DEIM Forum 2022 G31-5

# SNS における辺の活性度とハブとなるユーザへのフォローを考慮したユーザ推薦手法

山田　瑛平†　　田島　敬史††

† 京都大学大学院情報学研究科　〒 606–8317 京都府京都市左京区吉田本町
†† 京都大学大学院情報学研究科　〒 606–8317 京都府京都市左京区吉田本町
E-mail: †yamada@dl.soc.i.kyoto-u.ac.jp, ††tajima@i.kyoto-u.ac.jp

**あらまし**　ソーシャルネットワークサービス (SNS) において, 高性能なユーザ推薦を提供することは重要な課題の一つである. SNS の形態に依存しないこと, 投稿の頻度が少ないユーザにも適用できること, 未知のユーザの推薦が可能であることから, 本研究ではグラフ構造情報に着目する. 時間情報を用いることで推薦の性能が向上する可能性が考えられることから, SNS のグラフの時系列データを用いるユーザ推薦手法を扱う. 新しい辺により大きな重みを割り付ける手法について, 同時期に生成された辺の活性度が異なりうることに注目し, 辺の活性度を推測して更新する手法, ならびにソーシャルネットワークサービス上ではオーソリティ度とハブ度の高いユーザが存在し, どちらをフォロー対象とするかの傾向はユーザごとに異なることから, オーソリティ度とハブ度の重要度を学習する手法を提案し, それらの手法の性能を実験により確かめた. 実験の結果, 新規辺の重みを大きくすることでユー ザ推薦の性能が向上することが示され, さらに辺の活性度を適切に推定することによってさらなる性能の向上が見込めることが明らかとなった. また, オーソリティー度とハブ度の重みを学習する手法の実験結果は, ユーザごとのフォローの傾向を学習することが性能の向上につながることを示すものではなかった.

**キーワード**　ソーシャルネットワークサービス, 推薦システム, 時系列分析, ランダムウォーク, Twitter

## 1　背　　景

近年のインターネットの普及に追従して, Twitter. Inc の提供する Twitter[1], Meta Platforms, Inc. の提供する Facebook[2]や Instagram[3]といった, ソーシャルネットワークサービス (SNS) の利用が拡大している. SNS の用途は多岐にわたるものの, 一例として, 共通した興味や背景を持つ他者と友人関係を構築する, あるいは現実の友人とのより円滑なコミュニケーションを図る, 有名人や信頼のある情報源の発信する情報を得る, といった使われ方をすることが常である.

一般的な SNS に最低限備わっている機能は, 投稿および閲覧である. 例えば Twitter であれば, ユーザはテキストベース, 画像ベース, 動画ベースの 3 種類の投稿を行うことができる. ユーザは他ユーザの投稿を閲覧し, お気に入りやリツイート、リプライといったインタラクションが可能である. また, 一般的な SNS においては, フォローと呼ばれる行為により, 別のユーザの投稿が自身のタイムラインに表示されるようになる. 他ユーザの投稿を見て, フォローするに値すると考えたユーザはフォローを行う. SNS 上での友人関係の構築や信頼性のある情報源からの情報を取得するためにはフォローが必要であり, フォローは SNS において重要な部分を担う機能である.

SNS ではユーザが他ユーザを発見するのを助ける手段が提供されていることが多い. ここでは三種類の手段を例として取り上げる. 第一に, SNS に搭載されている, 他のユーザの投稿を自分をフォローしているユーザに向けて掲示する機能を使うことである. これは, Twitter においてはリツイートと, Facebook においてはシェアと呼ばれる. 他のユーザの投稿を自分の投稿として再掲することで, 自分をフォローしているユーザに投稿を推薦することができる. この機能は, ユーザが自分の知らないユーザに出会う手段を提供する. 第二に, ハッシュタグや検索機能を用いて未知のユーザの投稿に出会うことである. ハッシュタグは Twitter や Facebook に実装されている機能で, 投稿内に挿入することで, 同一のハッシュタグを挿入した全ての投稿を閲覧することができる. このような機能は, 未知のユーザへの出会いを促進する. 第三に, SNS がユーザ推薦システムを提供している場合がある. この機能を用いることで, ユーザはサービスから直接推薦を受けることができる.

上述した三種類のフォロー手段のうち, 一番目と二番目はユーザの意思決定に依存する. それに対して, ユーザ推薦システムは SNS の管理者がサービスとして提供することができる. 適切なユーザ推薦システムを提供することで, サービス管理者はユーザにより満足感のある SNS 体験を提供することができる. また, 推薦ユーザ群に広告欄を設けることができ, 高性能なユーザ推薦が提供されればそれだけ広告の効果は高くなる. よって, ユーザ推薦の質を高めることは管理者とユーザ双方に利益があり, SNS で高性能なユーザ推薦を提供することは重要な課題のひとつになっている.

実際に適用されている推薦システムの例に, 文献 [1] を挙げる

1：https://twitter.com/
2：https://www.facebook.com/
3：https://www.instagram.com/

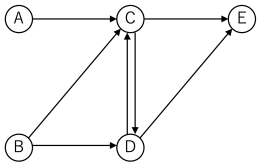


図 1　フォローグラフの例

ことができる．文献 [1] で示されているユーザ推薦アルゴリズムは以下の流れに沿うものである．まず, 当該ユーザをスタート地点とするランダムウォークを実行し, 通過したユーザのみを頂点集合にとる部分グラフ (Circle of Trust と呼ばれる) を形成する．次に, 部分グラフに SALSA [2] と呼ばれるランキングアルゴリズムを適用する．これによりユーザがどれだけ推薦に値するかを表したスコアが割り付けられ, ユーザが順序づけされることになる．

