Twitterにおける情報収集力に基づいたユーザスコアリング手法の提案

山川　衛† 田島　敬史††

† 京都大学工学部情報学科　〒 606-8501 京都府京都市左京区吉田本町
†† 京都大学情報学研究科　〒 606-8501 京都府京都市左京区吉田本町 36-1
E-mail: †yamakawa.mamoru.75s@st.kyoto-u.ac.jp, ††tajima@i.kyoto-u.ac.jp

あらまし　PageRank をはじめとした現在のソーシャルネットワークサービスにおけるユーザスコアリングアルゴリズムは，多くが各ユーザ個人の影響力に基づいたものとなっている．しかし SNS では個人の影響力が大きくはなくとも有益な情報を素早く収集することのできるアカウントの存在は重要なものであり，その特定は重要であると考える．本稿では，実際の SNS における情報伝播を表現するために，各ユーザのリツイートの行いやすさと各ユーザが発信する情報の質を用いて，ノードが受け取る情報量を表す Katz Centrality を拡張した新たなユーザスコアリングアルゴリズムを提案する．

キーワード　SNS，ランキングアルゴリズム，Katz 中心性，リツイート

1 はじめに

Twitter をはじめとした SNS では，今現在も多くのユーザが情報を発信し続けており，すべての情報を把握することは不可能に近く，有益な情報を効率よく得ることは非常に重要である．有益な情報を効率よく得るための手段の一つとして，ユーザを特徴量を用いてスコアリングし，良質なユーザを特定するというものがある．現在では Google の PageRank [8] をはじめ様々なスコアリングアルゴリズムが提案されている．

従来の手法の多くは「社会的影響力の高いアカウントは有益な情報を発信する」という前提でのユーザスコアリングを行っているが，Twitter では他のユーザの情報をさらに拡散するリツイートの存在により，ユーザ本人の社会的影響力が高くなかったとしても「有益な情報を多く早く収集していて，リツイートを行う頻度が高い」ユーザや，「直接影響力のあるユーザをフォローしていなくても他のフォロイーの拡散により社会的影響力のあるユーザの情報を多く集めることができる」情報収集力の高いユーザは非常に良質なユーザである．しかし，既存の情報収集力を指標としたランキングである Katz Centrality はユーザごとの社会的影響力，リツイート確率の差を，HITS アルゴリズムの Hub スコアでは 2 hop 先のノードを考慮していないという問題があった．そこで本稿では，ユーザごとのリツイート確率と社会的影響力を表す重みを用いて Katz Centrality を重みづけした新たな情報収集力を指標とする IGR_{dsl} と，親ノードから流れてきた情報のうち子ノードに伝える情報量を表す指標である IGR_{diff} を提案した．

提案手法が既存手法とは異なるランキングを生成できているか，また既存手法とどのような関係があるかを調べるために，IGR_{dsl}，IGR_{diff} の二つの提案手法と既存手法によって出力されるスコアと，それによるランキングについての考察を行った．結果として，提案手法は両者ともに既存手法とは異なるランキングを生成出来てはいたが，IGR_{dsl} は既存手法と非常に順位相関係数が高い結果となった．この結果から，提案手法は目的としていた有益な情報を素早く収集することのできるアカウントを発見するランキングを生成出来てはいない，もしくは目的としていた有益な情報を素早く収集する能力と社会的影響力には相関があるという二つの要因のいずれかが考えられる．また，ユーザのリツイート確率と PageRank をはじめとする既存手法に対しては負の相関がみられ，社会的影響力の高いユーザほどリツイート確率が小さい傾向にあるということが分かった．

提案手法が既存手法よりも優位な性能を発揮するかを調べるために，IGR_{dsl}，IGR_{diff} の二つの提案手法と既存手法を 5 種類のスコア入力とした場合と，5 種類のうち一つを除いた 4 種類を入力として多層パーセプトロンと決定木の二つの分類器の学習を行い，フォロワー数の予測タスクの性能比較を行った．性能比較の結果，IGR_{dsl} を除いた 4 種類で決定木を学習した際にその他の入力に比べて大きく性能が低下した．

2 関連研究

この章では，まず現在の社会的影響力を測定するユーザスコアリング手法がどのようなものであるか，その性質と比較を述べる．その後，本稿で提案する手法の考え方のもとになる社会的影響力の評価指標の詳細と，提案手法の計算に用いる行列，またその性質について述べる．

2.1 現在の社会的影響力の評価指標

社会的影響力を指標としたユーザスコアリング手法は，中心性，リンクトポロジーを用いたランキング指標，エントロピーなどが存在する．固有ベクトル中心性や Katz Centrality といった中心性は様々な基準に基づいて，ノードがどれだけネットワークの中心に位置するかを測定するものであり，PageRank や HITS といったリンクトポロジーを用いたランキング指標は中心性指標の多くが考慮していなかったノードによるスコアへの

寄与の差を考慮したものである [9].

現在の社会的影響力の評価指標に基づくユーザスコアリング手法の多くは何らかの形で PageRank に基づいており, PageRank はウェブページ用に提案されたアルゴリズムでありながら SNS における社会的影響力のあるユーザの特定に非常に有効である [10]. PageRank を Twitter 用に拡張したアルゴリズムは TwitterRank [11] や TURank [12] などが挙げられる. TwitterRank は Twitter 上に存在する homophily に基づいて PageRank を拡張したアルゴリズムで, ユーザの Twitter における一般的な社会的影響力や特定のユーザから見た影響力を算出することができる. また, TURank は PageRank をデータベース上のキーワード検索用に拡張した ObjectRank [1] を Twitter に当てはめたアルゴリズムである

影響力のあるユーザ特定の他にも, 別の指標に基づくユーザスコアリング手法も存在する. 例えば, 人気度を指標としたユーザスコアリング手法はフォローアップの関係に基づいており, 活発さを指標としたユーザスコアリング手法はリプライを考慮している [10]. また, 現在のユーザスコアリング手法において最も多く使われている特徴は「いいね」に関連するものである [10].

