Vol.2015-DBS-162 No.2
2015/11/26

公共データ間の関係性抽出のための属性分類手法の提案

近藤 拓也[1,a)]　遠藤 雅樹[1,2,b)]　江原 遥[1,c)]　廣田 雅春[3,d)]　横山 昌平[4,e)]　石川 博[1,f)]

概要：近年，各国政府や地方公共団体から，二次利用・機械判読に適した形式である Open Data として公開されている公共データが増加している．これら多種多様な機関・部署の公共データを統合することができれば，新しく有用な統計的関係性を抽出できる可能性がある．しかし，これらの公共データは，作成機関や部署によってデータセットの属性名に表記ゆれがある，もしくは属性・属性値の定義が異なる，といった問題がある．このため，現状では，異なる公共データの統合を行うためには，そのデータセット間の属性・属性値を人手で確認・修正する必要がある．本論文では，多種多様な公共データの統合を半自動的に行うために，データセットの属性を分類する手法を提案する．提案手法では，各データセットの属性を予め定めたクラスに分類し，複数の属性がそれぞれ同一のクラスと分類された場合に，その公共データは統合可能と判定する．そのうえで，実際に統合可能な公共データの組み合せから有用な関係性の発見を試みる．

キーワード：Open Data，表データ

1. はじめに

近年，多くの行政機関において，Open Data の利活用を推進する動きが高まっている．たとえば，2013 年 6 月の G8 サミットでは，データのアクセス，公開，および再利用の基礎などを「オープンデータ憲章 [*1] により規程した．また，日本でも，2012 年 7 月に公共データの利活用推進のための基本戦略である「電子行政オープンデータ戦略」[*2] を規程し，それぞれの省庁のデータを Open Data として公開した．それらの国や，地方公共団体の公開している Open Data のデータセットは，いくつかの Open Data のポータルサイト [*3][*4] にまとめられている．これら多種多様な機関・部署の公共データを統合し活用することができれば，新しく有用な統計的関係性を抽出し，その関係性を用いた新しいサービスの創出の促進に繋がる．たとえば，米国では気象情報や，農作物の収穫量の統計情報，土壌の水分量の統計情報などの Open Data を用いることで，農家向け収入保障保険ビジネスが創出された [*5]．さらに，各地方公共団体が公開している，予算，決算データから，払った税金が 1 日当たりどこにいくら使われているかを知る「Where does my money go?」[*6] のような市民主導のサービスも創出された．このように，気象情報や収穫量，土壌の水分量，または予算，決算データなどの統計的関係性を用いた，新たなサービスが様々創出されている．

Open Data の利活用は，現状，データのドメイン（分野）が限定されていれば，国・地方公共団体を超えた利活用が可能である．たとえば，税制分野においては，元々イギリスの「オープン・ナレッジ・ファウンデーション」が開発した「Where does my money go?」という税金の使用目的を市民が確認できるサービスが，日本に「Where does my money go〜税金はどこへ行った？」[*7] というサービス名で転用された例がある．しかし，ドメインを特定の分野に限定しない，多様な分野のデータセットを扱う場合には，作成機関や部署の違いによる属性名の表記ゆれや，属性の定義の違いによって，国や地方公共団体を超えたデータの利活用は難しい．ドメインが限定されていれば，それぞれの分野の専門家によって，人手でこれら表記・定義の問題を解消することができたのに対し，多様な分野のデータセットを扱う場合では，そのような戦略が取れないからである．

1 首都大学東京
2 職業能力開発総合大学校
3 大分工業高等専門学校
4 静岡大学
a) kondo-takuya@ed.tmu.ac.jp
b) endou@uitec.ac.jp
c) ehara@tmu.ac.jp
d) m-hirota@oita-ct.ac.jp
e) yokoyama@inf.shizuoka.ac.jp
f) ishikawa-hiroshi@tmu.ac.jp

*1 http://www.kantei.go.jp/jp/singi/it2/densi/dai4/sankou8.pdf
*2 http://www.kantei.go.jp/jp/singi/it2/pdf/120704_siryou2.pdf
*3 DATA.GOV:https://www.data.gov/
*4 DATA.GO.JP:http://www.data.go.jp/
*5 Total Weather Insurance:http://www.climate.com/
*6 http://wheredoesmymoneygo.org/
*7 http://spending.jp/