ここでは SNS のフォロー関係を表現するグラフを考えている．頂点を各ユーザ, 辺をあるユーザから別のユーザへのフォローを表現するものとしてみなし, 辺は有向辺として表現される．このようなグラフを本研究ではフォローグラフと呼ぶこととする．例えば図 1 のフォローグラフは, ユーザ A がユーザ C をフォローし, ユーザ B がユーザ C および D をフォローしている, といったフォロー関係を表現している．文献 [1] においては, ユーザ推薦において扱われる情報はユーザのグラフ構造のデータのみである．このようなグラフベースの推薦以外にも, ユーザの投稿内容からユーザ間の類似度を計算し推薦に用いるコンテンツベースの推薦手法, ユーザのインタラクションによって推薦ユーザを決定するインタラクションベースの推薦手法を考えることができる．

コンテンツベースの推薦は, ユーザの投稿内容の類似度によってユーザ同士の類似度を推定するものである．インタラクションベースの推薦は, Twitter であればツイートへのお気に入りやリツイート, リプライやプロフィールの閲覧といったユーザのインタラクションに基づくものである．この推薦はとても直感的な手法で, ユーザが投稿内容やプロフィールに興味を示したにもかかわらずフォローをしていない場合に推薦を行う．

このように SNS におけるユーザ推薦は様々な手法を考えることができるが, 本研究ではグラフベースの推薦を研究対象として扱う．これはまず第一に, グラフベースの推薦が SNS の形態によらないのに対し, コンテンツベースの手法は SNS の形態に依存することによる．投稿内容のテキストから類似度を求める手法は Twitter のようにテキストを中心とした SNS でのみ適用可能であり, 画像や動画を中心とした SNS では同様の手法を適用することができない．

第二に, いくつかの研究が示すように, 情報を収集するユーザは自身の投稿を行わないことによる．Twitter の利用状況を把握するために行われた研究 [3], [4] によると, 他のユーザと友人・知人といえるような相互関係を維持しているユーザは少なく, ほとんどのユーザは情報源または情報探索者として行動している．情報を探索するユーザは自分でツイートをしないため, コンテンツベースの推薦がうまく動作しない可能性がある．

第三に、インタラクションによる推薦はターゲットユーザに対して既知のものを推薦してしまう可能性が高い．例えば Twitter においてあるユーザの投稿にお気に入りやリツイートをしたことがあるのであれば, その時点でそのユーザのことを知っていると考えられる．ユーザ推薦が未知のユーザへの出会いを促進するものであると位置付ける限り, 既知のユーザを推薦する手法の価値は低くなる．

グラフ構造のデータは頂点の集合と有向辺の集合の組で表される．ただし, SNS におけるユーザのネットワークは, ユーザがあるユーザをフォローする, あるいはフォローしていたユーザをアンフォローするといったユーザの働きかけにより, 刻一刻と変化している．あるユーザがいつそれぞれのユーザをフォローしたか, どれだけの期間フォローしているか, 等の情報は, グラフ構造が時間経過に応じて変化するという SNS の特性を強く反映している．よって, SNS におけるユーザ推薦では, 時間的情報を用いることで性能の向上が期待できるのではないかと考えられる．そこで本研究では, 時間情報を加味した推薦システムを研究対象に扱うべく, 単一のグラフ構造のデータではなくグラフの時系列データを用いた推薦手法を扱う．まず新しい辺により大きな重みを割り付ける手法を考え, 同時期に生成された辺の活性度が異なりうることに注目し, 辺の活性度を推測して更新する手法を提案する．同時に, ソーシャルネットワークサービス上ではオーソリティ度とハブ度の高いユーザが存在し, どちらをフォロー対象とするかの傾向はユーザごとに異なることから, オーソリティ度とハブ度の重要度を学習する手法を提案する．本研究ではこれら二つの手法の性能を実験により確かめる．

## 2　関 連 研 究

本章では本研究に関係するいくつかの先行研究を取り上げる．2.1 節ではソーシャルネットワークにおける頂点の重要度の推定手法として, SALSA について取り上げる．2.2 節では時間情報を用いたユーザ推薦手法の関連研究を挙げるとともに, 推薦の評価に用いるべき指標を, 関連研究に触れつつ示す．

### 2.1　SALSA

SALSA [2] は PageRank [6] と HITS [7] から着想を得たランキングアルゴリズムである．SALSA は PageRank と同じくランダムウォークをシミュレートしたマルコフ連鎖モデルであり, HITS と同じく頂点のオーソリティー度とハブ度を考慮する．SALSA ではオーソリティー集合とハブ集合からなる二部グラフを構成する．Web ページにおける SALSA では, 検索クエリに適合したページの集合をオーソリティー, そのようなページを多くリンクしているページをハブとする．SALSA では, 適合していると判断されたページを多くリンクしているハブが存在するとき, そのハブがリンクしているページにオーソリティー度を割り振る．これにより, 検索クエリに適合しているにもかかわらず検索クエリに適合しないと判断されたページに適切に

オーソリティー度を割り当てることができる.

SALSA に用いられる二部グラフは, オーソリティー集合を決定した際にその集合内の頂点をリンクしている全ての頂点をハブ集合とすることによって得られる. オーソリティー集合内の要素をリンクしている頂点がオーソリティー集合内に存在しても良く, このため単一の頂点がオーソリティー集合とハブ集合の両方に属する可能性がある.

