情報処理学会論文誌　データベース　Vol.13　No.3　10–21 (July 2020)

表からの量的データ属性間の関係抽出

藤岡 周平[1,a)]　加藤 誠[2]　吉川 正俊[1]

受付日 2019年12月10日, 採録日 2020年4月1日

概要：本論文では，表中の数値を表す属性のなかから同一属性の対，および，上位下位関係にある属性の対を抽出する問題に取り組む．同一属性抽出においては，まず既存の表解釈手法を利用して，表中の各タプルが表すエンティティを特定する．異なる表中の 2 つの属性において，同一のエンティティを表すタプルの多くが同じ数値を含むのであれば，それらの属性は同一であると判定する．上位下位関係の抽出では，1 つの表中の量的データを表す属性の集合から，上位下位関係が成立している可能性が高い属性の対を抽出する．各タプルごとに，すべての下位属性候補の数値の和と上位属性候補の数値が一致すれば，それらに上位下位関係が成立していると判断する．提案手法の有効性を示すために，Web からクロールされ数値を含むと判断された，103,706 個の表を利用して実験を行った．提案手法および比較手法によって得られた同一属性の対を評価した結果，提案手法は適合率および再現率の面から優れた結果を示し，また，従来の属性名の類似性に基づく手法とはまったく異なる属性対が得られることを明らかにした．上位下位関係にある属性対の抽出についても実験を行い，比較手法よりも高い適合率と再現率を提案手法によって達成できることを示した．

キーワード：量的データ，情報統合，スキーママッチング

Extracting Relationship between Quantitative Data Attributes from Tables

Shuhei Fujioka[1,a)]　Makoto P. Kato[2]　Masatoshi Yoshikawa[1]

Received: December 10, 2019, Accepted: April 1, 2020

Abstract: In this paper, we tackle problems of finding identical attribute pairs and attribute pairs for which is-a relationship holds, from tables comprising numerical attributes. The identical attribute extraction firsts identifies entities represented by each tuple in tables by using an existing table understanding method. If most of the tuples representing the same entities contain the same value for two attributes in different tables, the two attributes are considered identical. The is-a relationship attribute extraction first extracts all the attribute pairs from tables for which is-a relationship can hold. We assume that is-a relationship holds for those attributes if the sum of the values of the sub attribute candidates equals to the value of the super attribute candidate. To demonstrate the effectiveness of our proposed approach, we conducted experiments with 103,706 tables that were crawled from the Web and were judged as including numerical values. Evaluating identical attributes pairs found by our proposed method and baseline methods, we found that the proposed method outperformed baselines in terms of precision and recall, and produced identical attribute pairs different from those found by a similarity-based method. We also evaluated is-a relationship attribute pairs and demonstrated that the precision and recall of the proposed method were higher than those of the baseline method.

Keywords: quantitative data, information integration, schema matching

1. はじめに

近年，官公庁や企業などによって様々な「オープンデータ」が公開されており，研究や調査，マーケティングなどの種々の目的のために，多様な統計データが利用可能になってきている．たとえば，政府統計の総合窓口（e-Stat）[*1]で

[1] 京都大学大学院情報学研究科
Graduate School of Informatics, Kyoto University, Kyoto 606–8501, Japan
[2] 筑波大学図書館情報メディア系
Faculty of Library, Information and Media Science, University of Tsukuba, Tsukuba, Ibaraki 305–8550, Japan
[a)] fujioka@db.soc.i.kyoto-u.ac.jp

[*1] https://www.e-stat.go.jp/

は，国勢調査や労働力調査といった統計調査の結果や，農林業センサス，工業統計調査といった産業統計などがCSVやExcel，PDFなどの形式で公開されている．また，国外においてもLondon Datastore[*2]やEurostat[*3]といった統計情報サイトが存在している．これに加えて，公開されている統計データを利用したサービスの開発もさかんに行われている．たとえば，米国のPredPol[*4]では，オープンデータとして公開されている，過去の犯罪発生データに基づいて，予測モデルを構築することで精度の高い犯罪予測を実現している．また，e-Statで公開されている統計データを引用した論文も多く見られることから，科学研究におけるオープンデータの有益性も見てとれる．

しかしながら，統計データはオープンデータとして公開されてはいるものの，多くのデータは統一的なオントロジーで記述されていない．そのため，同じ属性であっても，一貫した言語表現や構造によって記述されているとは限らない．この問題は，各々の企業や地方自治体，民間組織などによって集計された統計データを扱う際に顕著である．データを収集・公開する組織は，データを自組織のWebサイトに公開することが多く，組織が異なれば，自然と公開Webサイトも異なるため，属性表現の一貫性が保たれないことが多い．**表1**，**表2**に同じ属性が異なる言語表現によって記述されている場合の例を示す．表1はトヨタ，表2はホンダの車種ごとの燃費を表す．どちらも車種の燃費という同一の属性を表しているが，燃費を表す属性が表1では「燃費」，表2では「燃料あたりの走行距離」と記述されておりカラム名が統一されていない．

別の問題として，包含関係にある統計データが分割された形で公開されていることも問題となりうる．**表3**は各国の穀物消費量を，**表4**は各国の米消費量を表している．穀物消費量と米消費量の間には上位下位関係が成立しており，穀物消費量には米消費量が含まれている．

属性の同一性，および，上位下位関係が明らかではない場合には，統計データを検索するというシナリオ[*5]において様々な不都合が生じる．たとえば，表1と表2の例においては，これら表中のカラム名などで索引付けを行うのであれば，「燃費」というクエリで単純なブーリアン検索を行った場合に表2を得ることができない．同じように，表3と表4の例においては，「穀物消費量」というクエリで表4を得ることができない．理想的には，属性の同一性と上位下位関係を利用することによって，これらの表どうしは統合された形で検索結果として構成され利用者に提供されるべきである．

[*2] https://data.london.gov.uk/
[*3] http://ec.europa.eu/eurostat
[*4] http://www.predpol.com/
[*5] たとえば，情報アクセス評価のワークショップNTCIR-15では統計データの検索タスクが提案されている．https://ntcir.datasearch.jp/

表1　トヨタの車種別燃費
Table 1　Fuel consumption of each Toyota car.

車種	燃費
アクア	34.4
スペイド	22.2
タンク	21.8
パッソ	28.0
...	...

表2　ホンダの車種別燃料あたり走行距離
Table 2　Mileage per liter of each Honda car.

車名	燃料あたりの走行距離
アコード	31.6
ヴェゼル	27.0
NSX	12.4
オデッセイ	26.0
...	...

表3　各国の穀物消費量
Table 3　Amount of cereal consumption in each country.

国	穀物の食用消費量
アメリカ	33.8
中国	209.0
インドネシア	48.8
インド	185.8
...	...

表4　各国の米消費量
Table 4　Amount of rice consumption in each country.

国名	米の消費量
バングラデシュ	35.2
インド	98.1
フィリピン	13.2
日本	8.0
...	...

