第五論文

新井健介・河野和宏・馬場口登「推薦対象の属性から構築した階層構造を用いた TF-IDF 法による匿名化処理」電子情報通信学会技術研究報告 vol. 115、no. 479、EMM2015-81、2016 年、31〜36 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
EMM2015-81(2016-03)

[ポスター講演] 推薦対象の属性から構築した階層構造を用いたTF-IDF 法による匿名化処理

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　ユーザから収集したデータに対して，ユーザのプライバシーを保護しつつ，そのデータを情報推薦などに利活用する研究が活発になされている．我々はこれまで，TF-IDF 法を用いて，情報推薦に利用する個人データの属性に対して重みを付け，その属性に関する情報が残るよう匿名化する手法を提案した．本稿では，匿名化に利用される汎化規則を示す階層構造に対しても，推薦対象の属性をもとに構築し匿名化処理に組み合わせることにより，より情報推薦しやすい個人データに変換する匿名化処理を実現する．シミュレーションでは，疑似推薦情報を用意し，推薦対象に関する属性値を組み込んだ階層を構築して匿名化処理に活用可能なことを検証した．
キーワード　k-匿名化，TF-IDF 法，Mondrian アルゴリズム，CF-Tree，Normalized Web Distance，情報推薦

[Poster Presentation]Data Anonymization based on TF-IDF Method Using Hierarchies Built from Attributes of Recommended Targets

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 personal data anonymization method using hierarchies considering information recommendation for users. Using hierarchies based on attribute values about recommendation targets, our anonymization method based on TF-IDF method converts attribute values in personal data to anonymized values which are appropriate for recommendation. We examine through simulation that it is possible to build hierarchies which include attributes of recommended targets and create anonymized data by using the hierarchies.
Key words　k-Anonymization，TF-IDF Method，Mondrian Algorithm，CF-Tree，Normalized Web Distance，Information Recommendation

1. ま え が き

近年，個人データの不必要な開示によりプライバシーの侵害が発生しないよう，ユーザのプライバシーを保護するだけでなく，プライバシー保護された個人データを利用する需要も高まっており，プライバシーの保護と利活用の両立を考慮した研究が数多く行われている．例えば，馬場口らは，個人データの開示と，開示から個人が利得を得られる情報基盤 [1] を提唱している．この研究で考えられている情報基盤とは，特定の目的・用途を実現するために実世界を区切ることにより限定されたフィールド (複合商業施設，テーマパーク，駅，病院等) とサイバー空間をリンクさせる情報基盤である．この情報基盤では，センシングによって収集された個人データをプライバシー保護して開示する事により，フィールド内に登録された個人は，開示した個人データに応じた利得として，フィールドに応じた上質なサービスや有益な情報を得ることができる．

個人データを開示し，開示された情報に応じた利得を得られるシステムを考えた場合，利得をより多く得るためには開示される個人データに対して利得を得るために適したプライバシー保護を実施する必要がある．しかし，既存のプライバシー保護の研究では，匿名化により出力される個人データの情報損失の度合いのみを考慮しており，利活用に必要な個人データを考慮した匿名化処理を行っているとは言えない．

本研究の目的は，利活用として情報推薦を想定し，推薦情報

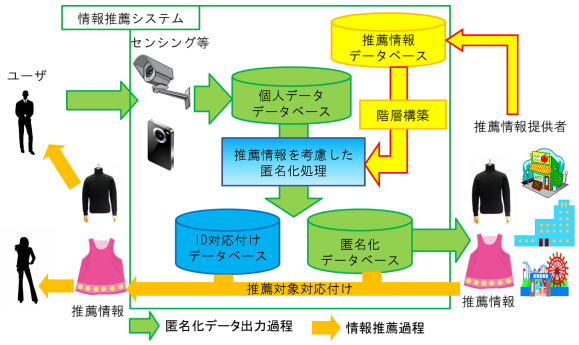

図 1　情報推薦を考慮した匿名化システムの概要図.

