1C-3

ドメイン知識を用いた異名義 RDF エンティティの同一性推定手法の検討

森田　大翼†　　飯塚　京士†　　村山　隆彦†　　赤埴　淳一†

日本電信電話株式会社　NTT 情報流通プラットフォーム研究所†

1　はじめに

近年，企業内の多様な情報を統合して分析し，企業戦略に役立てるニーズが高まっている．しかし，年度の移り変わりを期にデータの名称等が変更され，文字列比較による名寄せの手法だけではデータ統合がうまくできない場合が多い．本稿では，ドメイン知識を活用することにより，文字列比較による名寄せの手法だけでは解決不可能なデータ統合の問題の解決を図る．具体的には，各年度等の多様なデータを RDF に変換して統合し，データセット間のエンティティの同一性を，周辺情報とドメイン知識を利用して推定する手法を提案する．

2　データの同一性推定問題

2.1　背景

年度や四半期などの単位で，組織では多くの情報が生成，蓄積される．しかし，前後の期間の情報との整合性や連結性が欠如する場合が多い．

例えば，2007 年の「研究プロジェクト A」から，2008 年の「開発プロジェクト B」への名称変更の情報が管理されていない場合を想定する．「研究プロジェクト A」の文献として「文献 C」があり，「開発プロジェクト B」の成果物として「製品 D」がある場合，「文献 C」は「製品 D」の基礎技術を示した文献であると考えられるが，そのような関連付け情報の検索ができないという問題が起こる．

2.2　先行研究

文献[1]では，複数のデータ中の“Helen Hunt”と“H. M. Hunt”という記述に対し，実世界での同一性を推定する名寄せの問題に対して，特定ドメインにおけるデータ統合のための制約を利用するアプローチを提示している．例えば「著者 X と Y が，似た名前を持ち，且つ共通の共著者を 2 人以上持てば，X と Y は同一人物の可能性がある」というものである．しかし，この手法には，以下のような問題点がある．

- 制約には「2 人以上」という個数，「可能性がある」という確率的な表現がなされている．この値は対象データに依存し，ドメインの有識者であっても適切な設定は困難である．
- 同一性推定に用いられる周辺情報が，DB の同一レコードの値や文書の周辺の記述などに限られている．すなわち，周辺情報としての範囲が小さく，効果が限定的である

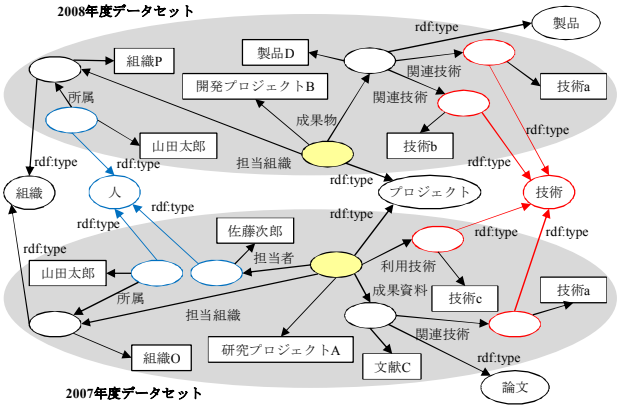

図 1: RDF のグラフ表現（一部プロパティの記述，及びリソースの URI は省略）

2.3　課題とアプローチ

ドメイン知識を用いたエンティティ（データの実体）の同一性推定問題には以下の課題がある．

- 同一性推定規則の対象データへの依存性の問題
- エンティティの周辺情報の取得の問題

この課題に対して以下のアプローチをとる．

A)　対象データの特性を考慮した，同一性推定規則を定義する

B)　エンティティの周辺情報として，複数エンティティ間の関連情報を利用する

3　RDF エンティティの同一性推定

3.1　対象データの前提条件

本稿の対象データの前提は以下のとおりである．

- 対象データ，及びデータ間の繋がりは RDF (Resource Description Framework)[1] に変換して表現される．RDF はリソースと呼ばれるエンティティに対してメタデータを記述できる．
- データセットは一定期間毎に発行される．
- 個々のエンティティには，RDF タイプが定義され，これをエンティティのタイプと呼ぶ．

2.1 節の例の RDF グラフ表現を図 1 に示す．

Estimating Sameness of Differently Named RDF Entities with Domain Knowledge

† Daisuke Morita
† Kyoji Iiduka
† Takahiko Murayama
† Jun-ichi Akahani
† NTT Information Sharing Platform Laboratories, NTT Corporation

