

ソーシャルメディア上の行為に基づく興味予測

Interest Prediction via Users' Actions on Social Media

則 のぞみ
Nozomi Nori

京都大学
Kyoto University
nozomi@ml.ist.i.kyoto-u.ac.jp

ボレガラ
ダヌシカ
Danushka Bollegala

リヴァプール大学
University of Liverpool
danushka.bollegala@liverpool.ac.uk

石塚 満
Mitsuru Ishizuka

早稲田大学
Waseda University
ishizuka@aoni.waseda.jp

keywords: social media, interest prediction, user modeling, graph, collaborative filtering

Summary

We propose a method to predict users' interests by exploiting their various actions in social media. Actions performed by users in social media such as Twitter and Facebook have a fundamental property: user action involves multiple entities - e.g. sharing URLs with friends, bookmarking and tagging web pages, clicking a favorite button on a friend's post etc. Consequently, it is appropriate to represent each user's action at some point in time as a higher-order relation. We propose *ActionGraph*, a novel graph representation to model users' higher-order actions. Each action performed by a user at some time point is represented by an *action node*. *ActionGraph* is a bipartite graph whose edges connect an action node to its involving entities, referred to as *object nodes*. Using real-world social media data, we empirically justify the proposed graph structure. We show that the prediction accuracy can be improved by adequately aggregating various actions. Moreover, our experimental results show that the proposed *ActionGraph* outperforms several baselines, including standard tensor analysis PARAFAC, a previously proposed state-of-the-art LDA-based method and other graph-based variants, in a user interest prediction task. Although a lot of research have been conducted to capture similarity between users or between users and resources by using graph, our paper indicates that an important factor for the prediction performance of the graph mining algorithm is the choice of the graph itself. In particular, our result indicates that in order to predict users activities, adding more specific information about users activities such as types of activities makes the graph mining algorithm more effective.

1. はじめに

1.1 ソーシャルメディア上での人々の様々な行為データの応用可能性

近年人々は情報を得たり共有したりする目的でソーシャルメディア上で他者と様々な関係を構築するようになっており、そのソーシャルネットワーク上で URL への言及やテキストの引用、再発信を行ったり、URL で特定されるような、画像やマイクロブログの投稿といったようなウェブ上の様々なリソースに対してタグ付けを行ったりと様々な行為を行うようになった。ユーザがハイパーリンクを用いてドキュメントを参照することでドキュメント同士が関連付けられていくことがウェブの大きな特徴であるが、現在のウェブが持っている面として、ソーシャルメディアの普及を背景として、より一般に人々の様々な行為によって様々なリソースが関連付けられるようになっていく点が挙げられる。

ユーザの行為は様々なリソースを関連付けるだけでなく、行為主体であるユーザの嗜好や興味を反映していると期待できるため、これらユーザの行為データを活用することでウェブの様々なアプリケーションを改善する取り組みがなされてきた。例えばユーザのタグ付けデータは個人化された検索 [Bao 07, Heymann 08, Xu 08] や人間関係の推論 [Schifanella 10]、新たなオントロジーの発見 [Mika 05] などのタスクに有効であることが指摘されている。

1.2 ユーザの興味予測とその課題

ユーザがあるリソースに対して何らかの行為を行ったとき、そのユーザがそのリソースに対して好意を抱いたのか疑問を抱いたのかなどは明らかではないが、少なくともそのリソースに何らかの興味を抱いたと捉えることができるだろう。したがって、本論文では、ユーザがどのようなリソースに対して行為を行うかを予測する問題

としてユーザの興味予測を定義する。

近年は、ソーシャルメディアの普及により、ユーザが様々なウェブサービスを用いて様々な行為を行うようになったことを背景に、ユーザの様々な行為を活用した上でユーザの興味を予測する研究が盛んに取り組みられている。例としては、ユーザがソーシャルネットワーク上で持つ他のユーザとの関係や様々なアイテムと持つ関係をもとにユーザの興味予測やアイテム推薦を行う研究 [Jiang 12, Ma 11, Yang 11] が挙げられる。これらの研究では、例えばユーザが言及するニュース記事を予測するのにユーザのソーシャルネットワークを用いるといったように、実際に予測する対象の行為と異なる種類の行為も予測に活用している。その背景には、異なる種類の行為であっても、対象（ユーザ）に対する相補的な性質を反映していると考えられるため予測に活用できるであろうという期待がある。

以上のように、ユーザの様々な行為を活用した上でユーザの興味を予測する研究は近年盛んに行われているものの、以下の二点の課題があると筆者らは考える。一点目は、(1) 様々なウェブサービス上の様々な行為データをユーザの興味予測に活用できるのか、それはどのような場合かといった問に答えるために定量的分析を行うことである。ユーザの興味予測に関する研究は数多くなされているが、ウェブサービスをまたぐような研究は少なく、特に、異なるウェブサービスをまたいだ異なる種類の行為を興味予測に活用できるのかといった点に答える定量的分析はほとんど行われていない。この背景には、従来は異なるウェブサービス間でユーザの同一性判定を行うことが難しく、異なるウェブサービスをまたいだユーザの興味予測という問題の応用可能性が限定されていたことなどがある。しかし、昨今では Twitter, Facebook, Foursquare など様々なサービス間で API 連携が行われており、異なるサービスを同時に利用するユーザも増え、ユーザの同一性判定が比較的行い易くなってきており、異なるウェブサービスをまたいだユーザの興味予測という問題の応用可能性が広がっていると考えられる。したがって、上記の問に答えるために定量的分析を行うことは重要な課題となる。

二点目は、(2) ユーザの行為データを多項関係データとして扱った上で、ユーザの興味を予測する必要があるということである。上記で言及した研究を含め、ユーザの様々な行為を活用した上でユーザの興味を予測する研究において一般的なアプローチは、ユーザの行為をユーザとアイテム間のような二項関係として捉えるものである。例えば、推薦システム等で普及している協調フィルタリング [Resnick 94, Shardanand 95] の基本的な枠組みは、あるユーザと似た興味を持つ他のユーザが気に入るアイテムをそのユーザに推薦することであるが、一般的にはユーザの興味はどのアイテムに（購入などの）行為を行ったかのみをもとに判断される。しかしながら、ユーザの

行為は、より一般には、ユーザ、リソース、キーワードなど複数のオブジェクトを巻き込む多項関係データとして捉えられるので、ユーザの様々な行為データを多項関係として活用した上でユーザの興味を予測する必要があるだろう。また、ウェブ上のユーザやリソースは大量にあるため、これらを巻き込む多項関係からユーザの興味を的確に予測するためにはデータ過疎に頑健な予測手法が必要となる。

1.3 アプローチ

一つ目の課題に対しては、*retweet* などの特定の種類の行為を予測するのに、各行為を特徴ベクトルで表現した際の特徴量を共有する別の種類（例えば *favorite*）の行為データを用いた場合にももとの種類の行為データを用いた場合と同程度の精度が実現できるのかそうでないかを評価した。ある種類の行為を予測するのに別の種類の行為データを用いても同程度に高い精度が実現できるならば、それら複数種類の行為データを組み合わせることで、ユーザの興味予測の性能向上が期待できるだろう。

二つ目の課題に対しては、まず、人々のウェブ上の様々な行為を多項関係として汎用的に表現した上で、多項関係の共起情報を欠損させることなく二項関係に変換することで、協調フィルタリングの基本的アイデアを今回の問題に沿うように拡張した。これを可能にするアクショングラフという、ユーザの様々な行為を多項関係として表現できる新しいグラフ表現を提案し、そのグラフを用いた予測手法を提案した。

1.4 本論文の貢献

以下で本論文の貢献について述べる。

(1) 様々なウェブサービス上の様々な行為データをユーザの興味予測に活用できるのか、それはどのような場合かといった問に答えるための定量的分析を行った。具体的には、ある種類の行為を予測するのに異なる種類の行為データを活用することで、ももとの種類の行為データを用いた場合と同程度の精度が実現できるのかそうでないかを評価する実験を実データを用いて行い、いくつかの知見を得た。(2) ユーザの様々な行為を多項関係として表現できる新しいグラフ表現としてアクショングラフを提案し、そのグラフを用いた予測手法を提案した。(3) 現実のソーシャルメディアのデータを用いた実験により、提案手法の有効性を既存の様々な手法と比較し評価した。特に、予測精度やデータ過疎への頑健性といった観点で評価した際に、提案手法はテンソルの標準的分解手法である PARAFAC や二項関係に対する LDA、構築しているグラフ構造のみ異なるが提案手法と同一のアルゴリズムを適用した様々なベースラインを上回る性能を示すことを確認した。(4) ユーザ間の類似度やユーザとリソースの間の類似度をグラフを用いて計算する手法は

