DEIM Forum 2021 B21-3

行動パターンを基にした異なるドメインに対するユーザ同定技術

草野　元紀†　　小山田昌史†

† 日本電気株式会社　〒 211–8666 神奈川県川崎市中原区下沼部 1753
E-mail: †{g-kusano,oyamada}@nec.com

あらまし　一般消費者は様々なサービスを利用するため、彼/彼女らの顧客データはそれらのサービス毎に別々に保管されている。このようなサイロ化した顧客データ群から同一ユーザを発見する取り組みは *User Identity Linkage* として知られており、広告配信業務や顧客分析などの様々な領域において応用されている。本論文では、異なるドメインの顧客データにおいてユーザの行動履歴を基にユーザ同定を行う方法を紹介する。本提案手法は自然言語処理と教師あり学習の手法に基づいて構成されており、行動履歴のみからユーザ同定を行うため、本名やメールアドレスなどの個人情報や、年齢や性別などのデモグラフィック属性が含まれていないデータに対しても適用可能である。

キーワード　User Identity Linkage, 行動履歴, 異種データ統合, ユーザ属性

1　序　　論

1.1　背　　景

ユーザの行動パターンを包括的に理解することは、顧客分析などのマーケティング業務において重要な課題となっている。顧客は毎日多くのオンラインおよびオフラインサービスを利用しているため、単一ドメインの顧客データを一つのみ調べるだけではユーザを深く理解することはできない。そこで、ユーザの横断的な行動を追跡するために複数のユーザデータを紐づけることに注目が集まっており、同一ユーザの発見に取り組む研究は *User Identity Linkage* (UIL) と呼ばれる。図 1 は UIL のユースケースを示している。

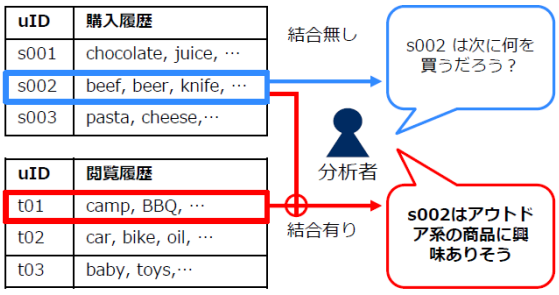

図 1　ターゲットユーザの購入履歴にウェブの閲覧履歴を連携できると、例えば、訪問したサイトに関連するカテゴリのアイテムを推薦したり、各ウェブ サイトの広告効果を測定したり、購入パターンを分析して訪問者にとってより魅力的なウェブサイトを改善したりするのに便利である。これらの応用では、ユーザの連携 $s002 = t01$ が不可欠である。

アプリケーションレベルで実装されている UIL としては、HTTP Cookie による Cookie Sync やアプリ連携許可のソーシャルログインシステムなどがある。ところが、EU 一般データ保護規則 (GDPR) を始めとした昨今の個人情報保護の観点による 3rd party Cookie の規制や、サービス乱立により全てのサービスに登録しているユーザは実質的に存在していないなどの課題がある [1]。

これらの課題を克服するために、分断された 2 つのユーザデータにおいて、片方のデータのユーザがもう片方のデータのどのユーザと連携するかを予測する機械学習を用いた UIL [Zafarani and Liu, 2009, Liu et al., 2014, Shu et al., 2016, Hadgu and Gundam, 2020] が研究されている。多くの研究では、ドメイン s のユーザ u がドメイン t のユーザ u' と同一であるかどうかを判断するために、それぞれのユーザの特徴量を計算し、それらの類似度を計算することで、高い類似度となるユーザペアを同一ユーザとして出力する。特徴量の計算方法はユーザに紐づくデータの種類によって異なり、例えば、デモグラフィック属性 (性別、年齢など) や行動データ（移動履歴、商品の購入履歴、SNS におけるコメント文章・タグ付け）などが特徴量計算の入力として与えられる。

1.2　課　　題

本論文が対象にするユーザデータとしては商品の購買履歴やウェブの閲覧履歴データなどの行動履歴データであり、その際に、従来法では今回考える問題設定を次の観点から対処できない。

1.2.1　デモグラフィック属性の不足

様々なデータソースから顧客関連のデータを収集する一般的なマーケティングプラットフォームである Data Management Platform (DMP) では、ユーザデータは匿名化してハッシュ値として保管されることが多い [Elmeleegy et al., 2013]。特に、企業間のデータを横断的に扱う場合においてはユーザの匿名化は必須であり、全てのデータに対して年齢や性別などのデモ

1：個人情報保護の話題に関して、GDPR は Cookie がユーザの同意なしに取得されることを問題視しており、総務省の情報銀行を包含する個人情報管理システム (PIMS) が最近提案されたように、匿名化などのプライバシー保護の下でユーザデータを活用することは可能であり、その際に Cookie を使用しない UIL は役立つ。

グラフィック属性が含まれているとは限らない。従って、ユーザのデモグラフィック属性を利用する UIL [Zafarani and Liu, 2009, Zafarani and Liu, 2013, Mu et al., 2016] は、匿名化されたユーザデータに適用することはできない。