2.2 行列の無限等比級数

対角成分が λ で n 次のジョルダン細胞を $J(\lambda, n)$ とする. 任意の固有値の絶対値が 1 未満である正方行列 A に対して, ある正則行列 P が存在して,

$$A = PJP^{-1}$$

と書ける. ここで J はジョルダン標準形であり,

$$J = \begin{pmatrix} J(\lambda_1, n_1) & & 0 \\ & \ddots & \\ 0 & & J(\lambda_k, n_k) \end{pmatrix}$$

のように書ける. また, A の任意の固有値の絶対値が 1 未満という仮定から, 行列 $(E - A)$ の任意の固有値は 0 でないので, $(E - A)$ は逆行列を持つ. このとき,

$$\begin{aligned} \sum_{k=0}^{n} A^k &= (E - A^n)(E - A)^{-1} \\ &= (E - PJ^nP^{-1})(E - A)^{-1} \\ &= P(E - J^n)P^{-1}(E - A)^{-1} \end{aligned}$$

となる. ここで, A の任意の固有値の絶対値が 1 未満なので,

$$\lim_{n\to\infty} J^n = O$$

となるので,

$$\begin{aligned} \sum_{k=0}^{\infty} A^k &= \lim_{n\to\infty} P(E - J^n)P^{-1}(E - A)^{-1} \\ &= (E - A)^{-1} \end{aligned} \tag{1}$$

となる.

2.3 ネットワークにおける隣接行列

有向ネットワークの隣接行列 A はノード i から j に枝が伸びている場合に ij 成分が 1 となり, そうでない場合は 0 となる行列である. また, A^n の ij 成分はノード i から j への n hop の経路の総数となる. また, 以下の等式が成り立つ.

$$(A^n)_{ij} = \sum_k A_{ik}(A^{n-1})_{kj} \tag{2}$$

この式は, ノード i からノード k までの 1 hop の経路数とノード k からノード j までの $(n-1)$ hop の経路数の積はノード i から j への n hop の経路の総数と等しいということを表している.

2.4 Katz Centrality

Katz Centrality は 1953 年に Leo Katz によって発表されたネットワークにおけるスコアリングアルゴリズムである [4]. Katz Centrality は以下のように定義される.

$$C_{\mathrm{Katz}}(i) = \sum_{k=1}^{\infty} \sum_j \alpha^k (A^k)_{ji}$$

ここで $C_{\mathrm{Katz}}(i)$ はノード i の Katz Centrality, A はネットワークの隣接行列, α は $0 \leqq \alpha < 1$ を満たすパラメータである.

α は A の絶対値最大の固有値を $\lambda_{\max}$ としたとき, $0 \leqq \alpha < 1/|\lambda_{\max}|$ を満たすように α を設定すると, Katz Centrality は以下のように計算できる [3].

$$C_{\mathrm{Katz}}(i) = ((I - \alpha A^T)^{-1} - I)\overrightarrow{1}$$

ここで, I は単位行列, $\overrightarrow{1}$ は全ての要素が 1 のベクトルである.

$C_{\mathrm{Katz}}(i)$ は, 直接つながっているノードだけでなくノード i までの経路が存在するすべてのノードを考慮したスコアであり [10], ノード i がどのくらいの情報量を他のノードから受け取るかを表す指標として解釈でき, その場合は以下の 3 点が前提となったネットワークにおいての指標といえる.

- 各ノードが発信する情報の重みが全て等しい
- 減衰率 α によって, 長い距離を伝ってきた情報ほど情報量が小さくなる
- 親ノードは全ての子ノードに確率 1 で情報を伝播させる

2.5 PageRank

PageRank はウェブページのスコアリングを目的としたアルゴリズムであり, 以下のように計算される.

1. 各ウェブページをノード, リンクをエッジとしたグラフの隣接行列を $A = (a_{ij})$ として, 以下のように遷移確率行列 $B = b_{ij}$ を定義する.

$$b_{ij} = \frac{a_{ij}}{\sum_k a_{kj}}$$

2. 行列 B の最大の固有値に対する固有ベクトルの要素が各ノードの PageRank に対応する.

PageRank は多くのページからリンクを受けているほど有用なノードであり, 有用なノードからリンクを受けているほど有

用であるという考えに基づいてノードをランキングしている．PageRank はウェブページの重要度を測定するほかにも，SNS におけるユーザの重要度，社会的影響力を測定する用途で用いることもできる．この場合はユーザがノード，フォロー関係がエッジとなる．PageRank を情報収集力の指標として考えた場合，親ノードが全ての子ノードに確率 $\frac{1}{n}$ で情報を伝えているとみることができる (n は子ノードの数)．

2.6 HITS アルゴリズム

HITS アルゴリズムはオーソリティスコアとハブスコアという二つのスコアを算出するアルゴリズム [6] であり，以下のように相互的に計算される．

$$\mathbf{a}^{(0)} = \mathbf{e}$$
$$\mathbf{h}^{(0)} = \mathbf{e}$$
$$\mathbf{a}^{(k)} = A^T\mathbf{h}^{(k-1)}$$
$$\mathbf{h}^{(k)} = A\mathbf{a}^{(k-1)}$$

ここで，$\mathbf{a}$ と $\mathbf{h}$ がそれぞれオーソリティスコアとハブスコアであり，$\mathbf{e}$ はネットワークのノード数と同じ次元で，すべての要素が 1 であるベクトルである．

オーソリティスコアはハブスコアの高いノードからリンクされていると高くなり，ハブスコアはオーソリティスコアの高いノードにリンクしていると高くなる．ネットワークにおいて HITS アルゴリズムのオーソリティスコアは影響力，ハブスコアは情報収集力の指標ととらえることができる．

2.7 ケンドールの順位相関係数

ケンドールの順位相関係数は Kendall によって考案された二つのランキング間の相関の強さを表す指標 [5] であり，以下のように計算される．

$$\tau = \frac{P - Q}{\binom{n}{2}}$$

ここで，τ は順位相関係数であり，P はそれぞれ二つのランキング $\{x_1, \cdots, x_n\}$，$\{y_1, \cdots, y_n\}$ から 2 項目を選んだときに順位関係が一致する組の個数，Q は不一致となる個数である．$\tau = 1$ の場合は順位が完全に一致しており，$\tau = -1$ の場合には完全不一致である．