Open Data の利活用には，公開されているデータセットの多様性が重要とされており，「オープンデータ憲章」においても，包括的な Open Data の公開を目的の 1 つとしていることからも，ドメインを限定せずに多種多様なデータの統合・利活用を可能にする手法が求められている．

そこで，著者らは様々な機関の公共データの統合を半自動的に行ったうえで，公共データの包括的な利用を可能とすることを目的に，属性の分類手法を提案する．本論文の構成は以下のようになっている．2 章に関連研究，3 章に提案手法，4 章にその実行例を示す．最後に 5 章では今回の研究のまとめを述べる．

2. 関連研究

2.1 Open Data の課題

Open Data，Open Government Data の有用性や，それらのデータを利活用する際の課題は，[1], [3] にまとめられている．たとえば，Open Data のライセンスについて論じているもの [4] や，Open Data の政策について論じているもの [5] がある．なかでも，新規に Open Data を作成する際の，使用語彙の統一や，データセット間のリンクが少ないといった問題で議論が盛んである．使用語彙の統一を図る目的で，山根ら [6] は，データ作成の際に，適切な語彙を選択するため，既存の Linked Data から語彙を推薦するシステムを示した．さらに，データ構造と表記の類似度から実体の同一性の判定を行い，キーワード抽出を行うことでリンクの付与を行った．しかし，使用語彙に関して，日本では語彙の再利用を行う「Linked Open Vocabularies (LOV)」[*8] に相当する取り組みは進んでいない．「5 star deployment scheme」[*9] により，Open Data のデータ形式でもっともよい形式とされる Linked Open Data(LOD) を構築し，データ間のリンクを作成する目的で，玉川 [7] らは，日本語 Wikipedia オントロジーのプロパティを用いた日本語語彙の構築を行った．

このように，新規の Open Data を作成する際の課題を扱った研究は多いが，現在公開されている Open Data を対象とする研究は少ない．実際，現在公開されている Open Data のほとんどは，仕様語彙の統一や他データセットへのリンク付けが行われていない [12]．一方，本研究では，Open Data を少ないコストで活用する手法について論じる．

2.2 データの統合

スキーマの定義や，その使用語彙の関係を用いた，データの統合については，様々な分野で議論が盛んである．たとえば，田仲ら [8] は，人手で表の構造に解釈を与え，その構造に着目し，表中のデータ間の関係性の獲得を行った．保田ら [9] は，2 種類の異なる農業情報のデータベースを統合するため，人手でスキーマの定義を行った．しかし，人手でスキーマの定義を本研究で扱おうとしている公共データのデータベースは，農業といった特定の分野に限定されていないため，人手でスキーマの定義を行う事は難しい．

異種データベースの統合について，石橋ら [10] は，データ検索・結合文脈に応じて確定されるデータ間の時間的・空間的関連性の計量を伴う異種データベース間時間的・空間的データ検索・結合方式を提案した．実験により，関連性記述方式との比較における提案方式の実現容易性・実用性を示している．武田ら [11] は，経済産業省の提供する工業統計調査の様々な表に対して，「The RDF Data Cube Vocabulary」[2] に従い，次元や測度，コード体系を定義することで，表をまたいで検索が可能となることを示した．一方で，複数の統計表を統合し活用するには，次元や測度，コード体系の定義のためのコストが高いことや，データの量が大きくなることを指摘した．そのため，データを包括的に利用するためには，人手で行う処理と，スキーマの定義や，その使用語彙の同一性の判定を行う処理に関して，コストを抑える必要がある．

