距離学習を用いたHuman-in-the-loop エンティティマッチングフレームワークの提案

大沢直史[1)] 伊藤寛祥[2)] 福島幸宏[3)] 原田隆史[4)] 森嶋厚行[2)]

1)知識情報・図書館学類 2)図書館情報メディア系 3)東京大学 情報学環・学際情報学府 4)同志社大学大学院 総合政策科学研究科

1.背景

エンティティマッチングとは，データベース中において同一の実体を参照しているレコードの集合を識別する問題である．エンティティマッチングは，データのクリーニングや複数のデータベースの統合において非常に重要であるが，エンティティ作成時の入力の揺れや欠損がしばしば生じ，完全にルールベースな手法ではすべてのエンティティに対して完全なマッチングを行うことはできないという問題がある．加えて，データベースがもつレコード数が大きい場合，すべてのレコードの組み合わせを参照し，それらが同じエンティティを参照しているか否かを判定する同定作業を行うためには膨大な計算時間がかかり，さらに同定にはその対象に合わせた正規化処理や同定キーの作成が必要になるといった問題がある．

2.目的

本研究ではHuman-in-the-loopに基づくアプローチに距離学習モデルと二値分類モデルを用いた手法を提案する．距離学習とは，データ間の関係を考慮し，同一のクラスに属するデータ同士は空間中で近い位置に埋め込み，異なるクラスに属するデータ同士は遠い位置に埋め込むようなマッピング関数を学習する手法である．適切な距離学習が行えれば，同一レコードの可能性がある候補のペア選択およびルールベースでは判定が難しいペアの選択について，アドホックでないアプローチを提供することが期待できる．

