ユーザグラフ上のランダムウォークに基づくクロスドメイン推薦

Recommendations by Randomly Walking on Cross Domain User Graph

中辻 真
Makoto Nakatsuji

NTT サービスエボリューション研究所
NTT Service Evolution Laboratories
nakatsuji.makoto@lab.ntt.co.jp

藤原 靖宏
Yasuhiro Fujiwara

NTT ソフトウェアイノベーションセンタ
NTT Software Innovation Center
fujiwara.yasuhiro@lab.ntt.co.jp

内山 俊郎
Toshio Uchiyama

NTT サービスエボリューション研究所
NTT Service Evolution Laboratories
uchiuyama.toshio@lab.ntt.co.jp

keywords: cross domain recommendation, random walk, graph mining

Summary

Content providers want to make recommendations across multiple interrelated domains such as music and movies. However, existing collaborative filtering methods fail to accurately identify items that may be interesting to the user but that lie in domains that the user has not accessed before. This is mainly because of the paucity of user transactions across multiple item domains. Our method is based on the observation that users who share similar items or who share social connections, can provide recommendation chains (sequences of transitively associated edges) to items in other domains. It first builds domain-specific-user graphs (DSUGs) whose nodes, users, are linked by weighted edges that reflect user similarity. It then connects the DSUGs via the users who rated items in several domains or via the users who share social connections, to create a cross-domain- user graph (CDUG). It performs Random Walk with Restarts on the CDUG to extract user nodes that are related to the starting user node on the CDUG even though they are not present in the DSUG of the starting user node. It then adds items possessed by those users to the recommendations of the starting node user. Furthermore, to extract many more user nodes, we employ a taxonomy-based similarity measure that states that users are similar if they share the same items and/or same classes. Thus we can set many suitable routes from the starting user node to other user nodes in the CDUG. An evaluation using rating datasets in two interrelated domains and social connection histories of users as extracted from a blog portal, indicates that our method identifies potentially interesting items in other domains with higher accuracy than is possible with existing CF methods.

1. は じ め に

多くのコンテンツプロバイダは、複数の異なるが関連性のあるサービスドメイン（以下、ドメイン）のアイテムを扱うようになっている。本論文では、ドメインをアイテムのタクソノミの上位に属する、粒度の粗いクラスと定義する。例えば、Amazon.com^{*1}では、書籍、音楽、映画、ファッションといった複数の異なるドメインのアイテムを取り扱っている。こうした環境では、複数のドメインを跨った効果的なアイテム推薦技術が必要となる。多くのコンテンツプロバイダは、協調フィルタリングに基づく推薦システムを構築している。協調フィルタリングは、“被推薦ユーザと同じアイテムを消費したユーザは、被推薦ユーザと類似する興味を持つ傾向にある”という仮定に基づく。一般的な協調フィルタリングでは、ユー

ザ間やアイテム間の興味の類似性を、アイテムに対する評価値の一致性で計算する。結果として、被推薦ユーザが過去に多く評価していたのと同じドメイン・ジャンル・性質のアイテムが推薦されることが多い。例えば、もしユーザがホラー映画のアイテムを高く評価していたとすると、一般的な協調フィルタリングでは、同じ監督により作られた映画や同じ役者が出ている映画、または同じホラージャンルの映画を推薦する事が多い。言い換えると、複数のドメインを跨ってアイテムを消費するユーザが少ないため、既存の協調フィルタリングでは、被推薦ユーザが評価をした事がないドメインのアイテムを推薦するのは困難である。また例え、ある程度の数のユーザが多くのドメインのアイテムに興味を持っていたとしても、ユーザの興味は特定のドメインに大きく偏ることが多い。そのため、既存の協調フィルタリングによる推薦は特定のドメインに偏った推薦となる事が多い。

*1 <http://www.amazon.com>

ドメインを跨る推薦は非常に重要となってきたが、現状ではクロスドメインを対象とする研究は少ない。文献 [Li 09b, Li 09a] では、複数のアイテムドメインを跨って同様の評価値の分布を行うユーザ集合を分析する。彼らの手法では、ユーザやアイテムが重複していない複数のドメインにおける評点データセットにおいて、同様の評価値の分布を持つユーザ集合の対応関係を知識として持つ。そして、映画データセットと書籍データセットを用いた検証において、複数ドメインから得られた知識を用い各ドメイン内での推薦精度を改善できる事を示した。しかし彼らの手法は、ユーザが過去に評価した事の無いドメインのアイテムの推薦を目的としたものではない。

本研究では、被推薦ユーザがまだ評価した事の無いドメインのアイテムを精度良く推薦可能とする手法を提案する。本手法を用いる事で、複数のドメインのアイテムを持つコンテンツプロバイダは、ドメインを跨る推薦を実現でき、ドメインを跨るユーザの回遊や購買を促進できる。特にあるドメインの評点データセットを多く持つが他ドメインの評点データセットを僅かしか持たない、ドメイン毎のユーザの購買数に偏りのあるコンテンツプロバイダには大きなメリットとなる。ユーザにとっても、例えば、ある音楽コンテンツプロバイダにおいて、邦楽のアイテムのみを消費していたユーザに洋楽の良いアイテムを多く推薦することができ、ユーザがそのアイテムに興味を持てれば、そのユーザは新たな興味を広げて、世の中に存在する様々な文化から喜びを享受できるようになる*2。

上記の目的を達成するため、本研究では以下の2つのアプローチを採る。

一つめのアプローチは、同じアイテムを共有するユーザや、ソーシャルネットワーク上でやり取りをした事のあるユーザの関係をグラフとして構築する事である。これにより、あるドメインのアイテムから異なるドメインのアイテムへの経路を構築でき、ドメインを跨るアイテムを推薦可能とするための透過的な相関 [Huang 04] を構築できる。邦楽アイテムと洋楽アイテムの評価サイトを例に挙げ説明する。邦楽と洋楽に属するユーザ集合は互いに異なる集合ではあるが、それらはより大きな同一の母集団からサンプルされた部分集合である(この仮定はある程度大きなWEBサイトに限定される) [Coyle 08]。このような特性を満たしている場合、つまり異なるドメイン間である程度ユーザ集合が重なっている場合*3に本アプローチは有効である。具体的には、提案手法ではまずドメイン特化ユーザグラフ (Domain-specific-user graph; DSUG) を構築する。DSUG は、ユーザをノードとしユーザ間の類似度を重みとして持つエッジからなり、各ドメ

イン毎に構築されるものである。また、ソーシャルベースユーザグラフ (Social-based user graph; SUG) も構築する。SUG は、ユーザをノードとしソーシャルネットワーク上においてユーザ間でコミュニケーションをしている場合にエッジを与える。次に、ドメイン毎に作られたDSUGとSUGを、複数ドメインに対し評点を与えているユーザやソーシャルネットワーク上で交流を行なっているユーザを介して連結し、クロスドメインユーザグラフ (Cross-domain-user graph; CDUG) を構築する。その上で、CDUG上を被推薦ユーザに対応するノードを起点とするRandom Walk with Restarts (RWR) [Lovasz 96] を行い、CDUG上のノードを、グラフ上を遷移確率に従い移動するトークンと定義されるwalk [穂坂 10] が通過する頻度を解析する。そして、被推薦ユーザが所属していないドメインではあるが、頻度良く通過されているユーザノードを被推薦ユーザに関連しているノードとして抽出する。最終的に、こうして抽出されたノードに対応するユーザの持つアイテムを被推薦ユーザへの推薦に組み込む。

2つめのアプローチについて述べる。各ドメインのユーザの類似度を計算する際に、ユーザによって共有されているアイテムの一致性のみを用いていれば、ユーザが所持するアイテム数が少ない場合、類似するユーザが少なくなる。こうした状況は、様々なドメインのアイテムを評価しているユーザに対応するノードを跨って、あるDSUGから異なるDSUG上のノードにwalkが遷移する際に頻繁に発生する。例えば、複数のドメインのアイテムを評価するユーザであっても、邦楽のアイテムは多く評価しているが、洋楽は僅かしか評価していないというユーザが多くいる。もしwalkが、そうしたユーザを経由して、邦楽ドメインから洋楽ドメインに遷移する場合、洋楽ドメイン内に遷移後、そのノードから先へのエッジがほとんど無く、洋楽内の多くのユーザノードに遷移できなくなる。こうした問題を回避し、被推薦ユーザノードから他のドメインに所属するユーザノードへの適切な経路を多く与えるため、本手法では、タクソノミベースのユーザ類似度計算法 [Ziegler 04, Nakatsuji 10a] を適用する。アイテムのタクソノミはサービスプロバイダによって、顧客が自身の好きなアイテムに容易にアクセスできるように設計されている。故に、あるアイテムを好むユーザは、そのアイテムが所属するクラスのアイテムをも好む傾向があると仮定できる。そこで、ユーザへのアイテムの評価値をそのアイテムを含むクラスへの評価値として反映させる。これにより、アイテムへの評価の一致性のみでなくクラスへの評価の一致性も使い、ユーザ間の類似度を計算でき、ユーザの所持するアイテム数が少ない場合においてもユーザ間の類似度を適切に計算できる。