そこで本論文では，表中の数値を表す属性の中から同一属性の対，および，上位下位関係にある属性の対を抽出する問題に対して取り組む．表の属性の同一性判定については，スキーママッチングの分野で広く取り組まれており，多くの既存研究が存在する[4], [5], [11], [13], [18]．既存の同一性判定手法では，**質的データ属性**（血液型やアンケートの段階評価といった名義尺度と順序尺度によって測られる値を表す属性）を対象とし，属性値を手がかりとして，その属性が何であるかを推定し属性の同一性判定を行っている．たとえば，Limayeら[7]の方法では，表1の「アクア」や表2の「アコード」などの値から，知識ベースなどを利用することで，これらのカラムが「車種」という属性を表しており，同一の属性であると判定することができる．しかしながら，本研究で対象とするような，**量的データ属性**（気温や体重といった間隔尺度または比例尺度に

よって測られる数量を表す属性）を扱う場合には，属性値を手がかりにしてその属性が何であるかを推定することは困難である．たとえば，表 1 の「34.4」や表 2 の「31.6」などの値から，既存の方法によってそれらの属性が「燃費」であることを推測するのは難しい．「燃費」や「燃料あたりの走行距離」などのカラム名どうしの類似性に基づく方法 [3], [4], [8], [10] も提案されているが，類似するカラム名が必ずしも同一の属性を表しているとはいえない．そこで，我々は量的データ属性であっても，同一であるような属性を抽出する方法として「異なる表中の 2 つの属性において，同一のエンティティを表すタプルの多くが同じ数値を含むのであればそれらの属性は同一である」という考えに基づいた方法を提案する．

量的データ属性間の上位下位関係の抽出方法としては，ある属性のすべての下位属性の値をタプルごとに合計すれば，各タプルにおけるその属性の値と等しくなるはずである，という考えを利用する．たとえば，「穀物消費量」と各種穀物の消費量が記載された表において，各タプルごとに米や小麦，トウモロコシの消費量などを合計し，それが同じタプルの「穀物消費量」の値と等しくなれば，「米消費量」や「小麦消費量」などは「穀物消費量」の下位属性であると判断する．

提案手法の有効性を示すために，Web からクロールされ数値を含むと判断された，103,706 個の表を利用して実験を行った．同一属性抽出の問題については，比較手法としてカラム名の類似度を用いた方法と，単に値が一致するタプル数の割合を用いた方法を採用し比較を行った．実験結果から，提案手法は適合率および再現率の面から優れた結果を示し，また，従来の属性名の類似性に基づく手法とはまったく異なる属性対が得られることが明らかになった．上位下位関係にある属性対の抽出についても実験を行い，比較手法よりも高い適合率と再現率を提案手法によって実現できることを示した．

本論文の貢献は以下のとおりである．(1) 量的データを表す属性から同一属性，および，上位下位関係にある属性を抽出する問題を提案した．(2) 量的データ属性の同一属性抽出に関し，数値の一致に基づく抽出方法を提案した．(3) 量的データ属性の上位下位関係抽出に関し，合計値の一致に基づく抽出方法を提案した．(4) 大規模な表データから属性間の同一関係，および，上位下位関係を抽出し，提案手法の有効性を評価した．

本論文の構成は以下のとおりである．2 章ではスキーママッチング，および，表の解釈に関する関連研究について述べる．また，同一属性抽出に利用する既存手法についても説明する．3 章では，各問題設定，および，提案手法の詳細について述べる．4 章では，実験で利用するデータ，各問題に対する実験内容，および，得られた実験結果を示す．最後に，5 章にて本論文の結論を述べる．

2. 関連研究

本研究の関連研究として，スキーママッチング，および，表の解釈についての研究があげられる．以下ではそれぞれの研究について，本研究との差異や関係性について述べる．また，表の解釈についての研究は，提案手法のなかで活用するため詳細に説明する．

2.1 スキーママッチング

2 つの異なるスキーマどうしにおいて，対応関係を発見することをスキーママッチングとよぶ．スキーママッチングの主な手法として，言語的マッチング，補助情報利用マッチング，インスタンスベースマッチング，構造ベースマッチング，制約ベースマッチング，ルールベースマッチングという 6 つの方法がある [1]．

量的データ属性に対しては，主にインスタンスベースマッチングの方法が提案されてきた．Sahay らはある属性の値について最大，最小，平均などの代表値を算出し，これらの類似性に基づいて属性の同一性判定を行っている [12]．Mehdi らは，量的データ属性の代表値などに基づいて，量的データの外形的パターンを正規表現で記述し属性間の同一性判定に利用している [9]．これらの 2 つの手法については，属性間の単位の違いやスケールが異なる場合には対応できない問題，および，表の間で出現するエンティティが異なる場合には同一だと見なされない問題がある．Chua らは，ある量的データ属性間の同一性を回帰によって判定する方法を提案している [2]．エンティティを利用する点や定量的データを対象にする点は本研究と共通するが，(1) エンティティの特定のために社員番号や ISBN など，広く用いられるキーが存在することを仮定している点，および，(2) ある属性が別の属性に基づく回帰によって説明される場合に同一と見なす点，が本研究とは異なる．

Doan らは，複数の学習器を組み合わせたメタ学習器を用いてスキーママッチングに取り組んだ [4]．メタ学習器を構成する学習器の 1 つである，単語の出現頻度に基づく学習器は，数量データに対してはうまく機能しないと論文中では説明されている．

Melnik らは，スキーマをラベル付きのグラフに変換し，そのグラフの各ノード間でラベルの文字列の類似度を算出し，その値に基づいて適切なマッチングを発見するというアルゴリズムを提案した [10]．この研究では表中の値そのものは利用しておらず，我々の手法とは異なる．

Madhaven らは，スキーマの要素の名前，データ型，制約，および，スキーマ構造をもとに，言語的マッチング，構造ベースマッチングを順に利用してスキーママッチングを行った [8]．しかし，Melnik らの手法と同様に，提案された手法は量的データ属性に特化しておらず，表中の値自体も利用しないという点で我々の手法とは異なる．

Do らは，スキーママッチングに対して 3 つの異なる手法を組み合わせて取り組んだ [3]．組み合わされた手法のうち，2 つはスキーマ要素の名前の類似性を主として利用し，残りの 1 つはスキーマ構造の類似性を主として利用している．すなわち，言語的マッチングと構造ベースマッチングを組み合わせたマッチングを行っている．この手法も他の手法と同様に，量的データ属性を対象としていない．

2.2　表の解釈

Yakout らの研究は，Web で公開されている表において，(1) 属性名を利用した表の空欄部分の補完，(2) 表内で与えられたエンティティを利用した表の空欄部分の補完，(3) 表内で表現されているエンティティを手がかりにした関連する属性の発見，という 3 つの操作を提案し，これらの自動化に取り組んでいる [16]．表の属性，および，表中の値を扱う研究ではあるが，主に質的データ属性を対象としており，属性値とエンティティが対応づけられることを仮定している．一方で，我々が扱う量的データ属性では，基本的には属性値とエンティティを対応づけられない．

Zhang らの研究は，表からの属性情報の自動抽出において，単位やスケールが異なる属性や時間依存であるような属性に対しても，正確に情報を抽出することを目的としている [17]．この研究では，事前知識として単位間の変換式などを想定しており，実験においても，会社の収入や国の面積など代表的な属性についてのみ評価が行われるなど，理想的な条件下を想定した手法を提案している．

Limaye らの研究では，表内のセル，属性，属性間の関係に対して，それぞれ，エンティティ，カテゴリ，関係を付与する機械学習手法を提案している [7]．カテゴリとはエンティティの種類を表すものであり，国，首都，日本の都市，果実などがあげられる．関係はエンティティ間の述語であり，兄弟関係や人物の出身地，会社の所在地などの例があげられる．これらはすべて知識ベースなどに記述されており，Limaye らの実験においては，YAGO [6] が知識ベースとして用いられた．

特に，カテゴリとエンティティの付与に着目したとき，Limaye らの手法は以下のように説明できる．

(1) ある表の属性 a に対し，あるカテゴリ c を割り当てる．
(2) 属性 a の各タプルにおける値をそれぞれ $x_1, x_2, \ldots, x_n$ とする．また，すべてのエンティティの集合を E とし，カテゴリ c に属するエンティティの集合を $E_c \subset E$ とする．各 x_k $(k = 1, 2, \ldots, n)$ に対して，

$$e_k^* = \underset{e \in E_c}{\operatorname{argmax}} \left(\phi_1(x_k, e)\phi_3(e, c)\right)$$

