D-002

異種データセット統合におけるコンテクスト情報を考慮したグラフベース埋め込み表現生成
Context-aware Graph Embedding for Integrating Heterogeneous Datasets

春木佑香 † 石倉茂 ‡ 出町和也 ‡ 早矢仕晃章 †
Yuka Haruki Shigeru Ishikura Kazuya Demachi Teruaki Hayashi

1 はじめに

デジタル化の急速な進展により，データ流通量は爆発的に増加し，学術界だけではなく，ビジネス分野においてもデータの利活用による価値創出に期待が高まっている．しかし，分析に利用されるデータセットは異なる組織やシステムから作成され，多様なデータ形式や表記が使用されていることが多い．複数のデータセットを連携・統合して分析に利用するためには，これらの異種のデータ間の不一致を解消する必要がある [1]．とりわけスキーママッチング (schema matching: SM) やエンティティ解決 (entity resolution: ER) はデータ統合において不可欠なタスクであり，従来は人手での作業が行われていた．しかし，人手での作業には時間的・金銭的コストがかかるため，自動化に関する研究が多く行われている [2]．とりわけ，データ構造をグラフ化して埋め込み学習を行うアプローチは，表データの構造情報を反映し高い性能でマッチングが行われることから，近年注目を集めている [3]．一方で，既存研究で性能検証に用いられるデータセットは，品質が高くドメイン固有性が低いなど，実務で使用されるデータセットとは性質に乖離があり，実務への応用には課題がある [4]．また，データセットの特性がマッチング性能に与える影響も十分に検討されていない．

本研究では既存手法である EmbDI [3] を応用して，データセットおよびこれに付随するコンテクスト情報を反映したグラフを構築し埋め込みを生成することを提案し性能を検証する．さらに，提案手法を多様な特性を持つ複数のデータセットに対して適用し，データの特性がマッチング性能に与える影響についても議論する．

2 関連研究

本章ではデータ統合に対するアプローチとして主要なナレッジベースおよびグラフベースの手法を紹介する．

ナレッジベースの手法は，オントロジーや辞書といった外部知識を活用してデータ統合タスクを行うアプローチである [4]．具体的には，WordNet などの語彙知識ベースを用いて同義語を識別したり，事前に定義されたマッピング辞書を活用して対応関係を学習したりする．また，事前学習済み埋め込み（Word2Vec や GloVe）を利用し，カラム名の意味的類似性を測定する手法もある．しかし，事前学習済みのモデルは，学習に用いた分野のデータセットに特化した埋め込みになっているので，多様なドメインのデータセットに対して汎用的ではなく，また，データセットの構造を考慮できていないという課題がある．

グラフベースの手法では，データ構造をグラフに変換したうえで埋め込みとして学習し，マッチングを行う．代表的な手法として，EmbDI [3] は三部グラフを構築し，ノード埋め込みを学習することでデータ統合タスクを実施する．REMA [4] は，ARC Graph を用いて属性情報をノードとして追加し，ランダムウォークを行うことでマッチングを実施する．これらの手法は，表データ特有の構造を反映できるという点で優れているが，その一方で計算コストの増加やデータ不足時の性能低下といった課題も挙げられる．

3 提案手法

以下に示す手順でカラムの埋め込み表現を学習した後に，カラムどうしのコサイン類似度を計算することでスキーママッチングおよびエンティティ解決を行う．

3.1 グラフの構築

EmbDI では，図 1 のような三部グラフを構築する．このグラフは，トークンノード（Token Node），レコード識別ノード（Record Identifier, RID），カラム識別ノード（Column Identifier, CID）の三種類のノードで構成される．提案手法では，複数（実験では 2 つ）のデータセットを入力とし，EmbDI と同様の三部グラフを構築する．さらに，データ構造のみを反映していた EmbDI にデータセットのコンテクスト情報を付加するため，カラムの類似度に応じて CID 間に重みつきのエッジを追加して四部グラフを構築する（図 1）．カラムの類似度はカラム名およびそれに付随するカラムの説明文をテキストとして入力し，BERT を 2 つのテキスト（カラム名とその説明文）が同じ概念かどうかを判定する二値分類モデルとして扱うことによって算出する．また，このモデルは 2 つのカラムが同一の概念を表しているか判定するタスクに関する追加学習をしてファインチューニングを行った．

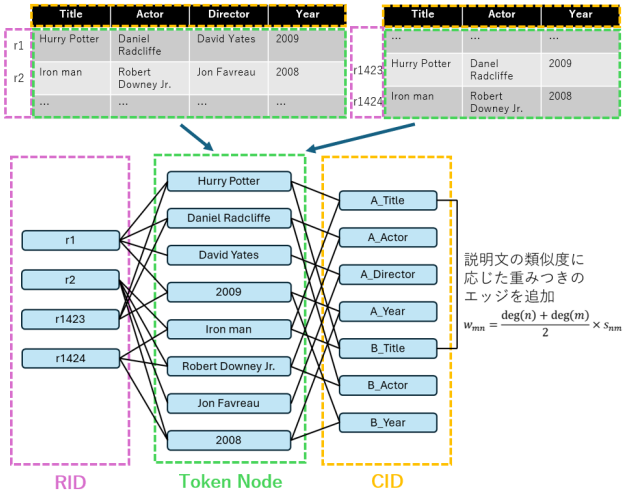

図 1 統合するデータセットから構築されるグラフ

そして，閾値 0.55 を超える CID のペアについて，類似度に応じた重みをもつエッジを CID 間に張る．CID であるノード n とノード m 間に張られるエッジの重み w_{nm} は，CID 間に張られるエッジ以外のエッジの重みを 1 としたとき，ノード n，m の次数の相加平均にカラム n と m の説明文の類似度 s_{nm} を掛け合わせたものである．