我々は、ブログポータル上の邦楽アイテムと洋楽アイテムへの書き込み頻度からなる評点データセットを用いた検証を実施した。さらに、同じブログポータルからユー

*2 洋楽と邦楽は日本国内のコンテンツプロバイダにおいて一般的にアイテムのタクソノミの上位で分岐する。例えば、本研究の検証で用いたlisten.jp (<http://listen.jp/>) のタクソノミにおいても、邦楽と洋楽は最上位で分岐している。

*3 アイテム集合は異なるドメイン間で重なっている必要は無い。

ザ間のコミュニケーションの履歴を取り出し、SUG を構築した。そして、邦楽データセットと洋楽データセットから構築された DSUG と SUG を連結し CDUG を構築した。結果として、邦楽の評点データセットと洋楽の評点データセットを混ぜた混合データセットから作られた DSUG 上での RWR よりも、提案する CDUG 上での RWR に基づきユーザのアイテムへの評価を予測した方が、被推薦ユーザがこれまで評価した事のないドメインのアイテムをより高精度に検出できることを確認した。さらに、DSUG を構築する際のユーザ類似度計算法として、アイテムの一致性のみに基づく手法よりも、タクソノミに基づく手法の方が、精度が大きく向上することも確認した。

本稿の構成について述べる。次章では関連研究について述べ、3 章では本研究の背景を述べる。次に、4 章ではクロスドメイン推薦を実現するために、DSUG と SUG を連結し、CDUG を構築する手法について説明する。また、アイテムのタクソノミに基づきユーザ類似度を計算する手法についても述べる。次に、5 章では 2 つの異なるドメインのアイテムを用い提案手法の検証を行い、最後に結論を述べる。

2. 関 連 研 究

複数のサービスドメインのアイテムを跨った推薦を行う研究として、ユーザの実世界上での行動をドメインに依存せずカテゴライズしてユーザの興味プロフィールを構築し、コンテンツベースフィルタリングに基づき興味に沿ったアイテムの推薦を試みる研究がある [Fukazawa 11]。また、ターゲットドメインとは異なるドメインから知識を転移することで、分類モデルやランキングモデルを学習する機械学習の研究がある [Li 09b, Li 09a, Li 11, Wang 09, Chen 08]。しかし、我々の知る限り、協調フィルタリングにおいて、書籍や映画等の関連するドメインを跨ったクロスドメイン推薦の研究をしているのは、前章でも述べた [Li 09b, Li 09a, Li 11] のグループと [Berkovsky 07a, Berkovsky 07b] のグループ、および我々の先行研究である [Nakatsuji 10b] らの研究のみである。しかし [Li 09b, Li 09a, Li 11] らの手法は、あるドメイン内の推薦の精度を向上するために、他ドメインのデータセットから得られる知識を活用するものであり、ユーザが過去に評価した事のないドメインのアイテムの推薦を目的としたものではない。彼らの手法は、書籍と映画などの関連するドメイン間で、ユーザがアイテムに与える評点分布を対応つける。この方法により推薦精度が改善される根拠は、書籍と映画はホラーやロマンスといったジャンルが共通したアイテムを取り扱っており、それが故に、ユーザの評点分布も同様の傾向になるという仮定に基づく。一方、本稿の提案手法では、僅かではあるが複数ドメインに属するユーザをバイパスとしてドメイン間を連結して

おり、例え評点分布が異なる傾向のあるドメインであっても有効性が期待できる。言い換えると、本手法は、僅かではあっても複数ドメインに属するユーザを必要とする。[Berkovsky 07a, Berkovsky 07b] のグループは、我々と同様に複数ドメインに属するユーザを必要とする手法である。彼らの研究は、ドメイン等のようなコンテンツの内容に沿って分割されたデータを用い、各分割されたデータセット内での協調フィルタリングの結果を組合せ、各データセットに沿った推薦結果を計算する。しかし、彼らの手法は各分割されたデータセット（例えば、ドメイン）内でのユーザ間やアイテム間の類似性は、同じアイテムを共有していたり、同じユーザを共有していない限り計算できず、結果としてクロスドメインの推薦結果を提示できないユーザも多く存在する事になる。我々の手法では、タクソノミに基づくユーザ類似度法や RWR を用いたグラフマイニング法を用いることで、ユーザ間で同じアイテムを共有していなくても間接的な類似性を計算できる点が異なる。我々の検証では、2 つのドメインのうち片方のドメインのアイテムのみを消費しているユーザを対象として、クロスドメインの検証を行なっているが、これは [Berkovsky 07a, Berkovsky 07b] のグループの研究では困難な事である。また [Berkovsky 07a, Berkovsky 07b] のグループは、本研究のようにソーシャルコネクションの履歴を用いていない。

グラフ上での Random Walk や RWR を用いて推薦を計算する研究がいくつかある [Gori 07, Konstas 09, Yildirim 08]。Konstas らは、last.fm^{*4}からクロールしたデータセットを用い、ユーザのアイテム視聴履歴のみでなく、アイテムへユーザ群が与えたタグやユーザ間のソーシャルコネクションの履歴等の種類の異なるデータを組合せてグラフを構築した。そして、こうした異種データを組み合わせることでユーザの興味を予測する際には、グラフモデルのアプローチが良いということを示した。本研究でも異なるアイテムドメインから得られる異なるアイテムの評点データセットを利用するため、グラフモデルのアプローチを採用する。我々の知る限り、被推薦ユーザが過去に評価した事のないドメインのアイテムを CDUG 上の Random Walk や RWR に基づき精度良く検出するのは本研究が初めてである。

アイテムのタクソノミ（もしくは、アイテムに対するシンプルなオントロジと呼ばれることもある）を用い推薦の精度向上を行う研究がある [Ziegler 04, Nakatsuji 10a]。彼らの手法は、ユーザのアイテム消費量が少ない場合に有効であるとされる。[Ziegler 04] の手法と [Nakatsuji 10a] の手法の大きな違いは、[Ziegler 04] は、クラスに対しユーザのアイテムの評点を反映した上で、ユーザ毎にクラスに対する消費頻度からなるベクトルを構築し、ベクトル間の Pearson 相関係数に基づきユーザ間のクラスに関する興味の類似度を計算している。それに対し、[Nakatsuji

*4 <http://www.last.fm>

10a] では、クラスの階層構造に着目し、各階層に所属する”クラス/子クラス”集合毎にユーザ間の興味の類似度を計算し、それを全階層の全”クラス/子クラス”集合に対し計算している。タクソノミ上のクラス構造を活用しているため、[Nakatsuji 10a] の検証では [Ziegler 04] の手法よりも [Nakatsuji 10a] らの手法の方が評点の予測精度が良いことを示している。なお本稿の貢献は、タクソノミに基づくユーザ間の類似度計算法の改善ではない。本手法で提案する CDUG は、複数ドメインのアイテムを消費するユーザをバイパスとして、異なる DSUG 間で walk が迂回することを許可している。しかし我々の調査によると、複数ドメインのアイテムを消費するユーザであっても、特定のドメインのアイテムは多く購入するが、他ドメインのアイテムはそれほど購入しないことが多い。そこで、アイテムの購入頻度が少ない場合においても頑健な推薦精度を示すタクソノミベースのユーザ類似度計算法により DSUG を構築することで、結果として、クロスドメイン推薦の精度を大幅に向上することを示したことが貢献である。なお実験においては、[Ziegler 04] と [Nakatsuji 10a] の手法の両方を用いて検証している。なお、[Nakatsuji 10a] を発展し、複数タクソノミをリンクさせ詳細にユーザ興味を管理する方法 [Nakatsuji 11] や、ユーザの時系列に沿った興味の変化に追従する方法 [Nakatsuji 12] が近年提案されている。