に関する情報を情報推薦提供者から，個人データに関する情報をユーザからそれぞれ収集し，得られた推薦情報を基に利活用を考慮したプライバシー保護処理を実現することである．本研究の概要図を図 1 に示す．我々はこれまで，匿名化する際，TF-IDF 法を用いて情報推薦に利用される属性に重みを付け，その属性に関する情報が残るよう匿名化する手法を提案した [2]．本稿では，匿名化に利用される汎化規則を示す階層構造に対しても，推薦対象の属性をもとに構築し，TF-IDF 法を用いた匿名化処理と組み合わせることにより，より推薦情報を考慮しつつ，個人を一意に特定できなくする匿名化処理を提案する．

2.　プライバシー保護の関連研究

データベースに収集した個人データに対して，なんらかの変換をかけることによりプライバシーを保護し，プライバシー保護された個人データを開示することを目的とする手法は Privacy Preserving Data Publishing [3](PPDP) と呼ばれる．本稿では，PPDP の技術に利活用を考慮した処理を組み込むことにより，利活用を考慮しつつユーザのプライバシーを保護する．

2.1　プライバシー保護条件

個人データを開示する場合，データと個人の対応を特定されないよう，一定の条件を満たすようにデータを変換する必要がある．匿名化された個人データが満たす必要のある条件として，k-匿名性 [4], [5]，l-多様性 [6]，t-近傍性 [7] 等が知られている．本稿では，最も基本となる指標である k-匿名性を用いる．

k-匿名性とは，開示する個人データと外部の個人情報を照らし合わせた時，外部の個人情報が個人を特定されないように開示した個人データ内の k 人と対応付く性質の事を言う．k-匿名性を得るためのデータ処理を k-匿名化と呼び，個人データ内に存在する全ての属性値の組み合わせに対して，他に k-1 人以上が同一の属性値の組み合わせを持つように変換する．

表 1 に示す個人データに対し，k-匿名化を実施した例を表 2 に示す．表 2 から，ID1 と ID4，ID2 と ID3 が同一の属性値の組み合わせを持つため，2-匿名性を満たしていることがわかる．

2.2　Mondrian アルゴリズム

k-匿名化を行うアルゴリズムのうち，本研究で用いる Mondrian アルゴリズム [8] に関して概略を述べる．Mondrian アルゴリズムとは，属性を次元とする多次元空間において，空間分割により生成される領域内に k レコード以上含まれ，かつこれ以上分割できなくなるまで，空間分割を分割後の領域に対して再帰的に繰り返す手法である．最終的にこれ以上分割できなくなった場合，各領域内のレコードをその領域を代表する属性値の組み合わせに置き換えることにより，k-匿名性を満たした個人データを出力することが可能となる．

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	橋本忠弘	男性	1966/04/10	自衛官	181cm
2	真田正一	男性	1987/12/15	研究者	186cm
3	笹本和夫	男性	1983/02/09	医師	180cm
4	村上市子	女性	1966/10/22	清掃員	158cm

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	*	*	1960 年代	*	*
2	*	男性	1980 年代	*	180cm 代
3	*	男性	1980 年代	*	180cm 代
4	*	*	1960 年代	*	*

2.3　匿名化後の個人データに関する評価指標

本稿において匿名化後の個人データの評価に用いる Precision Metric [5](PM)，Discernibility Metric [9](DM) に関して述べる．

PM は階層構造を用いた匿名化において，1 回の汎化による損失を階層構造から規定し，汎化回数から損失を計算する手法である．汎化処理によって情報損失を評価するため，その評価は階層の形状に依存するという特徴がある．PM の評価式は，匿名化後の個人データを T'，データに含まれるレコード数を N，匿名化す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．

$$Prec(T') = 1 - \frac{\sum_{i=1}^{N}\sum_{j=1}^{|A|}\frac{h^i_{A_j}}{|H_{A_j}|}}{N * |A|}. \quad (1)$$

DM は匿名化処理によって同一の属性値の組み合わせになったレコード集合のサイズを含まれるレコード数により評価する方法である．この集合のサイズは小さいほどより詳しく属性値を出力できているとみなせる．DM の評価式は，匿名化によって同一の属性値になったレコード集合を E，個人データ内の全てのレコード数を N とすると，以下の式で与えられる．

$$DM(T') = \sum_{\forall\ Es.t.|E| \geqq k} |E|^2 + \sum_{\forall\ Es.t.|E| \geqq k} |E| * |N| \quad (2)$$

3.　推薦対象の属性を基にした階層構造を用いた匿名化処理

推薦対象の属性値を基に個人データ内の属性値を汎化するための階層構造を構築し，情報推薦を考慮した匿名化処理と組み合わせることにより，情報推薦に用いやすいプライバシー保護された個人データを出力する匿名化処理を提案する．

3.1　推薦情報データ

推薦情報データは，推薦情報と，推薦したい対象の属性を記

表 3　推薦情報データの例.

