DEIM2022 E34-2

科学メタデータにおけるデータクリーニングのための不整合検出の効率化

中林　和也†　　清水　敏之††　　大手　信人†

† 京都大学大学院情報学研究科　〒 606—8501 京都府京都市左京区吉田本町
†† 九州大学附属図書館　〒 819–0395 福岡県福岡市西区元岡 744
E-mail: †nakabayashi.kazuya.83r@st.kyoto-u.ac.jp, ††shimizu.toshiyuki.457@m.kyushu-u.ac.jp,
†††nobu@i.kyoto-u.ac.jp

あらまし　科学データを管理するデータベースでは，各データセットに対してデータを説明するメタデータを付与し，メタデータを用いた検索が可能になっていることが一般的になっている．専門的な分野のメタデータは，専門性が高い単語を多く含み，複数人が自由記述形式で入力を行う場合もあるため，誤字や表記揺れが含まれやすく，自動的な修正が困難だと思われる不整合も多く含まれる．そのため，必要なデータの検索の際に弊害が生じることがあり，効率的なデータクリーニング手法が必要とされている．本研究では，既存手法によって検出された不整合候補を基に，不整合の種類ごとにグルーピングを行い，新たな不整合候補と合わせて提示する手法を提案する．提案した手法により，精度を大きく悪化させることなく，グルーピングによる効率的な提示が可能になることを確認した．

キーワード　関係データベース，データクリーニング ，科学データ，メタデータ，不整合検出

1　はじめに

自然科学など専門的な分野の研究を行う際，該当分野のデータベースから必要なデータを検索する機会が多く見られる．この際，データの概要や所有者，組織名といったメタデータを手がかりに検索を行う．しかし，科学データのメタデータ，すなわち科学メタデータには，欠損値や表記の揺れ，粒度の違いなどの様々な不整合が多く含まれ，必要なデータの検索に弊害を生じることがある．そのため，研究者や専門家の効率的な研究を実現するには，科学メタデータの不整合が修正された整備済みのデータベース提供することが重要である．このようなメタデータの不整合を修正することを，一般的にデータクリーニングと呼ぶ．

データクリーニングは，検出と修正の 2 段階に分けられる [1]．最も単純なデータクリーニング手法は，ユーザの目視で不整合を検出して修正を行う方法であるが，これは多大な労力と時間が必要になる．そのため，これらを削減するために，自動で不整合の検出や修正を行うデータクリーニング手法が多く提案されている [2]．例えば，一貫性制約 [3] や機械学習 [4] を用いて，データベース内の一貫性制約の違反やデータの重複，誤字といった不整合を検出し，それらを修正する手法がある．しかし，これらの手法は，科学メタデータのデータクリーニング手法として適さない．なぜなら，科学メタデータは，複数のユーザが自由記述形式でデータ入力を行うため，一貫性制約が乏しく，統計的な推測が難しい不整合を多く含み，学習データや外部辞書の作成コストが高いという特徴を持つからである．また，これらの手法の多くは，数値やカテゴリカルな値を対象としており，メタデータのような文字列データには対応していない．そのため，上述のような特徴を持つデータに対応したデータクリーニング手法が必要である．

そこで，科学メタデータのデータクリーニングに適した不整合の検出手法として，大森らの研究 [5] が提案する IDER(Inconsistency Detection based on Entity Resolution) 手法がある．この手法は，タプルのペアにエンティティ解決を適用することで，網羅的に不整合候補を検出し，ユーザ自身に不整合を修正してもらう手法である．そのため，一貫性制約や学習データの作成コストがかからず，科学メタデータのデータクリーニングに適している．しかし，網羅的な検出を行う IDER 手法 [5] は，適合率が低いため，大量の不整合候補がユーザに提示され，さらにその候補内には不整合ではない値も多く含まれる．そこで，本研究では，これらの問題を解決し，よりユーザが修正しやすいビューを作成する手法を提案する．具体的には，元の不整合候補から類似した不整合をグルーピングすることで一度に提示する不整合候補を少数に絞る．そして，不整合候補を少数に絞る過程で，元の不整合候補集合に含まれる不整合ではない値を取り除き，さらに IDER 手法で検出できなかった真の不整合の検出も行う．本研究における提案手法の導入により，修正に要する労力をさらに削減でき，より効率的な修正の実現が可能になる．

本論文の構成を以下に示す．第 2 章では，本研究の対象データとなる不整合候補を検出する IDER 手法 [5] について説明する．第 3 章では，本論文で提案する不整合の効率的な提示手法について説明する．第 4 章では，実際のデータに対して本研究で提案する不整合の効率的な提示手法を適用する実験とその結果について述べる．第 5 章では，本研究の提案手法における課題についての議論を行う．第 6 章では，まとめを述べる．

	title	owner	organization	year
t_1	title1	Yamada Taro	Kyoto univ.	2020
t_2	title2	Y.Taro	KU	2018
t_3	title3	Yamada Taro	-	2018
t_4	title4	Yamada Taro	Koto univ.	2021

(a) 1 枚の関係データに変換した科学メタデータの例

	着目属性	非着目属性		
	organization	title	owner	year
t_1	Kyoto univ.	$v_{t_1,\text{title}}$	$v_{t_1,\text{owner}}$	$v_{t_1,\text{year}}$
t_2	KU	$v_{t_2,\text{title}}$	$v_{t_2,\text{owner}}$	$v_{t_2,\text{year}}$

(c) (b) のペアの非着目属性を単語の分散表現に変換した例

	着目属性	非着目属性		
	organization	title	owner	year
t_1	Kyoto univ.	title1	Yamada Taro	2020
t_2	KU	title2	Y.Taro	2018

