

マイクロブログにおける対話手段と投稿内容に着目した ユーザ推薦に関する研究と分析

岡本 大輝[†] 豊田 正史^{††} 喜連川 優^{†††,††††}

[†] 東京大学情報理工学系研究科 〒 113-0033 東京都文京区本郷 7-3-1 東京大学 情報理工学系研究科

^{††} 東京大学生産技術研究所 〒 153-8505 東京都目黒区駒場 4-6-1 東京大学 生産技術研究所

^{†††} 国立情報学研究所 〒 101-8430 東京都千代田区一ツ橋 2-1-2

E-mail: †{hokamoto,toyoda,kitsure}@tkl.iis.u-tokyo.ac.jp

あらまし ソーシャルネットワークサービス等では、ユーザ推薦によって、他ユーザと効率的に、そして円滑に、コミュニケーションを取ることが可能になる。ユーザ推薦には、ソーシャルネットワークのグラフ構造に着目した手法や、投稿内容に基づく手法等が提案されているが、これらはソーシャルメディアの環境によって様々な併用方法が考えられる。本論文では、マイクロブログサービスの1つである Twitter 上において、ユーザ間の対話関係からグラフを構築し、対話の種類や回数、投稿内容の類似度を活用した Random Walk 手法を提案し、その精度と意義を検証した。
キーワード Micro Blog, Twitter, User Recommendation, Random Walk

1. ま え が き

近年、社会生活における web の重要性が増し、人々は web 上でも社会活動や経済活動を行うようになった。そうした中で、web ユーザに対する情報推薦は、web サービスにおいて非常に重要な技術となっており、その精度を高める研究は数多くなされている。例えばユーザに商品を推薦することで、購買行動が活発になる。また、ユーザにユーザを推薦することで、ユーザ間にコネクションが生じ、サービスの利用率を高めることが可能となる。

一方で、web 自体の肥大化と多様化に伴い、web ユーザの数が増加し、web の使い方も様々に変化・多様化している。そうした中で、推薦の手法もまた、様々なサービスやユーザの目的に適應する必要がある。このような環境の中で、これまで盛んに研究されてきた、ソーシャルネットワークのグラフ構造に着目した手法に加え、グラフに依拠しない様々な指標を併用し、状況や目的に適した推薦手法の提案が必要とされる。

本稿では、ソーシャルネットワークにおけるユーザ推薦を目的とし、マイクロブログサービスである Twitter の API を利用し、そのデータを用いた実験を行う。実験においては、Twitter のグラフ構造と、グラフに因らない特徴として、ユーザが投稿するテキストの内容の両方に着目し、それらを併用する。また、実験の結果や使用したデータを分析し、ユーザがサービスを利用する際の特徴や、グラフ構造との相関性を議論する。そして、ユーザの用途や特徴によって、最適な推薦手法にどのような違いが現れるのか、その考察を行う。

2. 関 連 研 究

近年、Web 上においても、様々なコンテンツやユーザを、効率的に、かつ的確に推薦する手法について、多種多様な研究が行われている。本章では、まず、広く推薦に用いられる手法と

して、コンテンツに着目した手法について論じる。次に、SNS におけるユーザ推薦をリンク予測問題へと帰着させることを示し、ソーシャルグラフに着目した手法について論じる。

2.1 特徴量に着目したユーザ推薦

広く一般的な推薦手法の一つとして、ユーザの特徴量に着目した手法が存在する。推薦の候補（ユーザ推薦の場合はユーザ）の特徴を元に、推薦の対象たるユーザとの親和性の高い候補を推薦するものである。

Hannon らは、Twitter におけるユーザ推薦において、ユーザのツイート内容と、フォロワー/フォローのリストを利用した content based の手法を提案した。彼らは、ユーザのツイートに含まれる名詞のベクトルと、ユーザのフォロワー/フォローのリストを、それぞれユーザの特徴量と見做し、機械学習によって特徴量に重み付けを行い、類似するユーザを推薦した [10]。

Wang らはさらに、マイクロブログの機能とユーザのアクティビティに着目し、他ユーザの投稿の拡散回数や拡散力を特徴量として、ユーザへのツイート推薦を行った [13]。

いずれの実験も結果的に、2-hop(隣接ユーザの隣接ユーザ)までは特徴量による作用が及ぶが、より遠方のユーザに対する特徴的作用に関しては言及していなかった。

2.2 ネットワーク構造を用いたユーザ推薦

SNS におけるユーザ推薦を始めとして、特定の環境下のユーザに対して、同一のサービスを利用するユーザを推薦する際には、一般的にそのネットワーク構造に着目したアプローチが用いられる [17]。ネットワーク構造に着目する場合、ユーザ推薦はネットワーク内のリンク予測問題に帰着する。

以下では、データ構造の表現に則り、推薦候補および被推薦の対象となるオブジェクトをノード、オブジェクト同士のリンク関係をエッジと表現する。また、ノード A とエッジを形成しているノードの集合を $\Gamma_{(A)}$ とする。

