DEIM Forum 2010 C3-4

マイクロブログにおける他者への影響を考慮した投稿者の重要度推定手法

吉本　和紀† 　鈴木　 優†† 　吉川　正俊††

† 京都大学工学部情報学科〒 606-8501 京都市左京区吉田本町
†† 京都大学大学院情報学研究科〒 606-8501 京都市左京区吉田本町
E-mail: †spzuka@db.soc.i.kyoto-u.ac.jp, ††{ysuzuki,yoshikawa}@i.kyoto-u.ac.jp

あらまし　本稿では投稿者のメッセージを用いて，投稿者メッセージが引用された回数と投稿者の選択関係を表すネットワークから，投稿者の重要度を算出する手法を提案する．メッセージの引用回数が多いということは，そのメッセージが他人に知らせたくなるほど重要であると考える投稿者が多いということであるため，重要度が高いといえる．また，ある投稿者をフォローするということは，その投稿者が有用であると判断することができるため，フォローの数から他の投稿者からの評価を判断することができる．本稿では二つの尺度を組み合わせることによって，投稿者に重要度を算出する方法を提案する．この手法により，利用者は全てのメッセージを読むことなく投稿者の重要度を測定することができる．

キーワード　マイクロブログ，投稿者ネットワーク，情報推薦

A calculation method of blogger's importance using influences to others in micro-blogs

Kazuki YOSHIMOTO†, Yu SUZUKI††, and Masatoshi YOSHIKAWA††

† Undergraduate School of Informatics and Mathematical Science, Faculty of Engineering, Kyoto University Yoshida-Honmachi, Sakyo, Kyoto, 606-8501 Japan
†† Graduate School of Informatics, Kyoto University Yoshida-Honmachi, Sakyo, Kyoto, 606-8501 Japan
E-mail: †spzuka@db.soc.i.kyoto-u.ac.jp, ††{ysuzuki,yoshikawa}@i.kyoto-u.ac.jp

Abstract　In this paper, we propose a novel method for assessing quality of twitter users using the number of retweets and followers. We have two assumptions that if a user submits messages which are important for the other users, these messages are frequently retwitted by the other users. Another assumption is that if a user submits important messages frequently, the user is followed by the other users. Moreover, if a qualified user retwits or follows a user, the quality score should be increased higher than the case of an unqualified user. We propose a quality score calculation method based on these two assumptions. In our experiments, we confirmed that our proposed method can calculate quality scores with high accuracy.

Key words　Micro-blog, Blogger network, Information recommendation

1. はじめに

近年，CGM(Consumer Generated Media:消費者生成メディア) の中でも，短いメッセージを書くブログであるマイクロブログが普及しつつある．投稿者は，自分の気に入った投稿者を選択することによってその投稿者のメッセージを読む．しかし，ある投稿者が書いた記事が他の投稿者の記事によって埋没してしまうため，どの投稿者を選択するかが重要になっている．

そこで本研究では，マイクロブログにおいてどの投稿者を選択するかを決める指針として，投稿者の重要度を測定するための手法を提案する．本提案における重要度とは，投稿者がどのような人物であるかは関係無く，どれほど重要なメッセージを投稿しているかだけに依存するものである．投稿者の重要度を算出する際に，我々はメッセージの引用回数，選択関係を表すネットワークの二つの観点を用いる．

一つ目は，メッセージの引用回数から重要度を算出する方法である．マイクロブログにおいてメッセージの引用は頻繁に行われている．ここで，メッセージの引用は，該当するメッセー

ジの内容が他の投稿者に知らせたい程重要である事を，引用した投稿者が表していると考えられる．そのため，引用される回数の多いメッセージは重要であると考えることができるため，この考え方を利用し，投稿者の重要度を推定する．重要なメッセージを多数投稿する投稿者は発信する情報の豊富な投稿者であるが，重要なメッセージを多数引用する投稿者も重要な情報を伝えているという意味で重要であり，豊富な情報を引用する投稿者であると言える．我々はこの考え方を元に，メッセージの重要度，発信する情報の豊富度，引用する情報の豊富度を再帰的に定義し，それらを求める．

二つ目は，投稿者の選択関係を表すネットワークから重要度を算出する方法である．多数の投稿者に選択されている投稿者は，多数の投稿者に気に入られていると言える．しかし有名人など，多くの利用者に既に知られている投稿者は選択されやすい傾向にあり，メッセージの重要度が反映しているとは考えにくい．そこで我々は，被選択数の多い投稿者が選択している投稿者は価値が高いのではないかという考えを元に，投稿者の重要度を算出する．多数の投稿者に選択されている投稿者は情報を広めている投稿者ではあるといえるが，情報元であるとは限らないので，それよりも情報元の投稿者の方が重要ではないかといえる．この指標は，そのような場合に情報元の投稿者を重要であると推定できると考えられる．また，人気のある投稿者がどのような投稿者を選択しているかという点は興味深いと思われる．