SALSA の計算に用いられる行列は, 通常の隣接行列 $L$ の各要素を行で正規化した行列 $L_r$ および各要素を列で正規化した行列 $L_c$ である. オーソリティー度は行列 $L_c^T L_r$ の定常ベクトルであり, ハブ度は行列 $L_r L_c^T$ の定常ベクトルである. すなわち, オーソリティー度 $a$ は $aL_c^T L_r = a$ および $a\mathbf{1} = 1$ を満たし, ハブ度 $h$ は $hL_r L_c^T = h$ および $h\mathbf{1} = 1$ を満たすように設定される. HITS では $a^{(1)} = L^T h^{(0)}$ が必ず満たされるのに対し, SALSA では二部グラフのオーソリティー集合およびハブ集合からスタートするランダムウォークを別々に考える.

### 2.2 時間的リンク予測

インタラクションを辺とするソーシャルネットワークにおける時間的リンク予測は, グラフの時系列データを用いたユーザ推薦と関係する. 時間的リンク予測とは, 時刻 1 から $T$ までのグラフのリンクデータが与えられたとき, 時刻 $T+1$ でのリンクを予測する問題である. 推薦とリンク予測とは, リンクが将来的に生成される確率を考慮する点において共通している. 文献 [8] は時系列データとして与えられる論文の共著ネットワークにおいてリンクを予測する手法が述べられている. 論文の共著ネットワークは, 頂点が研究者, 辺がその時刻において論文を共著したかを表す時系列データである. 時刻 $t$ におけるグラフのデータを $Z_t$, その推定値を $\hat{Z}_t$ とおくと, 以下のように表現される三つの手法が挙げられている.

- $\hat{Z}_t$ を, 直近の n データの平均とする. すなわち,

$$\hat{Z}_t = \frac{Z_{t-1} + Z_{t-2} + \cdots + Z_{t-n}}{n}$$

この手法は Moving Average と呼ばれる.

- $\hat{Z}_t$ を, 全データの平均とする. すなわち,

$$\hat{Z}_t = \frac{Z_{t-1} + Z_{t-2} + \cdots + Z_{t-T}}{T}$$

この手法は Average と呼ばれる.

- $\hat{Z}_t$ を, 過去の値の加重平均を取ることで予測する.

$$\hat{Z}_t = \alpha Z_{t-1} + (1-\alpha)\widehat{Z}_{t-1}$$

3 番目の手法は Simple Exponential Smoothing (SES) と呼ばれ, 以下のように書き換えられる.

$$\hat{Z}_t = \alpha Z_{t-1} + \alpha(1-\alpha) Z_{t-2} + \cdots + \alpha(1-\alpha)^{T-1} Z_{t-T} \quad (1)$$

$\alpha$ は減衰係数であり, $0 < \alpha < 1$ を満たす. SES は, グラフを集約する際に, 過去に遡ったデータであればあるほど, その情報としての価値を下げる手法である.

SES をユーザ推薦に用いることを考える. あるユーザに推薦すべきユーザは, 将来的にそのユーザがフォローする可能性の高いユーザである, という仮定のもとでは, リンク予測問題はユーザ推薦問題と等価になる. この仮定の是非については次章で扱う. SNS の利用者の興味や関心は絶えず変化するので, 過去のグラフ構造データとより現在に近い時刻のグラフ構造データとでは, 後者の情報をより重要視すべきである. よって, Moving Average や Average に比べて, SES はより SNS のユーザ推薦として適切な手法であると考えられる. ただし, 文献 [8] では扱うソーシャルネットワークの対象に論文の共著ネットワークが用いられており, 対象を SNS のフォロー関係を表すグラフとする本研究に同様の手法を用いることの妥当性は未知数である.

### 2.3 ユーザ推薦とリンク予測

リンク予測はリンクの生起確率を予測することのみが興味の対象である. 一方でユーザ推薦は, 推薦によってユーザの満足度を高めるという, より高次元的な目標を有している. そのため, リンク予測ではあるユーザが発見しやすいユーザを積極的に見つけることとなるが, ユーザ推薦はその逆であり, あるユーザが発見しにくいユーザを積極的に推薦すべきである.

ユーザ推薦システムが生起確率の高いリンクを推薦することは必ずしも良いことではなく, ユーザ推薦システムの評価に際してその精度のみを指標に用いることは適切ではないと言える. 文献 [9] はいくつかの推薦システムの評価尺度を提示している. それらを以下に示す. なお, 文献 [9] ではアイテム推薦を研究対象としており, ユーザ推薦を対象とする本研究とは性質が異なる. また, 研究 [9] は実験において, 映画に対してユーザが与えた評価値のデータセットを用いており, 評価という用語はこの評価値を指している.

- Unexpectedness. 推薦システムが提示したアイテムのうち, ユーザが知らなかったアイテムの割合. ユーザがあるアイテムを未知か既知かについてを直接調べることはできないので, 研究 [9] ではこの未知か既知かの判定を推定している.
- Serendipity. 推薦システムが提示したアイテムのうち, Unexpectedness に含まれ, かつユーザによって一定以上の評価を与えられたアイテムの割合.
- Coverage. 推薦システムが各ユーザに一度でも推薦したことのあるアイテムの総数.

Coverage を指標として用いることの背景には, 集中バイアスという問題がある. 協調フィルタリングを用いるような一般的な推薦システムは, 過去の売り上げや評価からアイテムを推薦する. そのため, 過去のデータが少ないアイテムは, たとえそのアイテムがどれだけ良いものであったとしても, 推薦対象に加わりにくい. 結果的には, 推薦システムが推薦するアイテムの総数が少なくなり, ユーザとアイテムとのマッチングの質の低下に繋がる [10]. Coverage が小さければ小さいほど, 集中バイアスが起こっていると言える. 逆に Coverage が大きければ, これは推薦システムが多様な推薦をしている証左である. 本研究では, 推薦システムの評価に, 精度だけでなくこれらのいくつかの指標を用いることで, 推薦システムを総合的に評価する.

## 3 提 案 手 法

本章では, 本研究で取り扱う手法の具体的な手法について説明する.