一方、[Schickel-Zuber 06] らは、アイテムタクソノミ上のクラスに対し、クラス配下のアイテム所属数に基づく先験的なスコアを事前に入れておき、さらにクラス間でのスコアの関係性を示す尺度を導入している。その上で、その関係性の尺度に基づき、あるユーザが消費したあるクラス配下のアイテムから、別のユーザがまだ消費していないクラスのアイテムへとスコアを伝搬させ、それによりアイテムの予測値を計算する。彼らの手法はユーザ間のアイテム消費の関係性などを考慮しないため、協調フィルタリングの枠組みの外にあり、彼らはオントロジフィルタリングと名付けている。しかし、彼らの手法はユーザがまだ消費したことのないドメインのアイテムを推薦するには適していない。それは、例えば、タクソノミ上のルートクラスで複数ドメインを結合しクロスドメインを可能とするタクソノミを構築したとしても、彼らの手法はタクソノミ上で距離の離れたクラスに属するアイテムを精度良く推薦するのは難しいからである。

3. 背景

提案手法では、既存の協調フィルタリングを拡張し、被推薦ユーザがまだ評点を与えた事の無いドメインのアイテムで、興味を持ちうる可能性のあるアイテムを検出し推薦する。

3.1 協調フィルタリング

従来の協調フィルタリングにおいてユーザ間の類似度を計算する際には、Cosine ベースアプローチ [Breese 98] と Pearson 相関係数に基づくアプローチ [Resnick 94] を用いることが多い。Cosine ベースアプローチにおいては、被推薦ユーザ a とあるユーザ u の間の類似度 $S(a, u)$ は、 a と u の評価ベクトルの Cosine 角度を計算することで求められる。形式的には、両方のユーザが評価を与えているアイテム集合を $\mathcal{M}$ とし、ユーザ u のアイテム I_i に対する評価値を r_{u, I_i} とすると、類似度 $S(a, u)$ は、以下の式で与えられる。

$$S(a, u) = \frac{\sum_i^{\mathcal{M}} (r_{a, I_i})(r_{u, I_i})}{\sqrt{\sum_i^{\mathcal{M}} (r_{a, I_i})^2} \sqrt{\sum_i^{\mathcal{M}} (r_{u, I_i})^2}}$$

一方 Pearson 相関係数に基づくアプローチでは、ユーザ間の類似度の計算において、ユーザの評価スキームはユーザ毎に異なるという考え方を採用しており、以下の式であらわされる。

$$S(a, u) = \frac{\sum_i^{\mathcal{M}} (r_{a, I_i} - \bar{r}_a)(r_{u, I_i} - \bar{r}_u)}{\sqrt{\sum_i^{\mathcal{M}} (r_{a, I_i} - \bar{r}_a)^2} \sqrt{\sum_i^{\mathcal{M}} (r_{u, I_i} - \bar{r}_u)^2}} \quad (1)$$

ここで $\bar{r}_u$ は、ユーザ u のアイテムへの評価値の平均値を示す。

Cosine ベースアプローチも Pearson 相関係数に基づくアプローチも、両方のユーザ a と u が評価を与えたアイテムのみに焦点を絞り計算を行っている点に注意が必要である。

最終的に $\mathcal{N}$ を被推薦ユーザ a と類似度の高いユーザ集合とすると、 a のアイテム I_i に対する予測値 P_{a, I_i} は、以下の式 (2) で与えられる。

$$p_{a, I_i} = \bar{r}_a + \frac{\sum_u^{\mathcal{N}} (r_{u, I_i} - \bar{r}_u) S(a, u)}{\sum_u^{\mathcal{N}} S(a, u)} \quad (2)$$

3.2 Random Walk with Restarts

グラフはなんらかの継承性や関連性のある構造的なデータを自然に表現できる。グラフ上ではオブジェクトはノード、オブジェクト間の関係はエッジ、関係の強さはエッジ上に与えられるスコアで表現される。こうしたグラフに対し Random Walk with Restarts (RWR) [Lovasz 96] を適用することで、グラフ上の 2 つのノード間の関連度 (relatedness) を計算できる。RWR においては、あるノード a を起点ノードとし、各遷移ステップにおいて、別のノードへのリンクをノード間のエッジ上の重みに基づく遷移確率に沿ってランダムに選択することで、 a と他のノードの関連度を計算する。また各ステップにおいて、確率 α に従って起点ノード a に戻る。 $\mathbf{p}^{(t)}$ をその要素 $p_u^{(t)}$ がステップ t における walk がノード u から引き継がれる場合の確率を示す列ベクトルとする。また $\mathbf{q}$ を a に対応する要素のみ値が 1 で、残りが 0 となる列ベクトルとする。さらに $\mathbf{A}$ を各列が正規化されたグラフ上のノード間の隣接確率行列であるとする。つまり $\mathbf{A}$ の各要素

$A(u, v)$ は、現在のノードを u とし次のステップにおけるノードが v となる確率を示す．そうすると各ノードに対する定常確率は、以下の式 (3) を収束するまで繰り返すことにより得られる．

$$\mathbf{p}^{(t+1)} = (1 - \alpha)\mathbf{A}\mathbf{p}^{(t)} + \alpha\mathbf{q} \quad (3)$$

こうして得られる定常確率は始点ノードに確率 α でバイアスをかけた場合の始点ノードから各ノードに対し長期間の間でどれだけ訪問しうるか、という確率を与える．それゆえ収束したのちの状態を l とした場合 $p_a^{(l)}$ は、ノード a と u の関連度を計算したものととらえられる．

4. 提案手法

本章では、まず 4.1 節でクロスドメインユーザグラフ (Cross-domain-user graph; CDUG) の構築について述べ、4.2 節において、被推薦ユーザが興味を持ちうるが、まだ消費したことの無いドメインのアイテム検出法について述べる．そして 4.3 節において、ドメイン特化ユーザグラフ (Domain-specific-user graph; DSUG) を構築する際の、ユーザノード間のエッジ上の重みの値に影響を与えるユーザ間類似度計算法について述べる．

4.1 クロスドメインユーザグラフの構築

本章では、DSUG とソーシャルベースユーザグラフ (Social-based-user graph, SUG) を結合し、CDUG を構築する方法について述べる．

まず、あるドメイン z においてユーザ i が評点行為を行なっている場合、ドメイン z におけるユーザ i を $u_i^{(z)}$ と表す．また、ユーザ i が他のユーザとソーシャルコネクションを持つ場合、ソーシャルコネクションを持つユーザ集合に属するユーザ i を $u_i^{(s)}$ と表す．ユーザは複数のドメインに属する事が可能であり、同時に他ユーザとソーシャルコネクションを持つ事も可能である．本稿では同一のユーザであるかどうかを右下の添字で示す．一方、同一のユーザであってもドメインが異なれば異なるユーザとして提案手法では計算処理をするため、右上の添字にドメイン ID を付与し、ドメインを跨る同一のユーザであっても異なるユーザとして識別できるようにする．同じく、ソーシャルコネクションを持つユーザ集合に属するユーザとしては、右上の添字に s を与え識別する．例えば、ドメイン 1 におけるユーザ u_1 は $u_1^{(1)}$ と表現され、ドメイン 2 におけるユーザ u_1 は $u_1^{(2)}$ と表現され、ソーシャルコネクションを持つユーザとしての u_1 は $u_1^{(s)}$ と表現される．

アイテムドメイン z における DSUG は、ユーザ集合 U_z におけるユーザをノードとし、 $u_i^{(z)}$ と $u_j^{(z)}$ の類似度を基にした重みを持つエッジをノード間に与えることで構築される．ユーザの類似度は、ドメイン内の評点データセットを用いて計算され、類似度計算法としては、前

章で述べたピアソン相関係数を用いても Cosine 類似度を用いても、4.3 節で述べるタクソノミに基づく手法やその他の類似度計算法を用いても良い．そして、ドメイン z の DSUG に対応する列毎に正規化された隣接行列 (column normalized adjacency matrix) $\mathbf{A}_z$ を準備する．ここで添字 z は対応するドメインの ID を示す．列内の値は、値の総和が 1 になるように類似度を線形変換する事で与えられる．一方、SUG についても同様にユーザをノードとし、ユーザノード $u_i^{(s)}$ から $u_j^{(s)}$ へ、ユーザ u_i が u_j にアクセスした回数に基づく重みを持つエッジを引く事で構築する．SUG における列毎に正規化された隣接行列 $\mathbf{A}_s$ は、DSUG と同様にして構築する．