これまで多々提案されてきたが、一連の評価実験により、グラフを用いたアルゴリズムの性能はそもそもどのようなグラフを構築するかに依存するという示唆を得た。また、ユーザの行動に關与する様々なエンティティの情報をグラフに付与することで、グラフマイニングのアルゴリズムを効果的に機能させることができるという示唆を得た。

1.5 本論文の構成

本論文の以降の構成について述べる。第二章では、様々なウェブサービス上の様々な行為データをユーザの興味予測に活用できるのか、それはどのような場合かといった問に答えるために行った実験について記述する。これは既存のロジスティック回帰を用いた分析である。続く第三章では、ユーザの興味予測のための手法を提案する。ユーザの様々な行為を多項関係として表現できる新しいグラフ表現としてアクショングラフを提案し、さらにアクショングラフを用いた予測手法について説明する。第四章では、ユーザの興味予測タスクにおいて提案のアクショングラフの性能を評価するために行った実験について記述する。第五章で関連研究について述べ、本研究の位置付けを明確にする。最後に第六章でまとめを行う。

2. ソーシャルメディア上のユーザの行為の分析

ここでは、様々なウェブサービス上の様々な行為データをユーザの興味予測に活用できるのか、それはどのような場合かといった問に答えるために、ある種類の行為を予測するのに、異なる種類の行為データを用いることができるかを評価する実験を行う。例えば、Delicious 上のタグ付け行為や Twitter 上の *retweet* 行為を使うことで Twitter における *favorite* 行為をどの程度正確に予測することができるだろうか。もし元々の *favorite* データを用いるのと同程度に良い精度で予測できるのであれば、これらの行為に反映されているユーザの嗜好には共通部分が存在すると考えられるので、これらの異なる行為データを用いることはユーザの興味予測に有効であると言えるだろう。実験を行った結論としては、異なる種類の行為データを活用することで、いくつかの場合において、もとの種類の行為のみを用いた場合と同程度の予測精度を実現できることが確認された。以下で実験条件と結果について述べる。

2.1 実験条件

データセット

Twitter^{*1}と Delicious^{*2}という二つのサービスからデータを収集した。我々は異なるウェブサービスやアプリケーションからのデータを組み合わせることを目指しているの

で、異なるウェブサービスやアプリケーションを“橋渡しする”もしくは“繋げる”ようなエンティティが必要である。URL はまさにそのようなエンティティとして考えることができる。加えて、URL はウェブ上の様々なリソースを参照することができるので、ユーザの興味の対象を表現するのもも適している。したがって、URL を含むような行為を分析の対象にすることが望ましいだろう。Twitter における主要な機能である *tweet*, *retweet*, *favorite* はこの要件を満たす。また *follow* はこれらの行為の背景にあるソーシャルネットワークを作る行為であるので、以上の四つを Twitter 上の行為として分析対象とした。以下でこれらの機能を簡単に説明する。Twitter 上ではユーザは *tweet* と呼ばれる短いテキストを投稿することができ、この投稿中には URL を含めることもできる。他のユーザを *follow* することでそのユーザの投稿を自身のタイムラインに表示することができる。他のユーザの投稿を *retweet* することで、その *tweet* を再発信することができる。ある投稿を *retweet* することで、そのユーザを *follow* しているユーザのタイムラインに元の *tweet* が表示されるようになる。ユーザはまた他のユーザの投稿に対して *favorite* することができ、これはお気に入りのような意味合いを持つ。

Twitter 上の特定のユーザを起点にし、フォローネットワークを 2 ホップ先まで辿ることで、ユーザの候補集合を得た。更に異なるアプリケーションのデータを橋渡しするためにはユーザの同一性判定も行えることが望ましいが、それは *FriendFeed*^{*3} というウェブアプリケーションのアグリゲートサービスを用いることで可能になる。したがって、上記で得られたユーザ集合の中で *FriendFeed* のアカウントを保有しているユーザを自動的に特定し、行為主体のユーザ集合として 3,356 人のユーザを得、これらのユーザが 2010 年の 8 月中旬に Twitter 上で行った各行為を分析対象のデータとした。この 3,356 人のユーザが、本実験で対象とした全ユーザ集合である。また、本実験では Twitter の他に Delicious と呼ばれるブックマークサービスにおけるタグ付け行為を分析の対象に含めた。Delicious におけるタグ付け行為は、あるユーザがウェブ上のリソースに対して任意のタグを付ける行為であり、リソースを参照するのに URL が用いられている。上記で得た 3,356 人のユーザ集合の中で Delicious におけるアカウントも保有し、かつ上記期間中に Delicious 上での行為が確認されたユーザ 151 人を特定し、Delicious における行為データも取得した。

各行為の特徴量と、そのデータサイズについては表 1 にまとめた。各行為の種類 *tweet*, *retweet*, *favorite*, *follow*, *tag* に対して、その簡略表現としてそれぞれ Tw, RT, Fav, Fol, Tag という表記を追記した。この簡略表現を以降の結果に関する図表中で用いる。*original tweet* は、*retweet/favorite* の行為における元の *tweet* の ID の文字

*1 twitter.com

*2 delicious.com

*3 friendfeed.com

表 1 ソーシャルメディア上のユーザの行為分析実験で用いたデータの概要

行為の種類	行為の数	エンティティ	エンティティ数
tweet (Tw)	213,929	ユーザ	3,032
		URL	147,686
retweet (RT)	22,680	ユーザ	1,908
		URL	16,931
		original tweet	20,557
favorite (Fav)	37,912	ユーザ	1,693
		URL	28,392
		original tweet	35,018
follow (Fol)	214,561	フォローしているユーザ	2,894
		フォローされているユーザ	3,198
tag (TAG)	21,530	ユーザ	151
		URL	7,938
		tag	3,033

列である．ユーザは一つのリソースに対して複数のタグを用いることができるので，複数のタグが用いられた場合には個々のタグを一つの特徴量として全体として複数の特徴量を用いた．ロジスティック回帰の入力には，表 1 のエンティティの他，各行為についてその行為の種類を特徴量として用いた．また，*tweet* と *retweets/favorite* で *original tweet* の特徴量が共有されるよう，*tweet* においては自身の *tweet* の ID の文字列も特徴量として用いた．また，各行為について行為数に関するユーザ分布を図 1 に示した．各図で，横軸は行為数を表し，縦軸は，そのような行為数であるユーザの数を表す．

予測設定

各行為は表 1 に示されている特徴量で表現されるものとする．30 日間のデータを 3 日ごとに区切った 10 個のスロットを作り，スロット $t[i](i \in [1, \dots, 9])$ のデータを訓練データとしてその行為が実際に行われるかどうかを判定する学習器を構築し，構築した学習器を用いてスロット $t[i+1](i \in [2, \dots, 10])$ の評価データを分類した．今回はユーザや URL を共有するようにデータセットを構築しているので，異なる種類の行為同士で一定の特徴量が共有されている．

以上の予測設定の下，ある種類の行為を予測するのに，同じ種類の行為を用いる場合と異なる種類の行為を用いる場合のそれぞれで予測精度がどう変わるかを評価した．例えば tag 行為について与えられた評価データを，評価データと同じ種類の行為 (tag) と異なる種類の行為データ (tweet, favorite, retweet, follow) それぞれを訓練データとして用いて構築した学習器で正例か負例かを判定する二値分類を行い，その精度を評価した．各訓練データおよび評価データについて 50% の負例をランダムに作成した．学習には L2 正則化ロジスティック回帰を採用し，最適化にあたっては Limited-memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) [Nocedal 80] を用いた．実装には Classias^{*4} と呼ばれるライブラリを用いた．正則化項の係数は 1 で統一した．行為の種類によってデータ量が異なるので，データ量の差によって生じる精度差の影響