を求める．ここで，$\phi_1(x_k, e)$ は値 x_k とエンティティ e のテキスト類似度を表し，編集距離，および，ユニグラムとバイグラムのジャカード係数の重み付き和によって定義される．また，$\phi_3(e, c)$ はカテゴリ c 自体の特殊性，および，エンティティ e とカテゴリ c の関連性を表し，それぞれ，E_c の小ささとエンティティ e とカテゴリ c 間の距離[*6]の逆数で近似され，それらの重み付き和で定義されている．そのため，e_k^* としては，値 x_k とのテキスト類似度が高く，カテゴリ c によく関連するようなエンティティが選ばれることになる．
(3) カテゴリ c ごとに下記の値を求め，最も高くなるようなカテゴリを採用する．また，採用されたカテゴリ c を用いたときの e_k^* が各値 x_k に割り当てられる．

$$\phi_2(a, c) \sum_{k=1}^{n} \phi_1(x_k, e_k^*)\phi_3(e_k^*, c)$$

ただし，$\phi_2(a, c)$ は属性 a とカテゴリ c のテキスト類似度を表し，$\phi_1(x_k, e)$ と同様に，編集距離，および，ユニグラムとバイグラムのジャカード係数の重み付き和によって定義される．

Limaye らの手法では，表の各属性に対してカテゴリが割り当てられるため，これによって属性の同一性を判定し，同一属性を抽出することが可能である．しかし，セルの内容が量的データである場合には，Limaye らの手法を適用することができない．より具体的には，まず数値へのエンティティの割当ができない．たとえば，表 1 の「34.4」に割り当てられるべきエンティティは YAGO などの知識ベースなどには存在していない．また，「34.4」に対するカテゴリも同様に定義されていない．ただし，Limaye らの手法によって表中の各タプルが表す数量が何のエンティティに関する値なのかが分かるため，これを手がかりとして量的データの同一属性抽出手法を提案する．

3.　提案手法

本章では，まず同一な量的データ属性の抽出手法に関する説明を行い，その後に，量的データ属性における上位下位関係の抽出手法に関する説明を行う．

3.1　同一な量的データ属性の抽出手法

質的データ属性 $a_i^{(l)}$ と量的データ属性 $a_i^{(n)}$ を含み，属性 $a_i^{(l)}$ が候補キーであるような表 T_i を考える．すなわち，任意のタプル $t, t' \in T_i$ について，$t[a_i^{(l)}] \neq t'[a_i^{(l)}]$ であるような表 T_i を前提とする．ただし，$t[a]$ はタプル t の属性 a の値を表す．同一な量的データ属性の抽出においては，ある表 T_i の量的データ属性 $a_i^{(n)}$ と別の表 T_j の量的データ属性 $a_j^{(n)}$ の同一性を判定し，もし同一であれば $(a_i^{(n)}, a_j^{(n)})$ を同一な量的データ属性の対として抽出する．

同一属性の抽出手法の概要を**図 1** に示す．処理 1 として，Limaye ら [7] の手法を表 T_i の質的データ属性 $a_i^{(l)}$ に適用することにより，表 T_i に含まれる各タプルの属性 $a_i^{(l)}$

[*6] 実験でも述べるとおり，カテゴリとエンティティは木構造で表現され両者の間の距離が計算可能である．

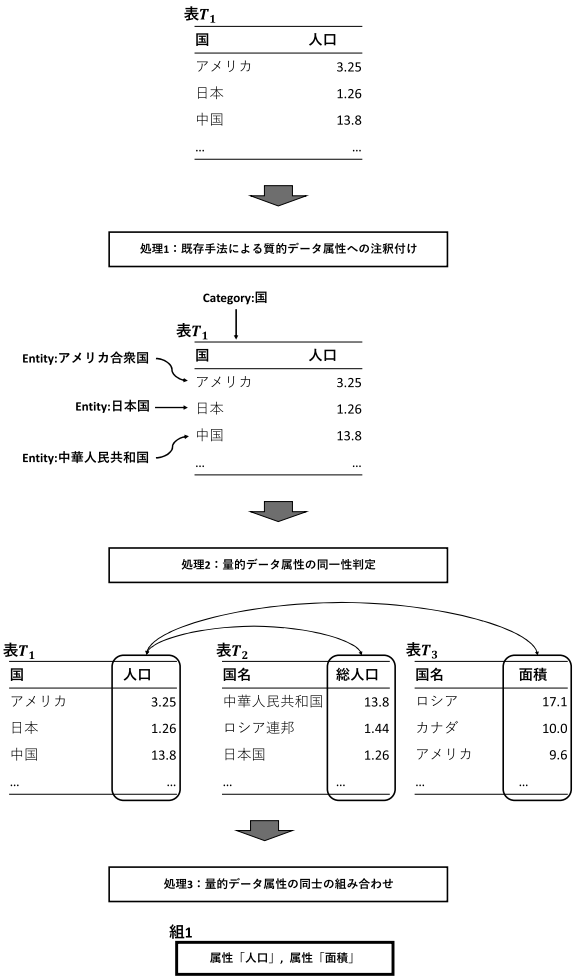

図 1　同一な量的データ属性の抽出手法の概要

Fig. 1　Overview of the extraction method for identical quantitative attributes.

の値に対してエンティティを，属性 $a_i^{(l)}$ に対してカテゴリを割り当てる．図の例では，表 T_1 の「国」と記述された質的データ属性 $a_i^{(l)}$ にカテゴリ「国」を割り当て，1 つ目のタプルの「アメリカ」という値に対してエンティティ「アメリカ合衆国」を割り当てている．次に，処理 2 として，質的データ属性 $a_i^{(l)}$，$a_j^{(l)}$ に対して同一のカテゴリが割り当てられた 2 つの表，T_i，および，T_j において，同一のエンティティが割り当てられたタプルに着目する．これらのタプルにおける量的データ属性 $a_i^{(n)}$，および，$a_j^{(n)}$ の値が，十分に近ければ属性 $a_i^{(n)}$ と $a_j^{(n)}$ は同一である可能性が高いと仮定する．図の例では，表 T_1 の「日本」という値を含むタプルと表 T_2 の「日本国」という値を含むタプルの，属性「人口」と属性「総人口」の値は同一である．また，「中国」および「中華人民共和国」という値を含むタプルにおいても同様である．そのため，表 T_1 の属性「人口」と表 T_2 の属性「総人口」は同一である可能性が高いと判断できる．一方で，表 T_1 の「アメリカ」という値を含むタプルと表 T_3 の「アメリカ」という値を含むタプルの，属性「人口」と属性「面積」の値は大きく異なる．もし，他のタプルについても同様にこれらの属性の値が大きく異なるのであれば，属性「人口」と属性「面積」は同一属性ではないと推定される．最後に，処理 3 として，同一である可能性が高い属性を組にして抽出する．たとえば，表 T_1 の属性「人口」と表 T_2 の属性「総人口」を同一な属性として抽出する．提案手法では，これら 3 つの処理を各表に対して行うことによって同一属性を抽出することになる．

以下では，同一な量的データ属性の抽出手法についてより厳密な説明を与える．前提として，Limaye ら [7] の手法によって，各表 T_i の質的データ属性 $a_i^{(l)}$ に対してカテゴリ c_i が割り当てられ，各タプル $t_i \in T_i$ の質的データ属性 $a_i^{(l)}$ の値 $t_i[a_i^{(l)}]$ について，エンティティ $e(t_i)$ が割り当てられているとする．

(1) 質的データ属性が同一であるような，すなわち，$c_i = c_j$ であるような，2 つの表 T_i と T_j において，質的データ属性の値に対して同一のエンティティが割り当てられたタプル対の集合を得る．すなわち，

$$P_{ij} = \{(t_i, t_j) \mid t_i \in T_i,\ t_j \in T_j,\ e(t_i) = e(t_j)\}$$

を求める．

(2) 同一のエンティティが割り当てられた値を含むタプル対のうち，表 T_i と T_j の量的データ属性 $a_i^{(n)}$ と $a_j^{(n)}$ における値が，**誤差を許容して一致**（後述）するタプル対を得る．すなわち，

$$P_{ij}^{+} = \{(t_i, t_j) \mid (t_i, t_j) \in P_{ij},\ t_i[a_i^{(n)}] \simeq t_j[a_j^{(n)}]\}$$