次に提案手法は、複数ドメインのアイテムを評価しているユーザや、ドメインを跨ってソーシャルコネクションを持つユーザを介して DSUG を結合し CDUG を構築する．以下、式 (3) で用いられる CDUG に対する列毎に正規化された隣接行列 $\mathbf{A}_c$ を構築する際の考え方を説明する．

初めに walk が、CDUG 上でドメイン z におけるユーザノード $u_i^{(z)}$ を訪問後に、そのユーザ u_i の SUG に対応するユーザノード $u_i^{(s)}$ の隣接ノードを訪問する確率を決定するパラメータ β を設定する．このパラメータはユーザにアイテムを推薦する際に、ソーシャルコネクションをどれだけ重視するかを決定する．すなわち、もしユーザ u_i がソーシャルコネクションを他のユーザと持つユーザであるとする、ユーザノード $u_i^{(z)}$ を訪問後に、ユーザ u_i とソーシャルコネクションを持つユーザノードに遷移する確率は β となる．そうすると、ユーザノード $u_i^{(z)}$ を訪問後に、ユーザ u_i とソーシャルコネクションを持つ他のユーザ u_j の SUG に対応するユーザノード $u_j^{(s)}$ を訪問する確率 $A_c(u_i^{(z)}, u_j^{(s)})$ は、以下の式で計算される．

$$A_c(u_i^{(z)}, u_j^{(s)}) = \beta \cdot A_s(i, j) \quad (4)$$

なお、もしユーザ u_i がソーシャルコネクションを他のユーザと持っていないとすると、 $A_s(i, j) = 0$ となるため、ユーザノード $u_i^{(z)}$ を訪問後にユーザノード $u_j^{(s)}$ を訪問する確率は 0 となる．

次に walk が、ドメイン z におけるユーザノード $u_i^{(z)}$ を訪問後に、 $u_i^{(z)}$ を介して、自ドメイン z または、他のドメインのユーザ u_i に対応するノードの隣接ノードを訪問する確率の計算方法を考える．ここでユーザ u_i が K 個の異なるドメインに属するアイテムを評価しておりかつ、他のユーザとソーシャルコネクションを持つとする．各ドメインが等しい確率で訪問されるとすると*5、次に

*5 議論の単純化のために、等確率でドメインを跨ると仮定した．なお、本研究の目的は、“あるドメインに興味を持つユーザに別ドメインの興味を持ちうるアイテムを推薦する”事である．本仮定は、あるユーザが複数ドメインに興味を持っている場合、例えばその中の特定ドメインに興味を傾いていたとしても別ドメインに属するユーザノードにも walk が遷移する機会を増やす効果がある．結果として、被推薦ユーザに別ドメインのアイテムも十分推薦可能にするため、有用と考える．

図1 CDUG に対する列毎に正規化された隣接行列 (column normalized adjacency matrix) の生成方法

walk が K 個あるドメインの内の 1 つに訪問する確率は、 $(1-\beta)/K$ となる．そうすると、ユーザノード $u_i^{(z)}$ を訪問後に、他のユーザ u_j のあるドメイン d に対応するユーザノード $u_j^{(d)}$ を訪問する確率 $A_c(u_i^{(z)}, u_j^{(d)})$ は、以下の式で計算される．

$$A_c(u_i^{(z)}, u_j^{(d)}) = \frac{(1-\beta)}{K} \cdot A_d(i, j) \quad (5)$$

もしユーザ u_i が、どのドメインのどのユーザともソーシャルコネクションを持たないとすると、次に walk が K 個あるドメインの内の 1 つに訪問する確率は、 $1/K$ となる．そうすると、ユーザノード $u_i^{(z)}$ を訪問後に、他のユーザ u_j のあるドメイン d に対応するユーザノード $u_j^{(d)}$ を訪問する確率 $A_c(u_i^{(z)}, u_j^{(d)})$ は、以下の式で計算される．

$$A_c(u_i^{(z)}, u_j^{(d)}) = \frac{1}{K} \cdot A_d(i, j) \quad (6)$$

なお、オープン ID^{*6}の普及に伴い、facebook^{*7}や google+^{*8}などのソーシャルネットワークサービスへのログインに用いる ID と同じ ID を用い、他のサイトへログインして様々なドメインのアイテムを購入する事ができるようになってきている．こうした流れを考慮すると、今後、本研究のようにソーシャルコネクション履歴を活用したクロスドメイン推薦は益々重要になると考えられる．一方、例えばソーシャルコネクションの履歴が入手出来ない場合においても、CDUG は構築できる．すなわち、パラメータ β を 0 に設定することで、SUG を組み込まない CDUG を構築できる．SUG を組み込む場合と組み込まない場合の結果については、検証実験において評価を行う．

2 つの小規模な DSUG と SUG における列毎に正規化された隣接行列を用いて CDUG を構築した例を図 1 に示す．この例では、 β を $1/3$ に設定している．例えば、walk がノード $u_1^{(1)}$ を通過後ノード $u_3^{(1)}$ を訪れる確率は、 $\frac{(1-1/3)}{2} \times 0.7 = 0.233$ として計算される．同じく、ノード $u_3^{(2)}$ を訪れる確率は、 $\frac{(1-1/3)}{2} \times 0.5 = 0.166$ として計算され、ノード $u_3^{(s)}$ を訪れる確率は、 $\frac{1}{3} \times 0.3 = 0.10$ として計算される．

ド $u_3^{(2)}$ を訪れる確率は、 $\frac{(1-1/3)}{2} \times 0.5 = 0.166$ として計算され、ノード $u_3^{(s)}$ を訪れる確率は、 $\frac{1}{3} \times 0.3 = 0.10$ として計算される．

4.2 他ドメインにおけるアイテムの検出

次に CDUG 上で RWR を実行する．グラフ上で RWR を実施することで、被推薦ユーザに対応するノード a から CDUG 上のノード u_i を walk が通過する確率を、式 (3) で示される RWR を収束するまで実行することで獲得できる．最終的に、CDUG 上での被推薦ユーザに対応するノード a と他ユーザに対応するノードとの間の関連性を計算できる．なお、式 (3) において α を小さく設定すると、被推薦ユーザ a が所属していない DSUG に所属するユーザを walk がより多く通過するようになる．後に述べる評価実験において、このパラメータを変化させた場合の効果の説明する．

被推薦ユーザ a と a が所属していないあるドメイン d に対する DSUG 上のユーザとの間の関連性を計算し、 a と他ユーザとの関連性を類似度の変わりに式 (2) に代入することにより、 a に対しドメイン d におけるアイテムの予測値を計算できる．

なおユーザ間で共有するアイテムが少ない状況では、共有アイテムの一致性のみに基づく Pearson 相関係数や Cosine 類似度に基づくユーザ間の類似度計算法では、類似ユーザが少なくなったり類似度の妥当性が落ちる．これは、協調フィルタリングにおけるスパースリティ問題として広く知られている [Sarwar 00]．そうした状況でユーザ間の類似度を計算し構築されたユーザグラフを用いて RWR を実施しても、グラフ上の walk は多くの適切なユーザノードを通過できなくなる．またこうした状況は、ドメインを跨って walk がグラフ上のノードを通過する際に多く観測される．例えば、多くの邦楽ドメインのアイテムを消費しているが、洋楽ドメインのアイテムは少ししか消費していないような、ドメイン間でのアイテム消費の割合が歪なユーザが多く存在するとする．こうした場合、邦楽ドメインに対して構築された DSUG から洋

*6 <http://ja.wikipedia.org/wiki/OpenID>

*7 <http://www.facebook.com/>

*8 <https://plus.google.com/>

図 2 類似度計算の例

楽ドメインに対して構築された DSUG へ、邦楽・洋楽を両方持つユーザを経由して walk が移動しようとしても、洋楽ドメインの DSUG に入ってからその先、洋楽ドメインの他の適切なユーザノードに移動しにくくなる。そこで本稿では、次節で述べるユーザが評価するアイテム数が少なくても頑健に類似度を計算できるタクソノミベースのユーザ類似度計算法 [Nakatsuji 10a] を適用する。