(b) タプル t1 とタプル t2 において，organization を着目属性とした例

	目的変数	説明変数		
	organization	title	owner	year
	$y(=0)$	$x_{t_1,t_2,\text{tite}}$	$x_{t_1,t_2,\text{owner}}$	$x_{t_1,t_2,\text{year}}$

(d) (c) のペアから作成した学習データの例

図 1: タプルの分散表現を用いた学習データの作成

2 既存手法

IDER 手法 [5] は，Muhammad らの研究 [6] を基に，タプルの分散表現とエンティティ解決を利用したデータクリーニング手法である．エンティティ解決とは，文脈やタプルなど周辺情報から，文字列の異なる単語や文章が同一の実体を表しているかを判定することである．エンティティ解決を利用することで，誤字脱字に加えて，同義語や語順が異なる単語のように同じエンティティを指しているが表記が異なる不整合の検出が可能になる．そのため，本研究で対象とする科学メタデータのように，一貫性制約が乏しく，統計的な推測が難しい不整合を含むデータセットに適している．

そして，IDER 手法 [5] は検出した不整合候補をユーザーに提示することでデータの修正を行う．このようなデータクリーニングを実現することで，機械的に判定が難しい科学メタデータの不整合に対応できる．そのため，網羅的な不整合候補の検出を実現する分類器を作成することが重要である．

2.1 不整合検出までの流れ

IDER 手法 [5] では，不整合候補の検出にタプルのペアのエンティティが一致しているかどうかを予測する二値分類器を用いる．

まず，IDER 手法 [5] では，正規化されている関係データから作成した図 1a のような 1 枚のテーブル $\boldsymbol{T}$ を入力値として使用する．次に，テーブル $\boldsymbol{T}$ における n 個の属性集合 $A_1, ..., A_n$ から不整合を修正したい属性 A_f(以下，着目属性) を選択し，テーブル $\boldsymbol{T}$ からタプルのペア集合

$$\boldsymbol{P} = \{(t_i, t_j) | i < j, \forall t_i, t_j \in T\}$$

を作成する．

図 1b は，A_f = “organization” とした場合の，タプルのペア (t_1, t_2) の例である．そして，図 1c のように，テーブル内の単語の共起性を利用したタプルの分散表現 [6] を利用して，非着目属性 A_k ごとに，d 次元の単語ベクトル $v_{i,k} \in [0,1]^d$ を得る．そのため，一般的に使用される学習済み単語ベクトル集合に含まれない専門的な単語のベクトルの取得が可能になる．最後に，図 1d のように，タプルのペア集合 $\boldsymbol{P}$ 内の各タプルのペアにおいて，非着目属性 A_k に関してコサイン類似度を用いた類似度ベクトル

$$x_{i,j,k} = \cos(v_{i,k}, v_{j,k})$$

を説明変数，

$$y = \begin{cases} 1, & \text{if} \quad t_i.A_f = t_j.A_f \\ 0, & \text{if} \quad t_i.A_f \neq t_j.A_f \end{cases}$$

を目的変数としたものを学習データとする．この学習データを 3 層のニューラルネットに適用し，二値分類器を作成する．

この二値分類器は，タプルのペアのエンティティが一致しているかどうかを判定する．そのため，表記は異なるがエンティティが一致しているペア，すなわち偽陽性と判断されたタプルのペア集合 $\boldsymbol{P_{\text{FP}}}$ を不整合候補として検出する．

2.2 IDER 手法の課題と研究の目的

IDER 手法 [5] は，タプルのペアにエンティティ解決を適用し，網羅的に検出した不整合候補を提示することで，ユーザ自身に不整合を修正してもらう手法である．そのため，学習データの作成コストが低く，一貫性制約を必要としない．しかし，網羅的な検出を行う IDER 手法 [5] は，適合率が低いため，大量の不整合候補がユーザに提示され，さらにその候補内には不整合ではない値も多く含まれる．

そこで，本研究では，これらの問題を解決し，よりユーザが修正しやすいビューを作成する手法を提案する．具体的には，類似した不整合をグルーピングすることで一度に提示する不整合候補数を少数に絞る．また，このグルーピングの過程で，元の不整合候補集合に含まれる不整合ではない値を取り除き，さらに IDER 手法 [5] で検出できなかった真の不整合の検出も行う．本研究における提案手法の導入により，修正に要する労力をさらに削減でき，より効率的な修正の実現が可能になる．