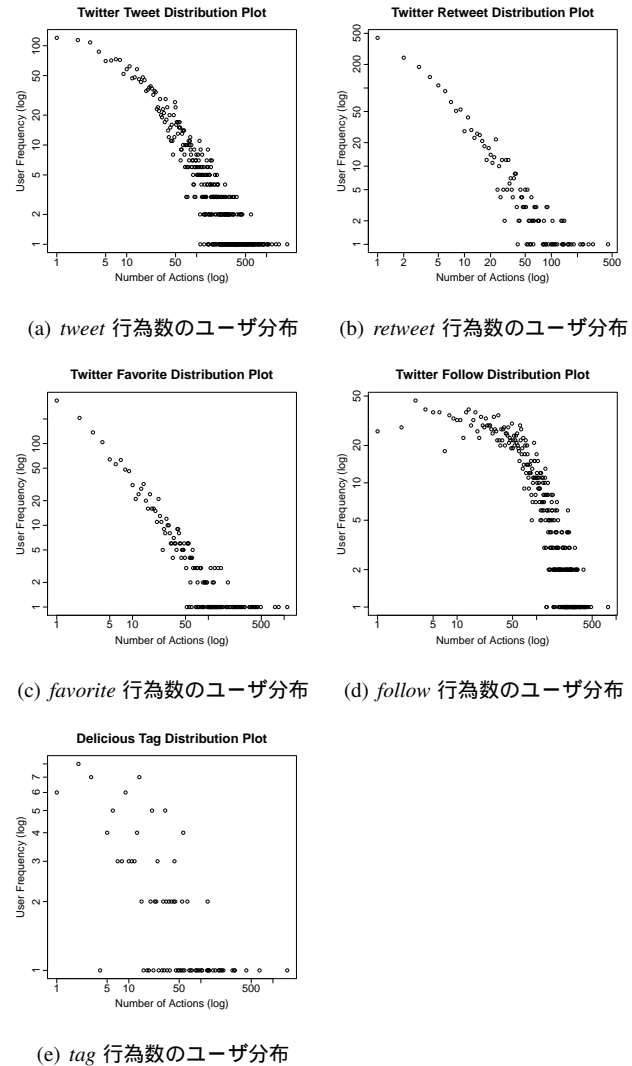


図 1 各行為について行為数に関するユーザ分布

を緩和するために，各スロットの各訓練データ/評価データの量 (学習/評価に用いる入力データ数) は異なる行為種類であっても同量になるように調整した．

2.2 結果

全結果を表 2 に記した．特に以下の考察で図 2, 図 3, 図 4 で表示した結果に対応する結果については表中で太字で表示している．以下でこの結果から得られるいくつかの考察を記す．

複数のウェブサービスにおける補足的な特徴量

図 2 は Twitter 上の行為を Delicious の *tag* 行為を用いて予測した場合と Twitter 上のもとの行為を用いて予測した場合の実験結果の比較を，図 3 は Delicious の *tag* 行為を Twitter の行為を用いて予測した場合と Delicious 上のもとの行為を用いて予測した場合の実験結果の比較を示している． $A - B$ で示される各棒グラフは，評価データの行為の種類が A であり訓練データの行為の種類が B であるような実験条件での予測精度の平均値を示しており，エラーバーは標準偏差を示している．この描画方法

*4 www.chokkan.org/software/classias/index.html

は以降の図で共通である。

例えば、図2の6本目のバーに対応するのは、Twitter上の *favorite* 行為を Delicious 上の *tag* を用いて予測する実験の結果であるが、これは、訓練データとして上記で説明した特徴量によって表現される *tag* のデータを用い、そのデータで学習した L2 正則化ロジスティック回帰のモデルで、上記で説明した特徴量で示される *favorite* の各行為を正例か負例かに二値分類する実験のことである。図2から、Deliciousの *tag* を用いることは、Twitterの *favorite* や *follow* の行為の予測に有効であることが確認できる。*favorite* 自身を用いて *favorite* を予測するよりも、*tag* を用いて *favorite* を予測する方が平均値としては高い精度を実現している。同様のことが *favorite* を *follow* に置き換えても言える。このことは、Deliciousの *tag* 行為は Twitterの *favorite* や *follow* 行為の代わりに用いることができること、すなわち *favorite* や *follow* のデータが過疎であるときに他の種類の行為データとして *tag* の行為データを活用することができることを示唆している。

一方で、逆に Twitter上の行為を用いて Delicious上の行為を予測する実験では、図3で示されているように上記のようなことは言えない。Twitter上の行為を用いて Delicious上の行為を予測した場合、Delicious上の行為を用いて予測した場合と比べて総じて低い予測精度となっている。

これは複数のドメイン（ウェブサービスやアプリケーション）における非対称な性質と言えるだろう。我々の実験では、Deliciousの *tag* 行為が Twitter上のいくつかの行為を予測するのに有効であった一方で、Twitterの行為データは Deliciousの *tag* 行為を予測するのに有効ではなかった。この理由としては、キーワード（タグ付け行為で使われたタグ）は、キーワードを含まない行為に反映されている嗜好に補足的な嗜好を反映していることが考えられる。例えば *favorite* に反映されている嗜好を説明するには、*favorite* 行為から直接的に観測できる特徴量（ここでは行為主体のユーザや言及されている URL）だけでは不十分であり、エンティティ（行為主体ユーザや言及される URL）を共有する *tag* 行為から観測される特徴量（キーワード）を用いることで、元の嗜好を補足することができると考えられる。以上のことから、異なるウェブサービス、アプリケーション間であってもユーザの行為データをアグリゲートすることが有効であることが示唆される。

類似した行為の種類

図4に、*tweet/retweet* から *tweet/retweet* を予測する実験結果を示した。図から、一方の種類の行為を予測するのに他方の種類の行為を用いても同等の結果になっていることが確認できる。このことは、上記で述べたのと同様に、一方のデータが過疎であるときに他方のデータを活用することができることを示唆している。ある種類の行為 *A* を予測するのに別の種類の行為 *B* を活用するこ

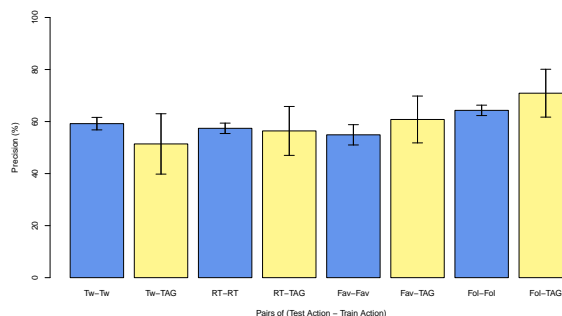


図2 Twitterの行為を Deliciousの *tag* 行為から予測する実験

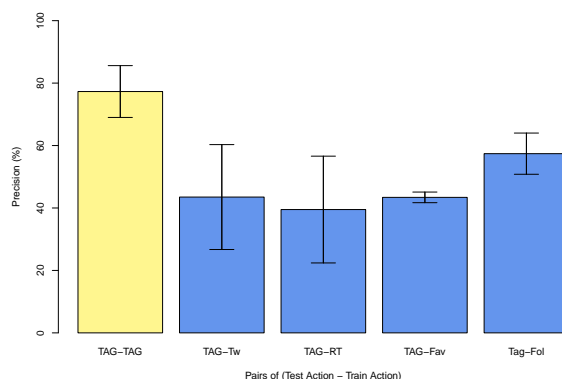


図3 Deliciousの *tag* 行為を Twitterの行為から予測する実験

とができるとき、行為 *A* と *B* に反映されている嗜好は共通部分を持つと考えられる。行為 *B* を予測するのに行為 *A* を活用することができるとき、行為 *A* と *B* は互いに類似していると言えるだろう。*tweet* と *retweet* の関係について言えば、これらの行為が相対的に類似していることは妥当と言えるだろう。今回対象とした行為の中で *tweet* と *retweet* はどちらも、自身をフォローしているユーザに対してある URL を発信していると言え、これは他の行為にはない性質だからである。

類似していない行為の種類

一方で表2中で、*follow* 行為を他の種類の行為を用いて予測した場合や、その逆に *follow* を用いて他の行為を予測した場合から確認できるように、他の種類の行為が有効に活用できない場合も存在する。上記の多くの場合において、他の種類の行為を用いても、元の行為を用いた場合よりも総じて低い精度しか実現できていない。*follow* 行為は他の行為よりも静的な行為であり、フォローネットワークは他の行為のプラットフォームとして考えられるので、*follow* 行為と他の行為は相対的に異なる行為なのであろう。

URLを言及する行為の類似性

表2から、Twitterにおける URL を言及する行為 (*tweet*, *retweet*, *favorite*) は相対的に互いに類似していることが確認できる。URL を言及する行為を予測するには、*follow* を用いるよりも他の URL を言及する行為を用いる方が

多くの場合で高い精度を実現する．これは，Twitterにおいて URL を言及する行為に反映されている嗜好は相対的に互いに類似しているためと考えられる．上記の考察と同様に，URL を言及する行為データが過疎であるときには他の URL を言及する行為データを活用できることが示唆される．