を求める．

(3) 2 つの表 T_i と T_j に含まれるタプルの，質的データ属性 $a_i^{(l)}$，および，$a_j^{(l)}$ の値に割り当てられた，すべてのエンティティの集合は

$$E_{ij} = \{e(t_i) \mid t_i \in T_i\} \cup \{e(t_j) \mid t_j \in T_j\}$$

と定義される．これらのエンティティのうち，量的データ属性の値が誤差を許容して一致するエンティティの集合は，

$$E_{ij}^{+} = \{e(t_i) \mid (t_i, t_j) \in P_{ij}^{+}\} = \{e(t_j) \mid (t_i, t_j) \in P_{ij}^{+}\}$$

である．

(4) 同一のエンティティが割り当てられたタプル対のうち，量的データ属性の値が誤差を許容して一致するタプル対の割合は $|P_{ij}^{+}|/|P_{ij}|$ であり，2 つの表 T_i と T_j に割り当てられたすべてのエンティティの集合のうち，量的データ属性の値が誤差を許容して一致するエンティティの割合は $|E_{ij}^{+}|/|E_{ij}|$ となる．2 つの割合が十分に高ければ，量的データ属性 $a_i^{(n)}$ と $a_j^{(n)}$ は同一であると判断される．より具体的には，

$$|P_{ij}^{+}|/|P_{ij}| \geq \tau_P \ \wedge\ |E_{ij}^{+}|/|E_{ij}| \geq \tau_E$$

を満たす場合に，属性対 $(a_i^{(n)}, a_j^{(n)})$ を同一属性として抽出する．ただし，τ_P，および，τ_P はハイパパラメータである．なお，$|E_{ij}^{+}|/|E_{ij}|$ が高ければ，属性対が同一である可能性がより高いと考え，実験においてはこの値が高い属性対を優先的に評価した．

上記において，量的データ属性の値が誤差を許容して一致するとは，表 T_i の量的データ属性 $a_i^{(n)}$ と表 T_j の量的データ属性 $a_i^{(n)}$ のそれぞれにおいて，両表において共通するエンティティを含むタプルについて，属性の値の平均が 0，分散が 1 になるように標準化を行ったうえで以下のように定義される．

$$t_i[a_i^{(n)}] \simeq t_i[a_j^{(n)}] \Leftrightarrow |t_i[a_i^{(n)}] - t_i[a_j^{(n)}]| < \theta_I$$

ただし，θ_I は誤差の閾値でありハイパパラメータである．ここで，標準化を行ったのには，主に 2 つの理由がある．1 つ目の理由は，単位がそろっていない場合についても属性の同一性が判定できるようにするためである．標準化によって，一方が他方に対して定数倍の関係になっている場合（円と千円や円とドルなど）についても，同一であると判定することができる．2 つ目の理由は，誤差自体を標準化するためである．たとえば，10^9 における 1 の誤差と 10^3 における 1 の誤差では，後者の誤差の方が大きいと見なされるべきであり，標準化を行った場合には，実際にそれぞれの誤差を比べると後者の方が大きくなる．

3.2　量的データ属性間の上位下位関係の抽出手法

量的データ属性の集合 $A_i^{(n)}$ を持つ表 T_i を考える．量的データ属性間の上位下位関係の抽出においては，ある表 T_i の量的データ属性の部分集合 $\tilde{A}_i^{(n)} \subset A_i^{(n)}$ と，ある量的データ属性 a $(a \in A_i^{(n)},\ a \notin \tilde{A}_i^{(n)})$ に上位下位関係が成立するかを判定し，もし成立する可能性が高ければ，$\{(a, \tilde{a}) \mid \tilde{a} \in \tilde{A}_i^{(n)}\}$ を上位下位関係が成立する量的データ属性の対として抽出することを考える．

上位下位関係の抽出手法の概要を**図 2** に示す．図中の表 T_i には，「総人口」，「15 歳未満」，「15〜64 歳」，「65 歳以上」の 4 つの量的データ属性が含まれている．この例では，量的データ属性の部分集合 $\tilde{A}_i^{(n)}$ として，「15 歳未満」，「15〜64 歳」，「65 歳以上」を考え，残りの量的データ属性「総人口」との上位下位関係を判定する．このとき，「北海道」という値を含む 1 つ目のタプルにおいて，属性集合 $\tilde{A}_i^{(n)}$ の値を合計すると，$577 + 3052 + 1656 = 5285$ のように計算され，この値は同タプル中の「総人口」の値と等しい．もし，他のタプルのにおいても同様に，属性集合 $\tilde{A}_i^{(n)}$ の値の和が，属性「総人口」の値と等しいのであれば，属性「総人口」が上位，属性「15 歳未満」，「15〜64 歳」，「65 歳以上」が下位であるような上位下位関係が推定される．

以下では，量的データ属性間の上位下位関係の抽出手法についてより厳密な説明を与える．

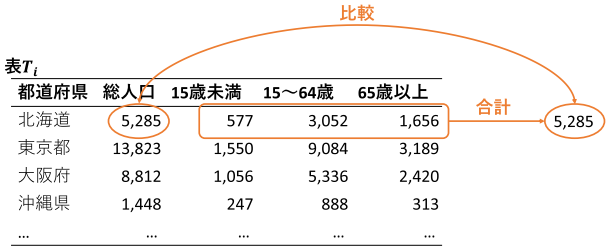

図 2　上位下位関係の各候補における妥当性の判定手法
Fig. 2　Overview of the validation method for is-a relationship candidates.

(1) 量的データ属性の集合 $A_i^{(n)}$ を持つ表 T_i を考える．各属性 $a \in A_i^{(n)}$ に対して，$|\tilde{A}_i^{(n)}| \geq 2$ を満たすような，部分集合 $\tilde{A}_i^{(n)} \subset A_i^{(n)}/\{a\}$ が属性 a の下位属性であるかを判定する．

(2) 属性集合 $\tilde{A}_i^{(n)}$ の値の合計値と属性 a の値が誤差を許容して一致するようなタプル，すなわち，

$$T_i^{+} = \left\{ t \;\middle|\; t \in T_i,\ t[a] \simeq \sum_{\tilde{a} \in \tilde{A}_i^{(n)}} t[\tilde{a}] \right\}$$

を求める．

(3) 属性集合 $\tilde{A}_i^{(n)}$ の値の合計値と属性 a の値が誤差を許容して一致するようなタプルの割合がある閾値以上ならば，すなわち，

$$|T_i^{+}|/|T_i| \geq \tau_T$$

であれば，属性 a が属性集合 $\tilde{A}_i^{(n)}$ 中の各属性と上位下位関係にあると推定する．ただし，τ_T はハイパパラメータである．

上記において，値 x と y が誤差を許容して一致するとは以下のように定義される．

$$x \simeq y \Leftrightarrow \frac{|x - y|}{\min\{x, y\}} \leq \theta_S$$

ただし，θ_S は誤差の閾値でありハイパパラメータである．同一属性の抽出で用いていた定義では標準化を行っていたが，属性値の合計を求める上位下位関係の抽出手法ではそのまま用いることができない．そのため，こちらの抽出方法では，誤差をより小さい方の値で割ることによって相対的な誤差を求めている．

ただし，事前実験において，単純に上記の手法を適用した場合の適合率が低かったため，下記のとおり，属性対に対する制約条件を提案し提案手法の適合率の改善を図った．

1 つ目の制約条件は，下位属性の特殊性に関する必要条件を利用したものである．任意の下位属性はある上位属性よりも特殊性が高いため，下位属性であることの必要条件として特殊性が十分高いことがあげられる．具体的には，提案手法によって得られた属性対における下位属性の言語