3 提案手法

この章ではまず提案手法に用いる特徴量の定義について述べ，その後提案するユーザスコアリング手法の式とその変形，改良について述べる．

3.1 情報収集力の算出方法

3.1.1 Twitter の特徴の考慮

Twitter は SNS の一つであり，各ユーザのタイムラインにはそのユーザがフォローしているユーザのツイート，リツイートが流れる．前項で述べた Katz Centrality を情報収集力の指標としてみた場合の前提とは異なり，枝 (i, j) が存在しても，i に到達した情報が必ず j へと伝播するわけではない．これは，枝 (i, j) と (j, k) が存在して枝 (i, k) が存在しないネットワークでは，k のツイートは必ず j のタイムラインには現れるが，i のタイムラインには j がリツイートを行わない限り現れないからである．

また，HITS アルゴリズムのハブスコアも情報収集力の指標となるが，ハブスコアはオーソリティスコアの高いノードと直接リンクが無ければ高くならないため，リツイートによって直接リンクがなくとも伝播経路が存在するなら情報を収集できる Twitter 上での情報収集力の指標とするには不十分であると考える．

以上の二つの情報収集力の欠点を踏まえて本稿で新たに情報収集力を定義するにあたり，情報の伝播について Twitter 特有の要素におけるリツイートの性質を考慮する．Twitter API の仕様上すべてのツイートを調べることができないため，各ユーザのリツイートの総数を以下のように推定する．

$$T_{\mathrm{RT}}(i) \approx \frac{|t_{\mathrm{RT}}(i)|}{|t(i)|} T(i) \tag{3}$$

ここで，$T_{\mathrm{RT}}(i)$ はユーザ i のリツイートの総数，$t(i)$ はデータセットに含まれるユーザ i のツイートの集合，$t_{\mathrm{RT}}(i)$ は $t(i)$ のうちリツイートであるものの集合，$T(i)$ はユーザ i のツイートの総数である．Twitter API では $T(i)$ の正確な値が収集可能となっている．この，$T_{\mathrm{RT}}(i)$ を用いて，ユーザ i がタイムライン上のツイートをリツイートする確率を以下のように定義する．

$$P_{\mathrm{RT}}(i) = \frac{T_{\mathrm{RT}}(i) + \varepsilon}{\sum_{\mathrm{j} \in \text{followees of i}} (T(j) + \varepsilon)} \tag{4}$$

ここで，$P_{\mathrm{RT}}(i)$ はユーザ i がタイムライン上のツイートをリツイートする確率，また，ε はスムージングのために用いる正の実数である．

ここで，n_1 から n_k までの k hop の特定の経路 $(n_1, n_2, \cdots, n_{k-1}, n_k)$ について，n_1 が発した情報が n_k まで到達する確率 P_{path} は，

$$P_{\mathrm{path}} = P_{\mathrm{RT}}(n_2) \times P_{\mathrm{RT}}(n_3) \times \cdots \times P_{\mathrm{RT}}(n_{k-1})$$

と考えることができる．これをもとに，$P_{\mathrm{RT}}(i)$ を用いて行列 P を以下のように定義する．

$$P = \left(A^T \begin{pmatrix} P_{RT}(1) & & 0 \\ & \ddots & \\ 0 & & P_{RT}(n) \end{pmatrix} \right)^T$$
$$= \begin{pmatrix} A_{11}P_{\mathrm{RT}}(1) & \cdots & A_{1n}P_{\mathrm{RT}}(1) \\ \vdots & \ddots & \vdots \\ A_{n1}P_{\mathrm{RT}}(n) & \cdots & A_{nn}P_{\mathrm{RT}}(n) \end{pmatrix}$$

ここで，A はネットワークの隣接行列，n はネットワークのノード数である．この行列 P の ij 成分は $A_{ij}P(i)$ であり，枝 (i, j) が存在する場合は確率 $P_{\mathrm{RT}}(i)$，存在しない場合は確率 0 で情報が伝播するということを表している．

この行列を n 乗したときの ij 成分 $(P^n)_{ij}$ を $P_{\mathrm{RT}}(i)$ で割った値は，存在するすべての n hop の i から j の経路についての P_{path}

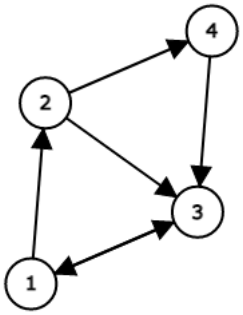


図 1　ネットワークの一例

の総和であり，i が情報量 1 で発した情報を j が n hop で得る際の情報量の期待値といえる．例えば，図 1 のネットワークでは，行列 P は

$$P = \begin{pmatrix} 0 & P_{\mathrm{RT}}(1) & P_{\mathrm{RT}}(1) & 0 \\ 0 & 0 & 0 & P_{\mathrm{RT}}(2) \\ P_{\mathrm{RT}}(3) & P_{\mathrm{RT}}(3) & 0 & 0 \\ 0 & 0 & P_{\mathrm{RT}}(4) & 0 \end{pmatrix}$$

となり，P の累乗の (1,2) 成分を見てみると，

$$\begin{aligned} P_{12} &= P_{\mathrm{RT}}(1) \\ (P^2)_{12} &= P_{\mathrm{RT}}(1)P_{\mathrm{RT}}(3) \\ (P^3)_{12} &= P_{\mathrm{RT}}(1)^2P_{\mathrm{RT}}(3) \\ (P^4)_{12} &= P_{\mathrm{RT}}(1)^2P_{\mathrm{RT}}(3)^2 + P_{\mathrm{RT}}(1)P_{\mathrm{RT}}(2)P_{\mathrm{RT}}(3)P_{\mathrm{RT}}(4) \\ &\vdots \end{aligned}$$

となり，ノード 1 からノード 2 までの 1 hop と 2 hop，3 hop の経路は一つ，4 hop の経路は二つ存在することが分かる．また，4 hop の経路において，一つ目の経路は確率 $P_{\mathrm{RT}}(1)^2P_{\mathrm{RT}}(3)^2 \times \frac{1}{P_{\mathrm{RT}}(1)}$ で，二つ目の経路は確率 $P_{\mathrm{RT}}(1)P_{\mathrm{RT}}(2)P_{\mathrm{RT}}(3)P_{\mathrm{RT}}(4) \times \frac{1}{P_{\mathrm{RT}}(1)}$ でノード 1 からノード 2 まで伝わり，その和がノード 2 がノード 1 が情報量 1 で発信した情報を 4 hop で受け取る際の情報量の期待値である．