なお複数ドメインで構築される DSUG を組み合わせて CDUG を構築すると、RWR を実行するグラフが巨大になり推薦の計算時間がかかる場合がある。計算量については、RWR では、グラフのエッジ数 m と式 (3) の繰り返し回数 t を用い $O(mt)$ となる。それに比べ、例えば推薦エンジンのコンテストで有名な Netflix Prize [Bennett 07] において多くの推薦システムで用いられた Singular Value Decomposition (SVD) [Golub 70] ではグラフのノード数 n を用い $O(n^3)$ となる。基本的には異なる変数に依存するため単純比較はできないが、著者らの経験上 RWR の方が高速である。なお、RWR に関しては RWR の高速化手法 [Fujiwara 12] を用いると高速化が $O(m+n)$ まで可能であり明らかに SVD よりも有利となる。

4.3 タクソノミに基づくユーザ間の類似度計算

次に、ユーザ a と u の類似度計算法 [Nakatsuji 10a] を説明する。まずアプローチを説明し、次にアルゴリズムを説明する。

§1 アプローチ

手法のアプローチを説明する。

- まずユーザ a とユーザ u のクラス C_i に対する興味の一致度 $S(a, u, C_i)$ を計算する。このスコアは、 r_{a, C_i} と r_{u, C_i} の内、小さい方の値を基に計算する。これにより興味の一致度の計算の際に、 C_i に対する興味が a より低いユーザのスコアを下げるができる。
- 次に a と u の間での、クラスに対する評価行為の類似度 $S_C(a, u)$ を $S(a, u, C_i)$ を用い計算する。その際 Jaccard coefficient の考え方を用いる。この Jaccard coefficient アプローチは”ユーザ a と u のどちらかのみが評価を与えているクラス集合”の内、”両方

が評価を与えているクラス集合”がどれだけの割合を占めているかを計算できる。これは言い換えると、タクソノミの各クラス毎 (例では、 C_i) での、ユーザの興味の”幅”の類似度を考慮しているといえる。これにより被推薦ユーザ a がクラス C_i 配下においてそれほど多くのクラスに評価を割り当てていない場合、クラス C_i 配下の多くのクラスに評価を割り当てているユーザ u を、 a から分離することができる。タクソノミ上の各クラスに対し上記処理を実施することで、タクソノミにおける各階層における各クラスでの興味の”幅”を考慮した詳細な興味の類似度計算を実施できる。一方 [Ziegler 04] らの手法におけるタクソノミのクラス集合への評価ベクトルの Pearson 相関係数を用いる方法では、各階層毎での興味の”幅”を考慮していなく、タクソノミの持つ階層構造を十分活かしていない。

- 次に、 a と u の間でのアイテムに対する評価行為の類似度 $S_I(a, u)$ を計算する。この場合は、我々は式 (1) で説明したアイテムに対する各ユーザの評価スキームの違いを考慮し計算を行う Pearson 相関係数に基づくアプローチを採用する。
- 最後に、上述した 2 つの類似度 $S_C(a, u)$ と $S_I(a, u)$ を組み合わせ、ユーザ間のクラス、アイテムに対する評価行為の類似度を計算する。

我々の実験では、タクソノミベースの手法としては本稿で説明した手法 [Nakatsuji 10a] と [Ziegler 04] らの手法を用いる。また、Cosine ベースアプローチや Pearson 相関係数に基づくアプローチとも比較し、CDUG を構築するに適切な類似度計算法について議論する (5.4.2 節参照)。

§2 アルゴリズム

[Nakatsuji 10a] の手法のアルゴリズムを以下に説明する。本アルゴリズムにおいては、 $C_i(a)$ をユーザ a が評価したクラス C_i の子クラス集合として定義する。

- まずユーザ a と u のクラス C_i に対する興味の一致度 $S(a, u, C_i)$ を、 $\min(r_{a, C_i}, r_{u, C_i})$ とする。
- 次に、クラスに対するユーザ間の評価行為の類似度 $S_C(a, u)$ を以下の式により計算する。

$$S_C(a, u) = \sum_C \frac{\sum_{C_j \in \{C_i(a) \cap C_i(u)\}} S(a, u, C_j)}{|C_i(a) \cup C_i(u)|}$$

- また、アイテムに対するユーザ間の評価行為の類似度 $S_I(a, u)$ を、Pearson 相関係数に基づくアプローチにより式 (1) に沿って計算する。
- そして、 $S_C(a, u)$ と $S_I(a, u)$ を全ユーザ集合を母集合としガウス分布に沿って標準化した上で、ユーザ間の類似度 $S(a, u)$ を $S_C(a, u) + S_I(a, u)$ として計算する。

§3 例

図 2 に示す例を用いアルゴリズムを説明する。

- (1) $S(a, u, C_2)$ は, $\min(6.0, 2.0) = 2.0$ となる. また, $S(a, u, C_3)$ は, $\min(0.0, 8.5) = 0.0$ となる.
- (2) $S_C(a, u)$ は, $(2.0/2) + (0.0/2) = 1.0$ となる.
- (3) $S_I(a, u)$ は, 式 (1) において, 両方のユーザが評価を与えているアイテムは I_2 のみであるため, $\frac{(4-3)(2-3.5)}{\sqrt{(4-3)^2} \sqrt{(2-3.5)^2}} = -0.667$ として計算される.
- (4) 最後にアルゴリズム (4) の手順で標準化をした上で, $S(a, u)$ は計算される.

5. 評価実験

ブログサイト上での洋楽ドメインと邦楽ドメインの2つのドメインへの書き込み頻度から獲得されたユーザの暗黙的な興味プロファイルと, 同一ブログサイト上でのソーシャルコネクション履歴を用い評価実験を行った. 評価としては, 洋楽ドメインと邦楽ドメインを混ぜた混合データセットより構築されたドメイン特化ユーザグラフ (Domain-specific-user graph; DSUG) 上で RWR を実施した場合と, 提案手法によるクロスドメインユーザグラフ (Cross-domain-user graph; CDUG) 上で RWR を実施した場合との精度を比較した. また, DSUG を構築する際のユーザ間類似度計測法として複数の手法を比較し精度への影響を議論した. 1 章でも述べたが, 洋楽と邦楽は一般的にアイテムのタクソノミの上位で分岐するため, 異なるドメインに属すると考える事ができる.*9

5.1 データセット

以下に述べるデータセットを用い, 提案手法の評価を実施した*10. ブログサイトなどのソーシャルネットワークサービス上での書き込みやソーシャルコネクション履歴を用いることで, 本研究の提案であるクロスドメイン推薦の検証を実施できる.

§1 洋楽に対する評価データセット

本データセットは, 3,545 ユーザからなる 48,695 の評価データセットからなる. 本データセットは, [Nakatsuji 10a] らの実験で用いたデータセットと同じであり, ブログポータル Doblog*11 における洋楽に対するユーザのブログ上での書き込み頻度を基にアーティストへの評価を与えている. また, 洋楽アーティストに対するタクソノミは, ListenJapan が提供するものであり, 279 のジャンルをクラスとし, 21,214 のアーティストをアイテムとする*12. ア

アイテムに割り当てられた評価の平均数は 6.3 である. 推薦に関する研究で良く用いられているデータセットである MovieLens のデータセット*13 に倣い, ユーザ毎に, 最も頻度の高いアイテムの評価を 5 とし, 最も頻度の低いアイテムの評価を 1 とし, 各ユーザの評価データセットを作成した. タクソノミにおけるクラス階層は深い. “Music” をルートとし, 平均的に 4 階層からなり, 時に 5 階層あり, 末端クラスは “Space rock” や “Acid jazz” といった粒度の細かいものが設定されている. そのためユーザの類似度を正確に計算できる.

§2 邦楽に対する評価データセット

さらに, 2,800 ユーザからなる 58,104 の評価からなる邦楽データセットも用いる. 本データセットも洋楽と同様に作られたものであり ListenJapan の 153 のジャンルをクラスとし 7,421 のアーティストを分類する邦楽タクソノミを用いている. このタクソノミにおけるクラス階層は, 洋楽のものと同様に深い. アイテムに割り当てられる評価の平均数は, 10.8 である.