推薦情報名	性別	生年月日	職業	身長
男性用雑誌	男性	*	*	*
CD（演歌）	*	1970 年代	*	*
指導用教科書	*	*	教員	*
服（L サイズ）	*	*	*	170cm～185cm

したデータである．推薦情報データは図 1 における推薦情報提供者から推薦情報データベースに保存される．

推薦情報データの例を表 3 に示す．表 3 では，男性用雑誌は性別が男性の人，CD（演歌）は生年月日が 1970 年代の人，指導用教科書は職業が教員の人，服（L サイズ）は身長が 170cm～185cm の人を対象に推薦されることを表している．

3.2　数値属性に関する階層構造の構築

推薦情報の対象となる属性が数値属性の場合，その属性値を CF-Tree [10] を利用してクラスタリングすることにより，数値属性に対する階層構造を構築する．CF-Tree とは，最初に非階層的クラスタリングによりクラスタを構築した後，クラスタの代表値に対して階層的クラスタリングを実施することにより階層構造を持ったクラスタリングを行う手法である．このとき，値域を持つ場合はその中点を代表値として用いる．

ここで，数値の値域の上限，下限に近い値のユーザへの情報推薦を考える．例えば表 3 では，身長が 170cm～185cm のユーザに服（L サイズ）を推薦したいが，仮に 170cm～185cm で区切る階層を作成した場合，169cm のユーザは範囲外となるため，本来その情報を求めていた可能性があるユーザに届かない場合がある．数値は連続であり，カテゴリカルなデータのように厳密に分類することは難しいため，本稿では確率モデルを導入し，厳密に区切るのではなく，確率的に分類する手法をとる．

図 2 に CF-Tree によるクラスタリングの手順を示す．まず，非階層的クラスタリングの一つである混合正規分布を用いた EM アルゴリズムにより，クラスタに分類する．混合正規分布を前提とすることにより確率密度を割り当てることが可能となる．その後，ウォード法を用いて階層的クラスタリングを行い，デンドログラムを作成する．ウォード法による階層的クラスタリングには混合正規分布の平均を入力とする．EM アルゴリズムより得られる正規分布を，属性を軸とする 1 次元空間に射影した後，その空間での属性の確率密度が一定以上となる範囲を値域とした葉ノードを作成し，デンドログラムと照らしあわせて階層構造を構築する．以下にアルゴリズムをまとめる．

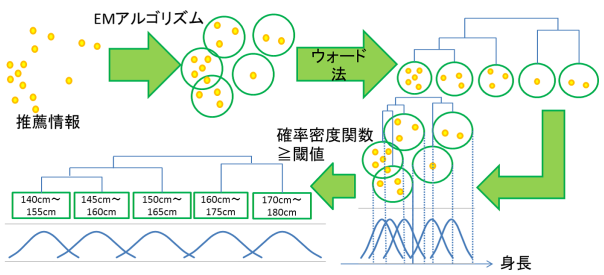

図 2　CF-Tree による数値属性の階層化.

Step1　推薦対象を選択するために用いる属性値を EM アルゴリズムによりクラスタリング
Step2　各正規分布の平均を基にウォード法により階層化
Step3　各正規分布の平均の下位ノードに確率密度が一定以上の数値を追加
Step4　階層的クラスタリングの結果に合わせて各正規分布の平均を併合した上位ノードを作成

3.3　カテゴリカル属性に関する階層構造の構築

カテゴリカル属性に対する階層構造を構築する手法は，オントロジーを利用する手法が有名であるが，従属関係に従い上下関係が構築されるため，必ずしも情報推薦に適しているとはいえない．そこで，カテゴリカル属性の場合も，推薦情報の対象となる属性値を利用し，階層を構築する．

カテゴリカル属性には，予め用意していたベースとなる階層構造に存在する属性値と推薦対象を選択するために用いる属性値の類似度を計算し，類似しているベースとなる階層構造の属性値の下位ノードとして追加することにより階層構造を構築する．本稿では，Normalized Web Distance [11](NWD) を用いる．NWD は任意の 2 単語間の非類似度を検索エンジンでの単語の検索ヒット数を用いて計算する方法である．この手法では，検索を利用するため，表記の多様性がある場合でも類似度を算出可能である．今回は検索エンジンとして Wikipedia を用いた．NWD は任意の 2 単語を x,y とし，検索エンジンに単語をクエリとして与えた場合のヒット数を返す関数を f として

$$NWD(x,y) = \frac{\max\{\log f(x), \log f(y)\} - \log f(x,y)}{\log N - \min\{\log f(x), \log f(y)\}} \quad (3)$$