3 提案手法

本研究で提案する手法は，類似した不整合値をグルーピングすることで一度に提示する不整合候補数を少数に絞る．また，

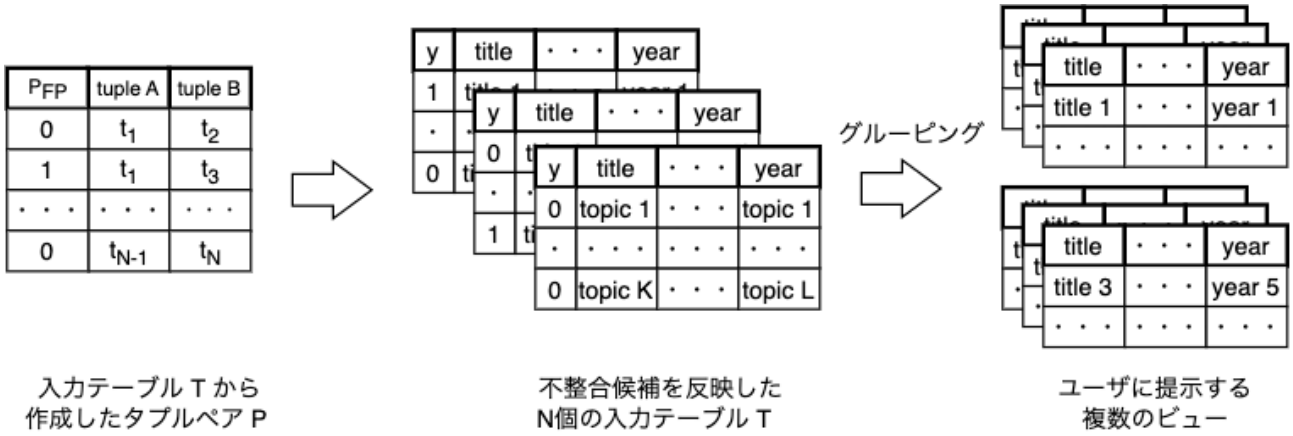


図 2: 入力テーブルのタプルをグルーピングする手法の流れ

この過程で，元の不整合候補集合に含まれる不整合ではない値を取り除き，真の不整合であるにもかかわらず真陰性 $\boldsymbol{P_{TN}}$ に分類された不整合，すなわち IDER 手法 [5] で検出できなかった不整合の検出も行う．

具体的には，図 2 に示すように，IDER 手法 [5] で検出されたタプルのペアの不整合候補集合 $\boldsymbol{P_{FP}}$ を予め入力テーブル $\boldsymbol{T}$ に反映し，入力テーブルのタプルをグルーピングを行う．この手法を “トピックに基づいたタプル分類手法” と呼ぶ．これは，科学メタデータ内における不整合はランダムに出現するのではなく，特定の組織名や所有者，分野など一定の条件を満たす際に出現するという考えに基づいている..

3.1 トピックに基づいたタプル分類手法

具体的な手法の流れについて説明を行う．この手法では，各タプルの属性ごとにトピック分類を行い，決定木を作成することで不整合候補の分類を行う．

まず，入力テーブル $\boldsymbol{T}$ の各値をトピックに変換する．具体的には，IDER 手法 [5] で算出した各属性内の値を d 次元の単語ベクトル $v_{i,k}$ を用いて，非着目属性ごとに x-means 法 [7] を適用する．x-means 法 [7] では，一般的に用いられるクラスタリング手法である k-means 法のようにクラスタ数を事前に与える必要がない．そのため，科学メタデータの各属性のように事前にクラスタ数が推定できないデータに適していると考えた．

次に，IDER 手法 [5] で検出されたタプルのペアの不整合候補集合 $\boldsymbol{P_{FP}}$ を予め大まかに分類し，入力テーブル $\boldsymbol{T}$ に反映する．具体的には，IDER 手法 [5] によって検出された不整合候補のタプルのペア集合 $\boldsymbol{P_{FP}}$ から，ペアの共通タプルをキーとして，N 個のグループ $\boldsymbol{g_n}(n = 1, ..., N)$ を作成する．そして，各グループごとに，トピック変換した入力テーブルに

$$y = \begin{cases} 1, & \text{if} \quad t \in \boldsymbol{g_n} \\ 0, & \text{if} \quad t \notin \boldsymbol{g_n} \end{cases}$$

とする二値ラベルを付与し，不整合候補として検出されたかどうかを反映する．例えば，タプルのペア集合 $\boldsymbol{P_{FP}} = [(t_i, t_j), (t_i, t_k), (t_l, t_m)]$ の場合，t_i をキーとした $\boldsymbol{g_1} = (t_i, t_j, t_k)$ と $\boldsymbol{g_2} = (t_l, t_m)$ の 2 個のグループが生成される．

最後に，上述で作成したデータを元に類似した不整合のグルーピングを行う．各非着目属性のトピックを説明変数，上述で付与した二値ラベル y を目的変数とした N 個の学習データを作成し，それぞれを用いた決定木 [9] を作成する．タプル集合からクエリを推測する Quoc らの研究 [8] を元に，事前に $y = 1$ を付与した不整合候補のタプルが含まれる決定木のノードに至る条件式を抽出し，クエリを生成する．これは，同じトピック構成を持つタプルは，同一エンティティである可能性が高いと考えられるからである．そして，生成したクエリから不整合を修正するためのビューを作成する．

4 実　　験

実際の科学メタデータを用いて，第 3 章で提案した手法によるグルーピング精度と検出精度の比較実験を行う．

4.1 入　　力

本実験では，入力データとして，長期生態学研究のデータベース JaLTER(Japan Long-Term Ecological Research Network)[1] とデータ統合・解析システム DIAS (Data Integration and Analysis System)[2] の 2 種類に科学メタデータを利用する．

両者ともに，自由記述形式の項目が多くあり，管理者以外の複数のメンバーが入力するため，多様なメタデータが存在し，不整合な値も多く存在すると考えられる．また，XML 形式によってデータが提供されているため，1 枚のデータテーブルに変換し，本研究における提案手法に適用する．それぞれのデータセットのデータ数，ペア数，採用属性をそれぞれ表 1a，表 1b に示す．両データともに，目視でデータの確認を行った際に不整合な値が多く存在した連絡組織名 (JaLTER では “ContactOrganizationName”，DIAS では “contactorganization”) を着目属性とした．また，JaLTER に比べて，DIAS は，データ数が多く，採用属性も多いという特徴がある．そして，どちらのデータセットに対しても，前処理として，日本語などの全

1：https://db.cger.nies.go.jp/JaLTER/
2：http://www.diasjp.net

表 1: 入力データセットにおけるタプル数，ペア数，採用属性

(a) JaLTER のデータ概要

入力タプル数	299
採用属性	'title', 'abstract', 'ContactName', 'ContactOrganizationName', 'keyword'
着目属性	ContactOrganizationName
ペア数	42591 件
一致ペア数	1206 件 (2.7%)
不一致ペア数	42385 件 (97.2%)