3.提案手法

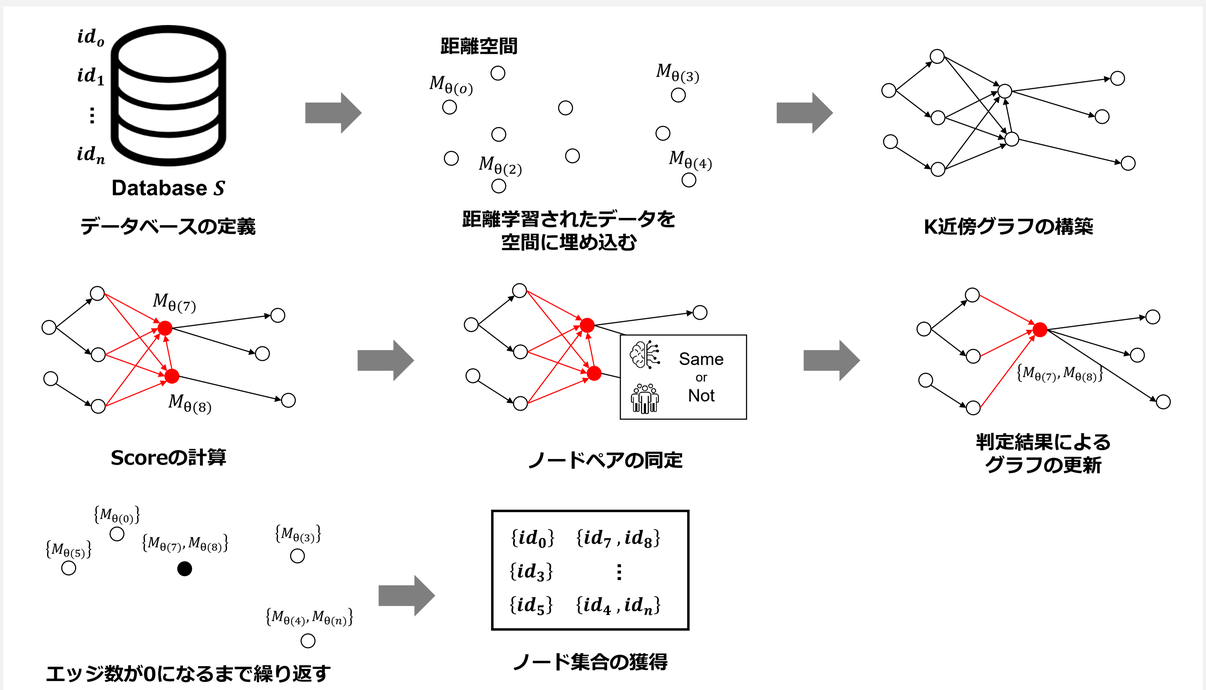

図1． 提案手法の手続き

1. データベースSが持つ書誌データを距離空間上に埋め込む．
   距離空間への埋め込みは4.実験で述べる距離学習モデルを用いて行われる．
2. 埋め込まれたデータをそれぞれノードとし，距離が最も近い上位K件のノードを接続するK近傍グラフを構築する．
3. 各ノードに対し*Score*を計算．

$$Score(v_i, v_j) = |N(v_i) \cap N(v_j)|$$

$N(v_i)$はノードv_iに対するエッジ集合である．基本的なアイデアは，K近傍グラフにおいて２つのノードにおいて共有する隣接ノード数が多いほど，その２つのデータが統合されたときに縮約されるエッジ数が大きくなり，結果として比較が必要になるデータのペア数が少なくなるというもの．

4. *Score*の高いノード同士を候補ペアとして二値分類モデルとクラウドソーシングによる同定を行う．ノード間がマッチした場合はノードが統合されることでグラフの縮約が行われ，一致しなかった場合はそのノード間のエッジを切断する．

以上の操作を全てのノードが独立するまで繰り返す．

機械学習モデルとクラウドソーシングにおけるタスクの割り当ては，**二値分類モデルの確率的な出力α**を閾値として行う．二値分類モデルの出力がαを下回った場合にその書誌データのペアをクラウドソーシングのタスクとする．

提案手法の優位性

- 全ての組み合わせを比較する必要がなくなるため，比較回数の削減が可能
- 人間が同定プロセスに介入することによって品質向上に期待

4.実験とその結果

本研究では国立国会図書館の総合目録によって作成された，表1に書誌データベースS内の書誌データマッチング問題を考える．実験は(1)二値分類モデルの構築(2)距離学習によるデータの表現(3)評価実験について行った．

表1.書誌データペアの例

タイトル	巻号	著者	出版社	ISBN	page	cm
若おかみは小学生!		亜沙美,令丈ヒロ子	講談社	4061486136	215p	18cm
若おかみは小学生! ： 花の湯温泉ストーリー１	[PART1]	亜沙美,令丈ヒロ子	講談社	4061486136	215p	18cm

(1)二値分類モデルの構築

書誌データの一致/不一致を判定する二値分類モデルの構築を行う．二値分類モデルは，表1に示すような書誌データがもつ各カラムの文字列から類似度を複数算出し，その類似度を学習する．

学習の結果，別な書誌の組み合わせを判定する書誌割れに関して大きい貢献するモデルを構築することが出来たが，反対に書誌誤同定に対しては改善の余地が見られた．

(2)距離学習によるデータの表現

本実験では書誌データ間で発生する関係性を距離空間上に表現することを目的とする．これは図1におけるM_θにあたる．距離空間上において，類似するデータは近く，非類似するデータは遠くなるように学習を行う．

(3)評価実験

提案手法の有効性を確認するための評価実験について述べる．実験は**(1)提案手法の比較回数の変化(2)提案手法によって獲得されるノード集合の再現率(3)フレームワークの同定能力とタスク数**についてそれぞれ確認を行う．同定能力の評価の指標は再現率と適合率の調和平均であるF1値を用いる．比較対象は書誌データが持つノードを書誌idの順に従って同定を行うナイーブな手法である．本実験では，クラウドソーシングは正しい判定をすることを仮定して行う．