と表される．

NWD を用いたカテゴリカル属性に関する階層構造の構築の手順を示す．ベース階層として，最上位に全ての属性値を表すことが可能な属性値を置いた，抽象度の高い属性値で構成される階層構造を用意する．追加する推薦対象に関する属性値とベース階層中の属性値の NWD による類似度を計算し，最も類似している属性値の下位ノードとして推薦対象に関する属性値を追加する．全ての推薦情報に対して同様の処理を実行した後，個人データ中の属性値と構築した階層構造の属性値の類似度を比較し，最も類似する属性値の下位ノードとして追加する．以上の階層構築の方法を以下のアルゴリズムにまとめる．

Step1　ベースとなる階層構造を入力
Step2　NWD によりベースとなる階層構造中の属性値と推薦情報が推薦対象の選択に用いる属性値の類似度を計算
Step3　類似度の値が最も近い属性値の下位ノードとして追加
Step4　全ての推薦情報に関して Step2 と Step3 を実行
Step5　個人データ中の属性値を同様にして階層構造に追加

3.4 推薦対象の属性を基に構築した階層構造と情報推薦を考慮した匿名化処理の組み合わせ

情報推薦を考慮した匿名化処理とは，TF-IDF 法を用いて情報推薦における属性の重要度を評価し，重要な属性をより詳しく出力する匿名化処理 [2] である．本稿では，この匿名化処理に推薦対象の属性を基にした階層構造を組み合わせる．

匿名化のためのアルゴリズムの詳細は次の通りである．まず，推薦情報から数値属性，カテゴリカル属性に対する階層構造を構築する．次に，数値属性に関しては，各正規分布の内，属する正規分布の決定を確率密度関数の値の大きさに従う確率によりランダムに割り当てる．

次に，Mondrian アルゴリズムに従った匿名化処理を個人データに実施する．まず，分割後に生成される全領域に対して各推薦情報がどれだけ重要かを TF-IDF 法により評価し，得られた TF-IDF 値の内，最大値となった推薦情報が各領域にとって最も固有で重要な推薦情報となる．その後，各領域にとって最も重要な推薦情報に関する TF-IDF 法の値を平均し，最も大きくなる次元を Mondrian アルゴリズムにおける分割次元に選択する．分割により形成された領域に対し，同様の処理を繰り返し，これ以上分割できなくなった場合，処理を終了する．階層構造の構築と TF-IDF 法による属性重要度を組み合わせた匿名化アルゴリズムをまとめると以下のようになる．

Step1　推薦情報を基にし，数値属性，カテゴリカル属性共に階層構造を構築
Step2　数値属性に対して，確率密度関数の値を基にどの正規分布に含まれるかをランダムに設定
Step3　与えられた領域において可能な分割の内，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 Step8 へ.
Step4　列挙した属性を基に分割した場合に生成される領域それぞれに対し，全推薦情報に関する TF-IDF 値の内の最大値を算出
Step5　算出した TF-IDF 法の最大値の平均を全次元に対して算出.
Step6　列挙した属性の内，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.
Step7　生成された領域に対し，再度 Step3 から分割を開始.
Step8　領域内の人物の属性値を代表値に置き換え.
Step9　他に Step8 を終えていない領域が存在するならその領域に対して Step3 を実行. 反対に全ての領域が Step8 を終えていれば，処理を終了.

4. 構築した階層と匿名化処理の検証

4.1 検証に用いたデータ

匿名化に用いた疑似個人データは氏名，性別，生年月日，職業，身長を含むデータとし，性別，生年月日，職業，身長に対して，k-匿名化処理を実施した．検証においては数値属性として身長を，カテゴリカル属性として職業を提案手法による階層構築に利用した．個人データは一様分布に従う属性値を持つ 100 人分の疑似個人データを 5 種類生成した．

検証に利用した推薦情報データは身長に関する推薦情報は表 4，職業に関する推薦情報は表 5 に示す．身長に関する推薦情報データは乱数により生成し，職業に関する推薦情報データはユーザ定義により生成した．