表現が閾値 θ_L 以上の長さを持つ場合のみ，その属性対を採用することにする．この条件を加えた提案方法を，実験では**提案手法+特殊性**と表記する．

2 つ目の制約条件は，1 つの下位属性が上位属性の値の大部分を占めない，という条件である．いい換えれば，上位属性の値に対し，下位属性はある程度均一な値をとるということである．1 つの値が上位属性の値の過半数以上を占めるような場合には，他の属性と釣り合いがとれておらず，上位下位関係が成立していない場合が多く見られたためこの条件が提案された．厳密には，提案手法によって表 T_i から属性対 $(a, \tilde{a})$ $(\tilde{a} \in \tilde{A}_i^{(n)})$ が，属性 a の値と属性集合 $\tilde{A}_i^{(n)}$ の合計値との比較によって得られた場合に，任意の属性 $\tilde{a}' \in \tilde{A}_i^{(n)}$ について，あるタプル $t \in T_i$ が存在して，

$$t[\tilde{a}']/t[a] \leq \theta_H$$

を満たす属性対 $(a, \tilde{a})$ のみを選定する．この条件を加えた提案方法を，実験では**提案手法+均一性**と表記する．

上記の 2 つの制約条件を加えた提案手法を，実験では**提案手法+特殊性&均一性**と表記する．なお，ハイパパラメータである θ_L，および，θ_H の値は，評価のためのデータとは異なるバリデーションデータによって決定された．詳細は実験にて述べる．

4. 実験

本章では，まず実験に使用したデータの詳細を述べる．次に，実際の実験において採用した設定について説明する．その後，同一属性抽出と上位下位関係抽出のそれぞれについて，実験設定を説明しその結果について議論を行う．

4.1 実験データ

実験では，Common Crawl [*7] にて公開されている，2018 年 10 月から 2019 年 4 月の間に収集された日本語 Web ページを利用した．実験用のデータとして，これらのページから，数値を含み，かつ，十分な大きさを持った 5,783,365 個の表を抽出した．十分な大きさを持つ，とは，行数と列数が 2 以上であること，セル数が 50 以上であること，すべてのセル中の文字数が 100 未満であること，さらに，空白なセルの数が 25%未満であることである．ただし，これらの条件を課しても，なお，レイアウトのためだけに用いられるような表が多く含まれていたため，さらなる選定を行ったうえで実験を行った．

最終的には，選定条件として，(1) 表の行数がヘッダ部分を含めて 10 以上であること，(2) 表の列数が 3 以上であること，(3) セルの内容と名前が完全一致するエンティティの数をカテゴリ別に数えた場合，いずれかのカテゴリにおけるエンティティの数が 4 以上であること，(4) 縦，横の両方においてセルの結合がないこと，(5) 表にヘッダが存在すること，(6) すべてのタプルにおいて列数が等しいこと，(7) 表内に質的データ属性と量的データ属性を少なくとも 1 つずつ含むこと[*8]，(8) 質的データ属性に対してカテゴリとエンティティを割り当てられること，という条件を設定した．これらの条件を適用した結果として，表の数は 103,706 個，それらの表に含まれる量的データ属性のカラム数は 360,361 個となった．

また，質的データ属性へカテゴリ，および，エンティティを割り当てる際に用いたデータは以下のとおりである．カテゴリは本論文の第 1 著者によって選定され，国，日本の都道府県，スポーツ，野菜など，89 個が用意された．次に，2010 年 6 月 24 日の Wikipedia のダンプデータに対して上位下位関係抽出ツール [14], [15], [19] を適用することによって語間の上位下位関係を得た．これらの関係は，木構造によって表現することが可能であり，カテゴリとして選ばれた語の子孫のうち，カテゴリからの距離が 5 以下であるようなすべての語を，そのカテゴリに属するエンティティとして採用した．エンティティとして扱われる語の例としては，アメリカ合衆国，京都府，サッカー，にんじんといったものがあげられる．なお，YAGO や DBPedia などの知識ベースを利用しなかったのは，これらの知識ベースの日本語版では，一部の代表的なカテゴリ（日本の都道府県など）などに対して，適切な上位下位関係が得られなかったためである．

4.2 同一な量的データ属性の抽出

まず，同一属性抽出に特有の前処理について説明する．その次に，比較手法，および，実験設定について述べ，最後に，同一属性抽出の実験結果について述べる．

4.2.1 前処理

同一な量的データ属性の抽出を行うために，いくつかの前処理を表に対して施した．(1)「順位」や「割合」を表す属性は表データに頻出し，得られる同一性が自明であることが多かったため，実験の合理性の観点から，これらを表す名称が用いられている属性を削除した．(2) 表内の量的データ属性のうち，量的データ属性の値が 1 から始まる連番になっている値を含む属性を削除した．(3) 表内の質的データ属性のうち，いずれかがすべてのタプルにおいて一意になっているか，すなわち，候補キーであるかを確認した．1 つの質的データ属性だけが候補キーであれば，その属性と表中の各量的データ属性のみからなる表を新たに作った．(4) (3) の結果得られた各表において，量的データ属性の値が欠損値になっている場合にはそのタプルを削除した．(5) 各表において，量的データ属性の値が 0 など特定の値に極端に偏っている場合，すなわち，ある値を持つ

[*7] https://commoncrawl.org/

[*8] 「前後に数文字の非数値が含まれることを許容する数値」を含むセルを持つ属性が量的データ属性として判定され，それ以外は質的データ属性として判定された．

タプルの割合が全体の 70%以上であるときに，その値をもつタプルを削除した．(6) 上記の処理の結果のうち，行数が 10 以上の表のみを採用した．

上記の前処理を施した結果，1 つの質的データ属性と 1 つの量的データ属性だけからなる，合計 19,893 個の表が得られた．

4.2.2　比較手法

同一属性を抽出する実験において，以下の 2 つの比較手法を採用し提案手法と比較を行った．

1 つ目の比較手法は，属性の言語表現の類似度に基づく手法である．これを**言語類似度手法**と表記する．この方法では，ある表 T_i の量的データ属性 $a_i^{(n)}$ と別の表 T_j の量的データ属性 $a_j^{(n)}$ が，

$$D(s(a_i^{(n)}), s(a_j^{(n)})) \leq \theta_C$$

を満たすようなすべての属性対を同一である属性として抽出する．ただし，$D(x, y)$ は文字列 x，y の編集距離であり，$s(a)$ は属性 a の言語表現（表のヘッダに書かれているカラム名）を表す．なお，$\theta_C = 2$ のとき，23,886,626 個の属性対が得られ適合率が著しく低くなる可能性があったため，本実験では $\theta_C = 1$ を採用した．

2 つ目の比較手法は，提案手法において，質的データ属性を考慮せず，単に量的データ属性の値が完全に一致する割合に基づいて同一性を判定する手法である．これを**数値一致度手法**と表記する．表 T_i の量的データ属性 $a_i^{(n)}$ と T_j の量的データ属性 $a_j^{(n)}$ における，すべての値のうち，両表において共通する値の割合が閾値 τ_B 以上であれば，すなわち，

$$\frac{|V_i \cap V_j|}{|V_i \cup V_j|} \geq \tau_B$$

を満たす場合に属性対 $(a_i^{(n)}, a_j^{(n)})$ を同一属性として抽出する．ただし，$V_i = \{t_i[a_i^{(n)}] \mid t_i \in T_i\}$ は表 T_i において量的データ属性 $a_i^{(n)}$ がとるすべての値であり，V_j も同様に定義される．得られた結果を見ながら閾値の調整を行い，実験では $\tau_B = 0.4$ を採用した．

4.2.3　実験設定

前述の前処理を施して得られた，1 つの質的データ属性と 1 つの量的データ属性だけからなる表に対して，提案手法を適用することにより同一であると推定される属性の組を抽出した．本実験では，ハイパパラメータ τ_P，τ_E をそれぞれ 0.80，0.40 に設定した．また，ハイパパラメータ θ_I は同一と判定されるべき属性値の差を確認しながら決定し，最終的に $\theta_I = 0.005$ を採用した．

次に，各手法で得られた属性対のすべてを評価するのはコストが大きいため，各手法から 430 対を選んで評価を行った．ただし，評価対象となる属性の表記ができるだけ重複しないように，また，評価対象となる属性のカテゴリにできるだけ偏りが生じないように属性対の選別を行った．さらに，言語表現が異なるような属性対の同一性を評価するために，まったく同じ言語表現を持つ属性の組も除いた．

表 5 手法ごとのカテゴリ別に得られた属性対の数
Table 5 Number of attribute pairs obtained by each method in each category.