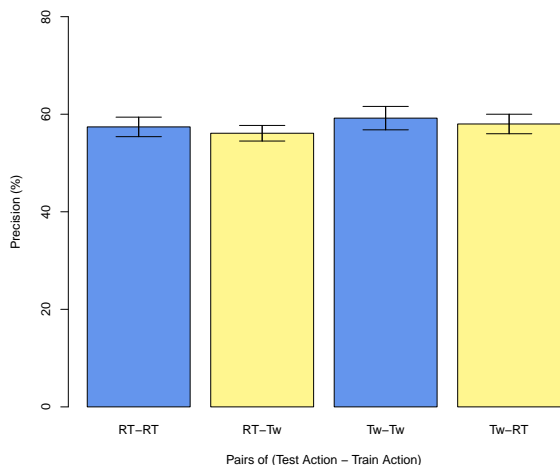


図4 *tweet/retweet* から *tweet/retweet* を予測する実験

3. 提案手法

二章で行った分析から，ユーザの様々な行為を興味予測に活用できることが示唆された．したがって，ユーザの興味予測においては，ユーザの様々な行為を活用する手法が有効であると期待される．本章ではユーザの様々な行為を活用したユーザの興味予測のための手法を提案する．

3.1 アクショングラフの構築

ここでは，ユーザの様々な行為を多項関係として表現できるグラフであるアクショングラフを提案する．以下で，例を用いてアクショングラフの構築方法について述べる．

ソーシャルメディア上でのユーザの行為はしばしば複数のエンティティを巻き込む．例えばあるユーザ A がある URL l で参照されるウェブサイトに対して“人工知能”というタグを付ける状況を想定してみる．この行為には四つのエンティティが関与していると考えられるだろう．すなわち，ユーザ A ，URL l ，行為の種類としての“タグ付け”，そしてキーワード“人工知能”である．この行為を表現するために，我々はこの行為それ自体を表現する一つのアクションノードを生成し，さらに上記四つのエンティティ，ユーザ A ，URL l ，キーワード“人工知能”，行為の種類“タグ付け”に相当する四つのオブジェクトノードを生成する．続いてこのアクションノードと，関連するオブジェクトノード全ての間にそれぞれ

表2 ソーシャルメディア分析実験における予測精度（各スロットでの平均と標準偏差）

評価に用いた行為	訓練に用いた行為	精度
TAG	TAG	77.3 ± 8.3%
TAG	Fav	43.4 ± 1.7%
TAG	Fol	57.4 ± 6.6%
TAG	RT	39.5 ± 17.1%
TAG	Tw	43.5 ± 16.8%
Fav	Fav	54.9 ± 3.9%
Fav	Fol	49.5 ± 2.4%
Fav	RT	53.0 ± 3.3%
Fav	Tw	51.6 ± 2.4%
Fav	TAG	60.8 ± 9.0%
Fav	Tw-Fol	50.5 ± 1.5%
Fav	Fav-Fol	53.3 ± 1.9%
Fav	Fav-RT	55.9 ± 2.9%
Fav	Fav-Tw	55.0 ± 3.1%
Fav	RT-Fol	51.2 ± 1.2%
Fav	RT-Tw	52.5 ± 2.5%
Fav	Fav-RT-Fol	54.7 ± 2.4%
Fav	Fav-RT-Tw	55.3 ± 3.0%
Fav	Fav-Tw-Fol	53.9 ± 1.8%
Fav	RT-Tw-Fol	51.2 ± 1.8%
Fav	Fav-RT-Tw-Fol	54.5 ± 2.2%
Fol	Fav	55.0 ± 3.6%
Fol	Fol	64.3 ± 2.0%
Fol	RT	52.4 ± 3.4%
Fol	Tw	53.2 ± 2.3%
Fol	TAG	70.9 ± 9.2%
Fol	Tw-Fol	62.9 ± 2.3%
Fol	Fav-Fol	63.5 ± 1.8%
Fol	Fav-RT	51.7 ± 2.9%
Fol	Fav-Tw	52.8 ± 2.5%
Fol	RT-Fol	63.2 ± 2.3%
Fol	RT-Tw	52.0 ± 2.8%
Fol	Fav-RT-Fol	62.4 ± 2.2%
Fol	Fav-RT-Tw	51.5 ± 2.8%
Fol	Fav-Tw-Fol	62.2 ± 2.1%
Fol	RT-Tw-Fol	62.2 ± 2.0%
Fol	Fav-RT-Tw-Fol	61.7 ± 1.8%
RT	Fav	53.7 ± 2.6%
RT	Fol	50.8 ± 1.5%
RT	RT	57.4 ± 2.0%
RT	Tw	56.1 ± 1.6%
RT	TAG	56.4 ± 9.4%
RT	Tw-Fol	52.7 ± 1.1%
RT	Fav-Fol	52.0 ± 1.1%
RT	Fav-RT	57.5 ± 2.3%
RT	Fav-Tw	54.6 ± 1.0%
RT	RT-Fol	55.5 ± 1.2%
RT	RT-Tw	58.2 ± 2.1%
RT	Fav-RT-Fol	56.2 ± 1.3%
RT	Fav-RT-Tw	57.9 ± 2.9%
RT	Fav-Tw-Fol	53.1 ± 1.5%
RT	RT-Tw-Fol	56.7 ± 1.9%
RT	Fav-RT-Tw-Fol	56.9 ± 1.9%
Tw	Fav	53.8 ± 3.8%
Tw	Fol	51.0 ± 2.5%
Tw	RT	58.0 ± 2.0%
Tw	Tw	59.2 ± 2.4%
Tw	TAG	51.4 ± 11.6%
Tw	Tw-Fol	56.0 ± 1.7%
Tw	Fav-Fol	51.5 ± 2.5%
Tw	Fav-RT	57.0 ± 2.0%
Tw	Fav-Tw	58.7 ± 2.0%
Tw	RT-Fol	53.7 ± 1.8%
Tw	RT-Tw	59.9 ± 2.2%
Tw	Fav-RT-Fol	53.5 ± 2.2%
Tw	Fav-RT-Tw	59.1 ± 1.7%
Tw	Fav-Tw-Fol	56.0 ± 1.8%
Tw	RT-Tw-Fol	56.6 ± 1.6%
Tw	Fav-RT-Tw-Fol	56.5 ± 1.4%

エッジを張る．

我々が提案するアクショングラフは重みなし無向二部グラフ $G = (V_{OB} \cup V_{AC}, E)$ として定義される．

エンティティの総数を m 、行為の総数を k とするとき、 V_{OB} 内の各ノード $v_{OB}^{(i)}$ ($i = 1, 2, \dots, m$) はオブジェクトノードによって表現される、何らかの行為に参与したある特定のエンティティに相当し、 V_{AC} 内の各ノード $v_{AC}^{(j)}$ ($j = 1, 2, \dots, k$) はアクションノードによって表現されるある特定の行為に相当する．エッジ $e^{(i,j)} \in E$ は、 $v_{OB}^{(i)} \in V_{OB}$ と $v_{AC}^{(j)} \in V_{AC}$ の間にはられる重みなしエッジである．すなわち、アクショングラフとは、アクションノードと、そのアクションノードが表現する行為に参与した全てのオブジェクトノードの間に重みなしの無向エッジが張られるような重みなし無向二部グラフである．図5にソーシャルメディア上でのユーザの行為の例と、対応するアクショングラフを示した．

アクショングラフの表現は元の多項関係に表現されている三つ以上の共起情報を保持している．なぜなら、ある行為に相当するアクションノードから出ているエッジを辿ることで、その行為に関わったすべてのエンティティを特定できるからである．

3.2 アクショングラフを用いたユーザの興味予測

続いて、上記で定義したアクショングラフを用いてユーザの興味を予測する手法について説明する．

推薦タスクにおいては協調フィルタリング [Resnick 94, Shardanand 95] と呼ばれる手法がよく用いられているが、協調フィルタリングにおいては、推薦タスクはユーザとアイテムの間の類似度を計算することでモデル化される．例えば、二人のユーザは関与したアイテム集合が類似していれば似ていると考えられ、あるユーザに対する推薦は、そのユーザと似たユーザが気に入ったアイテムの中でまだそのユーザが関与していないアイテム集合をもとに行われる．この基本的なアイデアは我々のアクショングラフに対しても適用できる．二人のユーザの行為が似ていればその二人は似ていると考えるのは自然である．従って、協調フィルタリングの枠組みに則り、ユーザのリソースへの興味を予測する問題を、ユーザの様々な行為を活用した上でユーザとリソースの間の類似度を計算する問題としてモデル化する．

以下で提案手法の詳細を説明する．

オブジェクトノードの総数が m 、アクションノードの総数が k のアクショングラフ $G = (V_{OB} \cup V_{AC}, E)$ が与えられたとき、 m 行 k 列の行列 B を、元のアクショングラフに対応する二値行列とする．行列 B の各要素 $B_{i,j}$ は以下のように定義される．

$$B_{i,j} \equiv \begin{cases} 1 & (e^{(i,j)} \in E \text{ の場合}) \\ 0 & (e^{(i,j)} \notin E \text{ の場合}). \end{cases}$$