### 3.1 ランキングアルゴリズム

ランキングアルゴリズムは単一の重み付き有向グラフからターゲットユーザごとに他のユーザがどれだけ推薦に値するかを表現した数値を得るものである. 本研究におけるランキングアルゴリズムは SALSA に類する手法を用いる. 通常の隣接行列 $L$ の各要素を行で正規化した行列を $L_r$, 各要素を列で正規化した行列を $L_c$ とすれば,

$$a^{(k)} \leftarrow (1-j)L_c h^{(k-1)} + jr$$
$$h^{(k)} \leftarrow (1-j)L_r^T a^{(k)} + jr$$
$$k \leftarrow k+1$$

という繰り返し計算を行うことによってオーソリティーベクトル $a$ およびハブベクトル $h$ を得る. $j$ は $0 < j < 1$ を満たすリスタート率であり, $r$ はターゲットユーザの位置の要素のみ 1 で他は 0 のベクトルである.

この繰り返し計算は以下のようなランダムウォークをシミュレートしたものであると言える.

- ターゲットユーザは自身のフォロイーからランダムに 1 ユーザ選択し, 遷移する.
- 選んだユーザのフォロワーからランダムに 1 ユーザ選択し, 遷移する.
- そのユーザのフォロイーから 1 ユーザランダムに選択し, 遷移する.
- 以上のような, 奇数回目の遷移はリンクと同方向に, 偶数回目の遷移はリンクと逆方向に遷移することを繰り返す. ただし, 各遷移において確率 $j$ でその遷移をストップし, スタート地点へと戻る.

この手法を採用する意図は以下のようにまとめられる. すなわち, 各ターゲットユーザは自身と同じようなユーザをフォローしているユーザに関心を持つ可能性が高い. これは, ユーザが自身と類似したトピックに興味を持つユーザ, すなわち自身との類似度が高いユーザをフォローしたいと考えることが前提となっている. また, 自身との類似度が高いユーザがフォローしており, 自分はフォローをしていないユーザをフォローしたいと考える可能性が高い. この 2 種類のユーザはそれぞれハブとオーソリティーに対応する. 本研究では基本的にオーソリティー度とハブ度を加算した値に基づいて推薦を行う. ただし, 良いオーソリティーをフォローする傾向にあるか良いハブをフォローする傾向にあるかはユーザごとに異なるため, 単純な加算は不適切であるとも考えられる. 節 3.3 ではオーソリティー度とハブ度をどう重みづけるかについての手法を提案する.

### 3.2 新規辺を優先する手法と辺の活性度の推定

2.2 節で説明した SES では, 複数時刻のグラフを一つの重み付きグラフに集約する. 各時刻における隣接行列 $P_t$ を式

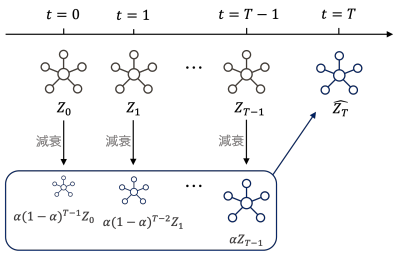


図 2 既 存 手 法

$$P = \alpha P_{t-1} + \alpha(1-\alpha)P_{t-2} + \cdots + \alpha(1-\alpha)^{T-1}P_{t-T}$$

にしたがって集約し, 単一の隣接行列を得る. この行列に対し, 2.1.1 項で示すようなランキングアルゴリズムを適用し, 推薦ユーザのリストを得る. 図 2 に, 既存手法である SES の概略を示す. SES は各時系列のグラフを過去に遡るに従って減衰し, それらの総和を取ることで時系列を集約した単一のグラフを得る手法である.

SNS のフォローグラフは, 論文の共著ネットワーク等のインタラクションを辺とするネットワークとは異なり, 一度生成された辺を消すためにはフォローを解除するという別のインタラクションが必要となる. この点で SNS のグラフは累積的な側面があると言える. グラフの時系列データにおけるリンク予測問題の研究 [8] では実験に論文の共著ネットワークが用いられているが, 論文の共著ネットワークは各期間ごとに論文を共著したかどうかを表現するものである. この共著ネットワークは, 新規に論文を共著しなければ一度生成された辺はその後の時系列のグラフで消滅する. 両者の性質が異なる以上, SES をそのまま SNS のフォローグラフに適用するのでなく, 累積的なグラフであることを念頭に入れて改善した手法を適用することにより, 性能の向上が期待できる.

SES を SNS のグラフにそのまま適用する場合, 辺が消滅しないという仮定の元では, 昔からある辺ほど大きい重みが割り付けられる. 例えば時刻 $t$ と時刻 $t+1$ で生成した辺 $e_1$ と $e_2$ があるとき, 辺 $e_1$ が時刻 $t+1$ で残存している限り $e_2$ よりも $e_1$ の方が大きい重みが割り付けられることになる. SES の時間的な減衰の効果はその差分が小さくなることのみである. しかし, 時間とともにユーザの興味が移り変わっていくという SNS の特性を反映させるためには, 新しい辺ほど大きい重みが割り付けられていて然るべきである. そこで, グラフの各辺について, 昔から存在する辺であればそれだけ重みを小さくしたような単一のグラフを構成することを考える. 具体的には, ある時刻で辺の存在した 2 頂点間の辺の重みを以下のように設定する.

$$w_{i,j} = \alpha^{(T-t_{min})} \tag{2}$$

ここで, $T$ は現在の時刻であり, $t_{min}$ は $i$, $j$ 間に辺が存在する最初のグラフの時刻である. $0 < \alpha < 1$ を満たす場合, $t_{min}$ が小さい値であればそれだけ $w_{i,j}$ は小さくなる. この重みを要素に持つ行列 $P$ に対してランキングアルゴリズムを適用し, 推薦リストを算出するような手法を考えることができる.