カテゴリ	提案	言語類似度	数値一致度
貨幣	48	21	10
企業	48	20	28
国	48	21	28
動物・生物・哺乳類・人体の部位・節足動物	48	102	125
日本の都道府県	48	21	27
惑星	48	21	0
人物	47	20	28
都市	45	21	27
公共サービス・教育・スポーツ・美術作品	36	60	34
物質・現象	14	10	30
映画	0	6	0
果物	0	7	8
魚類	0	21	10
金属	0	4	0
言語	0	2	0
作品	0	21	26
世界遺産	0	4	0
地方区分	0	2	0
調味料	0	8	14
天体	0	21	28
道具	0	2	3
農業	0	8	2
宝石	0	7	0
発酵食品	0	0	2

表 5 に，各手法で得られ選択された属性対の数を，質的データ属性に割り当てられたカテゴリ別に示す．ただし，一部のカテゴリに関しては含まれる属性対が少なかったため，複数のカテゴリをまとめて 1 つのカテゴリとして示している．いくかのカテゴリにおいて，提案手法が同一と判定した属性対がなかった理由としては，より厳しい条件による判定を行っていることと，当該カテゴリにおいて表間で重複するエンティティが少なかったことが考えられる．

各手法で抽出された属性対は，クラウドソーシングによって評価を行った．各属性対を 3 名のワーカに対して提示し，抽出元の表とともに提示して，同じ属性であるかどうかを判定してもらった．評価の品質を担保するために，10 個の属性対に対し 1 つの割合で，明らかに同一か同一でないかが判明している属性対を提示し，これらの属性対において，3 問以上，もしくは，5 割以上が不正解であったワーカを除外した．最終的に，3 名中 2 名のワーカが同一であると判定した属性対を正解だと判定した．

表 6　手法ごとの適合率，再現率，F 値
Table 6　Precision, recall, and F-measure of each method.

手法	正解数	適合率	再現率	F 値
言語類似度手法	106	0.247	0.133	0.173
数値一致度手法	190	0.442	0.239	0.310
提案手法	**307**	**0.714**	**0.386**	**0.501**

表 7　手法間の結果のジャカード係数
Table 7　Jaccard coefficient between the methods.

手法の組		ジャカード係数
提案手法	言語類似度手法	0.020
提案手法	数値一致度手法	0.186
言語類似度手法	数値一致度手法	0.069

4.2.4　実験結果

各手法において，正解とされた属性対の数，適合率，再現率，F 値を**表 6** に示す．ただし，すべての手法において得られた同一属性対を，全同一属性の対であると見なして再現率を計算した．提案手法は，適合率，再現率，F 値のすべての指標において，比較手法に比べて高い値を達成している．

また，提案手法と比較手法で得られた結果にどれくらい重複があるかを確認するために，各手法において得られ，正解だと判断された属性対のジャカード係数を求めた．これを**表 7** に示す．各手法間のジャカード係数はいずれも高くないが，提案手法と数値一致度手法のジャカード係数から，両手法が似た属性対を出力していることが分かる．実際に，数値一致度手法は提案手法で考慮していた質的データ属性や誤差の許容などを除いた方法であり，基本的な考え方は同一であるため，この結果は予期されていた結果である．一方で，提案手法と言語類似度手法の間の低いジャカード係数は，両手法から得られる結果が大きく異なることを示唆している．実際に，提案手法によって得られる属性対を確認すると，「負担重量」と「斤量」，「軌道半径（億 km）」と「太陽からの距離（m）」などの例が見られ，属性の言語表現が大きく異なり言語類似度手法では同一と判定されない属性対に対しても，提案手法であれば同一性を判定できているということが分かる．

手法ごとのカテゴリ別の適合率を**図 3** に示す．提案手法は多くのカテゴリにおいて，比較手法と比べて高い適合率を示している．特に，カテゴリ「日本の都道府県」と「人物」に関しては優れた性能を示している．提案手法によって得られる属性対のうち，質的データ属性に割り当てられたカテゴリが「人物」であるものを確認すると，同一の属性であり似たような言語表現がされていても，後ろに単位などが付与されることで編集距離が大きくなる場合が見られた．このような場合には，言語類似度手法では同一と判定できなかったと考えられる．また，人物の場合には，体重や競技記録など値が細かく変化していく属性が多く，同

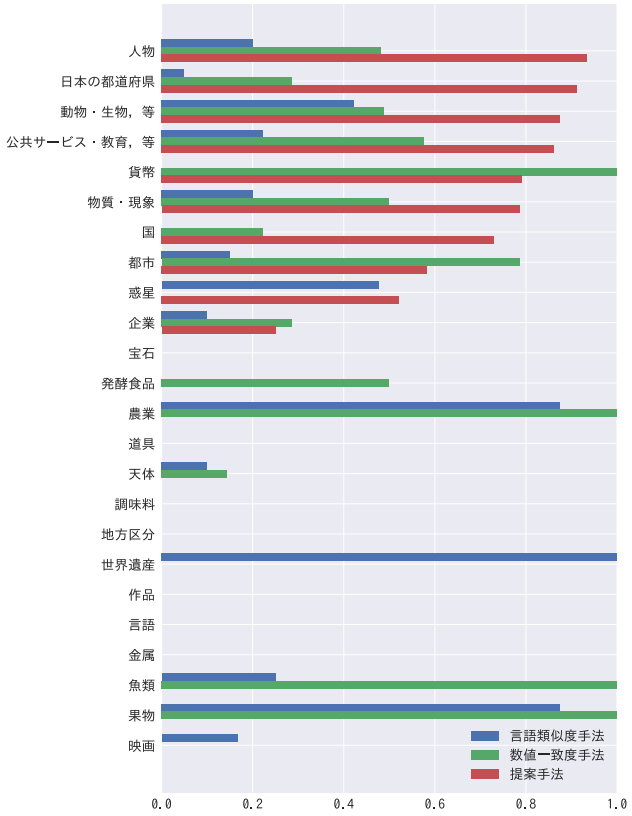

図 3　カテゴリ別の適合率
Fig. 3　Precision per category.

一属性であっても同一の値が出現しづらい．このため，完全一致のみで属性の同一性を判断する数値一致度手法は高い適合率を達成できなかったと思われる．カテゴリが「日本の都道府県」である場合には，カテゴリ「人物」と同様に「人口［人］」や「人口（単位：千）」など，単位が付加される例が多く見られた．また，同一属性であっても，異なる単位で記述されている場合もあり，そのような属性対は数値一致度手法では同一と判定されなかった．

一方で，表 5 からも分かるとおり，提案手法では属性対を得られなかったカテゴリも多く存在する．カテゴリ「世界遺産」においては言語類似度手法が，カテゴリ「農業」や「魚類」，「果物」においては数値一致度手法が，正解となる属性対を発見している．これらのカテゴリにおいて，提案手法が同一属性対を発見できなかった原因として，表間で共通するエンティティの割合が少ないことがあげられる．提案手法ではエンティティの重複率が 40%未満の表からは属性対を抽出しないため，これらのカテゴリにおいては同一属性対が抽出されなかったと推測される．提案手法では抽出されなかった同一属性対として，カテゴリ「世界遺産」の属性「登録年」と「登録」（言語類似度手法により抽出），カテゴリ「果物」の属性「糖質量（g）」と「100 g 中 糖質（g）」（数値一致度手法により抽出），などが例としてあげられる．

提案手法において，誤差を許容することにより，適合率

表 8　提案手法において誤差を許容しない場合との比較
Table 8　Comparison of the proposed method with/without allowing matching errors.