我々はこれらの二つの観点によって算出された重要度を合わせることによって，投稿者の重要度とする．一つ目の観点から投稿者の投稿するメッセージの重要さを考慮し，二つ目の観点から他の投稿者からの評価を考慮することができるので，これらを合わせることによってより妥当な推定ができると考えている．

本研究の位置づけを図 1 で示す．投稿者推薦システムは，個人の嗜好による部分と嗜好によらない部分に分けられると考えている．図 1 においては下の部分が個人の嗜好による部分である．この部分においてシステムの利用者は，自分の趣味や，検索したい問合せを入力する．そしてシステムは適合する投稿者のランキングを返す事が想定される．今回は図 1 の上の部分である，個人の嗜好によらない投稿者の重要度を推定する．これは個人の嗜好によらないので入力を必要とせず，重要度の高い順に並べた投稿者のランキングを返す．本研究に個人の嗜好を考慮したシステムを加えることによって，システムの利用者に合わせて重要度の高い投稿者を推薦するシステムになると考えている．

2. 関 連 研 究

2.1 ブログマイニング

投稿者の評価を行う研究は多数あり，用いるブログの指標も様々である．藤村ら [1] は，投稿者と記事のネットワークを用いて HITS アルゴリズムを元にした EigenRumor アルゴリズムを提案している．この論文によると，記事へのリンクは Web ページのリンクに比べて少ない．そのため，そのまま PageRank のような Web リンク解析の手法を用いることはできない．そこで EigenRumor アルゴリズムは，投稿者自身のスコアを投稿者の書いた記事のスコアに伝播させることによって，この問題に対処している．Kritikopoulos ら [2] は，類似した投稿者や，共通のタグを持つ記事等にリンクを張り，より密なネットワークを形成することでこの問題に対応している．これらはブログに関する研究だが，我々はマイクロブログを対象にすることでこの問題を回避すると考えている．つまり，一つ一つの記事へのリンクは少なくても，一人の投稿者の投稿する記事が多いので，投稿者の重要度を算出できると考える．

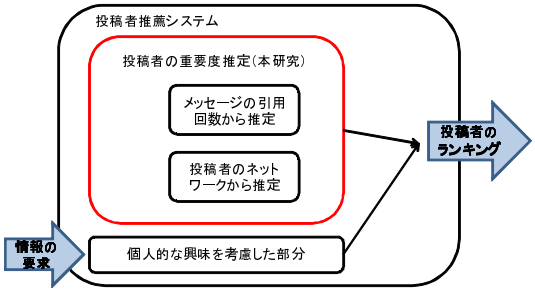

図 1　本研究の位置づけ

また，中島ら [3] はブログのエントリ数の増加や内容の変化などから重要な投稿者を発見している．そして Agarwal ら [4] は記事のネットワークから算出した重要度と投稿数との関係を調査している．我々はこれらの研究とは違い，メッセージと投稿者自身の周りからの評価を用いている．

2.2 Twitter

Bernardo ら [5] は Twitter の持つソーシャルネットワークの部分に注目し，メッセージ数とフォロワー数の関係などを調査をしている．Akshay ら [6] は Twitter における投稿者のネットワークが持つ様々な特性を調査している．また Owen ら [7] は Twitter を用いて流行の話題を推薦している．岩木ら [8] はブロガーの近接度やメッセージ内の単語から有用な記事の発見を行い，桑原ら [9] は投稿者のメッセージから共通の話題を持つ投稿者の推薦を行っているが，これらはシステムの利用者が投稿者としてある程度活動していることを前提としている．それに対して我々は，マイクロブログを始めたばかりの人にも利用できるよう，システムの利用者からの入力を必要としない．

また我々と同じく，Twitter において影響力のある投稿者を見つける事を目的とした様々なサービスが構築されている(注1)．これらは主に投稿者の評価としてその投稿者自身のメッセージやフォロワー数を用いている．すなわち，フォロワー数の多い投稿者ほど影響力の高い投稿者であるという考え方に基づいている．それに対して，我々は投稿者のネットワークに対して

(注1)：Retweetability http://www.retweetability.com/
TweetLevel http://tweetlevel.edelman.com/
Twib http://twib.jp/
Retweetist http://retweetist.com/
retweetradar http://www.retweetradar.com/
retweetrank http://www.retweetrank.com/
twittergrader http://twitter.grader.com/

