分割は，まず，ある属性の値を，パーティションと呼ばれる分割の基準として設定し，一つのグループをそのパーティションを基にして複数のグループに分割する．分割により生成された複数のグループに対してそれぞれ同様のグループ分けを繰り返し，次の分割によって生成される各グループ内に k 人存在せず，k-匿名性を満たさない場合，分割せずにグループ分けを終了する．分割終了後，最終的なグループ内の各属性値をグループの代表となる汎化した属性値に置き換えることにより，匿名化された個人データを生成する．パーティションの選択方法としては，属性値の中央点を選択する，人数がほぼ等しくなるように他の属性値と併合した属性値を構成する等の方法がある．

Mondrian アルゴリズムでは，グループに対して個別に分割を繰り返すため，含まれる属性値に偏りが存在する場合，偏った属性値をより詳しく細分化できるという特徴がある．また，トップダウン方式で分割し続けるため，分割する属性の選択に基準を設けることにより，基準に合わせた分割を実施することができる．

2.3 匿名化されたデータに対する情報損失の評価

表 2 や表 3 に示す通り，k-匿名性を満たすデータのパターンは数多く存在する．そこで，匿名化した場合，一般に情報損失が発生することから，情報損失を表す指標を用いて匿名化されたデータを評価し，最も情報損失が少ないデータを選択することが多い．例えば，よく用いられる評価値として，Prec 評価値 [7] や DM [8] などがあるが，本節ではシミュレーションによる比較対象としても用いる Prec 評価値を説明する．

Prec 評価値とは，1）一人の人物（一つのレコード）における一つの属性値に対し，汎化した回数をその属性において汎化可能な最大回数で正規化した値を計算した後，2）全人物の全属性での平均を取ることにより，得られる値である．汎化の回数を計算するためには，あらかじめ設定され，汎化のために用いられる階層構造を用いる．階層構造の例を図 1 に示す．階層構造は，上位に位置する階層の親ノードが下位の階層に位置する子ノードを包含した構造となる．

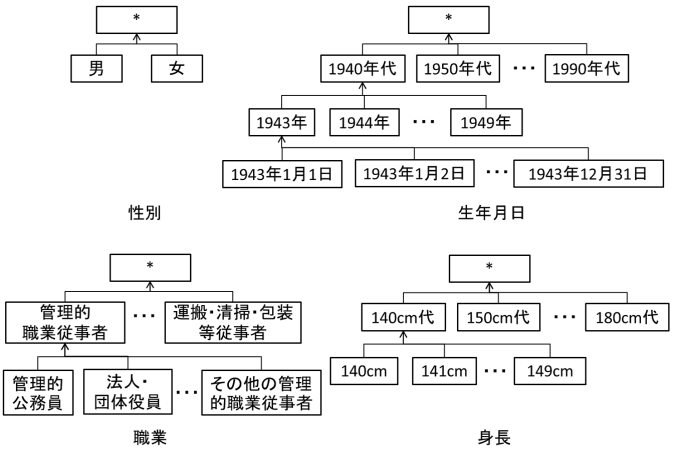

図 1　汎化規則の階層構造例.

Prec 評価値の評価式は，匿名化後の個人データを AT とした場合，その個人データに含まれる人数（レコード数）を N，汎化される属性の種類の総数を $|A|$，属性 A_j の階層構造の高さを H_{A_j}，i 番目の人物の属性 A_j が汎化された回数を $h^i_{A_j}$ とすると，以下の式で与えられる．

$$Prec(AT) = 1 - \frac{\sum_{i=1}^{N}\sum_{j=1}^{|A|}\frac{h^i_{A_j}}{|H_{A_j}|}}{N * |A|}. \quad (1)$$

3. 推薦情報を考慮した匿名化処理

k-匿名化アルゴリズムにより出力される匿名化されたデータは，情報損失に関する評価値を用いて評価され，最終的には情報損失の少ないデータが選択されることになる．しかしながら，情報推薦のような利活用を考えた場合，推薦に必要となる重要な属性値を汎化してしまう可能性があり，情報損失の少ないデータが必ずしも情報推薦に適切なデータになっているとは言い難い．そこで，TF-IDF 法を用い，情報の推薦のために必要となる属性を評価することにより，情報を推薦する際に必要な属性を汎化させずに残しつつ，他の重要でない情報を汎化することにより匿名化する手法を提案する．

3.1 情報推薦のために必要とされる属性値に関する規則

商業施設等から顧客への情報推薦を考える場合，一般的には推薦したい情報は決まっていることが多く，その情報を誰に推薦するかが重要となる．そこで，推薦される候補となる推薦情報と，その推薦情報を提供したいグループに関する規則を設定する．例えば，ある衣服の販売店が，表 4 のような衣服を推薦したい場合を想定する．表 4 に示す規則では，男性服 A は性別が男性で身長が 170cm 以下の人物に推薦したいことを表している．他の推薦情報に関しても，男性服 A と同様，表 4 に該当

表 4　情報を推薦するために必要とされる属性値に関する規則の一例.