検証に用いた属性に対する階層構造の詳細を述べる．提案手法による階層構造を用いなかった属性には，予め決定した階層構造を用いた．性別は情報なしと男女の二項目で高さが 2 の階層である．生年月日は情報なしと 10 年毎と 1 年毎と 1 日毎の高さ 4 の階層で，取り得る属性値が 1944 年 1 月 29 日から 1994 年 1 月 28 日までの 18,628 項目である．職業は情報なしと日本標準職業分類の大分類と中分類で構成される高さ 3 の階層で，取り得る属性値が日本標準職業分類中分類の 74 項目である．身長は情報なしと 10cm 毎と 1cm 毎の高さ 3 の階層で，取り得る属性値が 140cm から 190cm までの 51 項目である．

また，提案手法で推薦対象のカテゴリカルな属性から階層構造を構築する場合，ベースとなる階層構造が必要となる．本稿では，提案手法を適用する職業のベース職業として，一般のアンケートで用いられる職業や業種分類を利用した．ベース職業に関する階層構造を図 3 の黒枠に囲まれた属性値に示す．分類は，第一階層に情報なしの 1 項目，第二階層に職業 7 項目，第三階層として会社員の下位ノードに業種 19 項目により構成した．階層構築に追加する職業には，日本語 Wikipedia の職業一覧から 25 項目の職業を取り出した．

4.2 生成された階層構造

カテゴリカル属性に対する結果から先に述べる．表 5 を用いてカテゴリカル属性に対して階層構造を生成した結果を図 3 に示す．点線により囲われた属性値が追加された属性値である．カテゴリカル属性では，一部に関連がなさそうな職業の追加や上下関係の逆転が見られる．これは，Wikipedia に含まれる記事には有名人や企業などの固有名詞に関する記事が多く，個人の来歴や関連企業などにより関係しない単語が共起しやすくなっていること，単純な類似度では，どちらが上位でどちらが

表 4　身長に関する推薦情報データ.

推薦情報 ID	性別	生年月日	職業	身長
A	*	*	*	165cm～171cm
B	*	*	*	182cm～188cm
C	*	*	*	142cm～190cm
D	*	*	*	141cm～164cm
E	*	*	*	167cm～185cm

表 5　職業に関する推薦情報データ.

推薦情報 ID	性別	生年月日	職業	身長
A	*	*	医師	*
B	*	*	編集者	*
C	*	*	警察官	*
D	*	*	建築士	*
E	*	*	プログラマー	*

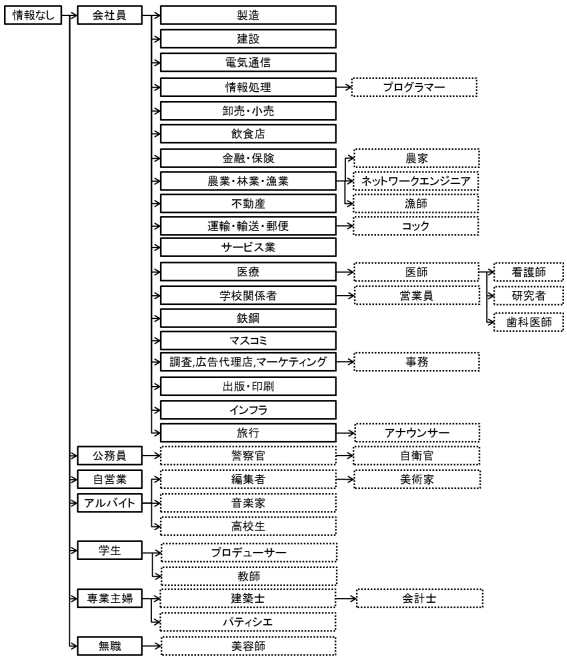

図 3　カテゴリカル属性の階層構築結果.

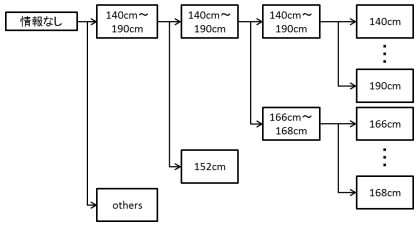

図 4　数値属性の階層構築結果.