3.1.2　情報収集力の定義

3.1.1 で定義した行列 P を用いて，ユーザ i の情報収集力 (Infomation Gathering Rank) を以下のように定義する．

$$IGR(i) = \sum_{t=1}^{\infty} \sum_{j} \left(\alpha^{t-1} (P^t)_{ji} \frac{w_j}{P_{RT}(j)} \right) \tag{5}$$

ここで，$IGR(i)$ はユーザ i の情報収集力である．また，w_j はユーザ j に割り当てる重みであり，α は P の絶対値最大の固有値を $\lambda_{\max}$ としたときに $0 \leqq \alpha < 1/|\lambda_{\max}|$ を満たすパラメータであり，Katz Centrality 同様 α によって遠くからの情報ほど情報量は小さくなる．各ユーザに割り当てる重み w_j は各ノードの情報源としての価値を表す重みであり，Twitter における一般的な情報収集力を算出したい場合は PageRank を，特定のトピックでのスコアを算出したい場合は TwitterRank や Topic-sensitive PageRank [2] を，特定のハッシュタグにおいては TRank [7] を，といったように用途に応じて使い分けることができる．

3.2　*IGR* の変形

(5) 式によって定義した *IGR* を変形する．まず，以下で定義する $IGR^n(i)$ を考える．

$$IGR^n(i) = \sum_{t=1}^{n} \sum_{j} \left(\alpha^{t-1} (P^t)_{ji} \frac{w_j}{P_{RT}(j)} \right) \tag{6}$$

ここで，$IGR^{n+1}(i)$ を変形すると，

$$\begin{aligned} IGR^{n+1}(i) &= \sum_{t=1}^{n+1} \sum_{j} \alpha^{t-1} \frac{(P^t)_{ji}}{P_{\mathrm{RT}}(j)} w_j \\ &= \sum_{t=2}^{n+1} \sum_{j} \alpha^{t-1} \frac{(P^t)_{ji}}{P_{\mathrm{RT}}(j)} w_j + \sum_{j} \frac{P_{ji}}{P_{\mathrm{RT}}(j)} w_j \end{aligned}$$

式 (2) より，$w_p(j) = \frac{w_j}{P_{\mathrm{RT}}(j)}$ とすると，

$$\begin{aligned} IGR^{n+1}(i) &= \sum_{t=2}^{n+1} \sum_{j} \left(\sum_{k} (\alpha P_{ki}) \alpha^{t-2} \frac{(P^{t-1})_{jk}}{P_{\mathrm{RT}}(j)} w_j \right) + \sum_{j} P_{ji} w_p(j) \\ &= \sum_{k} \left(\sum_{t=2}^{n+1} \sum_{j} (\alpha P_{ki}) \alpha^{t-2} \frac{(P^{t-1})_{jk}}{P_{\mathrm{RT}}(j)} w_j \right) + \sum_{j} P_{ji} w_p(j) \\ &= \sum_{k} \alpha P_{ki} \left(\sum_{t=2}^{n+1} \sum_{j} \alpha^{t-2} \frac{(P^{t-1})_{jk}}{P_{\mathrm{RT}}(j)} w_j \right) + \sum_{j} P_{ji} w_p(j) \\ &= \sum_{k} \alpha P_{ki} \left(\sum_{t=1}^{n} \sum_{j} \alpha^{t-1} \frac{(P^t)_{jk}}{P_{\mathrm{RT}}(j)} w_j \right) + \sum_{j} P_{ji} w_p(j) \end{aligned}$$

式 (6) の $IGR^n(i)$ の定義から，ネットワークのノード数を m とすると，

$$\begin{aligned} IGR^{n+1}(i) &= \sum_{k} \alpha P_{ki} IGR^n(k) + \sum_{j} P_{ji} w_p(j) \\ &= \sum_{j} \alpha P_{ji} IGR^n(j) + \sum_{j} P_{ji} w_p(j) \\ &= \alpha \begin{pmatrix} P_{1i} & \cdots & P_{mi} \end{pmatrix} \begin{pmatrix} IGR^n(1) \\ \vdots \\ IGR^n(m) \end{pmatrix} \\ &\quad + \begin{pmatrix} P_{1i} & \cdots & P_{mi} \end{pmatrix} \begin{pmatrix} w_p(1) \\ \vdots \\ w_p(m) \end{pmatrix} \end{aligned}$$

ここで，$\overrightarrow{IGR^n} = \begin{pmatrix} IGR^n(1) \\ \vdots \\ IGR^n(m) \end{pmatrix}$，$\overrightarrow{w_p} = \begin{pmatrix} w_p(1) \\ \vdots \\ w_p(m) \end{pmatrix}$ とすると，

$$IGR^{n+1}(i) = \begin{pmatrix} P_{i1} & \cdots & P_{mi} \end{pmatrix} \left(\alpha \overrightarrow{IGR^n} + \overrightarrow{w_p} \right) \tag{7}$$

となり，$IGR^{n+1}(i)$ は $IGR^n(i)$ で表せることがわかる．さらに式 (7) から，$\overrightarrow{IGR^{n+1}}$ は以下のように表せる．

$$\overrightarrow{IGR^{n+1}} = \begin{pmatrix} P_{11} & \cdots & P_{m1} \\ \vdots & \ddots & \vdots \\ P_{1m} & \cdots & P_{mm} \end{pmatrix} \left(\alpha\overrightarrow{IGR^{n}} + \overrightarrow{w_p}\right)$$

$$\overrightarrow{IGR^{n+1}} = P^T\left(\alpha\overrightarrow{IGR^{n}} + \overrightarrow{w_p}\right)$$

よって，$\overrightarrow{IGR^n}$ は

$$\begin{aligned}\overrightarrow{IGR^n} &= \alpha^n(P^T)^n\overrightarrow{IGR^1} + \left(\sum_{k=1}^{n}\alpha^{k-1}(P^T)^{k-1}\right)P^T\overrightarrow{w_p} \\ &= \alpha^n(P^T)^n\overrightarrow{IGR^1} + \left(\sum_{k=0}^{n-1}\alpha^{k}(P^T)^{k}\right)P^T\overrightarrow{w_p}\end{aligned}$$