1.2.2 行動パターンのドメイン間の類似性

デモグラフィック属性を用いずに UIL を行う関連研究 [Iofciu et al., 2011, Goga et al., 2013, Kong et al., 2013, Riederer et al., 2016, Feng et al., 2019] では、ドメイン間で共通して現れるオブジェクトに焦点を当てることで、ユーザの行動パターンの類似性を計算することができ、それによりユーザ連携を実施している [2]。ただし、実際のケースでは、異なるドメインに常に共通するオブジェクトがあるとは限らない。さらに、両方のドメインに共通するオブジェクトがある場合でも、ユーザの行動パターンは完全に異なっていることがある [3]。このようなユーザの行動パターンがドメイン毎に異なる場合、従来法では同一ユーザを発見することは困難である。

1.3 貢　　献

上記の課題に対処するために、本論文ではドメイン毎に異なる行動パターンの違いに対処する UIL の新しい枠組みを提案する。

提案法は、ベクトル機構、マッピング機構、選択機構、およびマッチング機構の 4 つの要素から構成される。ベクトル機構では、行動履歴を数値ベクトルに変換することで、ユーザ固有の特徴量抽出を行う。ここで、行動履歴は行動の系列データであり、行動を単語と見なすことで、自然言語処理の手法を用いることで効率的に数値ベクトルへ変換できる。マッピング機構では、ベクトル機構で得られた同一ユーザの各ドメインでの特徴量の対応関係を写像として捉え、ドメインを跨いだユーザの特徴量の変換を実施する。上記のベクトル機構とマッピング機構では、様々な選択肢の組み合わせが生じることになり、“良い” 組み合わせはデータセットに応じて変化する。選択機構では、そのベクトル機構とマッピング機構で用いられる手法の良い組み合わせを交差検証の流儀に倣って選択する。そのようにして選ばれた手法の組み合わせを用いて、最終的には、マッチング機構でまだ対応の取れていないユーザに連携するユーザを発見する。

本稿の貢献は以下のように要約される：

- デモグラフィック属性を用いずとも、同種とは限らない異なる行動パターンに対する UIL を達成する手法を提案している。
- ベクトル機構では自然言語処理を活用することで行動履歴からユーザに固有の特徴量抽出を、マッピング機構ではドメイン間のユーザの行動パターンの差異を変換することができる。
- ベクトル機構とマッチング機構において様々な手法の選択肢を許容し、各手法の組み合わせはデータセットに応じて最適なものを選択機構で選択できる。
- 数値実験では、3 つの種類の異なるオープンソースデータ Instacart, Click-Through Rates, Amazon データセット (食品の購入履歴、広告の閲覧履歴、アイテムのレビュー履歴) で検証を実施し、異なるドメイン間から同一ユーザの連携をそれぞれ 92.2%, 54.0%, 86.8% の精度で達成していることを確認した。

2 提　案　法

2.1 記号の準備

この論文では、U と I をユーザとアイテムの集合とする。ユーザに対する**行動履歴**は、アイテムの一連の系列として定義する。例えば、ユーザ $u \in U$ がバナナ 1 個、リンゴ 2 個、サーモン 1 匹を購入した場合、アイテムの系列 $b_u = (i_{banana}$、i_{apple}、i_{apple}、$i_{salmon})$ をそのユーザの行動履歴として扱う。

UIL の設定では、ユーザデータは 2 つ以上のドメインに存在しているため、この論文では (U^s, I^s, B^s) と (U^t, I^t, B^t) をドメイン s と t のユーザ、アイテム、および行動履歴の組とする。そして、ドメイン s と t の両方に登録されているユーザの集合は $U^{link} := U^s \cap U^t$ とし、何人かのユーザはドメイン間での対応が既に知られているものとする。この時の UIL の課題は、ドメイン t のどのユーザと紐づくかまだ分かっていないドメイン s のユーザの対応関係を発見することである。正確には、ドメイン s の行動履歴 $b_u^s \in B^s$ を入力すると、そのユーザのドメイン t での動作履歴を B^t の中から探し出すことに対応する。

この論文では、既に対応が判明しているユーザ集合を教師データ $U^{train} \subset U^{link}$ とし、ユーザの対応関係が判明していないユーザ集合をテストデータ $U^{test} \subset U^{link} \setminus U^{train}$ とする [4]。

2.2 方　　針

本研究の最終的な目標は、ドメイン s のユーザを与えたときに、対応する同一ユーザをドメイン t から見つけて、連携することである。ここでは、そのための中間目標を、二つの異なるドメインの行動履歴の類似度を計算する次の類似度関数 (1) を構築することと定める。

$$\mathrm{sim} : B^s \times B^t \to \mathbf{R},\ (b_u^s, b_{u'}^t) \mapsto \mathrm{sim}(b_u^s, b_{u'}^t) \qquad (1)$$

同一ユーザ $u = u'$ の場合に $\mathrm{sim}(b_u^s, b_u^t)$ が高い値を返すように類似度関数を構築できれば、探したいユーザのドメイン s での行動履歴を左辺に入れて、類似度が高くなるようなドメイン t の行動履歴を検索することで、対応するユーザを連携することができる。

ここではその類似度関数 (1) を得る方法の概要を紹介する。構成は大きく分けて、ベクトル機構とマッピング機構からなる。まず、ベクトル機構では写像 $v^s : B^s \to \mathbf{R}^{d_s}$ と $v^t : B^t \to \mathbf{R}^{d_t}$