まず、ユーザ同士の類似度や、ユーザとリソースの間の類似度など、エンティティ間の類似度を関与した行為から計算するために、エンティティ間の初期類似度行列として $M = BB^T$ を作成する． M の各要素 $M_{i,i'}$ はエンティティ $v_{OB}^{(i)} \in V_{OB}$ とエンティティ $v_{OB}^{(i')} \in V_{OB}$ が共起した行為の総数である．

続いて、行列 M で表現されるグラフ上でエンティティ間の類似度を伝播させることを考える．以下で具体例を用いて説明する．ユーザ A は URL l に対してまだどのような行為も行っていない状態で、ユーザ A が URL l にどれくらい興味を持つかを予測するために、ユーザ A と URL l の間の類似度を計算することを考える．「ユーザ A がユーザ B と似ており（何らかの行為において共起しており）、かつ、ユーザ B と URL l も似ている（何らかの行為において共起している）」ならば、ユーザ A と URL l も似ていると考えられるだろう．括弧内で示した前提は、行列 M で表現されるグラフ上でユーザ A と URL l の間にユーザ B を経由する経路長 2 のパスが存在することと同値である．ここでは簡単のために経路長 2 に限定したが、これはあらゆる経路長についても言え、また、経路長が短いほどユーザ A と URL l の類似度は高くなると考えられる．また、パスをたどる際のエッジの重みが大きければ、上記の類似度はより高くなると考えられる．ここではユーザ A と URL l の間の一つのパスについて考えたが、ユーザ A と URL l の間の類似度は、各パスにより付加される類似度の総和によって決まると考えられるだろう．

以上のような性質を持つ類似度を行列 M から作成するために、本論文ではグラフカーネル [Smola 03] を用いることを提案する．一部のグラフカーネルは、二つのノード間に存在するパスの数が多ければそのノード間の類似度が高くなり、パスの経路長が増えたとそのノード間の類似度が低くなるという性質を持つ．結果として、二つのノード間に短いパスが沢山あるほどその二つのノードは類似していると考えられる．

これまで数多くのグラフカーネルが提案されてきた [Smola 03] が、その中でも正則化ラプラシアンカーネルなどのラプラシアン行列に基づくカーネルが、関連したノードを見つけるのに有効であることが指摘されている [Ito 05]．行列 M の正則化されていないラプラシアン行列は $D - M$ として定義される．ここで D は $D_{i,i} \equiv \sum_j M_{i,j}$ として定義される対角行列である．このラプラシアン行列に対して $D^{-\frac{1}{2}}$ を左右から乗ずることで、対称化された正則化ラプラシアン行列 $L(M) \equiv D^{-\frac{1}{2}}(D - M)D^{-\frac{1}{2}}$ を得ることができる．本論文で用いる正則化ラプラシアン行列は下記のように定義される．

$$\sum_{k=0}^{\infty} (-\beta L(BB^T))^k = (I + \beta L(BB^T))^{-1} \quad (1)$$

ここで B は既に述べたように、オブジェクトノードと

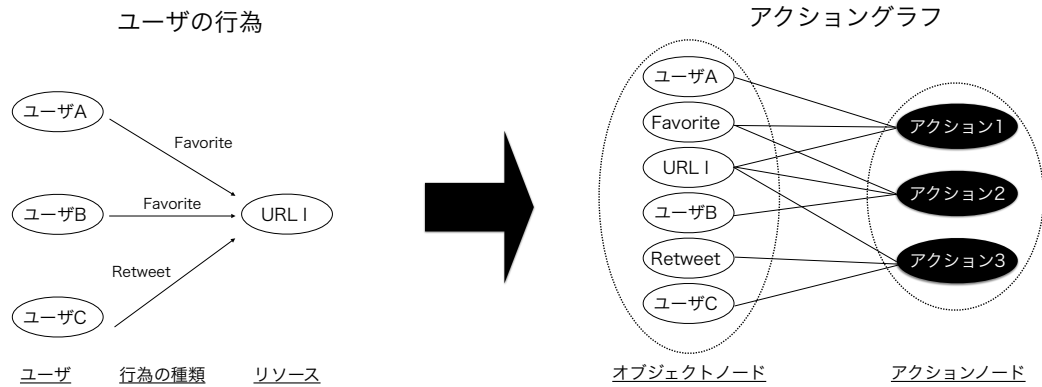


図5 アクショングラフの構築例

表3 ユーザーの興味予測実験で用いたデータの概要

行為の種類	行為の数	エンティティ	エンティティ数
Retweet (RT)	14,221	行為主体ユーザ	1,144
		言及ユーザ	7,935
		URL	11,335
Favorite (Fav)	22,755	行為主体ユーザ	1,125
		言及ユーザ	10,049
		URL	18,244

表4 データのスパースネス

行為の種類	スパースネス (テンソル)	スパースネス (アクショングラフ)
Retweet (RT)	1.4E-7	1.5E-4
Favorite (Fav)	1.1E-7	1.0E-4

アクションノードの総数がそれぞれ m, k であるようなアクショングラフ $G = (V_{OB} \cup V_{AC}, E)$ に対応する m 行 k 列の二値行列であり、 I は m 行 m 列の単位行列である。また、 β はノードペアの類似度を計算する際に離れたノードにどの程度重みをつけるかを調整するハイパーパラメータである。正則化ラプラシアンカーネルのほかにも、アクショングラフを用いてユーザーの興味を予測するのに適切であろうグラフカーネルは考えられるが、本論文では簡単のため正則化ラプラシアンカーネルに限定する。

4. 提案手法の評価実験

本章では、ユーザーの興味予測タスクにおいて提案のアクショングラフの性能を評価する実験について述べる。

4.1 実験条件

データセット

クローラの条件は二章「ソーシャルメディア上のユーザーの行為の分析」の実験と同一である。データセットは

公開されている^{*5}。表3にデータの概要を示した。各行為は「言及ユーザが投稿したURLを行為主体ユーザが *retweet* ないし *favorite* した」ことを示している。

また、表4に、テンソルを用いた場合、アクショングラフを用いた場合のそれぞれについて、データのスパースネスを表示した。スパースネスを、可能な組み合わせの中の実際の組み合わせの割合と考えると、テンソル、アクショングラフを用いた場合のスパースネスはそれぞれ、 $\frac{|V_{AC}|}{\prod_{k=1}^K |Entity^{(k)}|}$ 、 $\frac{|E|}{|V_{OB}||V_{AC}|}$ として定義される。ここで、 $Entity^{(k)}$ は、各エンティティのクラス(今回のデータでは、行為主体ユーザ/URL/言及ユーザの三種類)ごとにまとめたオブジェクトノードの集合を表す。今回のデータでは、 $Entity^{(1)}, Entity^{(2)}, Entity^{(3)}$ がそれぞれ、行為主体ユーザ集合、URL 集合、言及ユーザ集合である。 $||$ は集合のサイズを表す。今回の定義ではスパースネスの値が大きいかほどスパースネスが緩和されることになる。アクショングラフを用いた場合のスパースネスはテンソルを用いた場合のスパースネスの1,000倍となっており、アクショングラフを用いることでスパースネスが大幅に緩和されていることが確認できる。

また、今回のデータセットに対するアクショングラフの構築例を図6に示した。

予測設定

ユーザがウェブ上のあるリソースに対して何らかの行為を行ったとき、そのユーザはそのリソースに対して興味を抱いたと考えることができる。本実験では、ユーザーの興味予測を、ユーザが各リソースに対して抱く興味の度合い、すなわち、ユーザが各リソースに対してどれくらい *retweet* ないし *favorite* の行為を行いそうかの度合いをユーザとリソースの類似度をもとに予測する問題として定義する。

入力は関係データのタプルであり、タプルが与えられ

*5 ここでは https://dl.dropboxusercontent.com/u/1087723/TwitterActions2011_blind.zip のリンク先を示すが、データセットは一般公開されている。