新規辺の優先による推薦では, 辺の生成された時刻によって

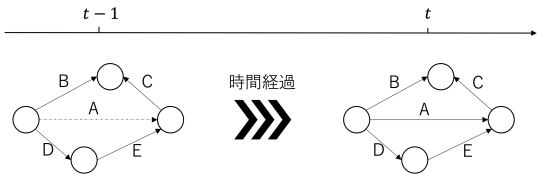


図 3　辺の活性度の推定

辺の重みが決定される．すなわち, 新規辺の優先による推薦は生成した時刻が古い辺を一律で不活性として扱っていることとなる．しかし実際のネットワークにおいて, 辺の生成時刻と辺の活性度とは必ずしも一致しない．ユーザ同士の関係性は時間の経過とともに変化しうるが, ユーザの興味が変化せず活発にインタラクションがされている古い辺の存在が想定される．そこで, 各辺が活性度の高い辺であるかどうかを推測し重みを更新することにより, より適切に辺に重みをつけることができるのではないかと考えられる．

グラフ構造のデータから活性度を推測する手法として, 各ユーザが新規にフォローしたユーザの情報を用いるものが考えられる．各ターゲットユーザが新規にあるユーザをフォローしたとき, そのフォローの理由はターゲットユーザと新規フォロイーの近傍に存在するユーザに見出すことができる．すなわち, あるユーザを媒介にターゲットユーザがその新規フォロイーを発見したと考え, その媒介となるユーザとターゲットユーザおよび新規フォロイーとを結びつけている辺は活性化していると見なせる．本研究では新規にフォローしたユーザへの距離 2 以下の全経路を活性化した辺として更新する手法を考える．各時刻 $t$ において, 時刻 $t-1$ に存在しなかった全ての辺 $e=(u_1,u_2)$ を走査する．時刻 $t$ の隣接行列 $P$ に対して $P_{u_1,x}=1$ かつ $P_{x,u_2}=1$ または $P_{u_1,x}=1$ かつ $P_{u_2,x}=1$ となる全ての $x$ を考え, このような頂点の組 $(u_1,x)$ および $(x,u_2)$ に対して最終的な重み付き行列 $P$ の重みを $\alpha^{(T-t)}$ で更新する．図 3 に辺の活性度の推定の例を示す．図 3 では, 辺 B, C, D, E が存在しており, 時間の経過により辺 A が生成した状況を考えている．辺 A の生成は辺の組 (B,C) および (D,E) に起因すると考え, A の生成時にこれら四つの辺が活性化していると仮定する．新規辺の優先による推薦では四つの辺は生成時刻に応じた重みが付加されていたが, 本手法では辺 A の生成された時刻に応じた重みに更新される．

### 3.3　オーソリティー度とハブ度の適切な重み付け

従来のオーソリティー度やハブ度を考慮する手法では, あくまで主体はオーソリティーであり, ハブはオーソリティーを見つけるための経由地点としての役割しか持たなかった．SNS のグラフでオーソリティー度だけでなくハブ度も考慮して推薦に反映させることの根拠は, 以下のようにまとめられる．

- Twitter におけるリツイートのように, 他ユーザの投稿を自身をフォローしているユーザのタイムラインに表示させる機能が提供されている SNS では, オーソリティーそのものよりもオーソリティーのコンテンツをフィルタリングし良いものだけ表示してくれるハブをフォローするほうが手っ取り早い場合がある．
- 良いハブを見つけることは, 良いオーソリティーを見つけることより難しい．

3.1 節で示したランキングアルゴリズムでは, 各ターゲットユーザにパーソナライズされたオーソリティーのスコアおよびハブのスコアが得られる．今までの節で述べた手法は全てそれらを単純に加算することにより最終的な推薦のスコアを得ることを前提としていた．しかし, ターゲットユーザがどれだけ良いオーソリティーをフォローする傾向にあるか, あるいはどれだけ良いハブをフォローする傾向にあるかの比率を考慮することにより, ユーザごとの特性に着目することができ, ユーザ推薦の性能が向上する可能性がある．

具体的には, $i$ 番目のユーザのオーソリティー度を $a_i$, ハブ度を $h_i$ とおくとき,

$$S_i = w_1 a_i + w_2 h_i$$

を最終的な推薦のスコアとすることを考える．$w_1$ および $w_2$ は重み付けの係数であり, $w_1>0, w_2>0$ を満たしていることが好ましい．オーソリティー度とハブ度の単純な和を推薦スコアとして利用する場合は, $w_1=w_2=0.5$ と表現される．

グラフの時系列データが用意されているとき, 過去のユーザのフォロー動作の履歴から最適な $w_1$ および $w_2$ を学習することが可能である．現在の時刻 $t$ における推薦ユーザを算出したいとき, 現在の一つ前の時刻 $t-1$ から現在にかけてターゲットユーザ $u$ がフォローしたユーザの集合 $S_t$ を用いて学習を行う．時刻 $t-1$ までのグラフの時系列データから各ユーザに対するオーソリティー度 $a$ とハブ度 $h$ が計算され, $a, h, S_t$ から適切な $w_1, w_2$ を得ることが目標となる．