2.2.1 グラフ上の近縁性を表す指標

ユーザ推薦においては、グラフ構造を利用した様々な手法が考案されている。

最もシンプルな推薦手法として、2つのノードに共通する隣接ノードの数を数えて、数が多い順に推薦するというものがある [2] [1]. また、共通する隣接ノードの数に、分母として両者の隣接ノードの集合の論理和をとった Jaccard's coefficient という指標がある [2]. 推薦候補に隣接ノードが多い場合、無用にスコアが高くなることを防ぐことができる。さらに、ノード A とノード B に共通する隣接ノードであるノード C の隣接ノードが少ないほど、C による A と B 近縁性が高まるとした Adamic Adar の評価指標がある [3]. Jaccard's coefficient と同じく、多数の隣接ノードを保有するノードによって、近縁性が不当に高まることを抑制することができる。

$$Score(A, B) = \sum_{\Gamma(A) \cap \Gamma(B)} \frac{1}{\log(\Gamma(C))} \quad (1)$$

推薦の対象となるノードから離れたノードも評価できる指標として、ノード A からノード B へ、距離 l で移動する経路の集合 $paths_{A,B}^{(l)}$ を用いた Katz の指標がある [11].

$$Score(A, B) = \sum_{l=1}^{\infty} \beta^l \cdot |paths_{A,B}^{(l)}| \quad (2)$$

一般的に、 β は 0 以上 1 未満の値をとり、 β が小さいほど、A, B 間の短距離経路が重視される。

2.2.2 Random Walk

推薦される対象 (target) となるノードを s として、 s の隣接ノードへとランダムで移動する動点の挙動を考える。動点はエッジに従ってさらに隣接ノードへの移動を繰り返し、最終的に動点が存在する確率が高いノードを s に推薦する [4] [16]. s と同じ距離に存在する複数の候補ノードの中でも、より s に近い対象を推薦する手法として、よく用いられている。

さらに Backstorm らは、facebook において友人関係をエッジとして、エッジの形成時間、インタラクションの回数、友人申請の方向などをエッジの特徴量とした。それらの特徴量を元に機械学習を行い、エッジにウエイトを付与してランダムウォークを行う Supervised Random Walk を提案した [7].

Random Walk を用いた推薦手法は、ネットワーク構造を用いて近接性の高いユーザをランキングすることにより推薦を行うが、これまでに提案された手法では、ユーザの投稿内容までは考慮されていなかった。また、投稿内容を用いた推薦手法では、ネットワーク構造が十分に考慮されていなかった。

3. マイクロブログにおけるユーザ推薦

本章では、マイクロブログの機能と特徴について説明し、本論文の実験の目的について述べる。

3.1 マイクロブログの機能

本論文で扱う Twitter は、マイクロブログと呼ばれるサービスの一つである。マイクロブログは、他のウェブサービスに無

い独特な機能をいくつか備えている。本稿ではそれらの機能について説明する。

• 短文投稿

アメリカの Twitter, 中国の weibo 等のマイクロブログサービスは、ユーザによる短文投稿を主な機能とする。投稿の内容、または投稿すること自体を **ツイート** と表現し、ユーザはあらかじめ決められた文字数の中で、自由かつ手軽にテキストを投稿することができる。

• ユーザのフォロー

マイクロブログのユーザが、特定の他ユーザの投稿を容易に閲覧できるようになる機能である。Twitter を利用した研究には、他ユーザのフォローを推薦するものもある [10].

• 投稿によるインタラクション

マイクロブログには、特定のユーザや投稿に対して特殊な投稿を行うことで、他ユーザとインタラクションを行う機能が存在する。本論文では、このインタラクションを用いたユーザ推薦を行う。

(1) リプライ

投稿文内に、@マークの後にユーザ ID を伴う (@USERID) ことで、そのユーザに対して言及を行う投稿のことを、リプライ (reply) という。リプライを受け取ったユーザは、常時見ているユーザの投稿とは別に、自分へのリプライとしてその投稿に注目することが可能となる。リプライは、ある投稿に対して (その投稿者に対して) 行われる場合や、相手のユーザの投稿と無関係に行われる場合などがあるが、いずれのリプライ投稿にも明確な相手が存在する。マイクロブログの短文投稿という特性から、リプライは同一ユーザ間で短時間に繰り返されることがある。

(2) リツイート

リツイート (retweet) とは、他のユーザの投稿を自らの投稿と同様の扱いで再投稿・拡散する機能である。あるユーザが別のユーザの投稿をリツイートした際には、投稿主のユーザに対してリツイートしたユーザを通知する仕組みがある。リツイートは、ユーザが興味を持った話題や意見を、自身をフォローするユーザに拡散する場合に行われるほか、投稿主に対する対話や意思表示の手段として行われる場合もある [14] [15].

3.2 ユーザ推薦における投稿内容の活用

マイクロブログは短文投稿を主な機能とするサービスである。それと同時に、投稿を用いて他のユーザとコミュニケーションを行うこともできる。他のユーザとのコミュニケーションが可能であるため、ユーザを推薦する技術も存在し得る。現在、Twitter では他ユーザを推薦する機能が実装・提供されている。

しかしながら、マイクロブログの主たる機能である短文の投稿を活用したユーザ推薦の研究は多くない。投稿内容の活用がマイクロブログニオケルユーザ推薦において有効であることは、Hannon らの研究 [10] でも示されているが、フォローユーザの推薦に留まり、限定的なものであった。

そこで本論文では、Twitter におけるユーザ推薦を目的とし、ソーシャルネットワークとユーザの投稿内容を併用したユーザ推薦の実験を行う。その後、実験の結果を検証し、投稿内容に