となる．ここで，α の定義から，行列 αP^T の任意の固有値の絶対値は 1 未満であるので，式 (1) から，

$$\begin{aligned}\lim_{n\to\infty}\alpha^n(P^T)^n &= O \\ \lim_{n\to\infty}\left(\sum_{k=0}^{n-1}\alpha^k(P^T)^k\right) &= \lim_{n\to\infty}\left(\sum_{k=0}^{n}\alpha^k(P^T)^k\right) \\ &= (E-\alpha P^T)^{-1}\end{aligned}$$

となるので，

$$\begin{aligned}\overrightarrow{IGR} &= \overrightarrow{IGR^{\infty}} \\ &= \lim_{n\to\infty}\overrightarrow{IGR^n} \\ &= \lim_{n\to\infty}\left(\alpha^n(P^T)^n\overrightarrow{IGR^1} + \left(\sum_{k=1}^{n}\alpha^{k-1}(P^T)^{k-1}\right)P^T\overrightarrow{w_p}\right) \\ &= O\overrightarrow{IGR^1} + (E-\alpha P^T)^{-1}P^T\overrightarrow{w_p} \\ &= (E-\alpha P^T)^{-1}P^T\overrightarrow{w_p} \qquad (8)\end{aligned}$$

と変形できる．

3.3 自己ループの削除

(8) 式で求めた IGR では各ノードが自分からも情報を受け取ることになり，Twitter の情報拡散を考えた場合不適切と考える．そこで，各ノードが自分から受け取る情報量を取り除いた IGR_{dsl} を以下のように定義する．

$$IGR_{\mathrm{dsl}}(i) = IGR(i) - IGR_{\mathrm{self}}(i) \qquad (9)$$

ここで，$IGR_{\mathrm{self}}(i)$ は以下のように定義する．

$$IGR_{\mathrm{self}}(i) = \sum_{t=1}^{\infty}\alpha^{t-1}\frac{(P^t)_{ii}}{P_{RT}(i)}w_i$$

ここで，α，P，$P_{RT}(i)$，w_i は $IGR(i)$ の算出に用いるものと同じである．以下 IGR_{self} の変形を行う．

$$\begin{aligned}IGR_{\mathrm{self}}(i) &= \sum_{t=1}^{\infty}\alpha^{t-1}\frac{(P^t)_{ii}}{P_{RT}(i)}w_i \\ &= \sum_{t=1}^{\infty}\alpha^{t-1}(P^t)_{ii}w_p(i)\end{aligned}$$

$P_{ii} = (P^T)_{ii}$ より，

$$\begin{aligned}IGR_{\mathrm{self}}(i) &= \sum_{t=1}^{\infty}\alpha^{t-1}((P^T)^t)_{ii}w_p(i) \\ &= \left(\sum_{t=1}^{\infty}\alpha^{t-1}(E\otimes(P^T)^t)_{ii}\right)w_p(i) \\ &= \left(\left(\sum_{t=1}^{\infty}\alpha^{t-1}(E\otimes(P^T)^t)\right)\overrightarrow{w_p}\right)_i\end{aligned}$$

ここで，$0<\alpha$ である場合，

$$\begin{aligned}&= \frac{1}{\alpha}\left(\left(E\otimes\left(\sum_{t=1}^{\infty}\alpha^t(P^T)^t\right)\right)\overrightarrow{w_p}\right)_i \\ &= \frac{1}{\alpha}\left(\left(E\otimes\left(\sum_{t=0}^{\infty}\alpha^t(P^T)^t - E\right)\right)\overrightarrow{w_p}\right)_i\end{aligned}$$

ここで，$\otimes$ はアダマール積であり，同じサイズの行列 A，B に対して

$$A\otimes B = (a_{ij}b_{ij})$$

で定義される演算である．αP^T の任意の固有値の絶対値は 1 未満なので，式 (1) より，

$$IGR_{\mathrm{self}}(i) = \frac{1}{\alpha}\left(\left(E\otimes\left(\left(E-\alpha P^T\right)^{-1} - E\right)\right)\overrightarrow{w_p}\right)_i$$

と書けるので，$\overrightarrow{IGR_{\mathrm{self}}} = \begin{pmatrix} IGR_{\mathrm{self}}(1) \\ \vdots \\ IGR_{\mathrm{self}}(m)\end{pmatrix}$ とすると，$\overrightarrow{IGR_{\mathrm{self}}}$ は，

$$\overrightarrow{IGR_{\mathrm{self}}} = \frac{1}{\alpha}\left(E\otimes\left(\left(E-\alpha P^T\right)^{-1} - E\right)\right)\overrightarrow{w_p} \qquad (10)$$

となり，(9) と (10) から，$\overrightarrow{IGR_{\mathrm{dsl}}}$ は

$$\overrightarrow{IGR_{\mathrm{dsl}}} = \overrightarrow{IGR} - \frac{1}{\alpha}\left(E\otimes\left(\left(E-\alpha P^T\right)^{-1} - E\right)\right)\overrightarrow{w_p} \qquad (11)$$

となる．

3.4 有用なユーザの特定

3.3 で定義した IGR_{dsl} は，ネットワークの情報を多く，早く集めることのできるユーザほど高くなるが，この IGR_{dsl} を用いてネットワークの情報を多く，早くフォロワーに提供するユーザを以下のように定義することができる．

$$\overrightarrow{IGR_{\mathrm{diff}}} = \overrightarrow{P_{\mathrm{RT}}}\otimes\overrightarrow{IGR_{\mathrm{dsl}}} \qquad (12)$$

ここで，$\overrightarrow{P_{\mathrm{RT}}} = \begin{pmatrix} P_{\mathrm{RT}}(1) \\ \vdots \\ P_{\mathrm{RT}}(m)\end{pmatrix}$ である．この IGR_{diff} は，ユーザが求める情報を多く，早く提供するユーザの推薦に用いることができると考える．

