品質を考慮したHuman-in-the-loop書誌同定手法の提案

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)情報学学位プログラム 2)図書館情報メディア系 3)慶應義塾大学 文学部4)同志社大学大学院 総合政策科学研究科

1.背景

ビッグデータの分野では，重複データの検出やデータクリーニング技術が注目されている．エンティティマッチングとは，あるデータベースにおいて同一のエンティティを参照するレコードを識別することを指し，これまでに多くの研究が行われている．エンティティマッチングは，データクリーニングや複数のデータベースの統合に不可欠な技術であるが，レコード作成時の入力の揺れや欠損にどう対応すべきかや，ルールベースのアプローチでは完全なマッチングを実現できないことが現状の課題である．

これらの解決策として，高速にマッチングが行える機械的な処理と，人間による精密な処理を組み合わせたHuman-in-the-loopによるアプローチが提案されている．

2.目的

本研究では不特定多数の人間に作業を依頼するクラウドソーシング技術とデータの統計量に基づき，そのレコードが同一であるか否かを判定するベイズ分類モデルを組み合わせた新しいHuman-in-the-loopエンティティマッチングフレームワークを提案する．提案手法は初めにベイズ推論によってエンティティマッチングを行い，次に判定が難しいようなレコードのマッチングを人間に依頼する．これによってレコードの誤同定を減らし，少ない人間の雇用回数によって高い品質のエンティティマッチングを実現する．

本年度は，昨年度の手法の課題点であった書誌判定モデルの精度向上を目標として活動に取り組む．

3.提案手法

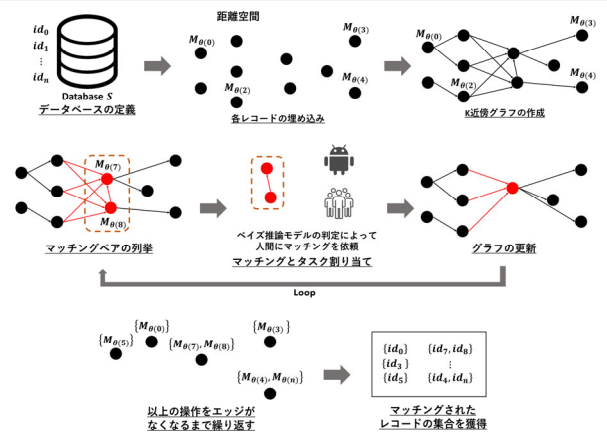

図1．提案手法の手続き

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である．

4. *Score*の高いノード同士を候補ペアとしてベイズ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．

3,4の操作を全てのノードが独立するまで繰り返す．

ベイズ分類モデルが行うクラウドソーシングへのタスクの割り当ては，ベイズモデルが回答した判定結果の他方のクラスに属する確率値を閾値τを上回った場合に行う．

例)ある書誌ペアに対し，ベイズ分類器が$P = 0.8$の確率で一致と判定したとする．$\tau = 0.1$とすると，$1 - P \geqq \tau$となるため，その書誌ペアはクラウドソーシングによって人間に再度判定を依頼する．

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に示すような書誌データベースS内の書誌データマッチング問題を考える．実験は(1)ベイズ分類モデルの構築(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,令丈ひ子	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,令丈ひ子	講談社	4061486136	215p	18cm

(1)ベイズ分類モデルの構築

書誌ペアの一致/不一致を判定するベイズ分類モデルの構築を行う．ベイズ分類モデルは大量の書誌ペアから複数の種類による文字列類似度を算出し，それらの確率分布を考えることによって構成される．類似度はfasttextによる単語分散表現，Jaro-Winker距離，Levenshtein距離，SequenceMatcherオブジェクトによる距離の4つを使用した．

(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)フレームワークの同定能力と割り当てられたタスクの割合(2)他手法とのタスク割り当て能力の比較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F-1値を用いる．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シミュレーション委によって行う．

(3-1)フレームワークの同定能力と割り当てられたタスクの割合

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を図2に示す．赤線と左縦軸が提案手法の同定能力を示し，青い線と右縦軸が全マッチングタスクの内，クラウドワーカに割り当てられたタスクの割合を示す．また，横軸はタスク割り当ての閾値τである．
閾値τを大きくし，人間へのマッチングを依頼した場合に提案手法のF-1値が向上した．また$\tau = 0.5$，すなわちどのペアのマッチングも人間に依頼しないベイズ分類器本来の性能はF-1値0.8となり，昨年度のモデル(F-1値0.5)に比べ，高い精度でのマッチングが可能となった．

(3-2)タスク割り当て能力の比較

図3に，タスク割り当て能力の比較を示す．赤線が提案手法，灰線がランダムにタスクを人間に割り当てた場合のフレームワーク全体の同定能力を示す．

提案手法はベイズモデルが回答した確率に基づく，不確実性の高いタスクを人間に割り当て，マッチングを行ってもらうことて判定の正解率を上げ，フレームワーク全体の能力を向上させている．また全体のマッチングタスクの内約30%のタスクを割り当てた場合にF-1値は1.0に近づく結果となった．同じ性能の場合，ランダムに割り当てる手法では約95%のタスクを人間が回答する結果となっている．

結果より，提案フレームワークは判定精度を保ちつつ，人間に割り当てるタスクを大きく減らすことができるため，金銭的コストやマッチングに要する時間を削減することが可能であると言える．

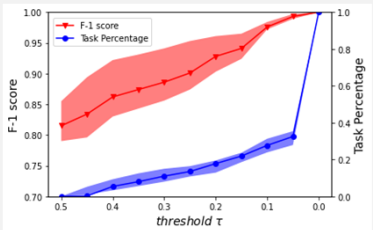

図2. τを変化させた場合のF-1値とワーカに割り当てたタスクの割合

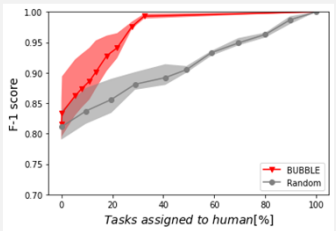

図3. タスク割り当ての割合と提案フレームワークの同定能力

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，ベイズ分類モデルとクラウドソーシングを組み合わせた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，ベイズ分類モデルにおいてはF-1値0.8の性能が得られ，また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タスク割り当てにおける閾値τを上げ，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，ベイズ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．

昨年度と比較し提案手法は性能が向上し，またさらに今後の発展が期待できる．