最も単純には $\Sigma_{s\in S_t} w_1 a_s + w_2 h_s$ を最大化するような $w_1, w_2$ を得ることが考えられるが, この場合の解は $w_1=1, w_2=0$ または $w_1=0, w_2=1$ のいずれかとなることが原理的に示される．本研究では最適化は行わず, $S_t$ 内の全ユーザにおけるオーソリティー度とハブ度の中央値を $w_1, w_2$ とする手法, および $S_t$ 内の全ユーザにおけるオーソリティー度とハブ度の上位のユーザの中央値を $w_1, w_2$ とする手法の二つを考える．後者は, $S_t$ 内の全ユーザのオーソリティー度のランキング $a_1, a_2, \cdots, a_{|S_t|}$ およびハブ度のランキング $h_1, h_2, \cdots, h_{|S_t|}$ が与えられたとき, その上位の $k$ 番目の値である $a_k$ および $h_k$ をそれぞれ $w_1$ および $w_2$ に用いる手法であり, 正確にはオーソリティー度とハブ度の上位 $2k-1$ 番目までの値の中央値ということになる．本研究における実験では $k=5$ に設定する．これらの値を重みに用いることに数学的な妥当性はないが, 例えばオーソリティー度がハブ度よりも全体的に高くなる傾向があれば $w_1$ を大きくし, より大きな重みをつけているという点においては妥当である．

## 4　実　　験

### 4.1　実験の流れ

データには代表的な SNS の一つである Twitter のフォローグラフを用いる．およそ一週間ごとに合計 5 回取得された有向グラ

フであり, 頂点数は 1,791,726, 辺の数は最終時点で 519,011,964 本である. 本研究では短期間の時系列と長期間の時系列それぞれについての実験を試みる. 前者では 2021 年 1 月中に取得した 5 時刻のデータのうち, 最後のデータを除く 4 回を推薦リストを得るための学習用データとし, 最後の 1 回を適合率の計算のためのテストデータとする. 長期間の時系列を想定した時系列データには, これらの短期時系列データを加工することによって得られる擬似的なものを用いる. 5 時刻のデータのうちの前から 4 番目の時刻のフォローグラフデータを, 以下の過程に従って生成する.

（1） 4 番目の時刻のフォローグラフのデータにおいて, 各ユーザのフォロイーのうちフォローが新しいユーザの 10%を削り, 残った 9 割のフォロイーはフォローしているものとしたフォローグラフを, 3 番目のフォローグラフとする.

（2） 同様にフォロイーから 10%ずつ削り, 2 番目および 1 番目のフォローグラフを生成する.

この過程は, 全ユーザが同一時刻に一定の割合でフォローをし続けて現在に至るという仮定を置いた上で得られる擬似的な時系列である. 得られた Twitter のフォローグラフのフォローにタイムスタンプが付与されておらず, 各ユーザがフォローした順番だけ得られているため, 本研究ではこの手法を採用した. 前者, すなわち生のデータをそのまま用いる手法を Twitter1, 後者を Twitter2 とする.

本研究で行う二つの実験の概要を以下に示す.

#### 4.1.1 実　験　1

既存手法および 3.2 で述べた二つの提案手法について, それぞれ以下のようにして推薦リストを得る.

- 既存手法. $P_t$ を式 1 に従って単一の隣接行列 $P$ に集約する. $P$ に対して 3.1 で述べたランキングアルゴリズムを適用し, 各ユーザのオーソリティ度とハブ度を得る. 最終的な推薦リストは, オーソリティ度とハブ度の和が高かった上位 $k$ ユーザである.
- 新規辺の優先による推薦. グラフを, グラフに登場した時刻が早い辺ほど軽い重みがつくように集約する. 具体的には, $P_t$ を式 2 に従って単一の隣接行列 $P$ に集約する. $P$ に対してランキングアルゴリズムを適用し, 各ユーザのオーソリティ度とハブ度を得る. 最終的な推薦リストは, オーソリティ度とハブ度の和が高かった上位 $k$ ユーザである.
- 辺の活性度の推定. 各ターゲットユーザの新規フォローについてのみ辺の更新を行ったものと, 全ユーザに対して辺の更新を行なったものの 2 手法について各指標を算出する. 前者では, 各時系列におけるターゲットユーザのフォロイーを一つ前の時系列におけるそれと比較し, ターゲットユーザから増加したフォロイーまでの距離が 2 以内の全ての経路における辺をその時系列に生起したものとして重みを更新する. 後者では, 各時系列における全ユーザのフォロイーを一つ前の時系列におけるそれと比較し, 同様のことを行う.

実験 1 では Twitter1 および Twitter2 を用いる. ランキングアルゴリズムにおけるリスタート率は $j = 0.5$ と設定して実験を行う. また, 各時系列の減衰に用いる係数は $\alpha = 0.5$ とする. ターゲットユーザ群は学習データからテストデータにかけて 100 人以上 105 人以下のユーザをフォローした 100 人のユーザとし, 推薦リストの総数は 100 ユーザとする.

#### 4.1.2 実　験　2

3.3 節で述べた, オーソリティー度とハブ度の重み付けによる手法の性能を評価する. 学習データにおける四つの時系列のうち, 前の三つの時系列から得られるオーソリティー度とハブ度のベクトルを $a$ および $h$ とする. $S$ は三つ目の時系列から四つ目の時系列にかけて増加したフォロイーとする. $w_1$ および $w_2$ の計算方法は 3.3 節で述べた通りである. Twitter1 のグラフの時系列データを用い, ターゲットユーザ群は学習データからテストデータにかけて 100 人以上 105 人以下のユーザをフォローし, さらに三つ目の時系列から四つ目の時系列にかけて 20 ユーザ以上をフォローしたユーザ 59 人とする. ターゲットユーザに制約を課しているのは, オーソリティー度とハブ度の重みの学習のためである.

### 4.2 評　　価

本研究の実験では, 以上の実験により得られた推薦リストに対して, 以下の尺度を計算する.