§3 洋楽・邦楽の混合データセット

洋楽と邦楽のタクソノミを組合せた混合タクソノミのアイテムに対する 4,825 ユーザによる 106,799 の評価からなるデータセットも利用した. 本データセットは, 上述した洋楽データセットと邦楽データセットをマージすることにより得られたものである. 混合タクソノミは 432 のジャンルをクラスとして含み, 28,635 のアーティストをアイテムとして含む. アイテムに割り当てられた評価の平均数は 7.4 である. このデータセットは混合データセットと名付けているが, 音楽というシングルドメインのデータセットと考えることもでき, 音楽ドメインに対する DSUG を作るために使われる. 本データセットは, 洋楽ドメインと邦楽ドメインそれぞれのドメインの DSUG を結合して出来る CDUG 上での RWR に基づく予測精度と, 混合ドメイン (音楽ドメイン) に対し構築された DSUG 上での RWR に基づく予測精度を比較するために用いるものである.

§4 ユーザのソーシャルコネクション履歴

また, 2006 年 4 月から 2008 年 4 月における Doblog におけるブログユーザのソーシャルコネクション履歴を分析した. ソーシャルコネクション履歴はユーザが他ユーザのブログサイトにアクセスすると, Doblog 上のデータベースに記録される. ソーシャルコネクション履歴の分析に用いられたユーザは, 洋楽と邦楽のアーティストのデータセットに含まれるユーザのみを対象としている. 結果として, 3,432 ユーザの 199,3716 のソーシャルコネクションを得た. 本履歴を用いて, ユーザをノードとし, ユーザ u のユーザ v のブログサイトへのアクセス頻度を u から v のエッジ上の重みとして持つ隣接行列を計算する.

詳細に分析結果を述べると, 洋楽データセットに含まれるユーザのうち他ユーザとの間にソーシャルコネクショ

*9 本検証で, まず洋楽と邦楽という比較のお互いに関連性の強いドメインのデータを用いて進めた理由は, Shlomo らの研究 [Berkovsky 07b] でも述べられている通り, 音楽と映画などお互いにより関連性の薄いドメインにおけるユーザ共通の実験データが入手しにくいためでもある.

*10 以下の各データセットは匿名のユーザ ID に変換されたものを用いている.

*11 <http://www.doblog.com/>; Doblog は 2009 年 5 月にサービスを終了している.

*12 音楽のタクソノミは ListenJapan のホームページよりアクセスできる.

*13 <http://www.grouplens.org/node/73>

ンも持つユーザの割合は 53% であり、邦楽データセットに含まれるユーザのうち他ユーザとソーシャルコネクションを持つユーザの割合は 52% であった。また、邦楽（洋楽）のみに興味を持つユーザが洋楽（邦楽）のみに興味を持つユーザとソーシャルコネクションを持つケースを調査したところ、全ソーシャルコネクションにおけるユーザペアの内 21% がそうしたケースである事が分かった。

5.2 検証基準

まずデータセット D を、予測元データセット T と予測先データセット P の 2 つに、 T と D の割合が 0.7 になるという条件の下で、ランダムに分割をした。実験ではこのようにして、 D から予測先データセットが重ならないようにして 2 パターンの検証データを生成した。そして、ユーザ毎にデータセット P においてユーザのアイテムへの評点データが存在する一方、そのアイテムの所属ドメインのアイテムはデータセット T 内のそのユーザの評点データには存在しないという条件を満たすユーザを集めた。そして集まったユーザ集合を対象として精度の検証を実施した。なお、DSUG や CDUG の構築においては、 T に存在する全てのユーザとその評点データを用いユーザ間の類似度を計算し、DSUG や CDUG を構築し、グラフ上でのユーザ間の関連性を計算した。

協調フィルタリングで標準的に用いられる検証基準に従い、データセット P に存在する評価値と手法が予測する予測値の誤差である、平均絶対誤差 (Mean Absolute Error; MAE) を用い評価をした。 n をデータセット P における評価値の数とし、 P_i と R_i をデータセット P における i 番目のアイテムに対する予測値および実際の評価値とすると、MAE は以下の式で表される。

$$MAE = \frac{\sum_{i=1}^n |P_i - R_i|}{n} \quad (7)$$

MAE に加えて、手法により予測されたアイテムに対する P に含まれるアイテムの割合であるカバレッジについても確認をした [Herlocker 04]。

$$\text{カバレッジ} = \frac{|\text{手法により予測されたアイテム}|}{|P|}$$

検証においては前述したとおり、ユーザ毎にデータセット P においてユーザのアイテムへの評点データが存在する一方、そのアイテムの所属ドメインのアイテムはデータセット T 内のそのユーザの評点データには存在しないという条件を満たすユーザ集合を対象として MAE とカバレッジを計算した。我々の 2 パターンの検証データでは $\frac{T}{D}$ を 0.7 とした際、そうしたユーザ集合はパターン毎に 126 人と 142 人存在する。

5.3 検証手順

本研究では、以下の 2 つの観点から検証を実施した。

(1) 複数のドメインを混ぜて DSUG を構築しその上で RWR を実施する場合と、ドメイン毎に DSUG を構築した上で提案法で CDUG を構築しその上で RWR を実施する場合とでの精度比較 (5.4.1 節に結果を示す)

(2) DSUG を構築する際の類似度計測法を複数試した上での精度比較 (5.4.2 節に結果を示す)

一つ目の精度比較として、我々は、洋楽データセットと邦楽データセットを用い構築された CDUG (以下、 $(CDUG(\text{邦}+\text{洋}))$ と記す) 上で RWR を実施した際の精度と、洋楽データセットと邦楽データセットと SUG を用い構築された CDUG (以下、 $(CDUG(\text{邦}+\text{洋}+s))$ と記す)、および音楽ドメインにおける混合データセットを用いて構築された DSUG (以下、 $(mixed)$ と記す) 上で RWR を実施した際の精度を比較した。

また、二つ目の精度比較として、我々は、DSUG を構築する際の類似度計算法として以下の複数の方法を用い、精度を比較した。

- Pearson 相関係数 (Pearson): ユーザ間の類似度は Pearson 相関係数で計算される。
- Cosine 類似度 (Cosine): ユーザ間の類似度は Cosine 類似度で計算される。
- Ziegler らの手法 (Ziegler): ユーザの類似度は Ziegler らにより提案された手法により計算される [Ziegler 04]。この手法は、2 章で説明した手法である。文献 [Ziegler 04] で用いられているパラメータ χ は 0.2 に設定した際に精度が最も良くなることを確認したためこの値を設定した。
- Nakatsuji らの手法 ($T(J\&P)$): 4.3 節で説明した類似度計算法である。

以下、それぞれの検証の結果について述べる。

5.4 検証結果

まず最初に、RWR を実施した場合の全メソッドのカバレッジが 80% に達し、かつ MAE が最も低くなるように設定するため、式 (2) におけるユーザ数 N を 100 から 1000 の間で振らせた。そうすることで、ある程度カバレッジがある場合における予測値の精度を分析できる。結果として両パターンの検証データにおいて N を 1000 に設定した。なお RWR を実施しない場合の各メソッドにおけるユーザ数 N は、カバレッジが 80% 以上になる事が無かったため N を 100 から 1000 の間で振らせた中で最も精度が良い場合の値を選択するようにした。

さらに、2 つの異なるアイテムドメインにアクセスしたユーザ数を調査し、4,825 ユーザの中で 1,520 ユーザ存在することを確認した。この人数から見て、邦楽と洋楽は同じ音楽であるため、ある程度相関が取れたドメインであると言える。

$\frac{T}{D}$ を 0.7 とした時の結果を表 1 に示す。この表において各テーブルの “Non” というのは、RWR を実行してい

表 1 $\frac{T}{D} = 0.7$ における MAE .