基づいた推薦が有効に作用する状況や用途について考察を深める。

4. 提案手法

本章では、マイクロブログにおいて投稿内容を活用したユーザ推薦の有用性を検証を目的として、今回行った実験の詳細を説明する。

4.1 Weighted PageRank

本論文では、Random Walk による推薦手法を基にして、投稿内容も考慮するように拡張した推薦手法を提案し、実験と検証を行う。

本手法では、推薦対象となるユーザ s を起点とし、他ユーザとのリンクをたどる Random Walk を行い、各ステップにおいて一定の確率で s に戻る、Random Walk with Restart 手法 [4] を基にする。また、遷移確率の決定にあたり、ユーザ間のインタラクション回数や投稿内容の類似度などを考慮して、遷移確率に重み付けを行う。遷移確率行列を M とすると、 k 回遷移後の存在確率のベクトル \vec{r}_k の更新式は、以下で表わされる。

$$\vec{r}_{k+1} = (1 - \alpha)M\vec{r}_k + \alpha\vec{r}_0 \quad (3)$$

ただし、 \vec{r}_0 は、ユーザ s のみ 1 で残りは 0 となるようなベクトルである。 α は、各ステップにおいて s に戻る確率を表し、今回の実験では Supervised Random Walk 手法 [7] において用いられていた 0.3 を値として用いた。

遷移確率の決定にあたり、インタラクションに係る 2 種類と、投稿内容の類似度に係る 3 種類の、計 5 種類の特徴量を用いる。それぞれから算出したウエイトと、適切に選んだ 2 種類の複合ウエイトを、各エッジのウエイトとして実験を行う。

4.1.1 インタラクション回数を用いたウエイト

ユーザ間のインタラクション回数は親密性を表すため、回数が多いほど高い確率で遷移するような重み付けを行う。ユーザ間で行われたインタラクションの回数を元に、全てのエッジにリプライ、リツイートに基づいた 2 種類の重みを設定する。

ユーザ u 、ユーザ v 間のリプライ回数を $n_{reply(u,v)}$ 、リツイート回数を $n_{retweet(u,v)}$ として、ウエイト $w_{i-reply(u,v)}$ と $w_{i-retweet(u,v)}$ を以下のように定義する。

$$w_{i-reply(u,v)} = 1 - \frac{1}{1 + n_{reply(u,v)}} \quad (4)$$

$$w_{i-retweet(u,v)} = 1 - \frac{1}{1 + n_{retweet(u,v)}} \quad (5)$$

4.1.2 投稿内容の類似度を用いたウエイト

訓練期間でリプライまたはリツイートによってエッジが形成された 2 ユーザで、両ユーザの $tf \cdot idf$ ベクトルのコサイン類似度を、投稿内容の類似度に基づくウエイト $W_{t(u,v)}$ として定義する。

$$w_{t(u,v)} = \cos(\vec{t}_{(u)}, \vec{t}_{(v)}) \quad (6)$$

なお、インタラクションの有無によって作成された 3 種類の

$tf \cdot idf$ ベクトルを、それぞれ式 6 の \vec{t}_u, \vec{t}_v とすることで、以下の 3 種類のウエイトを生成する。

- 全ての投稿の類似度に基づくウエイト $w_{t-all(u,v)}$
- リプライ投稿の類似度に基づくウエイト $w_{t-reply(u,v)}$
- リツイート投稿の類似度に基づくウエイト $w_{t-retweet(u,v)}$

4.1.3 複合ウエイト

上記のインタラクションを用いたウエイト 2 種類、投稿内容の類似度を用いたウエイト 3 種類から、それぞれ 1 種類ずつを用いた複合ウエイトを作成し、推薦に利用する。インタラクションを用いたウエイトを $w_{i(u,v)}$ 、投稿内容の類似度を用いたウエイトを $w_{t(u,v)}$ とすると、遷移確率行列の要素 $w_{(u,v)}$ は、以下の式で定義される。

$$w_{(u,v)} = \begin{cases} (1 - \beta)w_{i(u,v)} + \beta w_{t(u,v)} \\ 0 \quad (u, v \text{ 間にエッジが無い場合}) \end{cases} \quad (7)$$

こうして定義された遷移確率行列において、あるユーザがエッジを形成するユーザに対して、これらの重みに比例した確率で遷移を行うよう正規化を行う。

なお、今回の実験では、複合でない 5 種類のウエイトを用いた実験を行い、その結果を元に、 (W_i, W_t) の組み合わせを決定する。

4.2 データセット

Twitter 社が提供する API を用いて、ツイートを公開しているアカウントからツイートを収集した。著者の使用している Twitter アカウントを起点として、ソーシャルグラフ上で近傍に存在するユーザがフォローしているユーザを順次集めて、最終的に 250,000 ユーザの tweet データを取得した。その中で、特に 2013 年 8 月から同年 9 月まで (以下、訓練期間) のデータと、2013 年 10 月 (以下、テスト期間) のデータを抽出した。

4.2.1 グラフの構築

今回の実験は、マイクロブログサービスでのユーザ推薦の精度の向上と、着目するインタラクションの違いによる、投稿内容を活用した推薦の差異の分析を主目的とする。Twitter におけるユーザ同士のインタラクションは、いくつかの形態が存在するが、今回の実験では、リプライまたはリツイートのうち、いずれかが 2 ユーザ間に存在する場合に、エッジを形成するものとした。