推薦情報	性別	生年月日	職業	身長
男性服 A	男性	*	*	170cm 以下
男性服 B	男性	*	*	170cm 以上
女性服 A	女性	*	*	150cm 以下
女性服 B	女性	*	*	160cm～170cm

する属性をもつ人物に推薦したいことを表している．

3.2 TF-IDF 法における属性値の評価法

推薦情報を適切なユーザに提供できるよう，推薦情報の対象となる個人データの属性値に重きを置いて評価するために，本稿では TF-IDF 法を用いて属性値を評価する．TF-IDF 法を用いた場合，ある推薦情報が匿名化後のグループ内でどれだけ固有でかつ重要な情報であるかを評価することができるため，Prec 関数に代表される情報損失度合いを表す指標を用いた選択では難しい，推薦情報のために必要とされる属性値を汎化せずに出力することが可能となる．

TF-IDF 法における評価式は，ある文書中にある単語が出現する頻度を表す TF と，ある単語が他の全文書においても出現する汎用的な語かどうかを表す IDF の積算により表される．TF と IDF の評価式は，ある単語 t の文書 d 中での出現回数を $n_{t,d}$，文書 d 中での全単語 s の出現回数の合計を $\sum_{s \in d} n_{s,d}$，全文書数を N，ある単語 t が出現する文書数を $df(t)$ として，

$$tf(t,d) = \frac{n_{t,d}}{\sum_{s \in d} n_{s,d}}, \tag{2}$$

$$idf(t,N) = \log \frac{N}{df(t)}, \tag{3}$$

でそれぞれ表される．よって，TF-IDF 法による評価式は，次式で表される．

$$tf\text{-}idf = tf(t,d) \times idf(t,N). \tag{4}$$

以上をもとに，TF-IDF 法により各属性値の評価値を計算する方法を述べる．ここで，推薦情報に対して TF-IDF 法を利用するために，単語 t を推薦情報 r，文書 d を匿名化処理において分割されたグループ g と置き換える．

TF の値は，ある推薦情報 r のグループ g に推薦される推薦情報数に対する割合として考えられることから，TF の評価式は，ある推薦情報 r が匿名化グループ g を推薦対象としている場合に 1，推薦対象としていない場合に 0 となる値を $m_{r,g}$，匿名化処理によって分割されたグループ g を対象とする全ての推薦情報 s の数を $m_{s,g}$ とした場合，

$$tf(r,g) = \frac{m_{r,g}}{\sum_{s \in g} m_{s,g}} \tag{5}$$

と表される．次に，IDF はある推薦情報 r がどれだけのグループに推薦されるかの割合の逆数として表されることから，IDF の評価式は，全匿名化後のグループ数を N，ある推薦情報 r の推薦対象である匿名化後のグループ数を $gf(r)$ とした場合，

$$idf(r,N) = \log \frac{N}{gf(r)} \tag{6}$$

と表すことができる．そのため，TF-IDF 法による評価値は，TF の評価式と IDF の評価式より，以下の式で表現される．

$$tf\text{-}idf(r,g,N) = tf(r,g) \times idf(r,N). \tag{7}$$

3.3 TF-IDF 法に基づく Mondrian アルゴリズム

本節では，Mondrian アルゴリズムにおける分割する属性の選択に，選択した分割により生成されるグループそれぞれに対し，推薦情報に対する TF-IDF 法を用いた評価値を計算し，計算した TF-IDF 法の評価値の平均を取った値がより大きくなる属性を選択するアルゴリズムを提案する．