α	Pearson (mixed)	Cosine (mixed)	Ziegler (mixed)	T(J&P) (mixed)	Pearson (CDUG)	Cosine (CDUG)	Ziegler (CDUG)	T(J&P) (CDUG)	Pearson (social)	Cosine (social)	Ziegler (social)	T(J&P) (social)
Non	1.13	1.30	1.19	1.17	-	-	-	-	-	-	-	-
0.8	1.24	1.18	1.15	1.17	1.20	1.14	1.10	1.09	1.22	1.12	1.09	1.08
0.6	1.26	1.19	1.18	1.19	1.21	1.12	1.09	1.08	1.22	1.12	1.09	1.07

ない場合の結果を示す．また，“-”というのは，CDUG を用いた手法の場合，RWR を実行しないと他ドメインのアイテムを推薦不可能であるため，精度が測定不可能であるという事示す．なお，本結果は 2 パターンの検証データにおける結果の平均を示している．表 1 を用い以下で述べる 2 つの観点からの検証結果をそれぞれ説明する．

§ 1 CDUG 上での RWR による推薦の効果

表 1 に示すように全ての手法において，邦楽データセットと洋楽データセット毎に作成された DSUG を用い構築した CDUG 上で RWR を実施した場合（表 1 において (CDUG(邦+洋)) と標記し，かつ α の値が 0.8 または 0.6 の場合）の結果の方が，同じ手法で混合データセット上で RWR を実施した場合（表 1 において (mixed) と標記し，かつ α の値が 0.8 または 0.6 の場合）の結果に比べて精度が改善されることが分かる．特に，Cosine の $\alpha = 0.6$ ，Ziegler，T(J&P) については，ユーザを母集合とした T 検定において有意差 ($p < 0.05$) がある事も確認できた．以上の結果はユーザの類似度は単一ドメインのアイテムの評価データセットを用いた方が，混合ドメインのアイテムの評価データセットを用いた場合よりも適切に計算できることを示している．この結果から，他ドメインにおいて高く関連するユーザを CDUG 上の RWR により分析する提案手法の方が，複数ドメインのデータが混合された評価データセットを用い計算されたユーザの類似度に基づく DSUG 上での RWR による推薦手法よりも，良いアプローチであることが分かる．

なお，Pearson は (mixed) 上で RWR を実施しない場合 (α の値が Non の場合) の結果の方が良い．そこでカバレッジに着目したところ RWR を実施しない場合の Pearson はカバレッジが 50% を僅かに超える程度であることが分かった．これは Pearson は同じアイテムを共有しているユーザでないと類似ユーザと見做さないためユーザ数 N を大きく設定しても限界があること，および Pearson ではユーザの評価した全アイテムへの評価が同じ値である場合に他ユーザとの類似度を測れないことが原因である．結果として，カバレッジに限界があること，および CDUG 上で RWR を実施した場合の Pearson の結果が他の手法よりも悪いことから，既存の Pearson の性能には課題がある．

次に，邦楽データセットと洋楽データセット毎に作成された DSUG に加えて，ユーザ間のソーシャルコネクションを元に作成された SUG を結合し，CDUG を作成した場合の結果を議論する．ここで，4.1 節で説明したパラ

メータ β は最も精度が良くなる場合の値を用い，結果として 0.3 と設定した．この CDUG に対して計算された結果は，表 1 において (CDUG(邦+洋+s)) と示す．表は 2 パターンの検証データの結果の平均となっているためこの表からは明示的には確認できないが，両パターンの結果において，いくつかの手法により有意差は確認は出来ないが精度が向上することが確認された．よって，ユーザ間のソーシャルコネクションは被推薦ユーザがアクセスしたことの無いドメインのアイテムを精度良く検出するのに有効である可能性を持つことが分かった．

また，ソーシャルコネクションログを導入することにより，予測精度が向上したユーザを対象に，ソーシャルコネクションのデータの特徴を調査した．そして，“被推薦ユーザの予測先データセットに存在するアイテムを，被推薦ユーザとソーシャルコネクションを持つユーザや，SUG 上において間接的につながっているユーザが持っており，そうしたアイテムが精度良く予測できている”ことを確認した．このことから，ソーシャルコネクションを持つユーザ間は，ドメインに関わらず興味を持つアイテムが重なる傾向がある事が確認できた．ソーシャルコネクションを持つユーザの興味が重なる傾向がある事は文献 [Singla 08] の検証結果を裏付けている．

なお，このソーシャルコネクションの特徴を活かすことで，例えば，“あなたのお友達は，このようなアイテムにも興味を持っています．”という推薦根拠とともに，異なるドメインのアイテムを被推薦ユーザへ提示することもでき，ユーザの推薦への納得感の向上にも役立つ．このことは，精度の向上という点以外に，異なるドメインのアイテムに興味は明示的には無かったユーザに対しても新たな興味を喚起する可能性を持ち，有用である．

§ 2 タクソノミによる類似度計算法の効果

さらに，タクソノミに基づく 2 つの手法 (Ziegler と T(J&P)) と他の手法との違いを確認した．表 1 に結果を示す．これによると，(CDUG(邦+洋)) 上で RWR を実行した場合においても (CDUG(邦+洋+s)) 上で RWR を実行した場合においても，タクソノミを用いる手法の方が他の手法よりも高い精度を示すことを確認した．なお，表において太字になっている数値は，(mixed) データセットで最も精度が良かった Ziegler(mixed) よりも各手法が，ユーザを母集合とした T 検定において有意差 ($p < 0.05$) が認められたことを示している．

そこで高精度の原因を分析するため，2 つのドメイン両方を評価したユーザのドメイン間でのアイテム評価数の差を分析した．結果として，平均的に，あるドメイン

に対し評価をしたアイテム評価数は、もう一つのドメインに対し評価をしたアイテム評価数の約 4 倍にのぼることが分かった。言い換えると、ユーザはあるドメインのアイテムは多く評価するが、他ドメインのアイテムはそれほど評価しない傾向がある。Cosine と Pearson では、そうしたドメイン内でのユーザ類似度の結果がスパースティ問題により不適切になる。それに対し、タクソノミに基づく手法はユーザのアイテム評価数が少ない状況においても頑健に動くため、被推薦ユーザのノードから被推薦ユーザのノードが含まれないドメインのノードへと適切な経路を多く引くことができる。よって、他の手法に比べ、ユーザがアクセスをしたことが無いドメインのアイテムを検出するのに適切であり精度が良い。

これらの実験結果から、サービスドメインとして異なるが興味を持つユーザ層がある程度重なる場合に、2 つのドメインを混ぜて従来の協調フィルタリングを行うよりも、提案手法 (Ziegler(CDUG(邦+洋)) や $T(J\&P)$ (CDUG(邦+洋))) が有効である事を示している。そのことから、同様にドメインが異なっている程度ユーザ層が重なるようなデータセットであれば提案手法は有効と考えられる。さらに、ソーシャルネットワークのログを用いた場合、提案手法 (Ziegler(CDUG(邦+洋+s)) や $T(J\&P)$ (CDUG(邦+洋+s))) は、さらに精度を改善できる可能性がある事を示している。

6. 結 論

本論文は、ユーザが過去にアクセスしたことの無いドメインではあるが、興味を持ちうるアイテムを精度良く推薦可能とする手法を提案した。本手法では、まずタクソノミベースのユーザ類似度計測法を用いユーザ間の類似度を計算し、ドメイン毎にユーザをノードとしユーザ間の類似度を重みとするエッジを持つドメイン特化ユーザグラフ (Domain-specific-user graph; DSUG) を作成する。また、ユーザ間のソーシャルコネクションの頻度を基にした、ソーシャルベースのユーザグラフ (Social-based user graph; SUG) も作成する。そして DSUG と SUG を様々なドメインのアイテムを評価するユーザや、様々なドメインのユーザとソーシャルコネクションを持つユーザを介して結合し、クロスドメインユーザグラフ (Cross-domain-user graph; CDUG) を作成する。次に、CDUG 上で被推薦ユーザに対応するノードを始点とした RWR を実施し、被推薦ユーザに対応するノードを含まないドメインに対応する DSUG 上のユーザノードと、被推薦ユーザノードとの関連性を計算する。2 つの異なるドメインのアイテムに対するブログ上での書込み頻度に基づく評価データセットとブログ上でのソーシャルコネクション履歴を用いた評価実験により、提案手法は各ドメインの評価データを単にマージすることで作成された DSUG

上での RWR に基づく手法よりも精度が良いことを確認した。さらに、DSUG を構築する際のユーザ間の類似度計算法として、タクソノミベースの類似度計算法が他の類似度計算法よりも良い性能を与えることも確認した。

今後、関連性の高いドメイン間のみではなく、アパレルドメインと音楽ドメインのように、お互いの関連性がより低いドメインをも交えた 2 つ以上の異なるドメイン間に対して提案手法の適用性を検証する。提案手法は、ドメイン毎のユーザの類似度や、ユーザのソーシャルコネクションといったドメイン毎に独立した要素のみを用いるため、そうした環境においても適用可能であると考えられる。