図 1: 用語の説明

3. 提案手法

本研究では，分野を限定しないため，どの統計表にも出現しやすい時間，場所に注目して属性の統合を行う．図 1 で，本研究で使う用語について説明する．表は属性からなり，各属性には，属性名がついているものとする．各属性には，具体的なデータの中身であるカラムが紐づいており，カラム中に出現する属性値の集合を属性値集合と呼ぶ．また，属性の集合をスキーマと呼ぶ．図 2 に提案手法の概略を示す．まず，Open Data から時間，場所に関係する属性（とそのカラム）を抽出する．次に抽出した属性を，「クラス」に分類し，同じクラスに分類された属性は，統合可能と判定する．統合可能と判定された属性の集合を抽出し，実際に人手で確認することによって，有用な関係性を発見する．はじめに，3.1 節でデータセットの詳細について説明し，3.2 節で属性のクラス推定手法，3.3 節で推定した属性のクラスを用いた表の統合方法について述べる．

[*8] http://lov.okfn.org/dataset/lov/
[*9] http://5stardata.info/

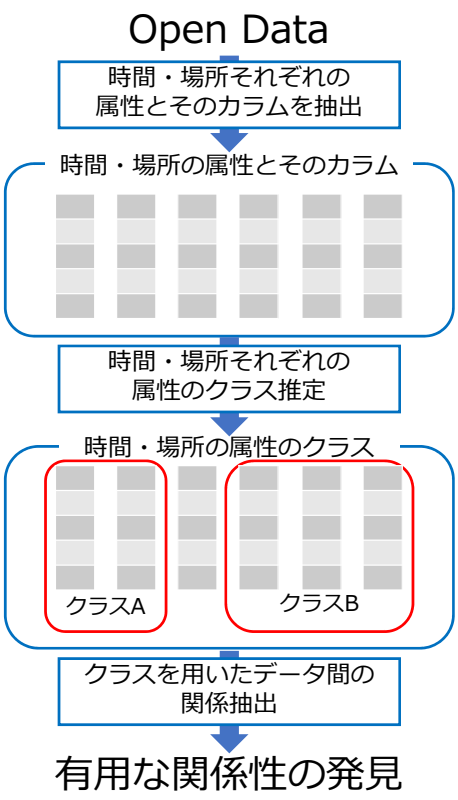

図 2: 分析の手順

表 1: E-stat から取得した表数

データの詳細	VALUE
E-stat から取得した表数	63,758
時間を表す属性を含む表数	38,896
場所を表す属性を含む表数	38,507
時間・場所を表す属性を含む表数	28,136

表 2: 取得した表の例
都道府県別，年齢階級別人口

時間軸 (年次)	全国都道府県	男女	年齢 5 歳階級	VALUE
1980 年	北海道	男	0-4 歳	11111
1980 年	北海道	女	0-4 歳	22222
1980 年	北海道	男女総数	0-4 歳	33333
1980 年	青森県	男	5-9 歳	44444
1980 年	青森県	女	5-9 歳	55555
1980 年	青森県	男女総数	5-9 歳	66666
…	…	…	…	…

3.1 データセットについて

政府統計のポータルサイトである E-stat[*10] から API[*11] を用いて収集したデータを表 1 に示す．また，取得データは表データを XML 形式に変換したものであるため，その XML 形式のデータを表 2 のような構造の表として扱う．取得期間は，2015 年 8 月 25 日から 2015 年 8 月 30 日である．取得できた表数は 63,758 件で，そのうち表 2 の “時間軸 (年次)” ような，時間を表す属性を含む表は 38,896 件，表 2 の “全国都道府県” のような場所を表す属性を含む表は 38,507 件，時間を表す属性・場所を表す属性どちらも含む表は 28,136 件であった．

[*10] https://www.e-stat.go.jp/SG1/estat/eStatTopPortal.do
[*11] http://www.e-stat.go.jp/api/

表 3: 取得した表の詳細

データの詳細	VALUE
表の作成団体数	7
取得した表が含まれる統計調査名数	43
取得した表が含まれる統計調査の集計数	874
時間を表す属性の属性名数	360
場所を表す属性の属性名数	2,367
スキーマの異なり数	29,589
時間を表す属性の属性値集合の異なり数	751
場所を表す属性の属性値集合の異なり数	1,099