(b) DIAS のデータ概要

入力タプル数	427
採用属性	'title', 'contactname', 'contactorganization', 'documentauthorname', 'documentauthororganization', 'datasetcreatorname', 'datasetcreatororganization'
着目属性	contactorganization
ペア数	90525 研
一致ペア数	8734 件 (9.6%)
不一致ペア数	81791 件 (90.4%)

角文字や括弧や米印といった記号などを取り除き，半角の小文字のアルファベットのみのデータに変換を行っている．

入力データセットと共に単語のベクトル化に使用する単語ベクトル集合も予め準備する．そのため，Glove [11] によって Common Crawl のウェブアーカイブから学習された単語ベクトル集合 [3] を学習済み単語ベクトル集合として入力として用いた．

4.2 評価方法

本研究で提案する手法と IDER 手法 [5] における不整合候補を比較するため，生成されたビューのグルーピング精度と検出精度の比較を行う．

4.2.1 グルーピング精度

JaLTER や DIAS といった科学メタデータの各属性には同一エンティティであるが表記揺れや誤字記入漏れ，粒度の違いなどによって表記が異なる値が多く含まれる．例えば，JaLTER の属性 “ContactOrganizationName” には 52 種類のエンティティ，DIAS の属性 “contactorganization” には 70 種類のエンティティが存在している．これらのエンティティ内に存在する値の分布を図 3a と図 3b に示す．JaLTER には，他のエンティティに比べ，“Field Science Center for Northern Biosphere Hokkaido University” や “Northern Forestry Research and Development Office Forest Research Station Field Science Center for Northern Biosphere Hokkaido Univeraity” といった entityID=0(北海道大学) の値が極端に多い．一方，DIAS には，“jamstec” や “japan agency for marine-earth science and technology” といった entityID=10(海洋研究開発機構) の値が最も多く含まれるが，entityID=0(東京大学) や entityID=51(気象庁) といった他のエンティティも同じくらい多く含まれている．以上から，JaLTER は出現する値のエンティティに偏りがあり，DIAS は出現する値のエンティティにばらつきがあるという特徴がわかる．このように，データセットごとに含まれるエンティティの分布に違いはあるものの様々なエンティティの不整合候補を含んでいる．

以上から，このような特徴を持つデータセットにおいて，IDER 手法 [5] で検出される不整合候補には，多くのエンティティが含まれる．そこで，1 つのビュー内に存在しているエンティティの数が少ないほど，修正が行いやすいビューであると考えた．そのため，グルーピング精度の評価では，不整合候補数ごとに，提示する各ビューに存在するエンティティ数の平均値を用いて比較を行う．

4.2.2 検出精度

検出精度の評価では，IDER 手法 [5] で検出された真の不整合値を取りこぼすことなく，新たな真の不整合値を検出できているかの確認を行う．

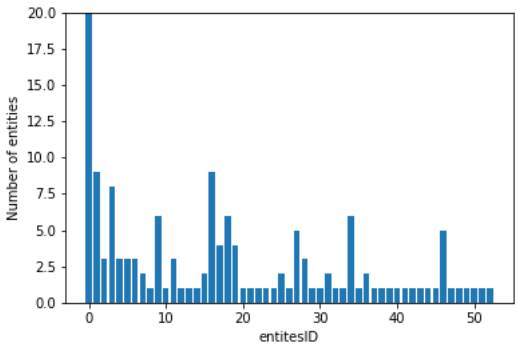
例えば，1 つのエンティティのみを含んでいる 2 つのビューを考える．このうち，1 つ目のビューは 1 種類の値のみを含んでおり，2 つ目のビューは 3 種類の値を含んでおり表記揺れが検出されているとする．これらは，4.2.1 節で述べたグルーピング精度の評価に従うと，どちらも同じくらい修正が行いやすいビューと言える．しかし，含んでいる値の種類数が異なる場合，グルーピング精度は同じであるが，検出精度は前者よりも後者の方が優れていると考えられる．このように，不整合候補数ごとに，ビューの中に何種類の値を含んでいるかを確認することで，検出精度の評価が可能である．また，手法同士の検出精度の比較においては，ビュー内に含まれる値の種類の平均値を用いるのではなく，生成された全てのビューに含まれる値の種類を用いて比較を行う．

本研究の実験で使用する JaLTER の属性 “ContactOrganizationName(連絡組織名)” における値の種類は 154 種類であり，そのうち真の不整合に該当する値の種類は 110 種類であった．同様に，DIAS の属性 “contactorganization(連絡組織名)” における値の種類は 115 種類，そのうち真の不整合に該当する値の種類は 58 種類であった．

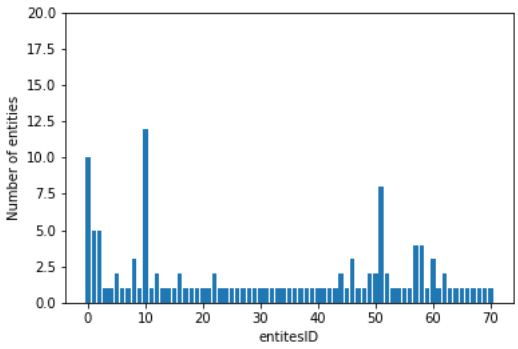
4.3 比較実験

1 節で述べた通り，事前の観察において不整合が多いと思われた属性を着目属性とし，JaLTER では，“ContactOrganization-

3：https://github.com/stanfordnlp/GloVe



(a) JaLTER の着目属性 “ContactOrganizationName”



(b) DIAS の着目属性 “contactorganization”