◇ 参 考 文 献 ◇

- [Bennett 07] Bennett, J., Elkan, C., Liu, B., Smyth, P., and Tikk, D.: KDD Cup and workshop 2007, *SIGKDD Explor. Newsl.*, Vol. 9, No. 2, pp. 51–52 (2007)
- [Berkovsky 07a] Berkovsky, S., Kuflik, T., and Ricci, F.: Cross-Domain Mediation in Collaborative Filtering, in *Proc. User Modeling'07*, Vol. 4511, pp. 355–359 (2007)
- [Berkovsky 07b] Berkovsky, S., Kuflik, T., and Ricci, F.: Distributed collaborative filtering with domain specialization, in *Proc. RecSys'07*, pp. 33–40 (2007)
- [Breese 98] Breese, J. S., Heckerman, D., and Kadie, C.: Empirical Analysis of Predictive Algorithms for Collaborative Filtering, in *Proc. UAI'98*, pp. 43–52 (1998)
- [Chen 08] Chen, K., Lu, R., Wong, C. K., Sun, G., Heck, L., and Tseng, B.: Trada: tree based ranking function adaptation, in *Proc. CIKM'08*, pp. 1143–1152 (2008)
- [Coyle 08] Coyle, M. and Smyth, B.: Web Search shared: Social Aspects of a Collaborative, Community-Based Search Network, Springer (2008)
- [Fujiwara 12] Fujiwara, Y., Nakatsuji, M., Onizuka, M., and Kitsuregawa, M.: Fast and Exact Top-k Search for Random Walk with Restart, *PVLDB*, Vol. 5, No. 5, pp. 442–453 (2012)
- [Fukazawa 11] Fukazawa, Y. and Ota, J.: User-centered profile representation for recommendations across multiple content domains, *KES Journal*, Vol. 15, No. 1, pp. 1–14 (2011)
- [Golub 70] Golub, G. and Reinsch, C. E.: Singular value decomposition and least squares solutions, *Numerische Mathematik*, Vol. 14, No. 5, pp. 403–420 (1970)
- [Gori 07] Gori, M. and Pucci, A.: ItemRank: A Random-Walk Based Scoring Algorithm for Recommender Engines, in *Proc. IJCAI'07*, pp. 2766–2771 (2007)
- [Herlocker 04] Herlocker, J. L., Konstan, J. A., Terveen, L. G., and Riedl, J. T.: Evaluating collaborative filtering recommender systems, *ACM Trans. Inf. Syst.*, Vol. 22, No. 1, pp. 5–53 (2004)
- [Huang 04] Huang, Z., Chen, H., and Zeng, D.: Applying associative retrieval techniques to alleviate the sparsity problem in collaborative filtering, *ACM Trans. Inf. Syst.*, Vol. 22, No. 1, pp. 116–142 (2004)
- [Konstas 09] Konstas, I., Stathopoulos, V., and Jose, J. M.: On social networks and collaborative recommendation, in *Proc. SIGIR'09*, pp. 195–202 (2009)
- [Li 09a] Li, B., Yang, Q., and Xue, X.: Can Movies and Books Collaborate? Cross-Domain Collaborative Filtering for Sparsity Reduction, in *Proc. IJCAI'09*, pp. 2052–2057 (2009)
- [Li 09b] Li, B., Yang, Q., and Xue, X.: Transfer learning for collaborative filtering via a rating-matrix generative model, in *Proc. ICML'09*, pp. 617–624 (2009)
- [Li 11] Li, B., Zhu, X., Li, R., Zhang, C., Xue, X., and Wu, X.: Cross-Domain Collaborative Filtering over Time, in *Proc. IJCAI'11*, pp. 2293–2298 (2011)
- [Lovasz 96] Lovasz, L.: *Random Walks on Graphs: A Survey*, Vol. 2, Bolyai Society Mathematical Studies, combinatorics edition (1996)

- [Nakatsuji 10a] Nakatsuji, M., Fujiwara, Y., Tanaka, A., Uchiyama, T., Fujimura, K., and Ishida, T.: Classical music for rock fans?: novel recommendations for expanding user interests, in *Proc. CIKM'10*, pp. 949–958 (2010)
- [Nakatsuji 10b] Nakatsuji, M., Fujiwara, Y., Tanaka, A., Uchiyama, T., and Ishida, T.: Recommendations Over Domain Specific User Graphs, in *Proc. ECAI'10* (2010)
- [Nakatsuji 11] Nakatsuji, M., Fujiwara, Y., Uchiyama, T., and Fujimura, K.: User Similarity from Linked Taxonomies: Subjective Assessments of Items, in *Proc. IJCAI'11*, pp. 2305–2311 (2011)
- [Nakatsuji 12] Nakatsuji, M., Fujiwara, Y., Uchiyama, T., and Toda, H.: Collaborative Filtering by Analyzing Dynamic User Interests Modeled by Taxonomy, in *Proc. ISWC'12 (to appear)* (2012)
- [Resnick 94] Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P., and Riedl, J.: GroupLens: An Open Architecture for Collaborative Filtering of Netnews, in *Proc. CSCW'94*, pp. 175–186 (1994)
- [Sarwar 00] Sarwar, B. M., Karypis, G., Konstan, J. A., and Riedl, J.: Analysis of recommendation algorithms for e-commerce, in *Proc. EC'00*, pp. 158–167 (2000)
- [Schickel-Zuber 06] Schickel-Zuber, V. and Faltings, B.: Inferring user's preferences using ontologies, in *Proc. AAAI'06*, pp. 1413–1418 (2006)
- [Singla 08] Singla, P. and Richardson, M.: Yes, there is a correlation: - from social networks to personal behavior on the web, in *Proc. WWW'08*, pp. 655–664 (2008)
- [Wang 09] Wang, B., Tang, J., Fan, W., Chen, S., Yang, Z., and Liu, Y.: Heterogeneous cross domain ranking in latent space, in *Proc. CIKM'09*, pp. 987–996 (2009)
- [Yildirim 08] Yildirim, H. and Krishnamoorthy, M. S.: A random walk method for alleviating the sparsity problem in collaborative filtering, in *Proc. RecSys'08*, pp. 131–138 (2008)
- [Ziegler 04] Ziegler, C. N., Lausen, G., and Thieme, L. S.: Taxonomy-driven computation of product recommendations, in *Proc. CIKM'04*, pp. 406–415 (2004)
- [穂坂 10] 穂坂 祐輔, 小野 廣隆, 山下 雅史: Multiple Random Walk of the Cover Time について (アルゴリズムと計算機科学の数理的基盤とその応用), 数理解析研究所講義録, Vol. 1691, pp. 85–90 (2010)

〔担当委員: 松尾豊〕

2012 年 3 月 29 日 受理

著者紹介

中辻 真(正会員)

2001 年京都大学工学部数理工学科卒業。2003 年京都大学大学院情報学研究科システム科学専攻修士課程修了。2010 年京都大学大学院情報学研究科社会情報学専攻博士課程修了。情報学博士。2003 年日本電信電話株式会社入社。現在、日本電信電話株式会社サービスエボリューション研究所勤務。セマンティックスを背景知としたマイニングの研究に取り組む。2007 年人工知能学会研究会優秀賞。電子情報通信学会、日本データベース学会各会員。

藤原 靖宏

2001 年早稲田大学理工学部電気電子情報工学科卒業。2003 年同大学大学院理工学研究科修士課程修了。2003 年日本電信電話株式会社入社。2011 年東京大学大学院情報理工学系研究科電子情報学専攻博士課程修了。時系列データ処理およびグラフマイニングの研究開発に従事。電子情報通信学会平成 19 年度論文賞、情報処理学会平成 19 年度論文賞、KDD 2008 best paper award runner-up、第 27 回テレコムシステム技術賞など受賞。電子情報通信学会、情報処理学会、日本データベース学会各会員。

内山 俊郎(正会員)

1987 年東工大・工・電気電子卒。1989 年同大学大学院修士課程了。同年(株)NTT データ入社。1991-1993 年南カリフォルニア大客員研究員。1999-2005 年通信・放送機構研究員、特別研究員。2006 年より日本電信電話株式会社サービスエボリューション研究所勤務。Web データマイニング、分光色再現の研究に従事。博士(工学)。電子情報通信学会、情報処理学会、IEEE 各会員。