下位かを判断出来ないこと，同じものに関わる上下関係にないものを判断できないことが原因である．対策としては，別の検索エンジンも併用して，ヒットする内容に偏りをなくす，類似する複数の属性値の下位ノードに推薦対象の属性値を追加する，上下関係をある程度記載した規則を追加する等が挙げられる．

表 4 の数値属性に対して階層構造を生成した結果を図 4 に示す．階層構築において，EM アルゴリズムにより生成されるクラス数は推薦情報数の半分を切り上げた数に設定した．EM アルゴリズムによる正規分布はそれぞれ，平均 180.42cm，分散 21.12 の正規分布，平均 167.00cm，分散 1.00 の正規分布，平均 152, 00cm，分散 $2.22 * 10^{-16}$ の正規分布となった．各正規分布はウォード法により，図 4 のように併合された．

以上の結果より，正規分布の分散の大小が激しく，元々の推薦情報の推薦対象属性値とは大きくかけ離れている．これは，推薦情報数が少なく，EM アルゴリズムを適用した場合に一つの正規分布を構成する値の数が少なくなってしまい，分散が大きい正規分布と小さい正規分布が混在したことが原因である．従って，EM アルゴリズムを実行する際のクラス数を推測する手法を追加し，一つの正規分布に含まれる値を多くする，または，分散の大きさを推薦情報から制限する必要がある．

4.3　匿名化処理結果

Precision Metric（PM）を基準とした Mondrian アルゴリズムの分割により生成した匿名化データ，TF-IDF 法による属性重要度を用いた Mondrian アルゴリズムによる匿名化データ，提案手法である推薦対象属性を考慮した階層を TF-IDF 法による属性重要度と併用した Mondrian アルゴリズムによる匿名化データを情報損失に関する評価値を用いて比較する．

表 6　Precision Metric の値を基にした匿名化結果.

性別	生年月日	職業	身長	人数
男女	*	*	10cm 毎	40
男女	*	大分類	*	43
男女	*	大分類	10cm 毎	9
男女	10 年毎	大分類	*	6
男女	*	中分類	*	2

表 7　提案手法により職業の階層を構築した場合の匿名化結果.

性別	生年月日	職業	身長	人数
男女	10 年毎	第三階層	*	5
男女	*	第三階層	*	24
*	10 年毎	第三階層	10cm 毎	4
男女	*	第一階層	1cm 毎	2
男女	*	第三階層	10cm 毎	12
*	10 年毎	第四階層	10cm 毎	2
男女	1 年毎	第四階層	*	2
男女	*	第一階層	*	8
男女	*	第二階層	10cm 毎	9
*	10 年毎	第三階層	*	3
男女	*	第二階層	*	14
*	*	第三階層	*	3
男女	10 年毎	第二階層	10cm 毎	4
*	*	第二階層	10cm 毎	3
*	10 年毎	第四階層	*	3
男女	10 年毎	第四階層	*	2

表 8　提案手法により職業の階層を構築した場合の匿名化結果.

性別	生年月日	職業	身長	人数
男女	10 年毎	*	第一階層	90
男女	10 年毎	大分類	第一階層	3
男女	*	*	第三階層	2
*	*	*	第四階層	3
男女	10 年毎	*	第四階層	3

まず，パラメータを $k = 2$ に設定し，PM の値を基にした匿名化による個人データを表 6 に，提案手法により職業と身長の階層構造を構築し，属性重要度を考慮した匿名化により個人データを匿名化した結果を表 7，表 8 に示す．

表 6 と表 7 より，属性重要度を考慮した匿名化により推薦対象の選択に用いられる属性を優先して詳しく出力していることが分かる．従って，推薦対象の属性を基に構築した階層構造は属性重要度を考慮した匿名化と併用でき，情報推薦に適した個人データを出力できていると言える．しかし，表 8 をみると，ほぼ全てのデータが第 1 階層の 140cm～190cm まで汎化されており，詳細な出力はできなかった．これは，もともと構築された階層構造に，140cm～190cm といった広い値域を持つ属性値が存在したこと，第 1 階層のもう一つの属性値である 152cm

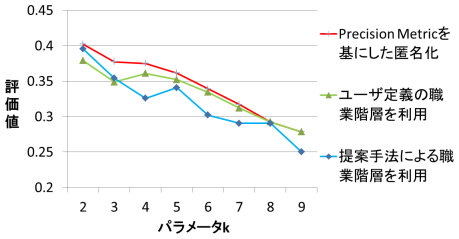

図 5　職業階層による匿名化の Precision Metric による評価.