4 実験，検証

これまでに提案したアルゴリズムの有用性を検証するために既存のアルゴリズムとの比較検証と実験を行う．

	データセット 1	データセット 2
起点としたユーザ	@univkyoto	@A_I_News
ノード数	40691	32739
エッジ数	509978	456483
平均 P_{RT}	2.5777926446798245e-06	9.430851368871925e-07

表 1　収集したデータセット

4.1　データセット

データセットは以下の方法で収集する．

1. ある特定のユーザのフォロワーを取得し，ユーザとフォロワーを併せた集合を S とする．
2. 任意のユーザ $s \in S$ について，フォロイーの ID とツイート数を取得し，以下の集合を作成する

$$E = \{(s_i,\ s_j)\,|s_j \text{ follows } s_i \wedge\ s_i,\ s_j \in S\}$$

3. 有向ネットワーク $G = (S, E)$ を生成
4. 任意のユーザ $s \in S$ について，直近のツイートを 100 件取得する．
5. 式 (3)，(4) を用いて P_{RT} を推定する．ここで，鍵アカウントは本稿で定義した P_{RT} が定義できないため，鍵アカウントでないアカウントの P_{RT} の平均を P_{RT} とする．
6. こうして生成した $(G,\ P_{\mathrm{RT}})$ が一つのデータセットである

この方法で二つのデータセットを作成した．それぞれのデータセットの情報を表 1 に記す．

4.2　比 較 検 証

ここでは 3.3 で定義した IGR_{dsl} と 3.4 で定義した IGR_{diff} によって生成されるランキングを以下の既存のスコアリングアルゴリズムによるランキングと比較検証する．

- Katz Centrality (Katz)
- HITS アルゴリズムにおけるハブスコア (Hub)
- PageRank (PR)

Katz Centrality と比較を行う理由は枝 (i, j) に $P_{RT}(i)$ で重みづけしたことによる変化が出るかどうかを検証するため，ハブスコアと比較を行う理由は直接リンクを張っていないノードからの影響を IGR が考慮できているかを検証するため，PageRank と比較を行う理由は IGR が有用であるか，つまり情報収集力の高いユーザと影響力の高いユーザとが一致していないかどうかを検証するためである．

なお，IGR_{dsl}，IGR_{diff} を算出するにあたり，以下のように各変数を設定した．

- $\alpha = \max\{0.8,\ \lambda_{\max}\}$
- $w_i = PR_i$

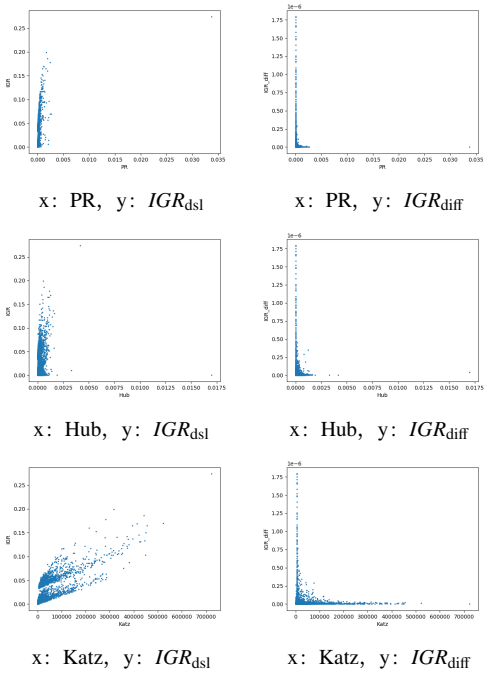
ここで，$\lambda_{\max}$ は式 (5)，(3.4) の IGR，IGR_{diff} を計算する際に用いる行列 P の固有値のうち，絶対値最大のもの，PR_i はノード i の PageRank である．

4.2.1　比 較 結 果

両データセットにおいて x 軸に既存手法，y 軸に提案手法をとった散布図は図 2 のようになった．

4.2.2　考　　察

表 2 から，順位相関係数が 1 ではないため，二つの提案手法



x：PR，y：IGR_{dsl}　　x：PR，y：IGR_{diff}

x：Hub，y：IGR_{dsl}　　x：Hub，y：IGR_{diff}

x：Katz，y：IGR_{dsl}　　x：Katz，y：IGR_{diff}

図 2　データセット 1 によるスコア

	IGR_{dsl}	IGR_{diff}	P_{RT}	PR	Hub	Katz
IGR_{dsl}	1.000000	0.746521	-0.475481	0.845105	0.308068	0.919916
IGR_{diff}	0.746521	1.000000	-0.216978	0.684089	0.245461	0.755400
P_{RT}	-0.475481	-0.216978	1.000000	-0.494364	-0.087246	-0.459095
PR	0.845105	0.684089	-0.494364	1.000000	0.326381	0.850073
Hub	0.308068	0.245461	-0.087246	0.326381	1.000000	0.332614
Katz	0.919916	0.755400	-0.459095	0.850073	0.332614	1.000000

表 2　データセット 1 の順位相関係数

	IGR_{dsl}	IGR_{diff}	P_{RT}	PR	Hub	Katz
IGR_{dsl}	1.000000	0.424273	-0.355303	0.756700	0.484674	0.888940
IGR_{diff}	0.424273	1.000000	0.221509	0.300288	0.129315	0.411511
P_{RT}	-0.355303	0.221509	1.000000	-0.434331	-0.297392	-0.356586
PR	0.756700	0.300288	-0.434331	1.000000	0.551747	0.777696
Hub	0.484674	0.129315	-0.297392	0.551747	1.000000	0.516667
Katz	0.888940	0.411511	-0.356586	0.777696	0.516667	1.000000

表 3　データセット 2 の順位相関係数

は既存手法と異なるランキングを生成していることが分かる．また図 2 から，既存手法が同程度のスコアを出力しているノード間でも提案手法では大きな差があるノードの組がみられるため，経路とノードの重みづけが機能していると考えられる．

IGR_{dsl} について，表 2，3 から，両データセットともに IGR_{dsl} と Katz に非常に高い順位相関がみられたため，IGR_{dsl} も情報収集力の指標として機能していると考えられるが，PR とも非常に相関が高く，実際，表 2，3 を見ると PR の高いユーザを IGR_{dsl} でも上位に推薦していることが分かる．これは，