(3-1)提案手法のエッジ数，比較回数の変化

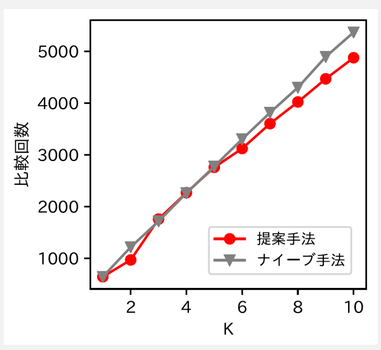

図1.エッジ数の変化

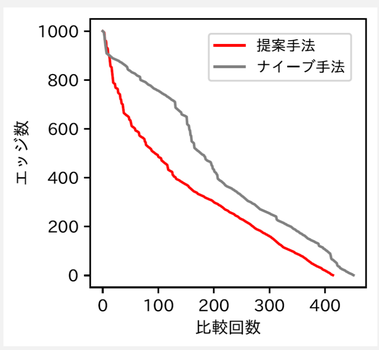

図2.比較回数の変化

図1ではScoreの高いノードからグラフを縮約することで，ランダムに選択したナイーブ手法と比較して比較回数を削減したことを示す．図2では近傍探索数Kを大きくした場合，累計の比較回数に有意差が生じた．

(3-2)ノード集合の再現率

本研究ではK近傍グラフのエッジに対して同定の判定を行う．より高精度に同定を行うためには，マッチングすべきレコードのペアがK近傍グラフのエッジにもれなく含んでいることが必要になる．図3はK近傍グラフ構築時におけるマッチングペアの再現率である．

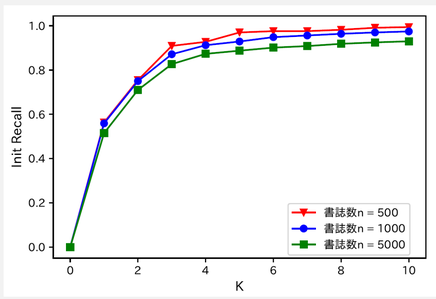

図3. K近傍グラフ構築時の同定ペアの再現率

(3-3)フレームワークの性能とタスク数

クラウドソーシングによる同定シミュレーションの結果，機械学習モデルの閾値αを大きくした場合に提案手法のF1値が向上した．これよりαを大きくした場合にタスク数が増加し，フレームワークの性能を向上させることができる．一方で一般にタスク数が増えるほど金銭的なコストが大きくなる．

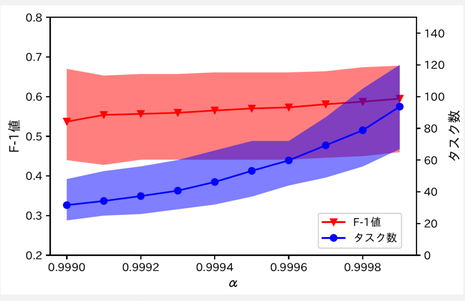

図4. αを変化させた場合のF1値とワーカに割り当てたタスクの数

5.まとめ

共通のスキーマを持つ複数のデータベースを対象としたエンティティマッチング手法として，距離学習を用いたHuman-in-the-loopエンティティマッチングフレームワークの提案を行った．提案手法の有効性を検討するため，国立国会図書館の総合目録を利用したいくつかの実験を行った．

実験の結果，二値分類モデルの学習においては70¥%の精度が得られ，距離学習を用いたブロッキングにおいては，K=10のK近傍グラフを構築することで，約90%の再現率で同定候補をあげることができることを示した．また，本フレームワークにおいて二値分類モデルの確信度の閾値を上げて，積極的にクラウドワーカーをタスクに取り入れることにより，二値分類モデルのみを用いた場合よりも精度が向上することを示した．さらに，共有ノード数に基づくスコアリングを行うことで，ランダムにエッジを選択して統合する場合と比較して比較回数が減少することを示した．