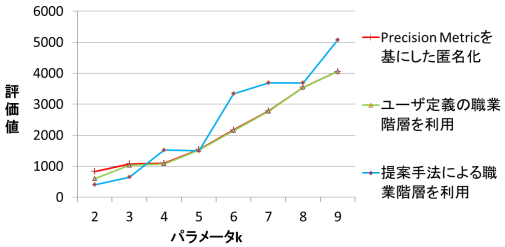

図 6　職業階層による匿名化の Discernibility Metric による評価.

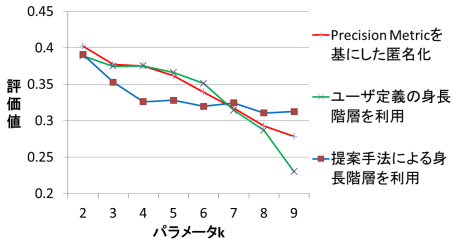

図 7　身長階層による匿名化の Precision Metric による評価.

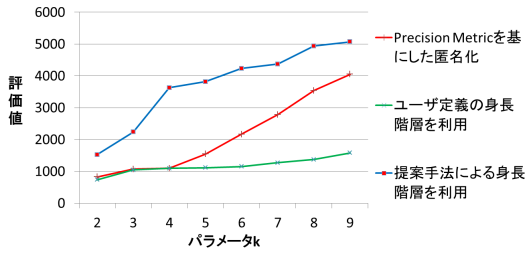

図 8　身長階層による匿名化の Discernibility Metric による評価.

を含むデータがほとんど存在しておらず，これ以上分割できなかったことが原因である．

次に，パラメータ k を 2〜9 に変更して，職業に関する階層構造をユーザ定義により構築した場合と提案手法により構築した場合の匿名化処理の結果の PM による評価を図 5，Discernibility Metric(DM) を図 6 に示し，身長に関する階層構造をユーザ定義により構築した場合と提案手法により構築した場合の匿名化処理の結果の PM による評価を図 7，DM を図 8 に示す．職業に関する階層については，どちらの評価値においても，提案手法による階層を使用したほうが損失は大きいが，ほぼ同値となる結果となった．身長に関する階層については，提案手法の方が DM による評価において，損失が大きくなった．

結果より，提案手法による職業に関する階層構造を用いた匿名化においては，どちらの評価値においてもユーザ定義による職業に関する階層構造を用いた匿名化と同等の情報損失になることが確認できた．また，提案手法による身長に関する階層構造を用いた匿名化においては，DM による評価において，ユーザ定義による職業に関する階層構造を用いた匿名化に比べ，情報損失が大きいことが確認できた．原因としては身長の階層構造に広い値域を持つ属性値が存在することにより，同一の属性値の組み合わせを持つデータが多く生成されたことが原因と考えられる．よって，階層構造の構築の際に正規分布の分散の大きさを制限する方法が対策として考えられる．

5. ま と め

本稿では，推薦情報が推薦される人物に関する属性値を基に汎化のための階層構造を構築し，個人データを情報推薦しやすい属性値に汎化する匿名化処理を提案した．検証では，推薦対象の属性値から構築した階層構造を利用することにより，数値属性は情報損失が増加し，カテゴリカル属性では情報損失があまり増加しないことを確認した．残された課題としては，分類クラス数の調節による数値属性に関する階層構築の改善，情報推薦における提案手法の有用性の評価などが挙げられる．

謝　　辞

本研究の一部は科学研究費補助金による．

文　　献

[1] 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.

[2] 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 EMM2015-10, Vol.115, No.38, pp.51-56, 2015.

[3] B. C. M. Fung, K. Wang, R. Chen, and 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan. “Mondrian Multidimensional k-Anonymity," *Proc. of International Conference on Data Engineering*, 11 pages, 2006.

[9] R. J. Bayardo and R. Agrawal. “Data privacy through optimal k-anonymity," *Proceedings of IEEE International Conference on Data Engineering*, pp. 217-228, 2005.

[10] T. Zhang, R. Ramakrishnan, and M. Livny. “BIRCH: An Efficient Data Clustering Method for Very Large Databases," *Proc. of the ACM SIGMOD International conference on Management of data*, pp. 103-114, 1996.

[11] R. L. Cilibrasi and P. M. B. Vitanyi. “Normalized Web Distance and Word Similarity," *ArXiv e-prints*, 2009.