図 3: 着目属性のエンティティごとの値の分布

Name(連絡組織名)”を，DIAS では，“contactorganization(連絡組織名)”を採用した．表 1 に記載のように，これらの着目属性における一致と不一致の比率は，JaLTER では 1206 件 (2.7%)：42385 件 (97.2%)，DIAS では 8734 件 (9.6%)：81791 件 (90.4%) である．両者ともにも不均衡なデータであるが，JaLTER よりも DIAS のメタデータの方が一致ペアの割合が多い．一般的に少数ラベルの比率が極端に少ないデータセットでは，ニューラルネットにおいてうまく学習できないため，JaLTER データでは，一致データと不一致データの比率が 1:9 になるようにアンダーサンプリング処理 [4]を施す．

これらの学習のデータを用いて，3 層ニューラルネットによって 20 エポックの学習を実行した．0 から 0.9 まで閾値を 0.05 ずつ変化させ不整合候補 $\boldsymbol{P_{FP}}$ を取得した．これらの不整合候補を用いて，IDER 手法 [5] と第 3 章で示した 2 つの手法の比較実験を行う．

4.3.1 グルーピングの精度の評価

本節では，4.2.1 節に従って，提案した手法のグルーピング精度の評価を行う．

本手法で作成された各ビューの中に含まれるエンティティ数の平均値を図 4 に示す．横軸は不整合候補の数を表しており，縦軸が各ビューの中に含むエンティティ数の平均を表している．ただし，IDER 手法 [5] ではグルーピングを行わないため，出力結果であるタプルのペアを 1 枚のビューに変換し提示するものとして扱う．

図 4 より，JaLTER と DIAS 共に，ほとんど全ての不整合候補数において，IDER 手法 [5] よりも，本研究で提案した手法の方が 1 つのビューに含まれるエンティティの種類が少ないことがわかる．一部において例外として，閾値 0 におけるトピックに基づいたタプル分類手法では，全ての不整合候補が 1 つのグループ $\boldsymbol{g_1}$ として分類されてしまったため，決定木 [9] で上手く分類できず，全ての不整合が 1 つのビューに分類されてしまった．

以上から，同じ不整合候補数を修正する際，IDER 手法 [5] と比較して本研究で提案した手法の方がエンティティを上手くグルーピングしており，修正が行いやすいビューを実現していると言える．

4.3.2 検出精度の評価

本節では，4.2.2 節に従って，提案した 2 つの手法の不整合検出の精度評価を行う．

各手法で検出された不整合候補に含まれる真の不整合の種類数を図 5 に示す．横軸は各手法ごとの不整合候補数を表しており，縦軸が真の不整合値の種類数を表している．

図 5a と図 5b の比較により，データセットごとに適している手法が異なることがわかる．JaLTER では，トピックに基づいたタプル分類手法よりも IDER 手法 [5] の方がやや良い精度である．一方，DIAS では，IDER 手法 [5] よりもトピックに基づいたタプル分類手法の精度が良い．

以上から，検出精度においては，データセットごとに適した手法が異なることが確認できた．JaLTER における一部の検出精度の悪化があったものの，本研究で提案した手法は，検出精度をほとんど落とすことなくグルーピングによる効率的な提示を行っていると言える．

5 提案手法の課題と改善

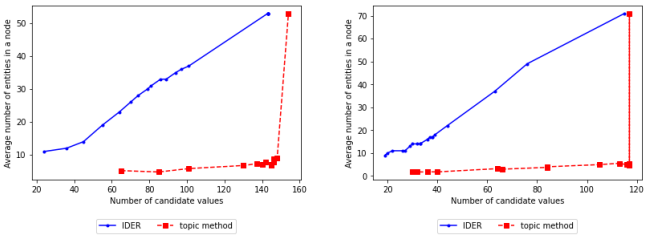
本研究における提案手法は，IDER 手法 [5] の検出精度をおおよそ保ったまま，ユーザが修正しやすいようなビューを実現手法である．しかし，データセットや閾値によっては新たな不整合を検出できず，検出精度が悪化しているものも見られた．例えば，図 5a のように JaLTER では IDER 手法 [5] の方が検出精度が良く，図 5b のように DIAS ではトピックに基づいたタプル分類手法の方が検出精度が良いという結果になった．これは，それぞれの手法において，以下の原因が考えられる．

5.1 課題の原因と改善

本研究における提案手法では，決定木に適用する学習データの作成方法が最も大きな課題である．

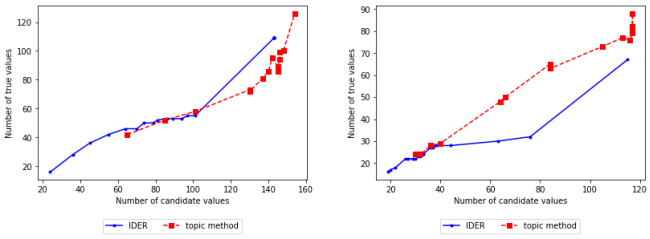
本手法では，決定木 [9] の学習データを作成するために，不整合候補のタプルのペア集合 $\boldsymbol{P_{FP}}$ から，ペアの共通タプルをキーとして作成した N 個のグループ $\boldsymbol{g_n}(n = 1, ..., N)$ を正解データとして目的変数に $y = 1$ の付与を行っている．そのた