4.2.2 ユーザの分類

構築したグラフをもとに、以下の条件を満たすユーザを推薦の候補ユーザとする。

- 自動投稿アカウントではない
- 訓練期間で 100 件以上のツイートを取得できた
- 訓練期間でデータセット内のユーザとインタラクションをしている

さらに、候補ユーザのうち、テスト期間において

- 新たに 10 以上のユーザとリプライを交わす
- 新たに 10 以上のユーザとリツイートを交わす

以上の2条件を満たすユーザを、積極的にインタラクションを行うユーザとして、ランダムに1000ユーザを抽出し、推薦の対象ユーザとした。

4.2.3 $tf \cdot idf$ ベクトルの作成

訓練期間において、各ユーザが投稿したツイートに形態素解析を行い、ユーザ毎に名詞セットを抽出した。リプライとリツイートは投稿によって行われることから、名詞セットにその特性を反映させるために、ツイートをインタラクションの有無で分類し、以下の3種類のそれぞれ名詞セットを抽出した。

- 全てのツイート内容から抽出した名詞セット
- リプライツイート内容から抽出した名詞セット
- リツイートの内容から抽出した名詞セット

また、各名詞セットから $tf \cdot idf$ ベクトルを生成し、各ユーザの特徴量とする。ユーザ v における名詞 $noun$ の $tf \cdot idf(v, noun)$ は、以下の式で表される。

$$tf \cdot idf(v, noun) = \frac{count(v, noun)}{\sum_N count(v, N)} \cdot \log \frac{|U|}{|u : u \ni noun|} \quad (8)$$

なお、 $count(v, noun)$ は、ユーザ v の一つの名詞セットに含まれる $noun$ の数、 $|U|$ は候補ユーザ数、 $|u : u \ni noun|$ は名詞 $noun$ を発していた候補ユーザの数である。

これにより、各ユーザごとに以下の3種類の $tf \cdot idf$ ベクトルを生成した。

- 全てのツイートから作成した $tf \cdot idf$ ベクトル \vec{v}_{all}
- リプライツイートから作成した $tf \cdot idf$ ベクトル \vec{v}_{reply}
- リツイートから作成した $tf \cdot idf$ ベクトル $\vec{v}_{retweet}$

4.2.4 グラフとユーザの分析

以上の条件で構築したグラフと、ユーザの分析を行った。

4.2.5 エッジに関する分析

図1に、訓練期間、テスト期間でそれぞれ構築したグラフのエッジ数を示す。実験においては、テスト期間に新たに生成されるエッジを正例エッジ、正例エッジに隣接することになるユーザを正例ユーザとして、推薦の精度の検証に用いる。

また、候補ユーザと対象ユーザのそれぞれで、インタラクションの種類別に隣接ユーザ数を算出した。その結果を表2に示す。

集計の結果から、

(1) テスト期間のリプライの6割以上は訓練期間でインタラクションを行ったユーザ同士のものである

(2) リツイートは交流の無かったユーザ同士でも多数行われている

という2点を述べるができる。また、選択した対象ユーザが平均より活発にインタラクションを行っていることもうかがえる。

4.2.6 対象ユーザを起点とした経路探索

訓練期間のグラフ上で、1000人の対象ユーザを基点として経路探索を行い、候補ユーザまでの距離と、正例ユーザまでの距離を分析した(図1, 表3)。

併せて、候補ユーザまでに至るまでの経路に存在するユーザ

表1 各期間のエッジ数

	訓練期間	テスト期間	テスト期間に新たに生成されるエッジ
ユーザ数	209,743		
リプライによるエッジ	2,569,214	1,450,987	454,361
リツイートによるエッジ	8,521,830	3,866,848	2,339,750
全てのエッジ(重複を除外)	10,134,638	4,889,299	2,552,946

表2 ユーザ分類別次数

	候補ユーザ	対象ユーザ
ユーザ数	161927	1000
リプライによる平均隣接ユーザ	27.45	86.04
リツイートによる平均隣接ユーザ	94.16	156.16
平均隣接ユーザ	111.10	210.45

の、リプライ回数とリツイート回数の平均値を計算し、さらに、対象ユーザから各候補ユーザに至る最短経路において、経路上の各エッジの両端に存在するユーザの $tf \cdot idf$ ベクトルのコサイン類似度も計算した(表4)。ユーザ A 、ユーザ B の $tf \cdot idf$ ベクトルのコサイン類似度は、以下の式で表される。

$$\cos(\vec{v}_{(A)}, \vec{v}_{(B)}) = \frac{\vec{v}_{(A)} \cdot \vec{v}_{(B)}}{|\vec{v}_{(A)}| |\vec{v}_{(B)}|} \quad (9)$$

なお、分析に用いる $tf \cdot idf$ ベクトルは、全てのツイートから作成したもの (\vec{v}_{all}) とした。2ユーザの $tf \cdot idf$ ベクトルのコサイン類似度は、ユーザ同士の投稿内容の類似度を示す。

図1と表3から、今回構成したグラフでは3hopに候補ユーザのピークが存在するが、リプライ、リツイートのいずれの正例ユーザも、訓練期間においては約90

また、表4より、正例ユーザへ至る経路上では、平均より多くのインタラクションが行われている傾向があり、特にリプライで顕著であることがわかる。さらに、正例ユーザへの経路では、平均より投稿内容類似度が高くなることがわかる。