次に，取得した表の詳細について表 3 に示す．表 3 に示した表の作成団体は，総務省，経済産業省，農林水産省，国土交通省，財務省，文部科学省，厚生労働省の計 7 団体であった．表 3 のスキーマの異なり数から，表 2 の [“時間軸 (年次)”，“全国都道府県”，“男女”，“年齢 5 歳階級”，“VALUE”] に該当する箇所の異なり数が 29,589 件あることから，多様な統計表がまとめられていることが分かる．また，表 3 の時間・場所それぞれの属性の属性名数と，表 2 の [“男”，“女”，“男女”] のような属性値集合の，異なり数から，属性名に対する属性値集合が集計団体ごとに異なることが推測できる．実際に，“時間軸 (年次)” の属性の属性値集合を確認すると [“1980”]，[“1993”]，[“2013 年 10 月”] などがあり，さらに属性名が “時間軸” であるものの属性値集合に [“1993”] があった．また “全国都道府県” のような都道府県を示す属性名として，他に “H25 地域”，“全国，都道府県”，“地域” などがあった．このように，表の作成団体が異なると，属性の定義も異なることが分かる．

3.2 属性のクラス推定について

今回対象とする表は，属性の定義が各団体ごとに異なることを 3.1 節で示した．このような異種データ間の統合では，2.2 節で示したように，属性を定義する，もしくは使用語彙の同一性の判定を行う必要があり，そのどちらも人手で行うにはコストが高い．そこで同種の属性を機械的にまとめるためクラスタリングを行い，生成されたクラスタから属性のクラス推定を行う．

本論文では，評価のため，表 1 の時間・場所を表す属性を含む全 28,136 件の表から，1,000 件をランダムに抽出し，この表の分析を通じて，クラス推定の可能性を評価する．

抽出した表の詳細について表 4 に示す．これらの表 4 の表に対し，スキーマと，場所を表す属性，時間を表す属性を特徴量に k-means クラスタリングを行う．次に，作成し

表 4: 属性の推定に用いる表の詳細

表の詳細	VALUE
表の作成団体数	4
取得した表が含まれる統計調査名数	21
取得した表が含まれる統計調査の集計数	303
時間を表す属性の属性名数	20
場所を表す属性の属性名数	204
スキーマの異なり数	958
時間を表す属性の属性値集合の異なり数	75
場所を表す属性の属性値集合の異なり数	171

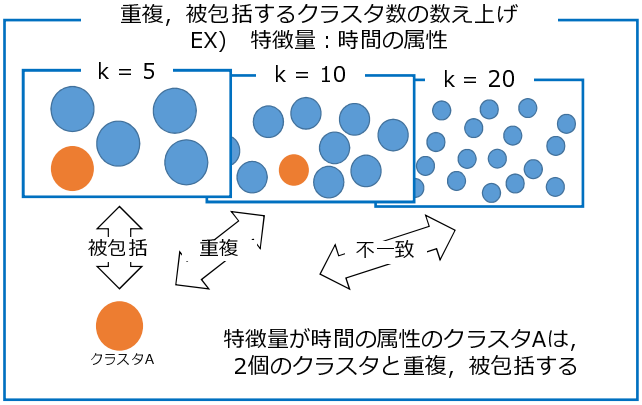

図 3: 複数回同一のクラスタとなるデータの抽出方法

たクラスタを用いた時間・場所それぞれの属性のクラス推定を行う．

3.2.1　クラスタリング

k-means クラスタリングは，クラスタ数 k をパラメタとして指定し，指定した k に対して，表を k 分割する手法である．特徴量を変えることで，スキーマを特徴量とする場合，場所を表す属性を特徴量とする場合，場所を表す属性を特徴量とする場合の 3 種類のクラスタリングを行った．k の値は，それぞれの種類の中で，特徴量の性質を考慮して決定した．

スキーマを特徴量に用いたクラスタリングでは，表 4 の取得した表が含まれる統計調査名数が 21 件であることから，少なくとも 20 分割できると考えたためである．また，取得した表が含まれる統計調査の集計数が 303 件であることから，最大でも 350 程度の分割で属性をまとめることができると考える．スキーマを特徴量に 1,000 件の表を分割する際の k を，$k \in \{5, 6, \ldots, 100\}$ とした．