[1] http://www.w3.org/RDF/

Algorithm: エンティティ間同一性推定

```
1:  D^t /* 期間 t のデータセット (τ_s ≦ t ≦ τ_t, t ∈ ℕ) */
2:  r_i^t  /* 期間 t のデータセットに含まれるリソース．
              下付き文字 i, j は連番を示す． */
3:  c_k /* タイプ．下付き文字 k は連番を示す */
4:  m_k /* タイプ c_k のエンティティの同一性推定関数 */
5:  for t in τ_s .. τ_t − 1 do
6:    for each r_i^t in D^t do
7:      for each r_j^{t+1} in D^{t+1} do
8:        if type(r_i^t) = type(r_j^{t+1}) then
9:          c_k ← type(r_i^t);
10:         m_k ← get_rule(c_k);
11:         if m_k(r_i^t, r_j^{t+1}) = true then
12:           create_metadata(r_i^t, r_j^{t+1});
13:         end if;
14:       end if;
15:     end loop;
16:   end loop;
17: end loop;
```

図 2: 同一性推定アルゴリズムの定式化

3.2 同一性推定ルール

ドメインの有識者によって作成された，あるタイプのエンティティ間の同一性を推定するルールを用いる．例えば以下のようなルールが考えられる．

「プロジェクト X と Y について，それぞれに関連する共通人物が P 人以上，関連する共通技術が Q 個以上あれば，X と Y は同一プロジェクトである」(*)

同一性推定ルールは以下の特徴を持つ．

a) ルールには変数を持たせることができる
b) 「関連する」という記述により，同一性推定に用いる周辺情報の範囲を広げている

a)の特徴は，2.3 節の A)の課題を解決する．b)の特徴は，例えば「関連する人」を，「エンティティから経由エッジ数が 2 以内で到達できる，人のタイプを持つエンティティ」として周辺情報の範囲を広げることで，B)の課題を解決する．図 1 の場合，「研究プロジェクト A」の“担当者”に加え，“担当組織”に“所属”する人物もまた関連人物とみなす．

3.3 アルゴリズム

図 2 に定式化されたアルゴリズムを示す．図 2 における *type* はエンティティのタイプを抽出する関数，*get_rule* は引数タイプのエンティティ間の同一性推定を行う関数を取得する関数，*create_metadata* は指定したエンティティ間に同一であることを示す RDF プロパティを付与する関数をそれぞれ示す．

4 実験

4.1 実験設定

ある研究機関における 2007 年度から 2009 年度の所員，組織，投稿論文，製品，プロジェクト，技術用語のデータを用いる．技術用語の情報は，投稿論文，製品，プロジェクトとの関連付けがあるが，網羅性に欠けている．

実験では，年度間のプロジェクトの同一性推定を(*)のルールを用いて行う．また，文献[1]の手法との比較も行う．プロジェクトのデータから得られる担当者の情報は最大 2 人のため，文献[1]の手法では制約を以下のように設定する．

「プロジェクト X と Y が共通の担当者を 1 人以上，且つ 1 つ以上共通技術を持てば，X と Y は同一プロジェクトの可能性がある (確率は 0.8 である)」(**)

実験の正解データとして，2007 年度から 2009 年度のプロジェクト間の 62 対の推移情報を利用するが，このデータもまた網羅性が不十分である．従って，本実験では再現率を重視し，評価指標に F_β-measure [2]を用い，β=2 と設定する．これは，再現率を適合率の 2 倍重視することを意味する．

表 1: 本提案手法の実験結果 (PT = Positive True, PF = Positive False, NF = Negative False)

P	Q	PT	PF	NF	適合率	再現率	F_2-measure
8	1	34	294	28	0.1156	0.5484	0.3137
6	1	38	377	24	0.1008	0.6129	0.3040
9	1	31	270	31	0.1148	0.5000	0.2992
7	1	35	344	27	0.1017	0.5645	0.2956
5	2	29	245	33	0.1184	0.4677	0.2941

表 2: 文献[1]の手法の結果

PT	PF	NF	適合率	再現率	F_2-measure
22	172	40	0.1279	0.3548	0.2619

4.2 評価

本稿の提案手法による実験結果のうち，F_2-measure の値の上位 5 つの結果を表 1 に示す．適合率は低いが，これは正解データの特性が主な要因である．今回のデータの場合，(*)のルールにおいて P=8，Q=1 と設定した時，良い結果が得られることが分かった．

また，文献[1]の手法による実験結果を表 2 に示す．表 1 と比較すると F_2-measure の値は下回っており，本稿の提案手法の有用性を示している．

5 まとめと今後の課題

本稿では，年度毎に発行されるデータセット間のエンティティの同一性推定問題に対し，周辺情報と，有識者によるドメイン知識を利用した同一性推定ルールを用いる手法を提案した．エンティティの同一性推定問題の課題を克服し，先行研究[1]より優れた評価が得られたことを実験的に示した．

本稿は同一性推定ルールの学習の可能性を示している．エンティティの同一性推定問題に対する正解データは，網羅性が低い傾向がある．今後は，少ない正解データから同一性推定ルールの変数，またはルール自体の学習手法を検討する予定である．

参考文献

[1] Shen, W., Li, X. and Doan, A., “Constraint-Based Entity Matching,” In Proceedings of the 20th national conference on Artificial intelligence pp. 862—867, 2005.
[2] van Rijsbergen, C. J., “Information Retrieval,” Butterworth, 1979.