2：例えば、Facebook(ドメイン s) ユーザが旅行中に特定の場所に関するテキストを投稿し、その場所にタグ付けされた写真を Instagram(ドメイン t) に投稿したとすると、この場合、共通して現れるオブジェクトは地名や緯度・経度などの位置情報ということになる。

3：ユーザの行動がクリック (ドメイン s) と閲覧 (ドメイン t) からなる、広告の効果測定タスクでは、クリックするという行動と閲覧するという行動は異なっている。一方、Facebook と Instagram の例では、ある場所を訪れたというそのユーザの背後の行動はドメイン間で共通している。

4：記号 $A \setminus B$ は集合 A, B の差集合を表す。

を構成する。次に、マッピング機構では同一ユーザのドメイン毎に得られるベクトル化した行動履歴 $v^s(b^s_u), v^t(b^t_u)$ の対応関係をつけるように写像 $f : \mathbf{R}^{d_s} \to \mathbf{R}^{d_t}$ を構成する。最後に、ベクトル機構での v とマッピング機構での f に基づいて、類似度関数を次のように定義する。

$$\mathrm{sim}_{v,f}(b^s_u, b^t_{u'}) := \cos(f(v^s(b^s_u)), v^t(b^t_{u'}))$$

ここで $\cos(\mathbf{v}, \mathbf{v}') := \langle \mathbf{v}, \mathbf{v}' \rangle / \|\mathbf{v}\| \|\mathbf{v}'\|$ $(\mathbf{v}, \mathbf{v}' \in \mathbf{R}^{d_t})$ はコサイン類似度を表す。ベクトル変換モデル v と写像 f の詳細については、以降の章で説明する。

2.3 ベクトル機構

ここでは、行動履歴をベクトルに変換する方法について紹介する。以下では、一般性を失うことなく、ドメインの表記を省略し、教師ユーザの行動履歴集合を $B^{train} := \{b_u \mid u \in U^{train}\}$ とする。この章では、ベクトル変換モデルを計 $3+1+4\times3+1 = 17$ 通り紹介する。これらの中から、後の選択機構で適切なベクトル変換モデルが選ばれる。

2.3.1 アイテム次元方式

行動履歴 $b_u = (i_1, \ldots, i_m) \in B$ に対して、行動履歴を構成するアイテム i を単語と見なすことで、行動履歴は文章と見なすことができる。このとき、文章と見なした教師ユーザの行動履歴集合 $B^{train} := \{b_u \mid u \in U^{train}\}$ で学習された Bag-of-Words, TFIDF [Jones, 2004], BM25 [Jones et al., 2000] で行動履歴をベクトルにすることができる。対応するベクトル変換モデルを $v_{bow}, v_{tfidf}, v_{bm25} : B \to \mathbf{R}^d$ と書くことにする。ここで、d は B^{train} に含まれる単語総数である。以後、各手法をまとめて v_* ($* \in \{$bow, tfidf, bm25$\}$) と書くことにし、$\mathbf{X}_*$ は v_* から定まるユーザアイテム行列とする。例えば、$* =$ bm25 の場合は、$\mathbf{X}_{bm25}[u, i] = v_{bm25}(b_u)[i]$ となる。

2.3.2 連結方式

上記の構成では、3 つのベクトル変換モデルを連結することも出来る。その変換モデルを $v_{concat}(b_u) := [v_{bow}(b_u), v_{tfidf}(b_u), v_{bm25}(b_u)]$ として定義する。

2.3.3 NMF 方式

ユーザアイテム行列は一般的にスパースであるため、それを圧縮する一つの方法として非負値行列分解 (NMF, [Lee and Seung, 2000]) を考える。NMF では、$\mathbf{X}_*[u, i] = \langle \mathbf{P}_*[u], \mathbf{Q}_*[i] \rangle$ なるユーザベクトル $\mathbf{P}_*[u] \in \mathbf{R}^k_{\geqq 0}$ とアイテムベクトル $\mathbf{Q}_*[i] \in \mathbf{R}^k_{\geqq 0}$ を求める。ここで、$\mathbf{P}_*[u]$ をユーザ u の特徴ベクトルとして扱い、対応するベクトル変換モデルを $(\mathrm{NMF}_k \circ v_*)(b_u) := \mathbf{P}_*[u]$ として定める。NMF のパラメータ k に関しては、$k = 10, 50, 100$ を試す。

2.3.4 Doc2Vec 方式

昨今の深層学習の発展により、文章をそのままベクトルに変換する手法は様々提案されている。ここでは、B^{train} から学習された Doc2Vec (Paragraph2Vec, [Le and Mikolov, 2014]) によるベクトル変換モデル $v_{doc2vec}$ を用いる [5]。今回の文書集合は行動履歴から構成されるものであり、厳密には我々が日常で使うような文章ではないので、公開されている事前学習済みモデルは転移させることが出来ず、モデルを一から学習させている。