場所を表す属性値集合を特徴量に用いたクラスタリングでは，$k \in \{5, 6, \ldots, 100\}$ とした．下限は，場所が世界，日本，地方，都道府県，市区町村の粒度で分類されている事が多いことを参考に，上限は表 4 の属性の属性値集合の異なり数を参考に決定した．時間をあわらす属性値集合を特徴量に用いたクラスタリングでは，$k \in \{5, 6, \ldots, 200\}$ とした．下限は，西暦，和暦，月，曜日，日に属性が大別できることを参考に，また，上限は表 4 の異なり数を参考に決定した．

3.2.2　属性のクラス推定

作成したクラスタを用いた時間・場所それぞれの属性のクラス推定を行う．まず，様々な分割数，特徴量で行ったクラスタリングの結果から，複数回，同一のクラスタとなる属性のまとまりを抽出する．抽出方法の例を図 3 に示す．図 3 では，特徴量として時間の属性の属性値集合を用いて，k を [5,6,..., 100] で分割した計 5,040 個のクラスタから 1 つのクラスタを選択し，ほかの分割数 k で分割されたクラスタに，重複，被包括するクラスタ数を数える．例えば，図 3 では，クラスタ A が 2 個のクラスタで重複，被包括している．クラスタ A に対して，この数値を $C_T(A) = 2$ とする．同様に，あるクラスタ B がスキーマを特徴量に用いたクラスタに重複，被包括する数を数え，その数値を $C_S(A)$ とし，場所の属性を特徴量に用いたクラスタに重複，被包括する数を数え，その数を $C_R(A)$ とする．C_T と C_S を，時間の属性値集合を特徴量として用いて $k = 5$ から $k = 100$ まで分割した，計 5,040 件のクラスタすべてについて計算した．この 5,040 件のクラスタの集合を $\mathcal{U}_T$ とおく．C_R と C_S を，場所を属性値集合を特徴量として用いて $k = 5$ から $k = 200$ まで分割した，計 20,090 件のクラスタすべてについて計算した．この 20,090 件のクラスタの集合を $\mathcal{U}_R$ とおく．

最終的に得たいクラスの集合は，時間に関するクラスの集合 $\mathcal{X}_T$ と，場所に関するクラスの集合 $\mathcal{X}_R$ の 2 種類である．これらは，集合の集合であるため，集合族となる．$\mathcal{X}_T$ は，以下のように定める．まず，$\mathcal{X}_T$ には，前述の C_S，C_T のそれぞれの指標が，しきい値 τ_S, τ_T 以上であるようなクラスタが含まれる．

$$\mathcal{X}_T \subseteq \{A | A \in \mathcal{U}_T, C_T(A) \geq \tau_T, C_S(A) \geq \tau_S\} \qquad (1)$$

ただし，$\mathcal{X}_T$ 中で，互いに重複・被包括の関係にあるクラスタがある場合は，それらのクラスタの和集合で置き換える．最終的に，互いに重複・被包括の関係にあるクラスタが含まれない，すなわち，以下の式が成り立つようにする．

$$\forall A, A' \in \mathcal{X}_T, A \neq A'; A \not\subseteq A' \qquad (2)$$

$\mathcal{X}_R$ についても，同様に，C_S, C_R のそれぞれの指標が，しきい値 τ'_S, τ_R 以上であり，かつ，互いに重複・被包括の関係にあるクラスタが含まれないクラスタの集合として定義する．最後に，しきい値 τ_S，τ_T，τ'_S，τ_R については，$\mathcal{X}_T$ に含まれるクラスタ数 $|\mathcal{X}_T|$，$\mathcal{X}_R$ に含まれるクラスタ数 $|\mathcal{X}_R|$ がそれぞれ最大になるように，定める．$\mathcal{X}_T$ を求める際の C_S に関するしきい値 τ_S と，$\mathcal{X}_R$ を求める際の，C_S に関するしきい値 τ'_S は異なる．

3.3 推定された属性クラスを用いた分類と表の統合について

3.2.2 節で推定した属性のクラスに基づき，属性を分類する手法について述べる．判定したい新しい属性 a' の属性値集合を $S_{a'}$ とする．$S_{a'} \subseteq \bigcup_{a \in C} S_a$ であるとき，属性 a' はクラス C に属すると判定する．次に，分類された属性を用いた表の統合について述べる．時間・場所を表す属性がそれぞれ同一のクラスと定義された表の組み合せを，統合可能な表の組み合せであると判定し，その表間の関係性について統計的に有用な関係性がないか人手で確認する．