† 東京大学大学院工学系研究科 School of Engineering, The University of Tokyo
‡ 株式会社インフォマート Infomart Corporation

表 1 実験の結果（**太字**は提案手法によって性能が改善した項目）

データの特性	Dataset	データセットの詳細	SM		ER	
			既存	提案	提案	提案
ドメイン固有性	Dataset A	先行研究と同様のデータセット	0.69	**0.85**	0.70	**0.80**
	Dataset B	Kaggle のデータセットを元に作成	0.67	**0.77**	0.63	**0.71**
	Dataset C	実務の企業内取引データをもとに作成	0.57	**0.65**	0.55	**0.70**
データセットサイズ	Dataset D	3000 rows & 3000 rows	0.70	**0.76**	0.68	0.68
	Dataset E	3000 rows & 100 rows	0.20	**0.26**	0.35	**0.43**
	Dataset F	100 rows & 100 rows	0.00	0.00	0.00	0.00
完全性 (欠損率)	Dataset G	欠損率 3% & 3%	0.70	**0.83**	0.64	**0.75**
	Dataset H	欠損率 3% & 20%	0.70	**0.83**	0.65	**0.69**
	Dataset I	欠損率 20% & 20%	0.70	**0.75**	0.60	0.60

3.2 ランダムウォークと埋め込み表現の学習

3.1 節にて構築した四部グラフ上でランダムウォークを行い，得られたシーケンスを用いて埋め込み表現を学習する．ランダムウォークは，四部グラフ内の各ノードを起点とし，確率的に隣接ノードへと遷移することで，ノードシーケンスを生成する．

次に，得られたシーケンスを用いて埋め込み表現を学習する．具体的には，生成されたシーケンスを文書コーパスとして扱い，word2vec の Skip-Gram モデルを適用することで，各ノードに対応する埋め込みベクトルを獲得する．なお，Skip-gram の埋め込み空間の次元数は 300，コンテクストウィンドウサイズは 3 で設定した．

4 実験

4.1 実験に用いたデータセット

データセットの特性が統合タスクの性能に与える影響を評価するため，表 2 に示した 3 つの評価項目を基に，異なる特性を持つ 9 つのデータセット A〜I を用意した．

表 2 データセットの特性

特性	定義
ドメイン固有性	データセットのカラム名や内容の独自性の高さによって定義される
データサイズ	データの行数によって定義される．なお，列数は 10 で統一している
完全性（欠損率）	空データとなっているトークンの割合で定義される

4.2 実験手順と評価手法

4.1 で言及した 9 つのデータセットに対して，既存手法の EmbDI と提案手法を用いて，それぞれスキーママッチングとエンティティ解決のタスクを行った．マッチング性能としては F1 Score を用いた．

5 結果と考察

データ統合タスクの結果は表 1 に示される通りである．まず，アルゴリズムの違いによる性能への影響について論じる．既存手法の EmbDI と提案手法を比較したところ，コンテクスト情報を取り入れた埋め込み学習を用いることで，マッチング性能は向上した．Dataset A〜C で比較すると，既存研究で用いられたものと同様の Dataset A では既存手法でも高性能を達成していたが，ドメイン固有性の高い Dataset B, C では性能向上が顕著であった．これは，EmbDI がカラム内のトークン内容に依存していたのに対し，提案手法はカラム名および外部知識を活用した BERT によるコンテクストを考慮したことが要因と考えられる．

次に，データセットの特性が及ぼすマッチング性能への影響について論じる．ドメイン固有性が高い，データサイズが小さい，または欠損率が高いという特性を持つデータセットにおいては，マッチング性能が低下する傾向が確認され，データセットの特性がマッチング性能に対して影響を与える可能性が示唆された．また，提案手法を適用することにより，性能が低下しやすい特性を持つデータセットにおいてもおおむね性能は向上し，提案手法が既存手法と比較して優れた性能を発揮する可能性が示唆された．

6 総括と今後の展望

本研究の新規性は，グラフベースとナレッジベースの 2 つのアプローチを融合した点，また，「ドメイン固有性」「データサイズ」「完全性（欠損率）」によって定義されるデータセットの特性がマッチング性能に与える影響を論じた点である．これによって，実験ではコンテクスト情報を考慮した埋め込み表現を用いることでマッチング性能が向上すること，またデータセットの特性がマッチング性能に影響を及ぼす可能性があることが示唆された．

今後の研究課題としては現在提案されているグラフベースの埋め込みアプローチについて，更なるアルゴリズムの改善を行う必要がある．例えば，トークンノードに対してもカラムノードと同様に言語的アプローチを適用することで，より包括的にコンテクスト情報を取り入れることが考えられる．

謝辞

本研究は株式会社インフォマートとの共同研究の成果です．心から御礼申し上げます．

参考文献

[1] Behzad Golshan, Alon Halevy, George Mihaila, and Wang-Chiew Tan. Data Integration: After the Teenage Years. ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems, 101 – 106 (2017).
[2] Doan, A., Halevy, A. Y., and Ives, Z., Principles of Data Integration, Morgan Kaufmann (2012).
[3] Cappuzzo, R., Papotti, P., and Thirumuruganathan, S., Creating Embeddings of Heterogeneous Relational Datasets for Data Integration Tasks, SIGMOD (2020).
[4] Koutras, C., Fragkoulis, M., Katsifodimos, A., and Lofi, C, REMA: Graph Embeddings-based Relational Schema Matching, EDBT/ICDT (2020).