以上から、ユーザ推薦の実験をするにあたり、今回構築したグラフの構造を参照し、近傍のユーザを推薦することの有意性が示された。また、ユーザ同士のインタラクションの回数と、投稿内容の類似度に着目することで、推薦の精度が向上する可能性があることが示された。

5. 実験と結果の考察

5.1 推薦の精度の評価方法

本稿では、Twitterのユーザに対し、その目的に応じて最適なユーザ推薦を行うことを目的とする。グラフを活用したユーザ推薦では、テスト期間のデータから作成されたグラフで新たに得られるエッジを正例エッジとして、手法に基づく予測と比較して精度を検証することが一般的であるが、今回の実験では、正例の条件を

- (1) リプライによるもの
- (2) リツイートによるもの

対象ユーザから見た候補ユーザの分布

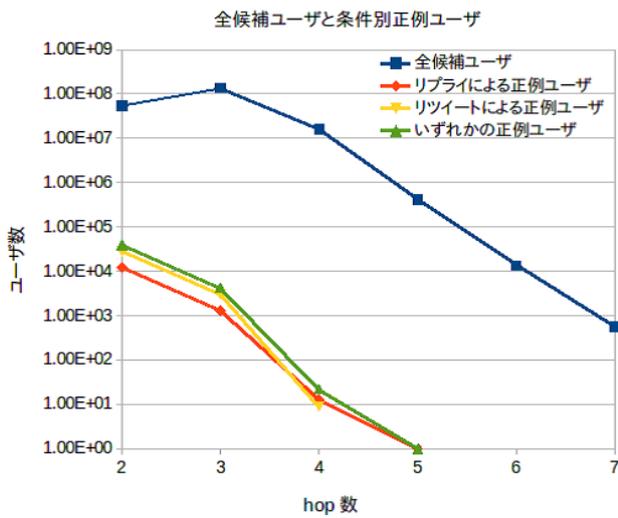


図 1 対象ユーザと候補ユーザの経路距離とユーザ分布

表 3 対象ユーザを起点とした候補ユーザの分布

	hop 数による累積率 (%)				平均総数
	2	3	4	5	
全候補ユーザ	24.343	92.206	99.795	99.993	209,742
リプライによる正例ユーザ	90.239	99.895	99.993	100	13.391
リツイートによる正例ユーザ	90.161	99.971	100	100	31.426
全ての正例ユーザ	89.999	99.946	99.998	100	42.428

表 4 候補ユーザへの経路上のエッジ特徴量

	全候補ユーザ	リプライによる正例ユーザ	リツイートによる正例ユーザ
リプライ回数	2.2428	7.5477	3.6453
リツイート回数	2.5136	2.7665	3.2995
投稿内容の類似度	0.0303	0.0429	0.0478

(3) それらのいずれか

の 3 パターンで変え、各ウエイトによる精度の変化と最適な推薦手法を検証する。

エッジに加える様々なウエイトを、それぞれ変化させながら、Prec@10, Mean Average Precision の 2 つの指標を用いて、推薦の精度を評価する。また、Mean Average Precision の説明の過程で、Average Precision にも言及する。

- Precision@10 (Prec@10)

1 ユーザに対する推薦候補として、スコアが高くなった上位 10 ユーザの中での精度。今回のデータでは、推薦のターゲットユーザがテスト期間に新たに 10 以上のエッジを形成するため、推薦の精度の評価としてこの値を使用する。

- Average Precision (AP)

1 ユーザに対する推薦候補のスコアランキングを上位から検証

表 5 各ウエイト単独による推薦 (Precision@10)

	正例の条件		
	リプライ	リツイート	いずれかのインタラクション
$W_{i-reply}$	0.0819	0.0490	0.1108
$W_{i-retweet}$	0.0233	0.0785	0.0940
W_{t-all}	0.0484	0.0739	0.1083
$W_{t-reply}$	0.0472	0.0539	0.0878
$W_{t-retweet}$	0.0375	0.0719	0.0979
$W_{baseline}$	0.0537	0.0624	0.1003

表 6 各ウエイト単独による推薦 (Mean Average Precision)

	正例の条件		
	リプライ	リツイート	いずれかのインタラクション
$W_{i-reply}$	0.1093	0.0444	0.0887
$W_{i-retweet}$	0.0347	0.0694	0.0740
W_{t-all}	0.0663	0.0659	0.0873
$W_{t-reply}$	0.0624	0.0502	0.0726
$W_{t-retweet}$	0.0511	0.0635	0.0779
$W_{baseline}$	0.0766	0.0616	0.0892

し、正例となる候補が出現した時点での精度の平均値。

$$AP_{(N_{TRUE})} = \frac{1}{N_{TRUE}} \sum_{K=1}^{N_{TRUE}} \frac{K}{rank\ of\ Kth\ TRUE} \quad (10)$$

但し、 N_{TRUE} は 1 ユーザが有する全ての正例の数である。

- Mean Average Precision (MAP)

推薦候補のスコアランキングを上位から検証し、正例となる候補が出現した時点での Average Precision の平均値。ランキングの上位に正例が集中していると高い値となる。

$$MAP = \frac{1}{N_{TRUE}} \sum_{K=1}^{N_{TRUE}} AP_{(K)} \quad (11)$$

5.2 実験結果

本項ではまず、前章で定めた 5 つのウエイトについて、それぞれ単独で用いて推薦を行った場合の精度を評価する。また、比較の為に全てのエッジのウエイトを均一としたウエイト ($w_{baseline}$) を用いた推薦の結果も記載する。