4. 提案手法の実行例

本章では，3.2.2 節の属性の推定を行った結果を 4.1 節で，3.3 節で統合した表についての結果と有用な関係性の発見について 4.2 節で述べる．

4.1 属性の推定結果について

時間を表すクラス集合 $\mathcal{X}_T$ に関しては，しきい値 $\tau_T = 31, \tau_S = 1$ のときに，$\mathcal{X}_T$ に含まれるクラス数が $|\mathcal{X}_T| = 48$ と最大となった．直感的には，大きいしきい値では，クラスタ数 k を変化させても，そのクラスタが残り続ける，すなわち，クラスタのクラスタ数 k に対する頑健性を示唆している．したがって，この得られたクラス $\mathcal{X}_T$ は，$\tau_T > \tau_S$ であることから，スキーマに関する特徴量を用いたクラスタリングより，時間の属性値集合を特徴量として用いたクラスタリングに対して，頑健なクラスになっていることが分かる．

$\mathcal{X}_T$ に含まれるクラスタを具体的に人手で確認したところ，“時間軸 (年次)” や，“時間軸” などの複数の異なる属性値集合を持つ属性に関しては，すべて同じ時間の属性値集合をまとめることができた．一方で，属性値集合に和暦が用いられたものに関して，そのデータ数が少なく，異なる時間であるにもかかわらず，すべて同じクラスタにまとまってしまった．

場所を表すクラス集合 $\mathcal{X}_R$ に関しては，$\tau_A = 162, \tau'_S = 6$ のときに，$\mathcal{X}_R$ に含まれるクラス数が $|\mathcal{X}_R| = 81$ と最大となった．このクラスを人手で確認したところ，大枠では，同種の属性値集合がまとめられていることが分かった．たとえば，都道府県をまとめた属性は，属性名の定義の違いなどが解消され，同一クラスにまとめられていた．さらに，場所に関するしきい値 $\tau_R = 162$ であることからも，場所に関して非常に頑健なクラスタが，クラスとして抽出されていることが分かる．

一方，外れ値・ノイズとみなせるような属性値集合を少量含むクラスも，ある程度存在していたが，すべてのケースがノイズとみなせるわけではない．たとえば，属性値集合が [“広島大都市圏”，“北九州・福岡大都市圏”，“中京大都市圏”，“阪神大都市圏”，“札幌大都市圏”，“仙台大都市圏”，“関東大都市圏”] であるような属性 11 件を含むクラス中に，属性値集合が [“中京大都市圏”，“京阪神大都市圏”，“全国・関東大都市圏”] の属性が 1 件のみ，同一のクラスタに含まれていた．これは，後者の属性値集合で欠損している，“広島大都市圏” といった要素も補って属性を分類した結果であるとも解釈できる．このような外れ値に相当する属性について，スキーマの定義が影響していると推測し，実際に人手でクラス内の属性のスキーマを確認したが，外れ値を生ずる原因と推測されるような箇所をスキーマの定義中で特定することはできなかった．しかし，$\tau_S = 6$ であることから，スキーマからの特徴量が，クラス集合 $\mathcal{X}_R$ に影響を及ぼしていることは確かである．

4.2 提案手法の実行例について

本研究で対象とした 1,000 件の表を，提案手法により，241 件のまとまりに統合することができた．この 241 件の統合された表を確認したところ，昭和 61 年の事業所・企業統計調査のサービス業の事業所数と，社会生活基本調査の生活行動に関する行動者数・行動率等をまとめた表が統合されていた．この統合された表を人手で確認し，統合は妥当に行われていることが分かった．したがって，たとえば，この統合された表から，サービス業の盛んな地域における人々の行動，といった新しい情報を取得することが可能である．

また，平成 19 年の全国物価統計調査の小売店の店舗数と，就業構造基本調査の全国での就業者数を示した表が統合されていた．この表の統合も，人手で確認したところ，妥当な統合であった．したがって，この統合された表から，地域別に，小売店の規模と就業者数の規模を結び付けた統計的分析が可能となった．このように，様々な新しい統計的関係性を発見できる可能性を秘めた，有意義な表の統合が得られていることを確認した．