4：https://github.com/scikit-learn-contrib/imbalanced-learn



(a) JaLTER (b) DIAS

図 4: 候補数ごとのノード内平均エンティティ数



(a) JaLTER (b) DIAS

図 5: 候補数ごとの真の不整合の数

め，誤ったペアの組み合わせ次第では，異なる 2 つのエンティティを 1 つの $\boldsymbol{g_n}$ に分類される恐れがある．例えば，$[(t_i, t_j) =$ (“Kyoto univ.”, “Hokkaido univ.”), $(t_i, t_k) =$ (“Kyoto univ.”, “KU”), $(t_j, t_l) =$ (“Hokkaido univ.”, ‘HU”)] といったタプルのペア集合があった場合，誤ったタプルのペア $(t_i, t_j) =$ (“Kyoto univ.”, “Hokkaido univ.”) が原因で京都大学のグループと北海道大学のグループが共通のグループとしてみなされてしまう．エンティティに偏りがある JaLTER では，少数のエンティティのグループが多数出現するエンティティのグループと同一であるとみなされてしまい，他の手法よりも精度が悪化したと考えられる．一方，エンティティにばらつきがある DIAS では，誤って同一のグループとみなした学習データが少なく，他の 2 手法よりも精度が良かったと考えられる．以上から，決定木に適用する学習データの作成における課題は，データセットによって精度が不安定になる最も大きな原因であると言える．この問題の解決案として，弱教師あり学習 [12] を用いてグループ $\boldsymbol{g_n}$ 内の異なったエンティティを取り除くといった方法がある．また，アクティブラーニング [13] を用いて，ユーザがインタラクティブにグループ $\boldsymbol{g_n}$ 内の異なったエンティティを取り除くといった方法も考えられる．

5.2 手法の改善実験

5.1 節で述べたように，学習データ内の異なったエンティティを取り除くことでトピックに基づいたタプル分類手法の精度が向上するかを確認する．

具体的には，同一エンティティのみを含む学習データを人手で作成し，トピックに基づいたタプル分類手法に適用する．そして，この理想的な学習データを用いた結果と，第 4 章で用いたトピックに基づいたタプル分類手法の結果，すなわち複数のエンティティを含む元の学習データを用いた結果の比較を行う．また，ベースラインの基準として，第 4 章と同様 IDER 手法 [5] で検出された結果とも比較を行う．

実験条件は，第 4 章と同様に，JaLTER では “ContactOrganizationName”を，DIAS では “contactorganization”を着目属性とし，0 から 0.9 まで閾値を 0.05 ずつ変化させた IDER 手法 [5] によって検出された不整合候補集合を入力として用いる．

5.2.1 グルーピング精度の評価

4.3.1 節で行った実験と同様に，4.2.1 節に従って，グルーピング精度の評価を行う．

各手法で作成された各ビューの中に含まれるエンティティ数の平均を図 6 に示す．横軸が各手法ごとの不整合候補の数を表しており，縦軸が各ビューの中に含むエンティティ数の平均を表している．図 6 より，JaLTER と DIAS のどちらのデータセットにおいても，元の学習データを用いた結果とほとんど同等である少数のビューの作成が可能である．また，元の学習データを用いた結果では，閾値 0 の際，全ての不整合候補が一つのグループとして分類されてしまったため，決定木 [9] が上手く機能せず，全ての不整合が 1 つのビューに分類されてしまった．しかし，理想的な学習データを用いた結果では，閾値が 0 においても適切なグルーピングが行われているため，いずれの閾値においても少数のエンティティを含むビューを作成することが可能であった．

以上から，トピックに基づいたタプル分類手法に同一エンティティのみを含む理想的な学習データを適用する方が，あらゆる閾値に対応でき，複数エンティティを含む元の学習データを適用した時と同等の結果が得られることがわかった．

5.2.2 検出精度の評価

次に，4.3.2 節で行った実験と同様に，4.2.2 節に従って，検出精度の評価を行う．

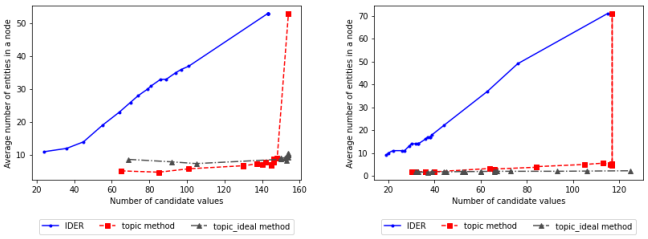
各手法で検出された不整合候補の中に真の不整合がどれくらい含まれているかを図 7 に示す．横軸は各手法ごとの不整合候補数を表しており，縦軸が真の不整合の種類数を表している．まず，4.3.2 節の実験で示したように，JaLTER を入力データとした際，元の学習データを用いた場合，IDER 手法 [5] に比べて検出精度が悪化している．一方で，理想的な学習データを用いた場合，図 7a より，ほぼ全ての不整合候補数において他の 2 つに比べて大幅に検出精度が改善している．次に，図 7b より，DIAS を入力データとした際，理想的な学習データを用いた結果は，IDER 手法 [5] よりは精度が向上したものの，元の学習データを用いた結果に比べて検出精度が悪化するという結果になった．

以上の結果を踏まえて，同一エンティティのみを含むグループを用いたとしても，必ずしも検出精度が上がるわけではないことがわかった．この原因に関して，次節で考察を行う．

5.2.3 検出精度悪化の原因調査

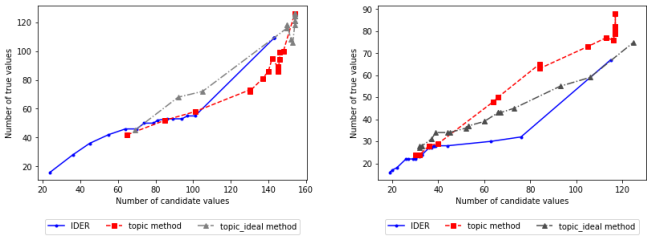
本節では，特定の不整合候補数におけるビューや真の不整合を用いて，5.2.2 節の実験で，DIAS を入力とした際に理想的な学習データを用いた精度が悪化してしまった原因を調査する．