表 5, 表 6 から、正例となるインタラクションの回数を利用したウエイトを用いることで、ベースラインより良い精度となることがわかる。また、投稿内容の類似度を用いた結果に着目すると、いずれの正例条件であっても全ての投稿内容を用いたウエイトで精度が高くなった。

実験の結果から、正例となるエッジの条件別に 2 種類の複合ウエイトを提案し、式 7 の β の値を変えながら精度を検証した。正例とするエッジの条件、および組み合わせたウエイトの種類は以下のとおりである。

- リプライによるエッジを正例とする場合

リプライの回数に基づくウエイト ($W_{i-reply}$) と、全ての投稿内容に基づくウエイト (W_{t-all}) を用いる。

- リツイートによるエッジを正例とする場合

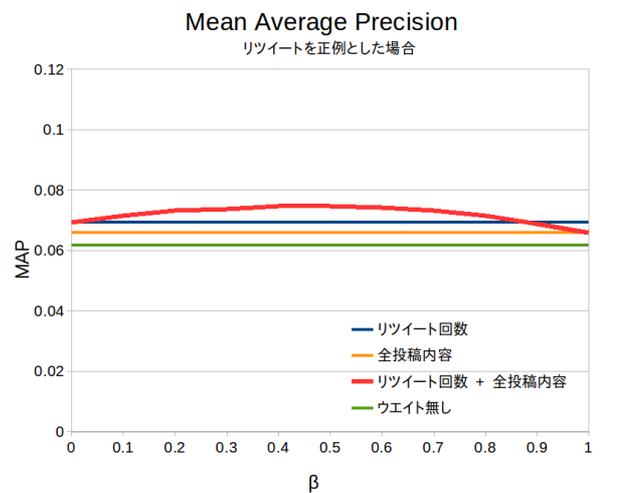
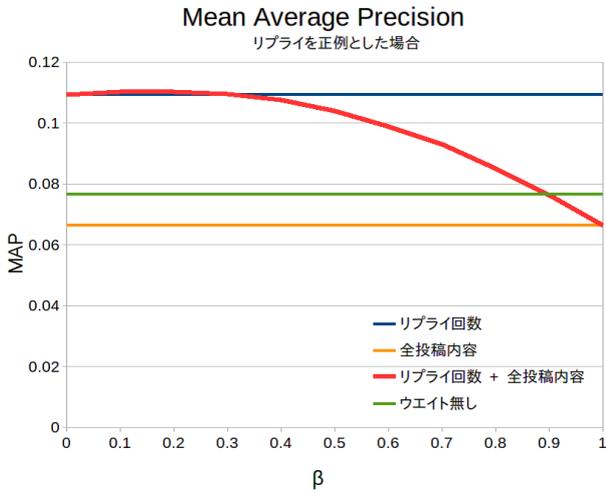
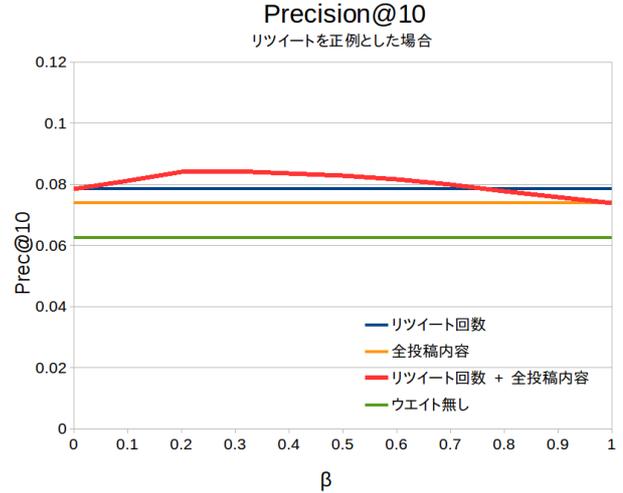
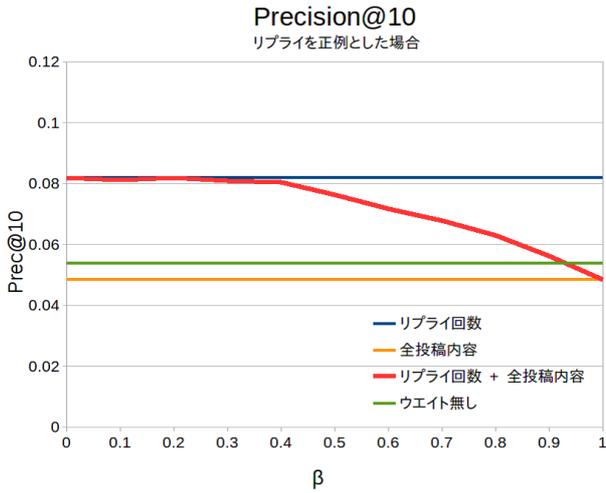


図 2 リプライによるエッジを正例とした際の精度

図 3 リツイートによるエッジを正例とした際の精度

リツイートの回数に基づくウエイト ($W_{i-retweet}$) と、全ての投稿内容に基づくウエイト (W_{t-all}) を用いる。

以下、図 2 にリプライによるエッジを正例とした際の実験結果を、図 3 にリツイートによるエッジを正例とした際の実験結果を、それぞれ示す。いずれも、 β が大きいほど投稿内容の類似度の重要性が高まる。