（1）提案手法が「直接影響力のあるユーザをフォローしていなくても他のフォロイーの拡散により影響力のあるユーザの情報を多く集めることができる」ユーザが上位に来るランキングは生成できていない

（2）提案手法は目的に合ったランキングを生成しているが，「直接影響力のあるユーザをフォローしていなくても他のフォロイーの拡散により影響力のあるユーザの情報を多く集めるこ

	All	$-IGR_{\mathrm{dsl}}$	$-IGR_{\mathrm{diff}}$	−PR	−Hub	−Katz
適合率 (MLP)	0.75	0.70	0.62	0.66	0.66	0.74
再現率 (MLP)	0.23	0.23	0.45	0.23	0.24	0.23
F_1 スコア (MLP)	0.22	0.23	0.48	0.21	0.23	0.21
適合率 (DT)	0.55	0.61	0.46	0.45	0.51	0.53
再現率 (DT)	0.40	0.40	0.30	0.35	0.32	0.38
F_1 スコア (DT)	0.43	0.45	0.34	0.36	0.36	0.41

表 4　訓練データ：データセット 1，テストデータ：データセット 2

	All	$-IGR_{\mathrm{dsl}}$	$-IGR_{\mathrm{diff}}$	−PR	−Hub	−Katz
適合率 (MLP)	0.67	0.68	0.60	0.70	0.80	0.67
再現率 (MLP)	0.38	0.37	0.37	0.36	0.41	0.36
F_1 スコア (MLP)	0.41	0.40	0.40	0.39	0.48	0.39
適合率 (DT)	0.71	0.75	0.48	0.60	0.59	0.57
再現率 (DT)	0.37	0.38	0.34	0.35	0.37	0.35
F_1 スコア (DT)	0.43	0.45	0.39	0.38	0.44	0.39

表 5　訓練データ：データセット 2，テストデータ：データセット 1

とができる」ユーザはそもそも PR における影響力の高いユーザである
という二つの原因が考えられる．

また，PR，Katz，Hub といった既存手法と P_{RT} に負の相関がみられるため，影響力のあるユーザはリツイート行動を行いにくい傾向があるのではないかと考えられる．

IGR_{diff} について，表 2，3 から，データセット 1 では PR，Katz とは強い相関がみられ，データセット 2 では弱い相関がみられた．また，P_{RT} とはデータセット 1 では正の，データセット 2 では負の相関がみられたため，IGR_{diff} は PR，Katz とはまた別の指標によるスコアリングができているが，P_{RT} に大きく依存しているスコアであると考える．

4.3 性能評価

4.3.1 評価方法

提案手法の性能を評価するため，以下の方法でフォロワー数による分類タスクを行う．

（1） データセット 1，2 の各ユーザのフォロワー数 F_1^i，F_2^i を取得し，それをもとに以下のように正解データ y_1，y_2 を作成する．

$$y_k^i = n \in \mathbb{Z} \quad (10^n \leqq F_k^i < 10^{n+1})$$

ここで，y_k^i は入力 x_k^i に対応する正解ラベルである $(k \in \{1,2\})$．

（2） 訓練データを $\{(x_1^1, y_1^1), \cdots, (x_1^n, y_1^n)\}$ とし，多層パーセプトロン (MLP) と決定木 (DT) の二つの分類器を学習する．

（3） テストデータを $\{(x_2^1, y_2^1), \cdots, (x_2^n, y_2^n)\}$ とし，分類器の性能を比較する．なお，性能の指標としては再現率，適合率，F_1 スコアを用いる．

（4） 訓練データとテストデータを入れかえて (2)，(3) を行う．

なお，今回は入力する特徴として IGR_{dsl}，IGR_{diff}，PR，Hub，Katz の五つのスコアを入力としたもの (All) と，All からどれか一つのスコアを除いた 4 種類の特徴を入力としたもの ($-IGR_{\mathrm{dsl}}$，$-IGR_{\mathrm{diff}}$，−PR，−Hub，−Katz) の計 6 種類を用意した．

4.3.2 評価結果

評価結果は表 4，5 のようになった．表 4，5 から，$-IGR_{\mathrm{diff}}$ は DT における適合率，再現率，F_1 スコア，MLP における適合率が All に比べ小さくなっており，MLP における再現率と F_1 スコアは All とほぼ等しい，または大きくなるという結果になった．また，−PR，−Hub，−Katz と比較しても性能の低下率が大きいため，IGR_{diff} はフォロワー数予測の観点で意味のあるスコアを算出していたと考えられる．一方 $-IGR_{\mathrm{dsl}}$ は All とほぼ変わらない結果となり，IGR_{dsl} はフォロワー数を計測するための特徴としては機能していなかった．これは，フォロワー数は自身の情報収集能力ではなく実際に子ノードにどれだけの情報を拡散するかが関係しているからではないかと考えられる．

5 考察と改善点

ここでは，4.2.2 で述べた提案手法によるスコアリングが期待したものとは異なる結果となった原因を考察し，改善点を述べる．

5.1 考　　察

原因 1 は今回用いた重みのうち P_{RT} によるものだと考える．今回，表 1 に記したように各ユーザのリツイート確率の平均がおおよそ 10^{-6} になっていた．各 $P_{\mathrm{RT}}(i)$ をおおよその 10^{-6} だと仮定して IGR のスコアを計算した際に，

$$\begin{aligned} IGR(i) &= \sum_{t=1}^{\infty} \sum_j \left(\alpha^{t-1} (P^t)_{ji} \frac{w_j}{P_{RT}(j)} \right) \\ &= \sum_j P_{ji} \frac{w_j}{P_{RT}(j)} + \sum_j \alpha (P^2)_{ji} \frac{w_j}{P_{RT}(j)} + \cdots \\ &= \sum_j A_{ji} w_j + \sum_j 10^{-6} \alpha (A^2)_{ji} w_j + \cdots \end{aligned}$$