Mondrian アルゴリズムは，本来分割するパーティションとして特定の階層構造を用いないが，取り得る値が名義尺度となるような，カテゴリカルな属性には順序関係が存在しないため，分割の候補が膨大となってしまう．分割後のグループそれぞれに TF-IDF 法を用いて評価値を算出する提案手法の場合，分割の候補が多いと計算量が極めて大きくなってしまうため，図 1 に示す階層構造を導入することにより，計算量を削減する．階層構造を用いて汎化する場合，分割によって生成されるグループは子ノードの何れかに該当する．加えて，我々が提案する，TF-IDF 法に基づく Mondrian アルゴリズムによって生成される匿名化個人データは，次元数，属性値が増えるにつれ組み合わせが膨大になり，全ての出力候補から TF-IDF 法による評価が最も大きいものを探索することは困難となる．そのため，従来の Mondrian アルゴリズムと同様，ヒューリスティックに探索するものとする．

以上をまとめると，提案する k-匿名化アルゴリズムは次のとおりである．なお，本アルゴリズムは空間分割により生成されるグループに対して，一つのグループを分割不可能になるまで分割し，その後，別のグループを分割する深さ優先探索に基づく分割となっている．

Step1　パラメータ k を設定. 初期グループとして全ての人物を含むグループを生成.
Step2　与えられたグループに対し可能となる分割の内，k-匿名性を満たす分割を持つ属性を列挙. 存在しなければ Step5 へ.
Step3　列挙した属性を基に分割した場合に生成されるグループそれぞれに対し，TF-IDF 値を計算し，その TF-IDF 値の平均を算出.
Step4　列挙した属性の内，TF-IDF 値の平均が最大となる属性を選択し，その属性の階層構造を用いてグループに分割.
Step5　生成されたグループに対し，再度 Step2 から分割を開始.
Step6　グループ内の人物の属性値をグループを代表する属性値に置き換え.
Step7　他に Step6 を終えていないグループが存在するならそのグループに対して Step2 を実行. 反対に全てのグループが Step6 を終えていれば，処理を終了.

4. シミュレーションによる検証

4.1 検証データと検証環境

検証には，100 人分のランダムに生成した個人データを利用し，グループの大きさ k は簡単化のため 2 とした．個人データの属性としては，氏名，性別，生年月日，職業，身長を用い，一様分布に基づき生成した．属性に対する処理としては，氏名

表 5　用いる属性に関する階層構造.

属性データ	第一階層	第二階層	第三階層	第四階層
性別	*	男女		
生年月日	*	10 年毎	1 年毎	1 日毎
職業	*	日本標準職業分類大分類	日本標準職業分類中分類	
身長	*	10cm 毎	1cm 毎	

表 6　用いる推薦情報に関する規則.

ID	推薦情報	性別	生年月日	職業	身長
1	男性用雑誌	男性	*	*	*
2	厄除けのお守り	*	1977 年〜1979 年	*	*
3	スパナ	*	*	生産工程従事者	*
4	服 A	*	*	*	165cm〜171cm
5	服 B	*	1974 年〜1979 年	*	167cm〜188cm

は一意に個人を特定可能なために抑圧，他の属性に関しては汎化もしくは抑圧を施した．

各属性に対する汎化のための階層構造を表 5 に示す．性別は男女 2 項目で 2 階層，生年月日は 1944 年 1 月 29 日から 1994 年 1 月 28 日までの 18,628 項目で 4 階層，職業は日本標準職業分類 [12] の 74 項目で 3 階層，身長は 140cm から 190cm までの 51 項目で 3 階層で構成した．

各推薦情報に対し，対象となる属性を記した規則を表 6 に示す．それぞれ ID1 の推薦情報は性別，ID2 の推薦情報は生年月日，ID3 の推薦情報は職業，ID4 の推薦情報は身長，ID5 の推薦情報は生年月日と身長に基づいて推薦対象を判別する．検証する際はこれらの推薦情報の内，一項目のみを基準として匿名化を実施した．

検証に利用したコンピュータは，OS は Windows 7 Professional 64bit，メモリは 8GB，プロセッサは Intel(R) Core(TM) i7-2600 CPU @ 3.40GHz であった．開発環境は Eclipse で開発言語は Java を使用した．

4.2 検 証 結 果

まず比較のため，従来手法である Prec 評価値を基準にして Mondrian アルゴリズムを適用した結果を表 7 に示す．Prec 評価値は,「汎化した回数/汎化のための階層構造」が大きいほど汎化による情報損失が大きいとする指標であるため，用意した階層構造の全体の高さが低い属性に対しては汎化することが情報損失が大きいと評価し，性別，職業，身長，生年月日の順に優先して分割する結果となった．この匿名化データでは性別を最初に分割した後，職業が汎化に用意した階層構造の偏りによって分割そのものが困難であったため分割せず，次に生年月日ではなく身長による分割を優先したため，ほとんどの人数が生年月日の情報を喪失した．この場合，ID2 の推薦情報を推薦することが困難になると想定される．