5.3 実験結果の考察と検証

本項では、前項までの実験結果と 4.2.4 で行った分析結果を元に、各ウエイトの特性や作用、および Twitter ユーザの特徴についての考察を述べる。

- インタクション回数によるウエイト

表 5、表 6 から、正例とするインタクションの回数をウエイトとして用いた場合に、精度が向上することがわかる。一方で、正例ではないインタクションのウエイトを用いても、それぞれ精度が低下してしまっている。

これにより、Twitter のユーザは、リプライとリツイートを明確に違う機能として活用し、頻繁にリプライを交わす相手のツイートを頻繁にリツイートするとは限らないと述べる事ができる。

また、表 5 で、リプライ、リツイートをそれぞれ正例エッジ

の条件とした場合の比較を行う。Precision@10 は一般的に、候補中に占める正例の数が多いほど高くなる傾向がある。表 3 によると、リツイートによる正例ユーザの数はリプライによる正例ユーザの 2 倍以上存在する。しかしながら、均一なウエイト ($W_{baseline}$) を用いた場合には、リツイートを正例とした場合の精度の比は 1.2 未満であり、リプライ回数によるウエイトでリプライを予測する場合と、リツイート回数によるウエイトでリツイートを予測する場合を比べると、リプライを用いた場合が高精度となっている。

このことから、リプライ関係によるユーザの親近性は、リプライ回数に顕著に反映され、将来のリプライを予測する際には、既知のリプライ回数を参照することが有効であると述べる事ができる。

- 投稿内容類似度によるウエイト

ツイート内容の類似度による推薦の結果にも着目する。正例とするインタクションによらず、ツイートを分類せずに全てのツイート内容の類似度を用いた場合に、精度が高くなる事がわかる。ここで、ベースラインとの比較を試みる。リプライによるエッジを正例とする場合には、どの類似度を用いてもベースラインより精度が低下した。一方で、リツイートによるエッジ

ジを正例とする場合には、全てのツイート内容、リツイートの内容のいずれの類似度を用いても、ベースラインより精度が向上している。表4によると、リツイートによるエッジで正例となるユーザへの経路では、リプライによる正例となるユーザへの経路より、隣接ユーザ同士の類似度が高い傾向があった。

以上のことから、Twitterのユーザは、他のユーザからリツイートを行う際には投稿内容を加味しており、その傾向はリプライよりも顕著であると推察できる。

● 複合ウエイト

まず、図2に着目する。いずれもリプライ回数によるウエイトと投稿内容類似度によるウエイトを結合した場合の結果を示しているが、リプライを正例として予測する場合には、投稿内容による改善が見られないことがわかる。一方、図3に着目する。リツイート回数と投稿内容類似度によるウエイトを結合し、リツイートを正例として予測した際の精度であるが、両者を単体で用いた場合より改善していることがわかる。

ここで、複合ウエイトを用いてリツイートを予測する際に見られた改善の有意性を検証するために、Wilcoxonの符号付順位和検定と、4分割交差検定を実施する。

(1) Wilcoxonの符号付順位和検定

各候補ユーザに対し、

- リツイート回数によるウエイトで推薦を行った場合
- 複合ウエイト ($\beta = 0.4$) で推薦を行った場合

以上の2回の推薦で得られた Mean Average Precision の、差の絶対値の順位と分布をもとに、有意性を評価する。

検定にあたり、Mean Average Precision の差による対象ユーザのヒストグラムを示す(図4)。

検定の結果、検定統計量 Z_0 は 8.065、正規分布に従った有意確率は $7.323e-16$ となる。これは、「複合ウエイトを用いた推薦の精度はリツイートの回数によるウエイトを用いた推薦の精度と差はない」とする帰無仮説を棄却するに値する低確率である。

(2) 4分割交差検定

1000人の対象ユーザをランダムに4分割し、3つを訓練事例として、ウエイトの結合割合 β を学習し、残る1つに適用する交差検定4回を行った。得られた結果を表7に示す。

Precision@10, Mean Average Precision のいずれでも、交差検定によって得られた平均値は、結合前のウエイトから得られた値より大きくなった。この検定においても、ウエイトの結合による精度の向上の有意性が示されている。

以上2つの検定から、リツイートを予測する場合に、リツイートの回数だけでなくユーザの投稿内容の類似度を用いると、精度が向上すると述べるができる。一方で、リプライを予測する場合の結果についても検定を行ったが、有意な精度改善は見られなかった。これにより、リプライを予測する際には、リプライの回数だけでなくユーザの投稿内容の類似度にも着目することで、精度が改善することが示された。