となり，2 hop 先のノードから得られる情報量は 1 hop 先のノードから得る情報量の $\alpha \cdot 10^{-6}$ となってしまう．今回，PageRank が最も大きいユーザは，最も小さいユーザの 5000 倍のスコアを持っていたが，$5000 \times 10^{-6} = 5 \times 10^{-3}$ であり，2 hop 先の最も影響力の高いユーザよりも 1 hop 先の最も影響力のないユーザの情報の方が非常に情報量が多いことになってしまう．また，1 hop 先に PageRank が p のユーザを 1 人持つユーザ a と，1hop 先に $P_{\mathrm{RT}}(i)$ が 10^{-6} での PageRank が 0 のユーザを 1 人，2 hop 先に PageRank が p のユーザを n 人持つユーザ b の IGR を比較しても，

$$\begin{aligned} \frac{IGR(a)}{IGR(b)} &= \frac{p}{0 + n \times 10^{-6} \times p} \\ &= \frac{10^6}{n} \end{aligned}$$

となり，$n > 10^6$ でないと情報量が逆転しない．これによって目的としていた「直接影響力のあるユーザをフォローしていなくても他のフォロイーの拡散により影響力のあるユーザの情報を多く集めることができる」ユーザのスコアを高めることがで

きず，1 hop 先に PageRank が高いノードを持っているユーザのスコアのみが高くなったのではないかと考える．このため，今回，各ユーザがタイムラインのツイートをリツイートする確率 P_{RT} を推定するために定義した式 (3)，(4) ではなく，新たな推定方法を考案する必要がある．

5.2 改 善 点

今回の式 (3)，(4) による P_{RT} の推定では，

$$\frac{(\text{ユーザ } i \text{ がこれまでにリツイートした回数})}{(\text{ユーザ } i \text{ のフォロワーの総ツイート数})}$$

という形で定義していたが，新たに「ユーザは面白い，有用なツイートのみを拡散する」と仮定し，新たなリツイート確率の推定方法を考える．

まず「面白い，有用なツイート」はいいね，もしくはリツイートによるインプレッションが存在するツイートであると定義する．ここで，ユーザ i が一定期間に $T'_{\mathrm{RT}}(i)$ 個のリツイートを行ったと仮定する．また，同じ期間にユーザ j が $T'(j)$ 個のインプレッション付きのツイートが投稿したとする．このとき，ユーザ i が有用なツイートをリツイートする確率は，

$$P'_{\mathrm{RT}}(i) = \frac{T'_{\mathrm{RT}}(i) + \varepsilon}{\sum_{\mathrm{j} \in \text{followees of i}} (T'(j) + \varepsilon)} \tag{13}$$

と定義できる．ここで，$P'_{\mathrm{RT}}(i)$ はユーザ i が有用なツイートをリツイートする確率，ε は式 (4) 同様スムージングに用いる正の実数である．

6 今後の課題

今後は，式 (13) で推定した新たなユーザのリツイート確率を用いたスコアの計算を行い，目的としたユーザを見つけることができるか，またリツイート確率の推定精度を上げるためにどれだけのツイートを収集すればよいかという検証を行っていきたい．

また，3.3 で定義した IGR_{dsl} は各ユーザが他ノードからどれだけ情報を早く，多く集めるかを指標としており，3.4 で定義した IGR_{diff} は各ユーザが集めた情報を実際に拡散する量を指標としていたが，IGR_{diff} をもとにして「実際にユーザが Twitter で発信し得る情報量」を指標としたスコア以下のようにを定義することができる．

$$IGR'(i) = IGR_{\mathrm{diff}}(i) + w_i$$

この IGR' の第一項は自分のもとに集まった情報量が多く，自分がリツイートを行いやすいほど大きくなり，第二項は自分の影響力が大きいほど大きくなる指標である．今後はこの IGR' が PageRank やその改良アルゴリズムよりも適切に Twitter における影響力を表すことができているかという検証も行っていきたい．

7 ま と め

本稿では Katz Centrality をもとにして，情報収集力を指標とした新たなユーザスコアリング手法の提案を行った．提案手法のうち一つはフォロワー数の予測に有用であることが分かったが，目的であった「直接影響力のあるユーザをフォローしていなくても他のフォロイーの拡散により影響力のあるユーザの情報を多く集めることができる」ユーザを発見するには至らなかった．今後は，今回用いたパラメータの推定方法を見直し，推定精度を上昇させることで，目的としたユーザを発見できるようにアルゴリズムを改善していきたい．

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] Andrey Balmin, Vagelis Hristidis, and Yannis Papakonstantinou. Objectrank: Authority-based keyword search in databases. In *VLDB*, volume 4, pages 564–575, 2004.

[2] Taher H Haveliwala. Topic-sensitive pagerank: A context-sensitive ranking algorithm for web search. *IEEE transactions on knowledge and data engineering*, 15(4):784–796, 2003.

[3] Björn H Junker and Falk Schreiber. *Analysis of biological networks*, volume 2. John Wiley & Sons, 2011.

[4] Leo Katz. A new status index derived from sociometric analysis. *Psychometrika*, 18(1):39–43, 1953.

[5] Maurice G Kendall. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93, 1938.

[6] Jon M Kleinberg, Mark Newman, Albert-László Barabási, and Duncan J Watts. *Authoritative sources in a hyperlinked environment*. Princeton University Press, 2011.

[7] Manuela Montangero and Marco Furini. Trank: Ranking twitter users according to specific topics. In *2015 12th annual IEEE consumer communications and networking conference (CCNC)*, pages 767–772. IEEE, 2015.

[8] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab, 1999.

[9] Sancheng Peng, Yongmei Zhou, Lihong Cao, Shui Yu, Jianwei Niu, and Weijia Jia. Influence analysis in social networks: A survey. *Journal of Network and Computer Applications*, 106:17–32, 2018.

[10] Fabián Riquelme and Pablo González-Cantergiani. Measuring user influence on twitter: A survey. *Information processing & management*, 52(5):949–975, 2016.

[11] Jianshu Weng, Ee-Peng Lim, Jing Jiang, and Qi He. Twitterrank: finding topic-sensitive influential twitterers. In *Proceedings of the third ACM international conference on Web search and data mining*, pages 261–270, 2010.

[12] Yuto Yamaguchi, Tsubasa Takahashi, Toshiyuki Amagasa, and Hiroyuki Kitagawa. Turank: Twitter user ranking based on user-tweet graph analysis. In *International Conference on Web Information Systems Engineering*, pages 240–253. Springer, 2010.