- 精度. 推薦リスト中のユーザのうち, 時刻 $T-1$ から $T$ にかけて実際にターゲットユーザがフォローしたユーザの数の平均.
- Coverage. 推薦システムが各ユーザに一度でも推薦したことのあるユーザの総数.
- 発見容易性. 推薦システムが提示した全てのユーザについて, その推薦されたユーザを発見する容易性の推定値.

発見容易性はグラフ構造のデータを用いて推定する必要がある. 各ユーザがターゲットユーザから見てどれだけ発見しやすいかを数値化する際, それは以下の主張を反映したものであるべきである.

- ターゲットユーザがフォローしている, あるいはターゲットユーザをフォローしているユーザは発見が容易である.
- ターゲットユーザから近い距離にいるユーザは, 比較的発見が容易である.
- フォロワーの多いユーザは発見が容易である.

これらを反映させる手法として, 本研究では発見容易性を以下のようなランダムウォークによって推定する.

- 時刻 $T-1$ におけるグラフの辺を全て無向辺に置き換えたグラフを用いる.
- ターゲットユーザを始点として, 各遷移で辺によって接続されたユーザから一人を等確率で選ぶ.
- 確率 $j = 0.5$ でターゲットユーザへと戻る.

このようなランダムウォークにおけるランダムサーファーの各頂点への分布は, 無向グラフにおける正規化された隣接行列 $P$ を用いて

$$x^{(k)} \leftarrow (1-j)Px^{(k-1)} + jr$$
$$k \leftarrow k+1$$

という計算を繰り返すことで計算される. $r$ はターゲットユーザ

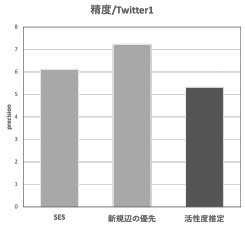


図 4　実験 1(精度, Twitter1)

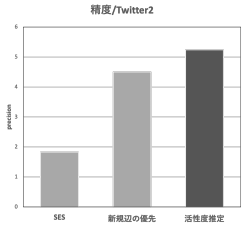


図 5　実験 1(精度, Twitter2)

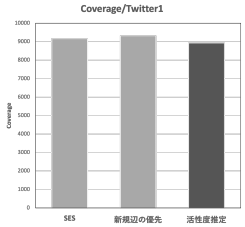


図 6　実験 1(Coverage, Twitter1)

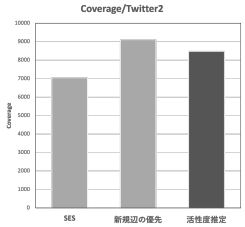


図 7　実験 1(Coverage, Twitter2)

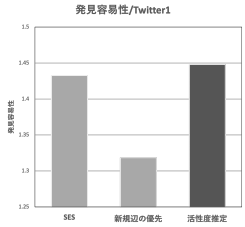


図 8　実験 1(発見容易性, Twitter1)

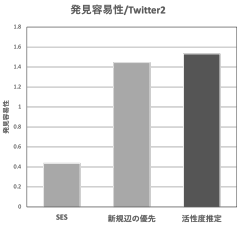


図 9　実験 1(容易性, Twitter2)

のみ 1 で他は 0 である列ベクトルである. 最終的な発見容易性はこの繰り返し計算が収束したときの $x$ の値である. この繰り返し計算によって得られるユーザごとの値は, ターゲットユーザのフォロイーおよびフォロワーであれば大きくなり, ターゲットユーザから遠ざかるにつれて小さくなるという点で一つ目および二つ目の主張を反映している. 三つ目の主張についても, フォロワーの多いユーザはランダムウォークで到達する確率が高くなることにより反映されると考えられる.

発見容易性は 2.3 項で述べた Unexpectedness や Serendipity と関連する. 精度が同じであっても, ターゲットユーザが発見しにくいようなユーザを推薦するシステムはそうでないシステムより価値があるものと考えられる. 推薦全体での発見容易性は, 本研究では推薦リスト内のユーザの発見容易性の総和とする. この値は, 発見が容易なユーザを推薦する傾向にあるか, あるいは発見が困難なユーザを推薦する傾向にあるかを表す値である. 基本的には発見が容易でないユーザを多く推薦する傾向のある推薦システムが好ましいものの, ターゲットユーザからの隔たりが大きいユーザを推薦する傾向にあることは精度の低下をもたらす. 発見容易性の推定値は 適切な小ささの値であることが望ましい.

## 5　実 験 結 果

### 5.1　実　験　1

図 4 から図 9 までに, 実験 1 の結果を示す.

二つのグラフにおける推薦リストについて, 新規辺の優先は既存手法を上回る精度を獲得している. 辺の活性度の推定については二つのグラフで実験結果に差が出ている. Twitter1 のグラフの時系列データを用いた推薦では, 活性度を推定することによる性能の向上は見られなかった. 一方で Twitter2 のグラフの時系列データを用いた推薦では, 他三つの手法と比較して辺の活性度の推定による手法が高い精度を獲得した. 辺の活性度を推定する手法が Twitter1 のグラフのデータではうまく作用しなかった一方で Twitter2 のグラフのデータで良い結果を残したことは, 二つのグラフのデータ間の期間に差があることに原因があると推察される.

辺の活性度の推定は, ある時刻で生成された複数の辺をいかに差別化するかという問題に対処したものである. 例えば時刻 $s$ から $s+1$ にかけて生成した 2 本の辺は, 他の手法であれば重みに差をつけることができない. 実際は 2 本の辺のうち 1 本は時刻 $s+1$ から継続的にユーザが興味を持ち続けている辺で, もう 1 本は長い時間が経過したためにユーザが興味を失ってしまっている, という場合がある. この両者を区別するべく重み付ける手法が辺の活性度の推定である.