2.4 マッピング機構

ドメイン s に対して、v^s を $B^{s,train} := \{b^s_u \mid u \in U^{train}\}$ から訓練されたベクトル変換モデル、$\mathbf{v}^s_u := v^s(b^s_u) \in \mathbf{R}^{d_s}$ を行動履歴 b^s_u のベクトル表現とし、ドメイン t でも同様の記法を用いる。この章の目的は、両方のデータセットに現れるユーザ u に対して、$f(\mathbf{v}^s_u) \approx \mathbf{v}^t_u$ となるような写像 $f : \mathbf{R}^{d_s} \to \mathbf{R}^{d_t}$ を構成することである。ベクトル変換モデルと同様に、この章では複数個の写像を考えるが、後に選択機構により適切な写像が選ばれる。

一つの構成方法は写像 $f : \mathbf{v}^s_u \mapsto \mathbf{v}^t_u$ を回帰として捉える方法である。この論文では、線形リッジ回帰 f_{linear}、RBF カーネルを使用したカーネルリッジ回帰 f_{kernel} 、および ℓ 層ニューラルネットワーク $f_{\ell nn}$ の回帰モデルを検討する。リッジ回帰の各パラメータは、グリッドサーチによって決定する。ニューラルネットワークに関しては、1 つの 100 次元隠れ層を持つ 2 層ニューラルネットワーク f_{2nn} と 2 つの 100 次元隠れ層を持つ 3 層ニューラルネットワーク f_{3nn} を考え、最適化アルゴリズムとして Adam [Kingma and Ba, 2015]、活性化関数として ReLU [Nair and Hinton, 2010] を使用し、バッチサイズを 100、エポックサイズを 100 に設定する。

別の写像として、2 つのベクトル $\mathbf{v}^s_u$ と $\mathbf{v}^t_u$ が同じ次元である場合、恒等写像 f_{id} も適用可能である。v_* ($* \in \{$bow、tfidf、bm25, concat$\}$) の場合、一般的に $d_s \neq d_t$ であるため、次元を調整するために各ベクトル変換モデルの座標を $I^s \cup I^t$ のすべてのアイテムにリセットすることで座標を揃えることができる。その際、$\mathbf{v}^s_u$ の拡張 $\tilde{\mathbf{v}}^s_u \in \mathbf{R}^d$ $(d = |I^s \cup I^t|)$ を $\tilde{\mathbf{v}}^s_u[i] = \mathbf{v}^s_u[i]$ $(i \in I^s)$; 0 $(i \in I^t \setminus I^s)$ として定める。

2.5 選択機構

これまで、ベクトル変換モデル v と写像 f を構成してきたが、この章ではデータセットに応じて最適な組み合わせ (v, f) を選択する方法について述べる。良さを測る指標の一つとして、平均逆順位 (MRR) を用いる。UIL において MRR を計算するためには、ドメイン s の同一ユーザを探す対象のテストユーザ集合 U^{test} とドメイン t の候補ユーザ集合 $(U^t \supset) U^{candi} \supset U^{test}$ を定める必要がある。ひとたび類似度関数 $\mathrm{sim}_{v,f}$ が定まると、U^{test} の各ユーザ u に対して、$\mathrm{sim}_{v,f}(b^s_u, b^t_{u_1}) \geqq \mathrm{sim}_{v,f}(b^s_u, b^t_{u_2}) \geqq \cdots$ となるように候補者ユーザ $u_1, \ldots, u_N \in U^{candi}$ を並び替えることが出来、$b^t_u = b^t_{u_{r_u}}$ の時に、ユーザ u の正解の順位を r_u で定める。このとき、MRR スコアは

$$\mathrm{MRR}_{v,f}(U^{test}, U^{candi}) := \frac{1}{|U^{test}|} \sum_{u \in U^{test}} \frac{1}{r_u}.$$

5：実験時には、Gensim モジュール (`https://radimrehurek.com/gensim/models/doc2vec.html` を用い、Doc2Vec の次元は 300、ウィンドウサイズは 5 に固定し、その他の言及していないパラメータは Gensim のデフォルトパラメータとする。

として定める [6]。

MRR を用いた最適な組み合わせ (v, f) の判定方法は K-fold 交差検証に従う [7]。まず、教師ユーザ集合を U^{train} を K 等分し、それらを $U_1, \ldots, U_K$ とする。ここでは、k を一つ固定し、$U^{tr}_{cv} := U^{train} \setminus U_k$, $U^{te}_{cv} := U_k$ とする。次に、U^{tr}_{cv} の行動履歴から v と f を構成し、$s_k(v, f) := \mathrm{MRR}_{v,f}(U^{te}_{cv}, U^{train})$ を計算する。この手順を $k = 1$ から K まで繰り返し、それらの平均 $\bar{s}(v, f) := K^{-1} \sum_{k=1}^{K} s_k(v, f)$ を計算する。最後に、その平均が一番高くなる (v, f) を選択する。疑似コードはアルゴリズム 1 で与えられる。