HITS アルゴリズムを用いて，またフォロワーのフォロワー数というものに注目している点で異なると考えている．

3. 重要度推定システム

Twitter において誰をフォローすればよいかという利用者の問題を解決するために，我々は重要度推定システムを提案する．このシステムは，投稿されたメッセージから各投稿者の重要度を算出する．利用者はシステムが計算した投稿者の重要度を閲覧することによって，誰をフォローしたら良いかという問題を解決する助けとなる．本システムでは利用者本人のメッセージを利用しないため，Twitter を初めて利用する時から利用できるという利点がある．

3.1 投稿者の重要度を示す値の算出方法

我々は投稿者を推薦する理由として，個人の興味による部分とそうでない部分に分けることができると考えている．例えば「○○が逮捕された.」というメッセージは，興味の有無にかかわらず価値のある情報を含んでいるといえる．しかし「お腹が減った.」というメッセージは，よほどその投稿者自身に興味が無い限り価値のある情報とはいえない．Twitter をどのような用途に用いるかは個人の自由であるが,「お腹が減った.」というメッセージを集めたいと思う投稿者は少ないと考えられる．また我々は，投稿者がどのような人かによる部分とよらない部分もあると考えている．有名人だからおもしろい，政治家だから重要だという考えもあるが，メッセージの内容が重要かどうかという事を判断していきたい．

我々は投稿者を二つの側面から判断する．一つは投稿者自身のメッセージから重要度を算出する方法であり，もう一つは投稿者をフォローしている投稿者の特徴を用いて重要度を算出する方法である．一つ目から投稿者がどのような人かとは関係ない，メッセージの重要度を推定し，そこから投稿者の重要度を算出する．そして二つ目からソーシャルネットワークにおける特徴を見いだし，投稿者の重要度を算出する．このように主観的判断と客観的判断を用いることによって，より正確に投稿者を判断することができるのではないかと考えている．

本稿で提案する投稿者の重要度を示す値の算出方法は以下の二つである．

- ReTweet に基づく重要度の算出アルゴリズム
- 投稿者のフォローの特徴に基づく重要度の算出アルゴリズム

以下でそれぞれの提案手法の詳細を述べる．

3.1.1 RT(ReTweet) に基づく重要度の算出アルゴリズム

投稿者の重要度を，その投稿者の投稿したメッセージから算出する．RT(ReTweet) とは Twitter の持つ機能であり，他の投稿者の投稿したメッセージを再投稿することである．つまり RT はメッセージの引用といえる．RT をするということは他の人に伝えたいメッセージであるため，そのメッセージ内容の重要度を示しているのではないかと考える．また重要な内容のメッセージを RT した投稿者は，重要な内容を他の投稿者に伝えているという意味で重要な投稿者であるといえる．我々はそのような投稿者を発見するために RT を用いて重要度を算出する．具体的には次のような考えを基にしている．

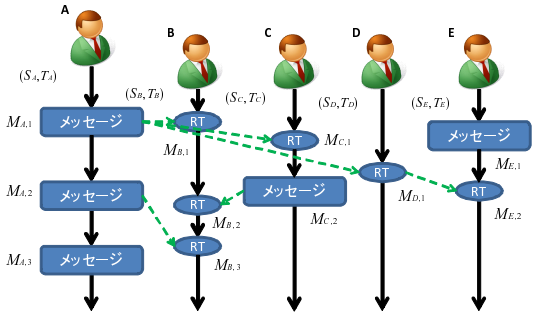

図 2 投稿者とメッセージの様子

- 多数の RT をされたメッセージは，重要なメッセージである
- 多数の重要なメッセージを書く投稿者は，重要な投稿者である
- 多数の重要な RT をする投稿者は，重要な投稿者である

図 2 では，投稿者とメッセージ伝達の様子を表している．投稿者 A が投稿した各メッセージの重要度を示す値を $M_{A,1}, M_{A,2}...M_{A,x}$ (x は A が投稿したメッセージの数)，投稿者 A の影響力を示す値を (S_A, T_A) (S_A は A の発信する情報の豊富さを示す値，T_A は A の引用する情報の豊富さを示す値) の二つ組で表す．