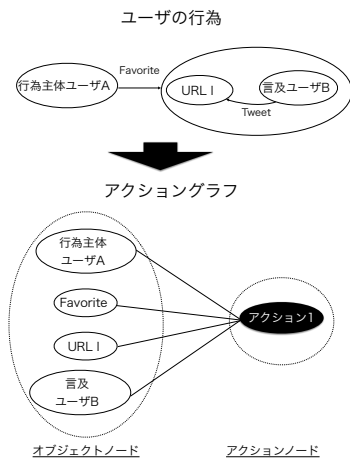


図6 実験で用いたデータセットに対するアクショングラフの構築例

ればアクショングラフを構築することができる。出力はエンティティ間の類似度である。我々は、あるユーザがあるリソースにどれくらい興味を持つかを、そのユーザとリソースの間の類似度をもとにして予測する。前述の実験と同様に30日間のデータを3日ごとに区切った10個のスロットを作り、スロット $t[i]$ ($i \in [1, 9]$) のデータを訓練データとし、スロット $t[i+1]$ ($i \in [2, 10]$) のデータを評価データとした。今回の実験では3日間でサンプリングしたデータを使うが、提案手法は特定のサンプリング間隔を前提としない。

比較手法

我々は提案手法を5つのベースラインと比較した。提案手法および5つの比較手法について表5にて整理した。比較の観点としては、関係モデルとして多項関係と二項関係があり、さらに手法の分類としてグラフに基づく手法、テンソル分解、確率的生成モデルの3つがある。関係モデルというのは、ユーザのリソースへの興味を予測するにあたり、ユーザの行為をユーザとリソースの二項関係として捉え予測に活用するのか、それ以外の情報（発信元ユーザや行為の種類）を踏まえた多項関係として捉え予測に活用するのかを比較するための項目である。手法としては大きくグラフに基づく手法、テンソル分解、生成モデルの3つを挙げているが、グラフに基づく手法に関してはどのようなグラフを構築したかで、アクショングラフ、二部グラフ、クリーク拡張の3つを挙げている。これらのグラフに基づく手法において適用するアルゴリズムは提案手法と同一である。これらの手法間の結果を比較することで、(a) ユーザの行為を多項関係として捉え予測に活用することで、二項関係として捉える場合よりも予測の性能を向上させることができるか、(b) 同じ情報を用いても構築するグラフによりアルゴリズムの性能が変わるかを確認できると期待できる。特に、提案手法と比較手法(3) AG-binomial で異なるのは関係モデルのみであるので、これら二つの結果を比較することで上記(a)を確認できる。また、提案手法と比較手法(2)

Clique Expansion で異なるのはグラフの構築方法のみであるので、これら二つの結果を比較することで上記(b)を確認できる。以下で各手法の詳細を説明する。

多項関係を扱うのに近年盛んに用いられているのがテンソル分析であり、その中でも標準的な手法の一つである(1) PARAFAC [Bader 06] を多項関係の分析手法の代表として採用した。PARAFACを用いることで各オブジェクトは潜在空間に射影される。本実験では、PARAFACによって得られたユーザとURLの潜在ベクトルのコサイン類似度によってユーザとURLの類似度を計算した。(2) Clique Expansion [Zien] は、元の N 項から成る多項関係を二項の全組み合わせに分解し、提案手法と同一のアルゴリズム(正則化ラプラシアンカーネル)を適用した手法である。これは、多項関係をペアワイズな二項関係に分解するのではなく、もともとの多項の共起を保存することが予測の精度やデータ過疎への頑健性といった性能の向上に実際に有効であるかどうかを評価するために採用した。(3) AG-binomial は、行為主体のユーザとURLのみからアクショングラフを構築し提案手法と同一のアルゴリズムを適用した手法である。これは、ユーザの行為をユーザとリソースの二項関係として捉えるのではなく、ほかのエンティティも巻き込む多項関係として捉えることが予測の精度やデータ過疎への頑健性といった性能の向上に実際に有効であるかどうかを評価するために採用した。(4) LDA [Blei 03] は、確率的生成モデルの一つであり、観測データはトピックと呼ばれる未観測の潜在変数によって説明される。あるユーザ u があるリソース (URL) r_i に興味を持つ確率は $P(r_i|u) = \sum_{j=1}^Z P(r_i|z_i = j)P(z_i = j|u)$ として計算される。ここで $P(r_i|u)$ は、ユーザ u が与えられた時の i 番目のURLの条件付き確率であり、 z_i はトピックである。LDAは二項関係を対象とする種々の手法の中で最高水準の手法の一つとして考えられるので、既存の二項関係の分析手法の代表として採用した。(5) Bipartite Graph は、ユーザとリソースの二部グラフに対して提案手法と同一のアルゴリズムを適用した手法であり、二項関係を対象とする既存手法の中の一つとして採用した。評価指標

予測精度の評価には R -Precision [Craswell 09] を、データ過疎への頑健性の評価には被覆率 [Ge 10] を用いた。以下で詳細を説明する。 R -Precision は、 R を正例のデータ数とした時に上位 R 個について評価した際の precision (精度) である。各スロットとユーザについて R -Precision の平均を計算し、 t 検定を行った。この際、行為数が0となるユーザは計算の対象外とした。各ユーザについての正例は、retweet/favoriteの行為を行ったURLである。被覆率は、評価データ中で類似度を計算できたユーザとURLのペアの割合を示す。推薦システムにおいては、推薦すべきアイテムを見落とさないことが重要である場面はしばしばあり、被覆率は特にそのような状況における推薦システムの性能の評価に用いられている。また、被覆率

表 5 実験で用いた手法の比較

	関係モデル	手法分類	グラフに基づく手法で構築しているグラフ		
			アクショングラフ	二部グラフ	クリーク拡張
提案手法	多項	グラフに基づく手法	✓		
(1) PARAFAC	多項	テンソル分解		NA	
(2) Clique Expansion	多項	グラフに基づく手法			✓
(3) AG-binomial	二項	グラフに基づく手法	✓		
(4) LDA	二項	確率的生成モデル		NA	
(5) Bipartite Graph	二項	グラフに基づく手法		✓	

は推薦システムのコールドスタート問題における対処の度合いを評価するのにも有効である。なぜかというところ、データが過疎なユーザやアイテムに対してはそもそも予測値が得られない状況がしばしば起こるため、そのような疎なデータに対しても予測値を得られること自体がデータ過疎に対する頑健性として考えられるからである。R-Precision と同様に、各スロットとユーザについて被覆率の平均を計算し、 t 検定を行った。

ハイパーパラメータ設定

全データの 20% をチューニング用データとしてランダムにサンプリングし、このデータを用いて R-Precision を評価指標として提案手法と比較手法のハイパーパラメータを調整した。 β を 0.01, 0.05, 0.1 の三つから調整した結果ほぼ同じ値となったが、 $\beta = 0.01$ のときが若干精度が高かったため、 β は 0.01 で固定した。正則化ラプリアンカーネルがハイパーパラメータ β に対して安定していることは指摘されており [Ito 05] 実際今回の実験でもそのことが確認された。PARAFAC においては潜在次元数を 200, 400, 600, 800 で調整した結果最も R-Precision の高かった 600 で固定した。LDA ではトピック数のほか α と β という二つのハイパーパラメータが存在する。LDA をドキュメントの生成モデルとして用いるとき α は各ドキュメントにおけるトピックのディリクレ事前分布のハイパーパラメータであり、 β は各トピックにおける単語のディリクレ事前分布のハイパーパラメータである。LDA においてはトピック数を 100, 200, 400, 600 で、 β を、0.01, 0.05, 0.1 で調整した結果最も R-Precision の高かったトピック数 400 と $\beta = 0.01$ で固定した。トピック数に応じて α を変化させることが有効であることが指摘されているので [Wallach 10]、今回の実験でも、ライブラリとして用いた $pLDA$ *6 の実装に従い各トピックについて α は $\frac{50}{topicNumber}$ として計算した。

4.2 結果

各スロットとユーザで平均した精度 (R-Precision) と被覆率の比較を表 6 に示した。提案手法と比較手法で $paired-t$ 検定を行い有意 ($p < 0.05$) であった手法を太文字で表示してある。

予測精度については提案手法が他の全ての手法よりも有意に高い精度を実現していることが確認できる。被覆率については、提案手法は *Clique Expansion* に対しては有意差のない結果となったが、それ以外の全ての手法に対しては有意に高い被覆率となった。特に、PARAFAC と *Clique Expansion* 以外のグラフに基づく手法 (*Bipartite Graph* と *AG-binomial*) に対しては提案手法はそれらを大きく上回る被覆率となった。提案手法を大きく下回る被覆率となった二つのグラフに基づく手法は、ユーザと URL の二項関係のみモデル化した手法である。提案手法がこれらの手法を大きく上回る被覆率を実現した理由として、行為の種類 (*retweet* や *favorite*) や *original-tweet-user*、すなわち URL を発信したもともとのユーザといった情報が有効に機能したことが考えられる。二項関係を予測する際にも元の多項関係を活用することで被覆率の向上が可能になることが示唆される。