Algorithm 1 選択機構

Input: 教師ユーザ集合 U^{train}、正の整数 $K \geqq 1$

Output: 選択されたベクトル変換モデル $\hat{v}$ と写像 $\hat{f}$

Initialization：初期化されたベクトル変換モデルの集合 $\mathcal{V}$ と写像の集合 $\mathcal{F}$

U^{train} を K 分割し、$U_1, \ldots, U_K$ とする。
for $k = 1, \ldots, K$ **do**
  $U^{te}_{cv} := U_k$, $U^{tr}_{cv} := U^{train} \setminus U_k$ とする。
  for v in $\mathcal{V}$ **do**
    v^s を $\{b^s_u \mid u \in U^{tr}_{cv}\}$ で学習させる。
    v^t を $\{b^t_u \mid u \in U^{tr}_{cv}\}$ で学習させる。
    for f in $\mathcal{F}$ **do**
      f を $\{(v^s(b^s_u), v^t(b^t_u)) \mid u \in U^{tr}_{cv}\}$ で学習させる。
      $s_k(v, f) = \mathrm{MRR}_{v,f}(U^{te}_{cv}, U^{train})$ を計算する。
    end for
  end for
end for
$\bar{s}(v, f) := K^{-1} \sum_{k=1}^{K} s_k(v, f)$ を計算する。
最適なモデルの組み合わせを選択する。

$$\hat{v}, \hat{f} = \underset{v \in \mathcal{V}, f \in \mathcal{F}}{\mathrm{argmax}}\ \bar{s}(v, f).$$

2.6 マッチング機構

アルゴリズム 1 では、MRR を用いてユーザ連携の性能評価を行った。別の性能評価の指標として、top-k 精度

$$\mathrm{Acc}_k := \frac{1}{|U^{test}|} \sum_{u \in U^{test}} \mathrm{id}(r_u \leqq k)$$

が UIL では広く用いられる。ここで、id は指示関数 [8]、r_u は u の順位を表す。特に、top-1 精度は、ユーザペアの一致精度を測定している：

$$\mathrm{id}(r_u = 1) = \mathrm{id}\left(u = \underset{u' \in U^{candi}}{\mathrm{argmax}}\ \mathrm{sim}_{v,f}(b^s_u, b^t_{u'})\right).$$

ただし、この類似度が最大のユーザをマッチングさせるやり方は、ドメイン t のある一人のユーザに対して、ドメイン s の異

6：ここでは、依存性を強調して $v, f, U^{test}, U^{candi}$ を MRR の記号に含める。

7：実験では、$K = 3$ とする。

8：事象 E に対して、E が正の場合に 1 を返し、そうでない場合は 0 を返す関数。

uID	Preference ordering
s001	**t02**, t03, t01, t04
s002	**t04**, t01, t03, t02
s003	**t02**, t03, t04, t01

uID	Preference ordering
t01	s002, s003, s001
t02	**s001**, **s003**, s002
t03	s001, s002, s003
t04	s002, s003, s001

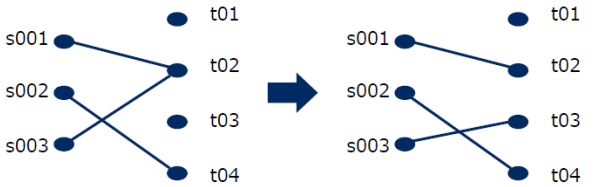

図 2　Gale-Shapley アルゴリズムの概要。この例において、最大類似度によるマッチングでは $s001$ と $s003$ が $t02$ に連携されている (左のグラフ)。各ユーザの好み順 (上の表) を見ると、$t02$ は $s003$ よりも $s001$ を好むため Gale-Shapley アルゴリズムでは、$t02$ には $s001$ を連携させ、余った $s003$ は次の好み順から $t04$ と連携される (右のグラフ)。

なるテストユーザを連携する可能性があり、これには現実世界では一人しかいないユーザが重複してしまうという問題がある。

上記の矛盾を避けるために、2 部グラフマッチングアルゴリズムの一つである Gale-Shapley アルゴリズム [Gale and Shapley, 1962] を UIL に使う (図 2)。

Gale-Shapley アルゴリズムを UIL に適用するには、$u \in U^{test}$ と $u' \in U^{candi}$ のすべてのペアに対して $\mathrm{sim}(b^s_u, b^t_{u'})$ を計算し、類似度の値で並べ替えられたドメイン s と t の行動履歴の系列 $(b^t_{u_1}, b^t_{u_2}, \ldots), (b^s_{u'_1}, b^s_{u'_2}, \ldots)$ を計算する。Gale-Shapley アルゴリズムは、各ドメインのユーザの好み順のリスト $\{(b^t_{u_1}, b^t_{u_2}, \ldots) \mid u \in U^{test}\}, \{(b^t_{u'_1}, b^s_{u'_2}, \ldots) \mid u' \in U^{candi}\}$ を入力することで、ユーザに重複のない行動履歴のペアの集合 $\{(b^s_u, b^t_{\hat{u}}) \mid u \in U^{test}\}$ を出力する。ここで、$\hat{u}$ は推測された u の連携ユーザである。Gale-Shapley アルゴリズムを応用したユーザ連携の疑似コードはアルゴリズム 2 で与える。

Algorithm 2 マッチング機構

Input: テストユーザ集合 U^{test}、候補ユーザ集合 U^{candi}、類似度関数 $\mathrm{sim}_{v,f}$

Output: 重複のないユーザペアの行動履歴集合 $\{(b^s_u, b^t_{\hat{u}}) \mid u \in U^{test}\}$

1: $\mathrm{sim}(b^s_u, b^t_{u_1}) \geqq \mathrm{sim}(b^s_u, b^t_{u_2}) \geqq \cdots$ に沿って、ドメイン s でのユーザ u の好み順 $(b^t_{u_1}, b^t_{u_2}, \cdots)$ を並び替える。
2: $\mathrm{sim}(b^s_{u'_1}, b^t_{u'}) \geqq \mathrm{sim}(b^s_{u'_2}, b^t_{u'}) \geqq \cdots$ に沿って、ドメイン t でのユーザ u' の好み順 $(b^s_{u'_1}, b^s_{u'_2}, \cdots)$ を並び替える。
3: Gale-Shapley アルゴリズムを上記のユーザの好み順に対して適用する。
4: $\{(b^s_u, b^t_{\hat{u}}) \mid u \in U^{test}\}$ を得る。