5. おわりに

本研究では，多種多様な公共データの統合を半自動的に行うために，データセットの属性を分類する手法を提案した．さらに公共データ間の属性をクラスに分類し，同種の属性をまとめ，そのクラスに基づき，本研究で対象とした表を統合することで，実際に統合可能な公共データの組み合せから，有用な関係の発見できることを確認した．属性の定義の曖昧性解消や，同一性の判別を人手で行うコストを，クラスタリングを応用した提案手法を用いて削減できることを示すことができた．提案手法により，1,000 件の表を 241 件のまとまりに統合することができた．統合された表を人手で確認することで，有用な関係性を発見できることが示唆された．

今後の課題としては，推定された属性のクラス間の関係性を考慮した統合を行うことや，属性のクラス推定の手法

の改善，属性のクラス推定の評価指標の確立，提案手法のデータセット全体への適応などがあげられる．

謝辞 本研究（の一部）は傾斜的研究費（全学分）学長裁量枠戦略的研究プロジェクト戦略的研究支援枠「ソーシャルビッグデータの分析・応用のための学術基盤の研究」による．

参考文献

[1] Jocelyn Cranefield, Oliver Robertson, and Gillian Oliver. Value in the MASH: exploring the benefits, barriers and enablers of open data apps. In *22st European Conference on Information Systems, Tel Aviv, Israel*, 2014.

[2] Richard Cyganiak, DERI, NUI Galway Dave Reynolds, and Epimorphics Ltd. *The RDF Data Cube Vocabulary*. World Wide Web Consortium, February 2014.

[3] Marijn Janssen, Yannis Charalabidis, and Anneke Zuiderwijk. Benefits, adoption barriers and myths of open data and open government. *Information Systems Management*, Vol. 29, No. 4, pp. 258–268, September 2012.

[4] Paul Miller, Rob Styles, and Tom Heath. Open data commons, a license for open data. In Christian Bizer, Tom Heath, Kingsley Idehen, and Tim Berners-Lee, editors, *LDOW*, Vol. 369 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2008.

[5] Anneke Zuiderwijk and Marijn Janssen. Open data policies, their implementation and impact: A framework for comparison. *Government Information Quarterly*, 2013.

[6] 山根昇平, 鵜飼孝典. オープンデータの linked data への変換〜ボキャブラリ統一とリンク付与の自動化〜. 第 34 回セマンティックウェブとオントロジー研究会, november 2014.

[7] 玉川奨, 香川宏介, 森田武史, 山口高平. 大規模 linked open data のための日本語語彙の構築. 人工知能学会論文誌, Vol. 29, No. 4, pp. 386–395, june 2014.

[8] 田仲正弘, 石田亨. 表構造の一般化に基づくオントロジの獲得. 情報処理学会論文誌, Vol. 47, No. 5, pp. 1530–1537, may 2006.

[9] 保田正則, 南石晃明. Gui ベース連邦型データベースの開発と農業情報データの空間的・時間的統合への適用. 農業情報研究, Vol. 10, No. 1, pp. 37–52, march 2001.

[10] 石橋直樹, 細川宜秀, 清木康. 時空間的文脈に応じた動的関連性計量機構を有する異種データベース間結合方式. 情報処理学会論文誌データベース（TOD）, Vol. 43, No. 2, pp. 128–145, march 2002.

[11] 武田英明, 加藤文彦, 小出誠二, 松村冬子, 大向一輝, 小林巌生, 岩山真, 浅野優, 濱崎雅弘. 統計データの lod 化とデータ間の関係の表現. No. 1N4-OS-10b-6. 人工知能学会全国大会 (第 27 回), june 2013.

[12] 西出頼継, 本間維, 永森光晴, 杉本重雄. 日本の open data 活用を目的としたデータセットのスキーマ分析とリンク関係の調査. 研究報告情報基礎とアクセス技術（IFAT）, Vol. 2013-IFAT-112, No. 4, pp. 1–8, september 2013.