また、提案手法が他のグラフに基づく手法の全てに対して総じて高い予測精度や被覆率を実現したことは特筆すべきである。グラフを用いてユーザとリソースの間の類似度を計算する研究はこれまで多々なされてきたが、本実験は、適用するアルゴリズムの性能はそもそも構築するグラフの選択に大きく依存することを示している。

以下で特に 4.1 節の実験条件にて述べた以下の二点：
(a) ユーザの行為を多項関係として捉え予測に活用することで、二項関係として捉える場合よりも予測の性能を向上させることができるか、(b) 同じ情報を用いても構築するグラフによりアルゴリズムの性能が変わるかについて結果を説明する。

提案手法と比較手法 (3) *AG-binomial* で異なるのは関係モデル (二項/多項) のみであるので、これら二つの結果を比較することで上記 (a) を確認できるが、予測精度と被覆率のどちらに関しても提案手法は (3) *AG-binomial* を有意に上回る結果となった。以上から、(a) ユーザの行為を多項関係として捉え予測に活用することで、二項関係として捉える場合よりも予測の精度や被覆率を向上させることができることが示唆された。また、提案手法と比較手法 (2) *Clique Expansion* で異なるのはグラフの構築方法のみであるので、これら二つの結果を比較することで上記 (b) を確認できるが、被覆率に関しては両者に

*6 www.code.google.com/p/plda

表 6 各手法における精度と被覆率の比較

	提案手法	PARAFAC	Clique Expansion	AG-binomial	LDA	Bipartite Graph
R-Precision (精度)	7.6%	3.4%	4.2%	4.2%	4.3%	2.0%
Coverage (被覆率)	99.8%	43.6%	99.8%	41.9%	99.0%	56.4%

有意差は確認できなかったものの、予測精度に関しては、提案手法が (2) *Clique Expansion* よりも有意に高い値となっていることが確認できる。以上から、同じ情報を用いても構築するグラフによりアルゴリズムの予測精度が変わることが示唆された。

5. 関連研究

多項関係を扱うことで三つ以上のエンティティの間の共起を捉えることができる。例えば二人の人間があるドキュメントに対して同一のタグを付けた場合でも、それぞれの人にとってそのタグで意図していることは異なるだろう。タグとドキュメントの二項関係だけでなくユーザを踏まえた三項関係(多項関係)を表現することで、上記の事実を反映した知見の抽出や予測モデルの構築が期待できる。規模の大きなデータに対して多項関係を扱う主要な方法は大きく、テンソル分析を用いた手法と多項関係をペアワイズの二項関係に分解して分析する手法の二つに分けることができる。二項関係をペアワイズに分解することで既存の行列やグラフに基づく手法による解析が可能となる。例としては、論文推薦を行うにあたって、著者、論文、会議の三項関係をペアワイズの二項関係に分解し、それぞれについて構築した行列に基いて最適化問題として定式化した研究 [Zhou 08] や、ユーザとリソースとタグの三項関係を一つの三部グラフや二つの二部グラフで表現しグラフアルゴリズムを適用した研究 [Hotho 06, Zhang 10] などが挙げられる。

しかし、これらのペアワイズな手法では元々の多項関係の共起情報が欠損してしまうという欠点がある。多項関係を保持した上で解析するアプローチとしてテンソル分解は近年盛んに研究されており [Kolda 09]、様々なテンソル分解手法が、CP 分解や Tucker 分解などの様々な分解モデル、カルバック・ライブラー情報量やユークリッド距離などを用いた様々な損失関数、そして様々な最適化手法とともに提案されている。特にウェブマイニングの分野では Symeonidis ら [Symeonidis 08] と Rendle ら [Rendle 09] が Tucker 分解に基づいたタグ推薦のためのテンソル分解手法を提案した。そのほか、Lin らは非負値行列分解をテンソルの場合に拡張し、ソーシャルメディア上でのコミュニティ抽出に適用した [Lin 09]。

以上のようにテンソル分解を用いた手法は近年盛んに研究されているが、本論文ではテンソルではなくグラフを用いて多項関係を解析しており、二項関係を対象としたグラフに基づく従来手法を基に、多項関係への拡張を

可能にするという点でそれらのアプローチと異なる。以下でこの点について述べる。ユーザにアイテムを推薦するタスクや、ソーシャルメディア上でのユーザの興味や行動を予測するタスクに対しては、協調フィルタリングの基になっている“自分と似たアイテムを気に入るユーザの気に入るアイテムは自分も気に入るだろう”といったような何らかの仮説を基に手法が構築されていることが多く、対象とするデータの特徴を踏まえた種々の仮説が考えられ、その仮説を基にした手法の有効性も検証されてきた。しかし、それらの多くは二項関係をモデル化したものであるか、多項関係を二項の全組み合わせに分解した上でのモデル化であった。基となっている種々の仮説には一定度の有効性があると期待できるので、それらの仮説を基に二項関係だけでなく多項関係もモデル化できるように拡張することは有意義であろう。特に、これまでの手法ではユーザとリソースの二項関係を主に対象としてきたのに対して、本論文ではエンティティと行為の二項関係という見方を導入することで多項関係をモデル化しグラフとして表現しているので、この見方を基にこれまで提案されてきたグラフに基づく手法を多項関係に対して拡張することも可能だろう。今回の例でいえば、我々は推薦システムで普及している協調フィルタリング [Resnick 94, Shardanand 95] のアイデアを拡張したと言える。協調フィルタリングの基本的アイデアは、人とアイテムの二項関係があるときに、一方(人同士もしくはアイテム同士)の類似度を他方の観点から評価することである。基になっているのは“似たようなアイテムを好む人同士は似ている”という仮説である。この仮説を基に、あるユーザと似た他のユーザが気に入るアイテムをそのユーザに推薦している。我々は、アイテムを行為に拡張し“行為が似ている人同士は似ている”という仮説を導入した上で、人と行為をグラフによって表現することで、グラフに基づくアルゴリズムを上記の拡張した仮説に沿う形で応用したと解釈できる。

最後に、本論文のベースとなっている著者らの研究 [Nori 11b, Nori 11a] と本論文の関係について説明する。[Nori 11b] では、本論文 1.5 節の「本論文の貢献」の (4) で述べた以下の点について独自の知見を得ている。すなわち、グラフを用いてユーザ間やユーザとリソース間の類似度を求めるためのアルゴリズムは多々提案されてきたが、本論文では、グラフを用いたアルゴリズムの性能はそもそもどのようなグラフを構築するかに依存すること、また、ユーザの行動に関する様々なエンティティの情報をグラフに付与することでグラフマイニングのアルゴリ

ズムを効果的に機能させることができることを示した点である。この点は本論文の主要な貢献となる知見である。[Nori 11a]では、本論文 1.5 節の「本論文の貢献」の(1)で述べた点に関して、ユーザの行動分析を網羅的に行った結果として独自の知見を得ている。また、[Nori 12]では、多項関係予測において課題となるデータの疎性に対処するために、接続行列を用いて本論文と同様に多項関係を二項関係に変換した上で、理論的大域解を保証する手法を提案している。

6. おわりに

本論文では、ソーシャルメディア上でのユーザの行為データからの興味予測という問題設定を行い、その意義や応用可能性を示した。上記のタスクを解くにあたって課題となる点として、データのアグリゲーションに関する知見を得る必要性と多項関係データを活用するデータ過疎に頑健な手法を構築する必要性の二点を述べた。前者の課題に対しては分析のための実験方法を提案し、複数のウェブサービスの複数の行為を含む実データを用いた実験を行うことでいくつかの知見を得た。後者の課題に対してはアクショングラフという、ユーザの様々な行為を多項関係として表現できる新しいグラフ表現を提案し、そのグラフを用いた予測手法を提案した。そして現実のソーシャルメディアのデータを用いて提案手法の予測精度やデータ過疎への頑健性を評価し、提案手法が様々なベースラインを上回る性能を実現することを確認した。グラフを用いてユーザ間やユーザとリソース間の類似度を求める研究は数多くなされてきたが、本実験の結果により、グラフを用いたアルゴリズムの性能はそもそもどのようなグラフを構築するかに依存すること、また、ユーザの行動に関与する様々なエンティティの情報をグラフに付与することでグラフマイニングのアルゴリズムを効果的に機能させることができることが示唆された。