重複のないユーザに対する連携精度として

$$\mathrm{Acc}_{\mathrm{GS}} := \frac{1}{|U^{test}|} \sum_{u \in U^{test}} \mathrm{id}(u = \hat{u})$$

を定義する。ユーザ重複の課題を解消していることから、$\mathrm{Acc}_{\mathrm{GS}}$ は Acc_1 よりも高くなることが予想される。

3 実　　験

3.1 データセット

本論文では、Instacart、Click-Through Rate (CTR)、および Amazon データセットを使用する。各データセットは 2 つのドメイン s と t に分割することで、UIL の設定を再現している。

- Instacart データセット [9]は通販サイト上の 1 年間分の食品の購入履歴からなるデータセットである。このデータセットでは、正確な購入日時は隠されており、代わりにユーザが初めて購入した日からの経過日数が T が与えられている。すべてのユーザは $T = 0$ で商品を購入するものの、$T > 300$ 後に購入するユーザ数は少ないため、行動回数のバランスをとって $T = 100$ の前と後の購入履歴で、ドメイン s と t を設定する。
- CTR データセット [10] はオンライン広告をユーザに提示したときにクリックしたかどうかのクリックログからなるデータセットである。このデータセットでは、device_ip と site_id をユーザ ID とアイテム ID として扱う。CTR データセットにはサイトがクリックされたかどうかのフラグがあるため、ドメイン s と t をクリック履歴と閲覧履歴の集まりとして設定する。
- Amazon データセット [11] [Ni et al., 2019] はユーザのレビュー履歴からなるデータセットである。ドメイン s と t として “5-core in Small subsets for experimentation” の “Books” と “Kindle Store” を設定する。

各データセットに関する数値データは表 1 で与えられる。表 1 の s, t の列の値は各ドメインの教師ユーザの行動履歴に含まれるアイテムの総数、Jac 列の値はドメイン間のアイテム集合 I^s, I^t の Jaccard similarity であり、Amazon データセットでは共通して現れるアイテムが存在していないことが示唆される [12]。

表 1　データセットの分析結果。

	Instacart	CTR	Amazon
s	28,356	1,907	83,423
t	30,338	2,749	209,762
Jac	0.67	0.69	0

3 つのデータセット全てにおいて、公平な比較をするという観点で、教師ユーザ、テストユーザ、候補ユーザの数を $5000, 500, 10000$ に固定する。ユーザの質を担保するために、教師ユーザは両方のドメインで 50 回以上行動しているユーザに限定し、その中でランダムに選ばれた 5000 人とする。教師ユーザと同様に、テストユーザはドメイン s で 50 回以上行動しているユーザの中からランダムに選ばれた 500 人とする [13]。候補ユーザは $U^t \setminus U^{train}$ から仮定をつけずにランダムに選ばれた 10000 人とする。

3.2 結　　果

表 2 は、提案アルゴリズム 1 で選択したベクトルと写像の組み合わせ、MRR スコア、アルゴリズム 2 による連携精度、top-k 精度 ($k = 1, 5, 10$) を示している [14]。

表 2　実験結果。MRR と Acc の単位は %。

Dataset	Models	MRR	$\mathrm{Acc}_{\mathrm{GS}}$	Acc_1	Acc_5	Acc_{10}
Instacart	(v_{concat}, f_{id})	93.4	92.2	92.0	95.0	95.8
CTR	(v_{bm25}, f_{id})	56.1	54.0	50.4	61.6	69.0
Amazon	(v_{bm25}, f_{linear})	76.2	86.8	65.0	91.0	95.6

結果としては、Instacart、CTR、Amazon データセットの連携精度 ($\mathrm{Acc}_{\mathrm{GS}}$) はそれぞれ $92.2\%, 54.0\%, 86.8\%$ であった。機械学習を用いずにアプリケーションレベルで作動する Cookie Sync が $62 - 73\%$ [Papadopoulos et al., 2019] と報告されていることを考えると、Cookie に依存せずに CTR で 54% の連携精度を示していることは十分な成果である。また、異なるドメインに共通して現れるアイテムが存在していないにもかかわらず、Amazon データを 86.8% の精度で連携出来ていることも注目に値する。Amazon データセットのドメイン間のギャップに関しては、回帰モデル f_{linear} が埋めていると思われる。$\mathrm{Acc}_{\mathrm{GS}}$ と Acc_1 を比較すると、ユーザの重複を除くことが連携精度向上に寄与していることが観測されている。