手法	正解数	適合率	再現率	F 値	取得件数
誤差許容	307	0.714	0.386	0.501	**4,040**
誤差不許容	**390**	**0.907**	**0.491**	**0.637**	3,312

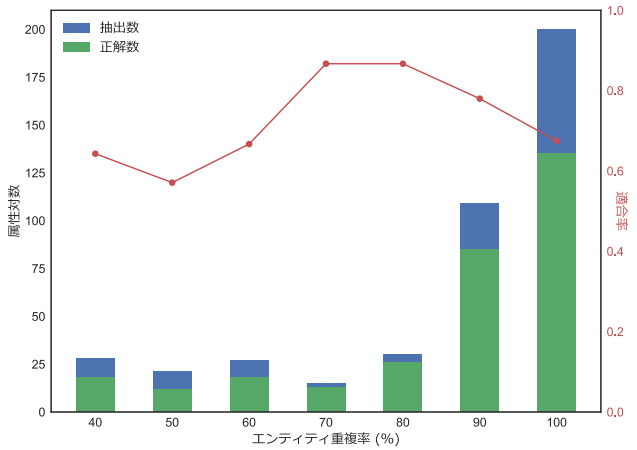

図 4　表間で重複するエンティティの割合と適合率
Fig. 4　Precision as a function of the number of overlap entities between tables.

を犠牲にしつつもより多くの同一属性を取得できるようになったのかを評価するために，誤差を許容する場合としない場合とを比較した．この結果を**表 8** に示す．誤差を許容しない場合の方が適合率，再現率，F 値が高いが，得られた属性対の数は誤差を許容する場合よりも少なくなることが示されている．誤差を許容しない方法で得られる属性対は誤差を許容する方法でも得られるため，より条件の厳しい，誤差を許容しない方法の適合率が高く，属性対の取得件数が少ないことは自然である．一方で，取得件数の増加率に対して，誤差を許容する方法の適合率が大きく下がってしまっていることから，誤差の許容度合いを決定するハイパパラメータの適切な設定が必要であるように思われる．

提案手法において，どの程度の重複するエンティティがあれば，高い適合率を達成できるのかを明らかにするために，エンティティの重複率と適合率の関係を分析した．これを**図 4** に示す．図中の横軸は 2 表の間で重複するエンティティの割合を，青色のバーは該当する属性対の数，緑色のバーはそのうちの正解数，赤色の折れ線グラフは各割合における適合率を表す．なお，エンティティの重複率が 40%未満の場合には提案手法により同一属性であると判定していないため，図中には重複率 40%未満の適合率は示されていない．この分析から，70%～90%の重複があるときに適合率が最も高くなること，重複率が 100%であっても適合率は低く，重複度が高くても適合度が高くない場合もあること，また，重複率 60%未満の場合には適合率がわずかに低下していく傾向があることが読み取れる．

表 9　正解・不正解であった属性対の比較
Table 9　Comparison of attribute pairs that were correctly/incorrectly estimated identical.

	数	有効数字桁	平均	標準偏差	尖度
不正解	123	6.44	4.46×10^{9}	1.91×10^{9}	11.45
正解	307	5.72	1.79×10^{9}	7.64×10^{8}	4.48

提案手法によって得られた属性対のうち，どのような属性の同一性判定が可能であったかを明らかにするために，得られた属性対の特徴を定量化し正解・不正解ごとに平均値を算出した．これを**表 9** に示す．2 つの属性について，有効数字の桁数，数量の平均，標準偏差，尖度をそれぞれ求め，それらの平均値を属性対の特徴としている．有効数字の桁数の差異からは，特に正確な数値を表すような属性について誤判定が多いことを示唆している．また，似たような傾向を表すと思われるが，平均値が大きい属性について正確な判定を行えていないことが分かる．標準偏差の差からは，エンティティ間での値のばらつきが大きい属性について困難であること，また，尖度の差からは，属性の値が平均に集中しているような属性の同一性判定が難しいことが分かる．

4.3　量的データ属性間の上位下位関係の抽出

上位下位関係の抽出では，同一属性抽出とは少し異なる実験設定が必要になる．そのため，まず，上位下位関係の抽出に特有の実験設定について述べ，その後に，上位下位関係抽出の実験結果について述べる．

4.3.1　実験設定

量的データ属性間の上位下位関係の抽出を行うための前処理として，同一属性抽出で行った前処理 (1)～(6) のうち，前処理 (3) 以外を，すべて，同じ順番で適用した．前処理 (3) は質的データ属性に関する前処理であり，上位下位関係の抽出においては質的データ属性を扱わないため不要な前処理である．最後に，少なくとも 3 つ以上の量的データ属性を含む表だけを選び出した．これらの前処理を施した結果として，103,706 個の表から 55,185 個の表が得られた．

上位下位関係抽出における比較手法として，属性の言語表現の包含関係に基づく手法（以下では，**言語包含関係手法**とよぶ）を用いる．この方法は，同じ表中の属性対において，一方の属性の言語表現が他方の言語表現を包含するときに，前者を下位属性，後者を上位属性として抽出する方法である．たとえば，「人口」という属性に対する下位の属性として，「年少人口」，「生産年齢人口」，「高齢者人口」などの属性が得られることになる．

実験では，提案手法のハイパパラメータ τ_T，θ_S をそれぞれ 0.85，0.10 に設定した．また，ハイパパラメータ θ_L，θ_H に関しては，下記で述べるバリデーションデータにおいて最も適合率の値が高くなる値を採用した．この結果，

それぞれ，$\theta_L = 4$，および，$\theta_H = 0.5$ と設定された．

バリデーションデータの作成方法は以下のとおりである．まず，提案手法を上記で用意した 55,185 個の表に対して実行し，得られた属性対 8,118 個の中から，上位属性の表記ができるだけ重複しないように 100 対をランダムに選択した．そして，後述するクラウドソーシングによる評価と同じ方法によって正解かどうかの判定を行った．なお，バリデーションデータは提案手法の適合度などを測るためには使用していない．

前述の前処理を施して得られた表に対して，各提案手法，および，比較手法を適用することによって，上位下位関係が存在すると推定される属性の組を抽出した．同一属性抽出に関する実験と同様に，これらの中から 300 個の属性対を選んで評価を行った．ただし，評価対象となる属性対の上位属性の表記ができるだけ重複しないように属性対の選別を行った．そして，選択された属性対それぞれにおいて，2 つの量的データ属性間に本当に上位下位関係が存在するか否かを，クラウドソーシングにて回答してもらった．3 人のワーカを同じ属性対に割り当て，2 名以上が上位下位関係を認めた場合のみ，属性間に上位下位関係が存在するものとした．また，同一属性の評価とまったく同じ条件でワーカを選別した．

4.3.2　実験結果

表 10 に各提案手法と比較手法の適合率，再現率，F 値を示す．比較手法である言語包含関係手法よりも，何も制約条件を加えていない提案手法が高い性能を示し，提案手法に対して 2 つの制約条件を加えたとき（**提案手法＋特殊性&均一性**）に最も高い適合率，再現率，F 値を達成した．提案手法が比較手法よりも性能が良いことから，量的データ属性の値の合計値を比較する方法の有効性を確認することができた．しかしながら，この方法だけでは必ずしも高い適合率，再現率，F 値を達成できたとはいえず，2 つの制約を加えることで初めて実用的な性能を達成できたといえる．また，2 つの制約を組み合わせたときとそれらを別々に用いたときの改善の度合いから，特殊性と均一性の 2 つの制約はおおむね独立に作用していることが読み取れる．

なお，**提案手法＋特殊性**と**提案手法＋均一性**の 2 つを比較した場合，**提案手法＋均一性**の方がより効果的であることが分かる．均一性の制約によって除去することができる，上位下位関係として不適当な例として，**表 11** の事例をあげる．この表において，「1 月生産量」と「1 月と 2 月の差」を合計すると「2 月生産量」の値となるため，「1 月生産量」と「2 月生産量」の組が提案手法では上位下位関係として抽出されてしまう．しかしながら，これは上位下位関係としては適当ではない例であり，均一性の制約によって除去することが可能である．

一方で，2 つの制約条件により除外されてしまった上位下位関係も存在する．特殊性の制約によって取り除かれた

表 10　各手法ごとの適合率，再現率，F 値
Table 10　Precision, recall, and F-measure of each method.