6. まとめ

本論文では、ソーシャルネットワークにおけるインタラクシ

リツイートによるウエイトと複合ウエイトの精度差とユーザ分布

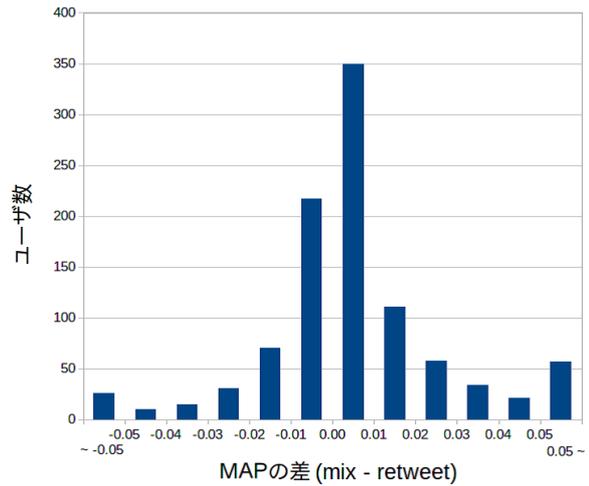


図4 Mean Average Precision の差によるユーザ数

表7 4分割交差検定による結果と比較

	Prec@10	MAP
$W_{i-retweet}$	0.0784	0.0693
W_{t-all}	0.0738	0.0659
$\beta W_{i-retweet} + (1 - \beta) W_{t-all}$ (4分割交差検定)	0.0839	0.0745
$W_{baseline}$	0.0624	0.0616

ンに基づくユーザ推薦の精度の向上を目的として、インタラクションの回数とユーザの投稿内容の類似度に着目した Random Walk に基づく推薦手法を提案した。この手法では、ユーザのインタラクションを元にソーシャルグラフを構築し、インタラクション回数と投稿内容の類似度をエッジのウエイトとして、ウエイトを遷移確率として Random Walk によるユーザ推薦を行った。また、推薦するユーザとの関係がリプライによるものか、リツイートによるものかを区別し、ウエイトに用いる特徴量が各条件下で推薦の精度に与える影響を検証した。

実験の結果、リプライを正例として予測する際には、リプライの回数によるウエイトのみを利用し、投稿内容の類似度によるウエイトを加味することによる精度の向上が見られないことが判明した。一方で、リツイートを予測する際には、リツイートの回数によるウエイトのみではなく、投稿内容の類似度によるウエイトを併せることで、精度が向上した。また、Wilcoxonの符号付順位和検定を行うことで、精度向上が有意であることを示した。

以上のことから、Twitterにおいてユーザが他者とインタラクションを行う際には、インタラクションに応じた特徴があることが示された。まず、ユーザ同士のリプライ関係の強さは、その回数に顕著に表れ、リプライを予測する際には回数を活用することで、精度が向上する。一方、リツイートを予測する際には、その回数のみならず、ユーザの投稿内容の類似度を活用することで、精度が向上する。これにより、ユーザは Twitter において、他者とリプライを交わす際には、相手のユーザとの既存の関係性が重視され、他者のツイートをリツイートする際

には、既存の関係性に加えて投稿の内容が重視されている、と述べることができる。

今回の研究により、インタラクションに応じた特徴量の選択と活用が、よりユーザの需要に即した推薦が可能であることが示された。

文 献

- [1] M.E.J.Newman, Clustering and preferential attachment in growing networks, *Physical Review E - Volume 64*, 2001.
- [2] G. Salton, M. J. McGill, *Introduction to modern information retrieval*, 1986.
- [3] Lada A Adamic, Eytan Adar, Friends and neighbors on the Web, *Social Networks*, 2003.
- [4] Sergey Brin, Lawrence Page, The anatomy of a large-scale hypertextual Web search engine, *Computer Networks and ISDN Systems*, 1998.
- [5] David Liben-Nowell, Jon Kleinberg, The Link-Prediction Problem for Social Networks, *The American Society for Information Science and Technology*, 2007.
- [6] Jonathan L. Herlocker, Joseph A. Konstan, Lorget G. Terveen, John T. Riedl, Evaluating collaborative filtering recommender systems, *ACM Transaction on Information Systems*, 2004,
- [7] Lars Backstorm, Jure Leskovec, Supervised random walks: predicting and recommending links in social networks, *WSDM*, 2011,
- [8] Fouss F., Random-Walk Computation of Similarities between Nodes of a Graph with Application to Collaborative Recommendation, *IEEE Transactions on Knowledge and Data Engineering*, 2007.
- [9] Jure Leskovec, Daniel Huttenlocher, Jon Kleinberg, Predicting Positive and Negative Links in Online Social Networks, *WWW*, 2010.
- [10] John Hannon, Mike Bennett, Barry Smyth, Recommending twitter users to follow using content and collaborative filtering approaches, *RecSys*, 2010.
- [11] Katz L., A new status index derived from sociometric analysis, *Psychometrika* 18(1), 1953.
- [12] Glen Jeh, Jennifer Widom, SimRank: a measure of structural-context similarity, *KDD*, 2002.
- [13] Beidou Wang, Can Wang, Jiajun Bu, Chun Chen, et al., Whom to Mention: Expand the Diffusion of Tweets by @ Recommendation on Micro-blogging Systems, *WWW*, 2013.
- [14] Danah Boyd, Scott Golder, Gilad Lotan, Tweet, Tweet, Retweet: Conversational Aspects of Retweeting on Twitter, *System Sciences (HICSS)*, 2010.
- [15] Haewoon Kwak, Changhyun Lee, Hosung Park, Sue Moon, What is Twitter, a social network or a news media?, *WWW*, 2010.
- [16] Sergey Brin, Lawrence Page, The PageRank Citation Ranking: Bringing Order to the Web, 1998.
- [17] Lise Getoor, Christopher P. Diehl, Link mining: a survey, *SIGKDD*, 2005.