◇ 参考文献 ◇

- [Bader 06] Bader, B. W. and Kolda, T. G.: *Algorithm 862: MATLAB tensor classes for fast algorithm prototyping*, in TOMS 32(4), pp. 635–653 (2006)
- [Bao 07] Bao, S., Xue, G., Wu, X., Yu, Y., Fei, B., and Su, Z.: *Optimizing web search using social annotations*, in Proceedings of the 16th International Conference on World Wide Web, WWW '07, pp. 501–510 (2007)
- [Blei 03] Blei, D. M., Ng, A. Y., and Jordan, M. I.: *Latent Dirichlet Allocation*, Journal of Machine Learning Research, Vol. 3, pp. 993–1022 (2003)
- [Craswell 09] Craswell, N.: *R-Precision.*, in Encyclopedia of Database Systems, p. 2453 (2009)
- [Ge 10] Ge, M., Delgado-Battenfeld, C., and Jannach, D.: *Beyond Accuracy: Evaluating Recommender Systems by Coverage and Serendipity*, in Proceedings of the fourth ACM Conference on Recommender Systems, RecSys '10, pp. 257–260 (2010)
- [Heymann 08] Heymann, P., Koutrika, G., and Garcia-Molina, H.: *Can social bookmarking improve web search?*, in Proceedings of the first International Conference on Web Search and Web Data Mining, WSDM '08, pp. 195–206 (2008)
- [Hotho 06] Hotho, A., Jäschke, R., Schmitz, C., and Stumme, G.: *FolkRank: A Ranking Algorithm for Folksonomies*, in LWA, pp. 111–114 (2006)
- [Ito 05] Ito, T., Shimbo, M., Kudo, T., and Matsumoto, Y.: *Application of kernels to link analysis*, in Proceedings of 11th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD 2005), pp. 586–592 (2005)
- [Jiang 12] Jiang, M., Cui, P., Wang, F., Yang, Q., Zhu, W., and Yang, S.: *Social Recommendation Across Multiple Relational Domains*, in CIKM '12, pp. 1422–1431 (2012)
- [Kolda 09] Kolda, T. G. and Bader, B. W.: *Tensor Decompositions and Applications*, SIAM Review, Vol. 51, No. 3, pp. 455–500 (2009)
- [Lin 09] Lin, Y.-R., Sun, J., Castro, P., Konuru, R., Sundaram, H., and Kelliher, A.: *MetaFac: community discovery via relational hypergraph factorization*, in Proceedings of 15th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD 2009) (2009)
- [Ma 11] Ma, H., Zhou, D., Liu, C., Lyu, M. R., and King, I.: *Recommender Systems with Social Regularization*, in WSDM '11, pp. 287–296 (2011)
- [Mika 05] Mika, P.: *Ontologies Are Us: A Unified Model of Social Networks and Semantics*, in Proceedings of the fourth International Semantic Web Conference, Vol. 3729 of ISWC '05, pp. 522–536, Springer (2005)
- [Nocedal 80] Nocedal, J.: *Updating quasi-Newton matrices with limited storage*, Mathematics of computation, Vol. 35, No. 151, pp. 773–782 (1980)
- [Nori 11a] Nori, N., Bollegala, D., and Ishizuka, M.: *Exploiting User Interest on Social Media for Aggregating Diverse Data and Predicting Interest*, in Proceedings of the fifth International Conference on Weblogs and Social Media, ICWSM '11, pp. 241–248 (2011)
- [Nori 11b] Nori, N., Bollegala, D., and Ishizuka, M.: *Interest Prediction on Multinomial, Time-Evolving Social Graphs*, in Proceedings of the 22nd International Joint Conference on Artificial Intelligence, IJCAI '11, pp. 2507–2512 (2011)
- [Nori 12] Nori, N., Bollegala, D., and Kashima, H.: *Multinomial Relation Prediction in Social Data: A Dimension Reduction Approach*, in AAAI '12, pp. 115–121 (2012)
- [Rendle 09] Rendle, S., Balby Marinho, L., Nanopoulos, A., and Schmidt-Thieme, L.: *Learning optimal ranking with tensor factorization for tag recommendation*, in Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '09, pp. 727–736 (2009)
- [Resnick 94] Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P., and Riedl, J.: *GroupLens: An Open Architecture for Collaborative Filtering of Netnews*, in Proceedings of the 1994 ACM conference on Computer supported cooperative work, pp. 175–186 (1994)
- [Schifanella 10] Schifanella, R., Barrat, A., Cattuto, C., Markines, B., and Menczer, F.: *Folks in Folksonomies: social link prediction from shared metadata*, in Proceedings of the third ACM International Conference on Web Search and Data Mining, WSDM '10, pp. 271–280 (2010)
- [Shardanand 95] Shardanand, U. and Maes, P.: *Social Information Filtering: Algorithms for Automating “Word of Mouth”*, in Proceedings of the SIGCHI conference on Human factors in computing systems, pp. 210–217 (1995)
- [Smola 03] Smola, A. J. and Kondor, R.: *Kernels and regularization on graphs*, Springer (2003)
- [Symeonidis 08] Symeonidis, P., Nanopoulos, A., and Manolopoulos, Y.: *Tag recommendations based on tensor dimensionality reduction*, in Proceedings of the second ACM Conference on Recommender Systems, RecSys '08, pp. 43–50 (2008)
- [Wallach 10] Wallach, H. M., Mimno, D., and McCallum, A.: *Rethinking LDA: Why Priors Matter*, in Proceedings of the 23rd Annual Conference on Neural Information Processing Systems (NIPS 2009) (2010)
- [Xu 08] Xu, S., Bao, S., Fei, B., Su, Z., and Yu, Y.: *Exploring folksonomy for personalized search*, in Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '08, pp. 155–162 (2008)
- [Yang 11] Yang, S.-H., Long, B., Smola, A., Sadagopan, N., Zheng, Z.,

and Zha, H.: *Like Like Alike: Joint Friendship and Interest Propagation in Social Networks*, in Proceedings of the 20th International Conference on World Wide Web, WWW '11, pp. 537–546 (2011)

[Zhang 10] Zhang, Z.-K., Zhou, T., and Zhang, Y.-C.: *Personalized recommendation via integrated diffusion on user-item-tag tripartite graphs*, Physica A: Statistical Mechanics and its Applications, Vol. 389, No. 1, pp. 179–186 (2010)

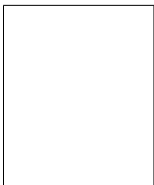
[Zhou 08] Zhou, D., Zhu, S., Yu, K., Song, X., Tseng, B. L., Zha, H., and Giles, C. L.: *Learning multiple graphs for document recommendations*, in Proceedings of the 17th International Conference on World Wide Web, WWW '08, pp. 141–150 (2008)

[Zien] Zien, J. Y., Schlag, M. D. F., and Chan, P. K.: *Multilevel spectral hypergraph partitioning with arbitrary vertex sizes*, Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on

{ 担当委員 : × × }

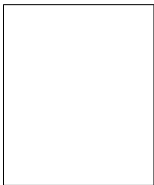
19YY 年 MM 月 DD 日 受理

—— 著 者 紹 介 ——



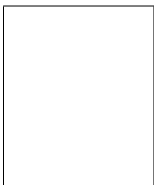
則 のぞみ(学生会員)

2010 年東京大学工学部システム創成学科卒 . 2012 年東京大学大学院情報理工学系研究科創造情報学専攻修士課程終了 . 2014 年から京都大学情報学研究科知能情報学専攻博士課程 . 機械学習 , データマイニングおよびヘルスケアへの応用に興味を持つ .



ボレガラ ダヌシカ(正会員)

2005 年東京大学工学部電子情報工学科卒 . 2007 年同大学院情報理工学系研究科修士課程修了 . 2009 年同研究科博士課程修了 . 博士 (情報理工学) . 同研究科・助教 , 講師を経て現在は University of Liverpool (Department of Computer Science) の Senior Lecturer (Associate Professor) . 自然言語処理に興味を持つ . WWW, ACL, ECAI などの会議を中心に研究成果を発表 .



石塚 満(正会員)

1971 年東京大学工学部卒 , 1976 年同大学院工学系研究科博士課程修了 . 工学博士 . 同年 NTT 入社 , 横須賀研究所勤務 . 1978 年東京大学生産技術研究所・助教授 (1980-81 年 Purdue 大学客員准教授) . 1992 年同大学工学部電子情報工学科・教授 . 2001 年同大学院情報理工学系研究科・教授 . 2013 年定年退職 , 東京大学名誉教授 . 2014 年早稲田大学基幹理工学研究科・教授 (エジプト日本科学技術大学担当) . 研究分野は人工知能 , Web インテリジェンス , 意味計算 , 生命的エージェントによるマルチモーダルメディア . IEEE , AAAI , 電子情報通信学会 , 情報処理学会等の会員 , 本会元会長 .

意味計算 , 生命的エージェントによるマルチモーダルメディア . IEEE , AAAI , 電子情報通信学会 , 情報処理学会等の会員 , 本会元会長 .