次に，TF-IDF 値を基準にして Mondrian アルゴリズムを適用した結果を表 8 から表 11 に示す．表 6 における規則の内，性別による分類を持つ推薦情報である ID1 を用いた場合が表 8，生年月日による分類を持つ推薦情報である ID2 を用いた場合が表 9，職業による分類を持つ推薦情報である ID3 を用いた場合が表 10，身長による分類を持つ推薦情報である ID4 を用いた場合が表 11 である．提案手法では，それぞれ推薦情報の推薦対象とする人物に関する属性が変わると，本稿の手法により生成される匿名化後の個人データにより詳しく出力される属性も変化していることがわかる．

表 7　Prec 評価値を用いた Mondrian アルゴリズムによる結果.

性別	生年月日	職業	身長	人数
男女	*	*	10cm 毎	92
男女	10 年毎	*	10cm 毎	8

表 8　性別による分類を持つ推薦情報を用いた際の提案手法によるグループ分割の結果.

性別	生年月日	職業	身長	人数
男女	10 年毎	*	*	39
男女	*	*	10cm 毎	53
男女	10 年毎	*	10cm 毎	8

表 9　生年月日による分類を持つ推薦情報を用いた際の提案手法によるグループ分割の結果.

性別	生年月日	職業	身長	人数
男女	10 年毎	*	10cm 毎	23
男女	10 年毎	*	*	58
男女	10 年毎	大分類	*	5
*	10 年毎	大分類	*	12
*	10 年毎	*	10cm 毎	2

表 10　職業による分類を持つ推薦情報を用いた際の提案手法によるグループ分割の結果.

性別	生年月日	職業	身長	人数
男女	10 年毎	*	*	92
男女	10 年毎	*	10cm 毎	8

表 11　身長による分類を持つ推薦情報を用いた際の提案手法によるグループ分割の結果.

性別	生年月日	職業	身長	人数
男女	*	*	10cm 毎	73
*	10 年毎	*	10cm 毎	12
男女	10 年毎	*	10cm 毎	15

さらに詳しく見ると，表 9 の生年月日による分類を持つ推薦情報を用いた場合，全ての人物に生年月日の属性に関するデータを残すことができた一方，職業による分類を持つ推薦情報を用いた場合は表 11 に示す通り職業に関するデータを残すことができなかった．これは，職業の汎化に用意した階層構造に偏りが存在したため，分割そのものが困難であったと考えられる．また，表 8 に示す性別による分類を持つ推薦情報を用いた場合と表 11 に示す身長による分類を持つ推薦情報を用いた場合では，より詳しく出力したい属性に対して処理した後に可能な限り分割した結果，Prec 評価値によるデータに近づいたものと考えられる．

さらに，複数の属性の組み合わせによって推薦対象を分類す

表 12　生年月日と身長による分類を持つ推薦情報を用いた際の提案手法によるグループ分割の結果.

性別	生年月日	職業	身長	人数
男女	*	*	10cm 毎	92
男女	10 年毎	*	10cm 毎	8

表 13　各手法による匿名化データの Prec 評価値.

手法	Prec 評価値
Prec 評価値を用いた Mondrian アルゴリズム	0.382
性別による分類を持つ推薦情報を用いた際の提案手法	0.365
生年月日による分類を持つ推薦情報を用いた際の提案手法	0.351
職業による分類を持つ推薦情報を用いた際の提案手法	0.343
身長による分類を持つ推薦情報を用いた際の提案手法	0.367
生年月日と身長による分類を持つ推薦情報を用いた際の提案手法	0.382

る推薦情報である ID5 の推薦情報を用いて提案手法により匿名化した結果を表 12 に示す．Prec 評価値を基準にした場合は性別を最初に分割しており，今回の複数の属性を組み合わせた推薦対象では，身長を最初に分割した点において匿名化処理のプロセスは異なっていたが，分割を繰り返すことにより，最終的には Prec 評価値を用いた場合と同様の匿名化データを出力した．