Twitter1 のグラフの時系列は約 1 週間ごとに収集したデータであるため, 各時刻ごとのユーザの興味に差はそこまで生じず, 最初の時刻で活性化されていた辺の活性度が失われてしまうことが考えにくい. Twitter2 のデータは長期間の時系列データを想定したものであり, 長期間の時系列データにおいてはかつて生成された辺への興味が現在では失われてしまっている可能性が十分に考えられる. Twitter1 のデータと Twitter2 のデータにおける辺の活性度の推定手法の精度の差は, 以上の時系列データとしての性質の差を反映したものではないかと推察される.

Coverage は多様性を数値で表現した指標である. SES および辺の活性度の推定と比較して, 新規辺を優先する手法の推薦リストは優れた Coverage を示している. 発見容易性は推薦リストの意外性を表現する値であり, 値が低ければそれだけ意外な推薦をしていることとなる. 発見容易性に関しては, Twitter1 のデータでは新規辺の優先は良い値を示している一方, Twitter2 のデータでは新規辺の優先・辺の活性度の推定は SES より悪化している.

### 5.2　実　験　2

図 10 から図 12 までに, それぞれ精度, Coverage, 発見容易性を示す.

オーソリティー度とハブ度を加算する通常の手法と比較して, それぞれの中央値を重みとする手法は精度がわずかに下がり, 5 番目の値を重みとする手法は精度がわずかに改善した. ユーザにどれだけ適合した推薦が可能であるかという点で, 単純な加算よりオーソリティー度とハブ度の重み付けを工夫する手法の方が優れている可能性が示唆されたと言える. しかし, 実験 1 で得られた結果の差と比較すると, これらの差はいずれも小さいものである. オーソリティー度とハブ度を適切に重み付ける

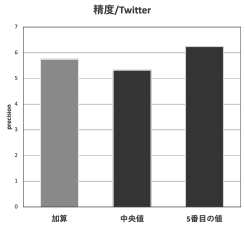


図 10　実験 2(精度)

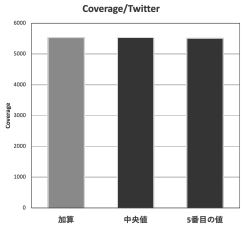


図 11　実験 2(Coverage)

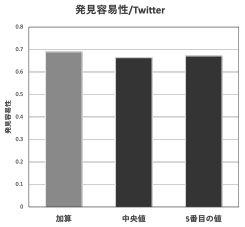


図 12　実験 2(発見容易性)

ことによって性能の向上が期待できる可能性は残されているが, 本実験では重みを考慮しない方法と比較してほとんど差が生じなかったと結論づけられる.

## 6 終わりに

本研究では SNS におけるグラフ構造のみを用いたユーザ推薦手法をいくつか提案し, 二つの SNS におけるグラフの時系列データを用いて実験を行なった. 新規辺を優先する手法は既存手法である SES より優れた性能を示した. また, 辺の活性度を推定する手法を提案したが, こちらは一部のデータにおいてのみ優れた性能を示した. 辺の活性度の推定手法が一部のデータでのみ性能の向上が見られたのは, 時系列データのグラフの取得時間の幅によるものであると推察される. 同じく提案手法であるオーソリティー度とハブ度の重みを学習する手法の実験結果は, ユーザごとのフォローの傾向を学習することが性能の向上につながることを示すものではなかった. 多様性や意外性については, 辺の活性度の推定が最も優れた値となった. 新規辺の重みを大きくすることでユーザ推薦の性能が向上することが示され, さらに辺の活性度を適切に推定することによってさらなる性能の向上が見込めることが明らかとなった.

## 7 謝　　辞

本研究は, JST CREST (JPMJCR16E3), JSPS 科研費 21H03446 の支援を受けたものである.

## 文　　献

[1] P. Gupta, A. Goel, J. Lin, A. Sharma, D. Wang, and R. Zadeh, Wtf: the who to follow service at twitter, Proceedings of the 22nd international conference on World Wide Web (WWW ’13), pages 505-514, 2013.

[2] R. Lempel and S. Moran, The Stochastic Approach for Link-Structure Analysis (SALSA) and the TKC Effect, Proceedings of the Ninth International World Wide Web Conference (WWW9), pages 387-401, 2000.

[3] A. Java, X. Song, T. Finin, and B. Tseng, Why we twitter: understanding micrblogging usage and communities, Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis (WebKDD/SNA-KDD ’07). ACM, pages 56-65. 2007.

[4] B. Krishnamurthy, P. Gill, and M. Arlitt, A few chirps about twitter, Proceedings of the first workshop on Online social networks (WOSP ’08). ACM, pages 19-24. 2008.

[5] Jon M. Kleinberg, Authoritative sources in a hyperlinked environment, Proceedings of the 9th ACMSIAM Symposium on Discrete Algorithms, pages 668- 677, 1998.

[6] L. Page, S. Brin, R. Motwani, and T. Winograd, The PageRank citation ranking: Bringing order to theWeb, Technical report, 1999.

[7] J. Kleinberg. Authoritative sources in a hyperlinked environment, Journal of the ACM, pages 604-632,1999.

[8] P. R. D. S. Soares and R. B. C. Prudencio, Time series based link prediction, Proc. Int. Joint Conf. Neural Netw., pages. 1-7, 2012.

[9] P. Adamopoulos , A. Tuzhilin , On unexpectedness in recommender systems: or how to expect the unexpected, Workshop on Novelty and Diversity in Recommender Systems, at the 5th ACM International Conference on Recommender Systems, pages. 11-18, 2011.

[10] D. Fleder and K. Hosanagar, Blockbuster culture ’ s next rise or fall: The impact of recommender systems on sales diversity. Manage. Science, 55(5), pages. 697-712, 2009.