理想的な学習データを用いた結果よりも元の学習データを用



(a) JaLTER (b) DIAS

図 6: 元の学習データと理想的な学習データを使用した際の不整合候補数ごとのノード内の平均エンティティ数



(a) JaLTER (b) DIAS

図 7: 元の学習データと理想的な学習データを使用した際の不整合候補に対する真の不整合候補値の種類数

いた結果の方が大幅に良くなった例として，66 件の不整合候補数が検出された際を考える．この 66 件のうち真の不整合値の種類数は，元の学習データを用いた結果では 50 件，理想的な学習データを用いた結果では 43 件であった．そのため，理想的な学習データを用いた場合と比べて，元の学習データを用いた場合の方が真の不整合が 7 件多く，検出精度が良いという結果であった．

次に，66 件の不整合候補数が検出された際のビューごとのエンティティ数の分布を確認する．この図を，図 8 に示す．横軸はビュー内に含むエンティティ数を示しており，縦軸は該当するビューの数を示している．この図から，どちらの学習データを使用した場合もエンティティを 2 種類か 3 種類を含むビューが多く生成される．しかし，元の学習データでは，24 種類のエンティティを含むビューが 1 つ生成されている．元の学習データでは，正確に分類されなかった不整合候補がこのような雑多なエンティティを含むビューに分類されていることがわかる．

以上を用いて，片方のみで検出された真の不整合値とそれを含むビューを調査する．元の学習データのみで検出された真の不整合値は 14 件，理想的な学習データのみで検出された真の不整合値は 7 件であった．元の学習データのみで検出された 14 件中 12 件は，上述した雑多なエンティティを含むビューに含まれる．このような雑多なエンティティを含むビューは，ユーザが修正しやすいビューとは言い難い．以上から，元の学習データを用いた結果では，検出精度が向上して新たな真の不整合が検出されたわけではなく，雑多なエンティティを含むグループによって検出された修正に適さないビューに偶然含まれていた値であったと考えられる．一方，理想的な学習データのみで検出された真の不整合値は，全て少数のエンティティを含むビューに含まれている．そのため，これらの新たな不整合は，検出精度の向上によるものであると考えられる．

このように元の学習データで生成される修正に適さないビューは，IDER 手法 [5] に適用する閾値を緩めるほど，増加する傾向があると思われる．これは，入力に使用する不整合候補が増加するため，誤ったグループ間の結合が生じるからである．そのため，ユーザに不整合を提示する際には，何らかの基準を定め，このような修正に適さないビューは省く必要がある．

5. 2. 4 不適当なビューの削除における検出精度の変化

5. 2. 3 節において，元の学習データを使用した例では，修正

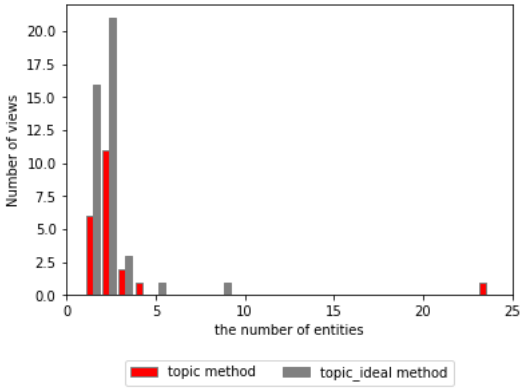
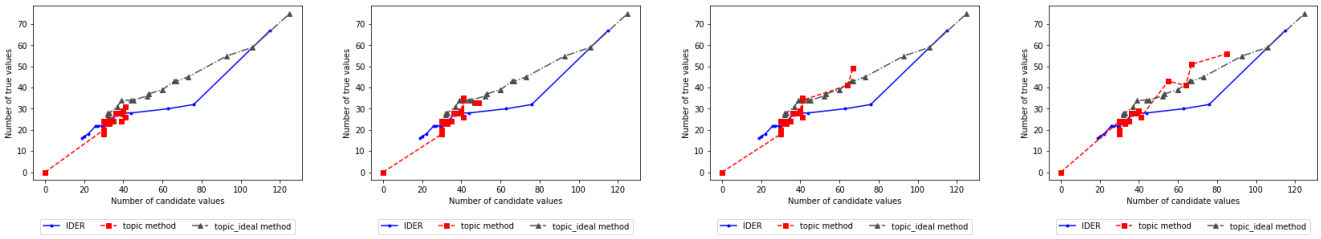


図 8: 66 件の不整合候補数で生成されるビューごとのエンティティ数の分布

に適さないビューが生成されることが確認できた．そこで，本節では，5. 2. 3 節と同じ条件下において，5 種類，10 種類，15 種類，20 種類以上のエンティティ数を含むビューを省いた検出精度の比較を行う．この結果を，図 9 に示す．図 7 と同様，横軸は各手法ごとの不整合候補数を表しており，縦軸が真の不整合の種類数を表している．

まず，少数のみのエンティティを含むビューのみを集計した図 9a と図 9b では，全ての不整合候補数において理想的な学習データを使用した結果が精度が良い．しかし，10 種類から 20 種類のエンティティを含むビューも集計に入れた場合の図 9c と図 9d では，一部の不整合候補数において，元の学習データを使用した結果が理想的な学習データを使用した結果を上回っている．これにより，5. 2. 3 節でも述べたとおり，雑多なエンティティを含むグループによって検出された修正に適さないビューによって，検出精度が良くなったように見えてしまうことが確認できた．