点線の矢印は矢印の元のメッセージを矢印の指している先のメッセージが RT したことを示している．例えば投稿者 A が投稿したメッセージ $M_{A,1}$ は多く RT されているので，重要なメッセージといえる．また，投稿者 B はよく RT をしているので，引用する情報が豊富であるといえる．このような特徴を (S_A, T_A) という値で示したいと考えている．我々は投稿者の重要度を算出するために，まずメッセージの重要度を考える．そしてそのメッセージから投稿者の発信する情報の豊富さや引用する情報の豊富さを判断する．これらの値は以下の式で導出される．

$$M_{A,t} = V + \frac{\sum_i (M_{i,j} \cdot T_i)}{F_A} \tag{1}$$

$$S_A = \frac{\sum_{k=1}^{x} M_{A,k}}{x} \tag{2}$$

$$T_A = \frac{\sum_l (M_{l,t} \cdot S_l)}{x} \tag{3}$$

$M_{i,j}$ は $M_{A,t}$ を RT したメッセージ，F_A は A をフォローする人の数，$M_{l,t}$ は A が RT をしたメッセージの元のメッセージ，V はメッセージの本来持つ値であり，定数である．これらの式について説明する．

式 (1) では，引用する情報が豊富な投稿者による RT の方が価値があると考えて，T_i を掛けている．そしてフォロワー数によって正規化している．メッセージを書いているかは重要なので，全てのメッセージにある価値として V を含めている．ただ (S_A, T_A) を求める際にメッセージの数で割るので，メッセージが多いほど値が高くなるということはない．式 (2) では，投稿

Algorithm 1 User-value

1: set all $M_{A,t}$ to V
   set all S_i to 1 and $preS_i$ to 0
   set all T_i to 1 $preT_i$ to 0
2: **while** $|S_i - preS_i| > \epsilon$ and $|T_i - preT_i| > \epsilon$ **do**
3: $preS_i \leftarrow S_i$
   $preT_i \leftarrow T_i$
4: calculate $M_{A,t}$ by expression (1)
   calculate S_i by expression (2)
   calculate T_i by expression (3)
5: **end while**

したメッセージの重要度を表す値の和で発信する情報の豊富さを表している．ただメッセージの投稿数が多い投稿者が重要であるわけではないので，投稿したメッセージの数で割っている．式 (3) では，発信する情報が豊富な投稿者のメッセージを RT する方が価値があると考えて，S_l を掛けている．そしてメッセージの数によって正規化している．S_A は 0 になることはないが，T_A は RT しているメッセージがない場合に 0 になる．

これらの値は Web リンク解析における HITS アルゴリズムと同じように相互再帰的に定義してあり，初期値を与えて十分収束した値になるまで計算する．詳細を Algorithm1 に示す．このアルゴリズムの流れは次のようになっている．

（1） ユーザ，メッセージに初期値を与える

（2） メッセージ毎に，RT の数に応じた値を付ける（RT は，RT 元のメッセージの値を考慮した値になる）

（3） 各ユーザについて，そのユーザのメッセージの値の和と，そのユーザのつながりから値を算出する

（4） 十分に収束した値になるまで，2 に戻る

このアルゴリズムにおいて ϵ は閾値であり，終了条件を決定する．これにより投稿者の重要度を利用者に提供する．利用者は投稿者の発信する情報の豊富さと引用する情報の豊富さを同時に見ることができ，フォローする投稿者を決める際の手がかりになると考えている．

3.1.2 投稿者のフォローの特徴に基づく重要度の算出アルゴリズム

投稿するメッセージの内容とは別に，我々はフォロワー数も投稿者の重要度を示すと考えている．ところがフォロワー数が多い投稿者だけを推薦してしまうと，フォロワー数が多ければ多いほどさらにフォロワー数が増加する傾向になり，逆にフォロワー数が少ない投稿者はフォローされる可能性がさらになくなる．つまりフォロワー数は投稿者の人気を表すものといえる．我々は人気と重要度は異なると考えているため，フォロワー数だけが投稿者の重要度を表す指標ではない．

そこで我々はフォローの特徴を用いて，新たな投稿者の重要度を示す値の算出方法を考える．これは，多数のフォロワーを持つ投稿者がフォローする投稿者は重要な投稿者ではないかという考えを基にした方法である．図 3 において，ノードは投稿者，エッジはフォロー関係を表し，エッジの元の投稿者がエッジの指している投稿者にフォローしていることを示している．