手法	正解数	適合率	再現率	F 値
言語包含関係手法	33	0.111	0.092	0.101
提案手法	92	0.307	0.258	0.280
提案手法＋特殊性	108	0.362	0.303	0.330
提案手法＋均一性	140	0.470	0.392	0.427
提案手法＋特殊性&均一性	**177**	**0.592**	**0.496**	**0.540**

表 11　上位下位関係として不適当な属性対が得られる例
Table 11　Example of an is-a relationship that was incorrectly identified.

国	1 月生産量	2 月生産量	1 月と 2 月の差
アルジェリア	105	105	0
アンゴラ	166	164	−1.8
エクアドル	53	53	−0.4
ガボン	20	19	−0.7

例として，

- 上位「総人口」—下位「男性」,「女性」
- 上位「貿易」—下位「輸入」,「輸出」

など，特に語義のうえでは上位下位関係にない属性が取り除かれている事例が見られた．均一性の制約では，

- 上位「世界人口」—下位「アジア」,「ヨーロッパ」,「北アメリカ」

などのように，下位属性の値に偏りがあるような上位下位関係が取り除かれる事例があった．今回の実験では，適合率を重視してハイパパラメータを設定したが，偽陽性を許容しハイパパラメータをより小さい値に設定する方が良い場合もあると思われる．

5.　結論

本論文では，表中の数値を表す属性のなかから同一属性の対，および，上位下位関係にある属性の対を抽出する方法について提案を行い，定量的な評価を行った．提案手法および比較手法によって得られた同一属性の対を評価した結果，提案手法は適合率および再現率の面から優れた結果を示し，また，従来の属性名の類似性に基づく手法とはまったく異なる属性対が得られることを明らかにした．上位下位関係にある属性対の抽出についても実験を行い，比較手法よりも高い適合率と再現率を提案手法によって達成できることを示した．また，いくつかの制約を加えることにより，提案手法の性能が向上することが確かめられた．

同一属性の抽出については，言語表現の類似度を併用する方法や，同一と判定された属性対を訓練データとして，同一である属性対を発見する方法などが今後との課題として考えられる．また，和以外の数値間の関係，たとえば，「人口密度」は「人口」を「面積」で割った値になっている，などのように，属性間に成立する多様な数量的関係を

発見することも可能ではないかと考えられる．

謝辞　本研究は JSPS 科研費 18H03244，18H03243，および，JST さきがけ JPMJPR1853，筑波大学研究基盤支援プログラムの助成を受けたものです．ここに記して謝意を表します．

参考文献

[1] Bernstein, P.A., Madhavan, J. and Rahm, E.: Generic schema matching, ten years later, *VLDB*, Vol.4, No.11, pp.695–701 (2011).
[2] Chua, C.E.H., Chiang, R.H. and Lim, E.-P.: Instance-based attribute identification in database integration, *The VLDB Journal*, Vol.12, No.3, pp.228–243 (2003).
[3] Do, H.-H. and Rahm, E.: Coma: A system for flexible combination of schema matching approaches, *Proc. 28th International Conference on Very Large Data Bases*, pp.610–621, VLDB Endowment (2002).
[4] Doan, A., Domingos, P.M. and Levy, A.Y.: Learning source description for data integration, *WebDB*, pp.81–86 (2000).
[5] Doan, A., Madhavan, J., Dhamankar, R., Domingos, P. and Halevy, A.: Learning to match ontologies on the semantic web, *VLDB*, Vol.12, No.4, pp.303–319 (2003).
[6] Fabian, M., Gjergji, K., Gerhard, W., et al.: YAGO: A core of semantic knowledge unifying wordnet and wikipedia, *16th International World Wide Web Conference, WWW*, pp.697–706 (2007).
[7] Limaye, G., Sarawagi, S. and Chakrabarti, S.: Annotating and searching web tables using entities, types and relationships, *VLDB*, Vol.3, No.1–2, pp.1338–1347 (2010).
[8] Madhavan, J., Bernstein, P.A. and Rahm, E.: Generic schema matching with cupid, *VLDB*, Vol.1, pp.49–58 (2001).
[9] Mehdi, O., Ibrahim, H. and Affendey, L.: An approach for instance based schema matching with google similarity and regular expression, *International Arab Journal of Information Technology*, Vol.14, pp.755–763 (2017).
[10] Melnik, S., Garcia-Molina, H. and Rahm, E.: Similarity flooding: A versatile graph matching algorithm and its application to schema matching, *Proc. 18th International Conference on Data Engineering*, pp.117–128, IEEE (2002).
[11] Rahm, E. and Bernstein, P.A.: A survey of approaches to automatic schema matching, *VLDB*, Vol.10, No.4, pp.334–350 (2001).
[12] Sahay, T., Mehta, A. and Jadon, S.: Schema matching using machine learning, ArXiv, abs/1911.11543 (2019).
[13] Shvaiko, P. and Euzenat, J.: A survey of schema-based matching approaches, *Journal on Data Semantics IV*, pp.146–171 (2005).
[14] Sumida, A. and Torisawa, K.: Hacking wikipedia for hyponymy relation acquisition, *Proc. 3rd International Joint Conference on Natural Language Processing: Volume-II* (2008).
[15] Sumida, A., Yoshinaga, N. and Torisawa, K.: Boosting precision and recall of hyponymy relation acquisition from hierarchical layouts in wikipedia, *LREC* (2008).
[16] Yakout, M., Ganjam, K., Chakrabarti, K. and Chaudhuri, S.: Infogather: Entity augmentation and attribute discovery by holistic matching with web tables, *SIGMOD*, pp.97–108 (2012).
[17] Zhang, M. and Chakrabarti, K.: Infogather+: Semantic matching and annotation of numeric and time-varying attributes in web tables, *SIGMOD*, pp.145–156 (2013).
[18] Zhang, Z.: Towards efficient and effective semantic table interpretation, *ISWC*, pp.487–502, Springer (2014).
[19] 隅田飛鳥，吉永直樹，鳥澤健太郎：Wikipedia の記事構造からの上位下位関係抽出，自然言語処理，Vol.16, No.3, pp.3_3–3_24 (2009).

藤岡 周平

2018 年京都大学工学部情報学科卒業．2020 年同大学大学院修士課程修了．情報検索の研究に従事．

加藤 誠　（正会員）

2012 年京都大学大学院情報学研究科博士後期課程修了．博士（情報学）．現在，筑波大学図書館情報メディア系准教授．情報検索の研究に従事．電子情報通信学会，日本データベース学会，ACM，ACM SIGIR 東京支部各会員．

吉川 正俊　（正会員）

1985 京都大学大学院工学研究科博士後期課程修了，工学博士．現在，京都大学大学院情報学研究科教授．データベース等の研究・開発に従事．電子情報通信学会フェロー．ACM 各会員．本会フェロー．

（担当編集委員　戸田 浩之）