最後に，情報損失の観点から分析するため，Prec 評価値による Mondrian アルゴリズムと本稿の提案する TF-IDF 法による Mondrian アルゴリズムを用いて，出力した匿名化後の個人データに対する Prec 評価値の結果を表 13 に示す．Prec 評価値による Mondrian アルゴリズムの出力した個人データが最も評価値の高いデータであることが分かるが，これは提案する TF-IDF 法による Mondrian アルゴリズムの出力する個人データは推薦情報の推薦対象かどうかの判断に用いられる属性を優先して分割した結果である．ただし，Prec 評価値による Mondrian アルゴリズムと提案手法は分割が出来なくなるまで，分割を繰り返すため，Prec 評価値に大幅な差は発生しないことが分かった．

以上の結果より，提案する TF-IDF 法による Mondrian アルゴリズムは情報損失の観点からは従来の手法より多少劣るが，用意した推薦情報を推薦するために利用される属性をより多くの人数に関して詳しく出力している匿名化データを生成した．そのため，この匿名化データを用いると，詳しく出力した属性から推薦情報を推薦する対象である人物に対しては推薦でき，推薦対象でない人物には推薦しないと判断することができる．そのため，匿名化データを利用した情報推薦においては，提案手法の方が優れているといえる．

5. ま と め

本稿では，TF-IDF 法を用いて，k-匿名化処理によって作られるグループにとって推薦情報がどれだけ重要かの重要度を評価し，推薦情報を推薦する場合に必要とされる属性をより詳しく出力する Mondrian に基づいた k-匿名化アルゴリズムを提案した．結果，用意された推薦情報に応じて出力される匿名化後の個人データに詳しく残される属性が選択されることを確認した．

残された課題としては，推薦情報の規則の構築方法の確立，実際の推薦情報と個人データに対する適用，プライバシー保護に段階を持つ手法 [13] との併用，提案手法が有効かどうかに関する評価の定量化等が挙げられる．

謝　　辞

本研究の一部は科学研究費補助金による。

文　　献

[1] B. C. M. Fung, K. Wang, R. Chen, P. S. Yu. “Privacy-Preserving Data Publishing: A Survey of Recent Developments,” *ACM Computing Surveys*, Vol. 42, No. 4, Article 14, pp. 1-53, 2010.

[2] L. Sweeney. “k-Anonymity: A Model for Protecting Privacy,” *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, Vol. 10, No. 5 , pp. 557-570, 2002.

[3] A. Machanavajjhala, D. Kifer, J. Gehrke, M. Venkitasubramaniam. “L-diversity: Privacy beyond k-anonymity,” *ACM Transactions on Knowledge Discovery from Data*, Vol. 1, No. 1, Article 3, pp. 1-52, 2007.

[4] N. Li, T. Li, S. Venkatasubramanian. “t-Closeness: Privacy Beyond k-Anonymity and l-Diversity,” *IEEE International Conference on Data Engineering*, pp. 106-115, 2007.

[5] K. LeFevre, D. J. DeWitt, R. Ramakrishnan. “Mondrian Multidimensional k-Anonymity,” *Proceedings of the 22nd International Conference on Data Engineering*, 2006.

[6] K. LeFevre, D. J. DeWitt, R. Ramakrishnan. “Incognito: Efficient Full-Domain k-anonymity,” *Proceedings of ACM SIGMOD international conference on Management of data*, pp. 49-60, 2005.

[7] L. Sweeney. “Achieving k-anonymity Privacy Protection using Generalization and Suppression,” *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, Vol. 10, No. 5, pp. 571-588, 2002.

[8] R. J. Bayardo, R. Agrawal. “Data privacy through optimal k-anonymization,” *IEEE International Conference on Data Engineering*, pp. 217-228, 2005.

[9] 馬場口 登. “[特別講演] マルチメディア処理とプライバシー保護技術 —実フィールドにおけるプライバシー情報開示と新たなサービスに向けて—,” **電子情報通信学会技術研究報告** *EMM2012-95*, Vol. 112, No. 226, pp. 33-38, 2012.

[10] 北研二, 津田和彦, 獅々堀正幹. “情報検索アルゴリズム,” **共立出版**, 東京, 2002.

[11] J. L. Lin, M. C. Wei. “An Efficient Clustering Method for k-anonymization,” *Proceedings of the 2008 international workshop on Privacy and anonymity in information society*, pp. 46-50, 2008.

[12] 総務省 日本標準職業分類: *http://www.soumu.go.jp/*

[13] 新井 健介, 河野 和宏, 馬場口 登. “ユーザの主観に適応した ID 種別毎のプライバシー保護,” **電子情報通信学会** *2014* **年総合大会**, D-21-3, p.194, 2014.