以上から，生成されるビューに含まれるエンティティ数や検出される新たな不整合候補を考慮すると，DIAS を入力値とした場合であっても，理想的な学習データを用いることが望ましい言える．そのため，5. 1 で述べたように，弱教師あり学習 [12] やアクティブラーニング [13] を用いて，グループ内の異なったエンティティを取り除く方法は精度向上に適している．また，修正に適さないビューに含まれる不整合候補を省く基準やビュー内のエンティティ数を考慮した精度指標の作成が必要である．



(a) $K = 5$ (b) $K = 10$ (c) $K = 15$ (d) $K = 20$

図 9: K の種類以上のエンティティ数を含むビューを省いた検出精度の比較

6 ま と め

本研究の対象である科学メタデータには，一貫性制約が乏しく，学習データの作成コストが高いため，既存の自動的なデータクリーニング手法の適用が難しい．そこで，大森らの研究 [5] によって，網羅的に不整合候補を検出し，ユーザ自身に修正を行ってもらう IDER 手法 [5] が提案されている．しかし，適合率が低いため，大量の不整合候補がユーザに提示され，さらにその候補内には不整合ではない値も多く含まれる．本研究では，これらの課題を解決し，よりユーザが修正しやすいビューの作成を目的とし，“トピックに基づいたタプル分類手法” を提案した．具体的には，類似した不整合をグルーピングすることで一度に提示する不整合候補を少数に絞り，元の不整合候補集合に含まれる不整合ではない値の除去や検出できなかった真の不整合値の検出を行う．

実際の科学メタデータを用いて，本研究における提案手法で生成されるビューの評価を不整合値のグルーピング精度と真の不整合の検出精度の 2 つの観点から行った．本研究で提案した両手法共に，IDER 手法 [5] と比較して，1 つのビューあたりのエンティティ数を大幅に絞ることができ，修正に効率的なビューが作成できることを確認した．しかし，真の不整合の検出精度に関しては，データセットごとにばらつきがあり，新たに検出できた真の不整合もあったものの，IDER 手法 [5] で検出した真の不整合の取りこぼしも見られた．以上から，本研究で提案した手法は，精度をおおよそ保ったまま，グルーピングによる効率的な提示が可能になることを確認した．

次に，上記の実験において，真の不整合の検出精度の安定性を改善するための実験を行った．この原因として，学習データの作成の際に，不純物となるデータが混入することが考えられる．そこで，人手で不純物を取り除いた理想的な学習データを作成し，元の学習データを適用した結果と比較実験を行い，理想的な学習データを使用した際の精度の向上が確認できた．そのため，本研究で提案した手法の検出精度を向上させるためには，弱教師あり学習やアクティブラーニングといった手法によって学習データ内から異なったエンティティを予め取り除くことが望ましいと考えられる．

謝辞 本研究の一部は JSPS 科研費 JP18K11315 の助成を受けたものです．

文 献

[1] X.Chu and I.F.Ilyas and S.Krishnan and J.Wang.: “Data Cleaning: Overview and Emerging Challenges, ” In Proceedings of the 2016 International Conference on Management of Data, SIGMOD Conference 2016, San Francisco, CA, USA, June 26 - July 01, pp. 2201-2206, 2016.

[2] X.Chu and I.F.Ilyas.: “Trends in Cleaning Relational Data: Consistency and deduplication, ” Foundations and Trends in Databases,vol.5(4), pp. 281-393, 2015.

[3] P.Bohannon and W.Fan and F.Geerts and X.Jia, and A.Kementsietsidis.: “Conditional functional dependencies for data cleaning,” 2007 IEEE 23rd international conference on data engineering, IEEE, pp. 746–755 2007.

[4] A.Heidari and J.McGrath and I.F.Ilyas and T.Rekatsinas.: “HoloDetect: Few-Shot Learning for Error Detection, ” Proceedings of the 2019 International Conference on Management of Data, SIGMOD ’19, New York, NY, USA, Association for Computing Machinery, pp. 829-846, 2019.

[5] 大森 弘樹, 清水 敏之, 吉川 正俊.: “エンティティ解決手法を応用したデータクリーニングのための不整合検出,” DEIM2020, 2020

[6] M.Ebraheem and S.Thirumuruganathan and S.Joty and M.Ouzzani and N.Tang.: “Distributed representations of tuples for entity resolution,” Proceedings of the VLDB Endowment, Vol. 11, No. 11, pp. 1454–1467, 2018.

[7] D.Pelleg and A.Moore.: “X-means: Extending Kmeans with Efficient Estimation of the Number of Clusters. ” Proceedings of the 7th International Conference on Machine Learning (ICML), pp. 727-734 2000.

[8] Q.T.Tran and C.Y.Chan and S.Parthasarathy.: “Query by Output.” Proceedings of the 2009 ACM SIGMOD International Conference on Management of Data, SIGMOD ’09, pp. 535-548, 2009.

[9] J.R.Quinlan: “Induction of Decision Trees. Machine Learning,” Mach Learn, 1, pp. 81-106 1986.

[10] A.Arasu and M.Götz and R.Kaushik.: “On active learning of record matching packages.” In Proceedings of the 2010 ACM SIGMOD International Conference on Management of data (SIGMOD ’10). Association for Computing Machinery, New York, NY, USA, pp. 783–794 2010.

[11] J.Pennington and R.Socher and C.Manning.: “Glove: Global Vectors for Word Representation,” Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, Association for Computational Linguistics, pp. 1532-1543 2014.

[12] C.Northcutt and L.Jiang, and I.Chuang.: “Confident learning: Estimating uncertainty in dataset labels.” Journal of Artificial Intelligence Research, 70:pp. 1373–1411, 2021.

[13] B.Settles.: “Active learning literature survey.” Computer Science Technical Report 1648, University of Wisconsin-Madison, January 2009.