表 2 の Models 列を見ると、選択したモデルが各データセットで異なるということは、データセットに適した類似度関数を設計することが出来ていることを示唆している。興味深い考察として、選択された方式は圧縮しないベクトル表現と線形な写像という複雑ではない組み合わせで良い連携が達成されている。この研究を開始した当初は、ユーザの移動履歴に特化した最先端の UIL 手法 [Zhou et al., 2018, Feng et al., 2019] がディープラーニングを用いた方法を提案していたため、ディープラーニングベースのモデル ($(v_{doc2vec}, f_{nn})$ を加工したような手法) を考えていたがあまり機能しなかった。提案法内の選択機構は、各データセットで最適なモデルを検索することで、複雑なモデルが最適値を達成するはずであるという偏見を防ぐのにも役立っている。

実は、従来手法である [Iofciu et al., 2011] と [Goga et al., 2013] は、今回の提案手法における (v_{bm25}, f_{id}) と (v_{tfidf}, f_{id}) の組み合わせを選択していることと見なすことが出来る。従って、過学習が起こらない限り、データセットに応じて (v_{bm25}, f_{id}) と (v_{tfidf}, f_{id}) よりも優れた組み合わせを選択できる提案法が理論的には良い連携を実現する。表 3 は、関連研究 [Iofciu et al., 2011, Goga et al., 2013] との比較実験の結果を示している。提案法は他の方法よりも高い MRR を示し、特に Amazon デー

9：https://www.kaggle.com/c/instacart-market-basket-analysis/data

10：https://www.kaggle.com/c/avazu-ctr-prediction/data

11：https://nijianmo.github.io/amazon/index.html

12：Jaccard similarity は、2 つの集合 A, B の間の類似性を $\mathrm{Jac}(A, B) := |A \cap B|/|A \cup B|$ として計算する。

13：これから連携を取ろうとしているドメイン t 内のテストユーザの行動は不明であるため、テストユーザのドメイン t での仮定を課すことはできない。

14：今回の実験では正解ペアを探す候補ユーザ数は 10000 であるため、ユーザの連携精度のチャンスレートは 0.01% である。

タセットの場合、提案法の MRR は 54.49 であったのに対し、両方のドメインで共通して現れるアイテムの存在を仮定した手法である [Iofciu et al., 2011, Goga et al., 2013] は連携することに失敗している。

表 3　平均 MRR の詳細。

Method	Instacart	CTR	Amazon
Our method	**99.49**	**62.57**	**54.49**
[Iofciu et al., 2011]	99.42	**62.57**	0.42
[Goga et al., 2013]	98.04	37.14	0.42

4 関 連 研 究

UIL は様々な分野で研究されており、各提案手法はユーザに関連するデータの種類によって異なる。サーベイ論文 [Shu et al., 2016] ではユーザプロファイル、ユーザネットワーク、およびユーザコンテンツと大きく 3 つユーザデータを分類している。ユーザプロファイル情報はユーザ名やメールアドレスなどのプロフィール情報や、学歴、年齢、性別などのデモグラフィック属性に対応し、ユーザネットワーク情報は Twitter のフォロー/フォロワー関係などのユーザのソーシャルグラフに対応し、ユーザコンテンツ情報は文章、写真、移動などの行動履歴に対応する。データの種類がユーザプロファイル情報のみが与えられる場合 [Zafarani and Liu, 2009, Zafarani and Liu, 2013, Mu et al., 2016], ユーザネットワーク情報のみが与えられれている場合 [Man et al., 2016, Zhou et al., 2018, Zhong et al., 2018]、ユーザプロファイル、ネットワーク、コンテンツ情報の全ての組み合わせを使える場合 [Liu et al., 2014, Hadgu and Gundam, 2020] など、状況に応じて提案される UIL の手法は異なっている。

ユーザコンテンツ情報の一例としての行動履歴に対処する研究の多くは移動履歴に焦点を当てている [Goga et al., 2013, Kong et al., 2013, Riederer et al., 2016, Feng et al., 2019]。これらの手法のほとんどは緯度・経度や郵便番号などのドメインに固有の追加情報を必要とするため、本研究の設定に位置特化の手法を適用することはできない。

特定のドメインに特化しない汎用的な行動履歴に対する UIL は [Iofciu et al., 2011] および [Goga et al., 2013](の TFIDF 版) で提案されている。これらの手法では、二つの異なる SNS データにおいてユーザが付与したある種のタグに対して、タグの出現頻度のドメイン毎の違いに着目してユーザ連携を実施している。しかし、暗黙のうちに両方の SNS に共通して現れるタグ (アイテム) の存在を仮定しているため、共通するタグが現れない場合ではうまく機能しない。

5 結　　論

本研究の貢献の一つは、デモグラフィック属性が利用できず、同じユーザの行動パターンがドメイン毎に異なっている場合にユーザ連携を行う新しい枠組みを提案したことにある。提案法は、ユーザの固有の特徴を抽出するベクトル機構と同じユーザの 2 つの異なる行動パターン間のギャップを埋めるマッピング機構で構成されており、様々なモデルの選択肢の中からデータセットに最適な組み合わせを選択機構によって自動的に選択する。実験では、Instacart、CTR、および Amazon データセットのテストユーザを、それぞれ 92.2%, 54.0%, 86.8% の精度で連携することに成功している。特に、ドメインの共通部分に注目する関連研究ではユーザ連携に失敗している Amazon データセットに対し、本提案法の連携精度が高いことはその優位性を示している。

謝　　辞

本研究については東京大学の Hao Jia 氏、NEC の董于洋氏に有益なコメントを頂いた。ここに謝意を表す。

文　　献

[Elmeleegy et al., 2013] Elmeleegy, H., Li, Y., Qi, Y., Wilmot, P., Wu, M., Kolay, S., Dasdan, A., and Chen, S. (2013). Overview of turn data management platform for digital advertising. *Proc. VLDB Endow.*, 6(11):1138–1149.

[Feng et al., 2019] Feng, J., Zhang, M., Wang, H., Yang, Z., Zhang, C., Li, Y., and Jin, D. (2019). DPLink: User identity linkage via deep neural network from heterogeneous mobility data. In *WWW*, pages 459–469. ACM.

[Gale and Shapley, 1962] Gale, D. and Shapley, L. S. (1962). College admissions and the stability of marriage. *The American Mathematical Monthly*, 69(1):9–15.

[Goga et al., 2013] Goga, O., Lei, H., Parthasarathi, S. H. K., Friedland, G., Sommer, R., and Teixeira, R. (2013). Exploiting innocuous activity for correlating users across sites. In *WWW*, pages 447–458. International World Wide Web Conferences Steering Committee / ACM.

[Hadgu and Gundam, 2020] Hadgu, A. T. and Gundam, J. K. R. (2020). Learn2link: Linking the social and academic profiles of researchers. In *ICWSM*, pages 240–249. AAAI Press.

[Iofciu et al., 2011] Iofciu, T., Fankhauser, P., Abel, F., and Bischoff, K. (2011). Identifying users across social tagging systems. In *ICWSM*. The AAAI Press.

[Jones, 2004] Jones, K. S. (2004). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 60(5):493–502.

[Jones et al., 2000] Jones, K. S., Walker, S., and Robertson, S. E. (2000). A probabilistic model of information retrieval: development and comparative experiments. *Inf. Process. Manag.*, 36(6):779–840.

[Kingma and Ba, 2015] Kingma, D. P. and Ba, J. (2015). Adam: A method for stochastic optimization. In *ICLR*.

[Kong et al., 2013] Kong, X., Zhang, J., and Yu, P. S. (2013). Inferring anchor links across multiple heterogeneous social networks. In *CIKM*, pages 179–188. ACM.

[Le and Mikolov, 2014] Le, Q. V. and Mikolov, T. (2014). Distributed representations of sentences and documents. In *ICML*, volume 32 of *JMLR Workshop and Conference Proceedings*, pages 1188–1196. JMLR.org.

[Lee and Seung, 2000] Lee, D. D. and Seung, H. S. (2000). Algorithms for non-negative matrix factorization. In *NIPS*, pages 556–562. MIT Press.

[Liu et al., 2014] Liu, S., Wang, S., Zhu, F., Zhang, J., and Krishnan, R. (2014). HYDRA: large-scale social identity linkage via heterogeneous behavior modeling. In *SIGMOD Conference*, pages 51–62. ACM.

[Man et al., 2016] Man, T., Shen, H., Liu, S., Jin, X., and Cheng, X. (2016). Predict anchor links across social networks via an embedding approach. In *IJCAI*, pages 1823–1829. IJCAI/AAAI Press.

[Mu et al., 2016] Mu, X., Zhu, F., Lim, E., Xiao, J., Wang, J., and Zhou, Z. (2016). User identity linkage by latent user space modelling. In *KDD*, pages 1775–1784. ACM.

[Nair and Hinton, 2010] Nair, V. and Hinton, G. E. (2010). Rectified linear units improve restricted boltzmann machines. In *ICML*, pages 807–814. Omnipress.

[Ni et al., 2019] Ni, J., Li, J., and McAuley, J. J. (2019). Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In *EMNLP/IJCNLP (1)*, pages 188–197. Association for Computational Linguistics.

[Papadopoulos et al., 2019] Papadopoulos, P., Kourtellis, N., and Markatos, E. P. (2019). Cookie synchronization: Everything you always wanted to know but were afraid to ask. In *WWW*, pages 1432–1442. ACM.

[Riederer et al., 2016] Riederer, C. J., Kim, Y., Chaintreau, A., Korula, N., and Lattanzi, S. (2016). Linking users across domains with location data: Theory and validation. In *WWW*, pages 707–719. ACM.

[Shu et al., 2016] Shu, K., Wang, S., Tang, J., Zafarani, R., and Liu, H. (2016). User identity linkage across online social networks: A review. *SIGKDD Explorations*, 18(2):5–17.

[Zafarani and Liu, 2009] Zafarani, R. and Liu, H. (2009). Connecting corresponding identities across communities. In *ICWSM*. The AAAI Press.

[Zafarani and Liu, 2013] Zafarani, R. and Liu, H. (2013). Connecting users across social media sites: a behavioral-modeling approach. In *KDD*, pages 41–49. ACM.

[Zhong et al., 2018] Zhong, Z., Cao, Y., Guo, M., and Nie, Z. (2018). CoLink: An unsupervised framework for user identity linkage. In *AAAI*, pages 5714–5721. AAAI Press.

[Zhou et al., 2018] Zhou, F., Liu, L., Zhang, K., Trajcevski, G., Wu, J., and Zhong, T. (2018). DeepLink: A deep learning approach for user identity linkage. In *INFOCOM*, pages 1313–1321. IEEE.