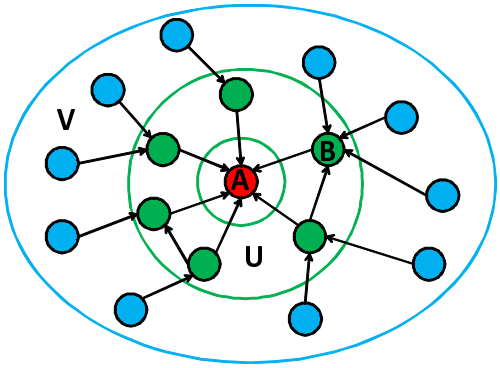

図 3　投稿者の関係を表すグラフ

まず，投稿者 A をフォローしている投稿者の集合を U とする．U の要素数は A のフォロワー数といえる．これ以降 U の要素数を $n(U)$ と表す．次に U の要素である投稿者をフォローしている投稿者から，U の要素である投稿者を除いた投稿者の集合を V とする．つまり，V の投稿者の中に A をフォローしている投稿者はいない．

ここで，多数のフォロワーを持つ投稿者 B は多数の目についている投稿者といえるが，B 自身が重要なメッセージを多く発信しているとは限らず，別の情報源がある可能性がある．もし A からの引用を多く用いていた場合，重要なメッセージを発信している投稿者 A の方が重要であるといえる．フォロワー数の多い投稿者は，多くの投稿者に情報を提供しているため影響力があるといえる．しかし我々は影響力がある投稿者は重要であるという考えではなく，著名人ほど重要であるとも思わない．それよりは，影響力のある投稿者がどのような投稿者をフォローして，どのような情報を得ているかの方が重要であると考えている．この指標を扱うことによって，我々は情報源の方が高い数値がつくという可能性があると考えている．以下でこの考えを基にした重要度の算出方法を示す．

U と V の要素数を考えたときに，我々は $\frac{n(V)}{n(U)}$ という値を A の重要度を示す値として提案する．つまり $\frac{n(V)}{n(U)}$ が大きいほど，A は重要な投稿者といえるのではないかと考える．$n(V)$ は，$n(U)$ が増えていくにつれ増加する．ところが，$\frac{n(V)}{n(U)}$ が小さくなるという場合は，U に比べて V が比較的少ないか，U の中で相互にフォローしている投稿者が多いかで起こることである．U に比べて V が比較的少ない場合，U の投稿者は比較的フォロワー数の少ない投稿者であるといえる．よってフォロワー数の多い投稿者がフォローしている投稿者が重要であるという考えを基にすると，$\frac{n(V)}{n(U)}$ が小さいほど重要度は小さくなることになる．U の中で相互にフォローしている投稿者が多い場合，A と U の投稿者は密接につながっているといえる．つまり A と U の投稿者は閉鎖的なコミュニティを形成しているといえる．ここで重要な投稿者は，時間が経てば閉鎖的なコミュニティの中に収まらないネットワークを形成すると考えると，$\frac{n(V)}{n(U)}$ が小さくなるほど，A の重要度を示す値が小さくなることになる．また，V の投稿者は U の投稿者が A のメッセージを RT した時に読む人々であるので，$n(V)$ が小さいと影響力は小さいと考えるのは自然である．本来多いほど良いとされて

Algorithm 2 User-value2

Require: Blogger set X

for all Blogger A in X **do**
  set U_A to $\emptyset$
  set $n_A(U)$ to 0
  set $n_A(V)$ to 0
  for all B such that B is a follower of A **do**
    $U_A \leftarrow U_A \cup \{B\}$
    $n_A(U) \leftarrow n_A(U) + 1$
    for all C such that C is a follower of B and $C \notin U$ **do**
      $n_A(V) \leftarrow n_A(V) + 1$
    end for
  end for
  calculate $\frac{n_A(V)}{n_A(U)}$
end for

いた $n(U)$(A のフォロワー数) を分母に持ってくることで，フォロワー数の順番とは全く違った結果になると考えられる．

Algorithm2 は以下のような流れになる．

（1） タイムラインから十分な数のメッセージを取得

（2） 取得したメッセージの各投稿者に対して，その投稿者をフォローしている投稿者 (U の要素) のフォロワー数と，さらにその投稿者をフォローしている投稿者のうち U に含まれない投稿者 (V の要素) のフォロワー数を取得

（3） 各投稿者に対して，$\frac{n(V)}{n(U)}$ を計算する

こちらの手法は前述の RT を用いた手法と違い，投稿者に付与される値は一つである．この数値によって投稿者の重要度を表す．これ以降この値を FF 値と呼ぶことにする．

3.2 二つの重要度を合わせた混合手法

我々は 3.1.1 節において，投稿したメッセージの RT された数から投稿者の重要度を推定した．そして 3.1.2 節では，投稿者のフォロワーのフォロワー数から投稿者の重要度を推定した．投稿者自身のメッセージから判断するのは主観的な判断であり，投稿者のフォロワーのデータから判断するのは客観的な判断といえるので，これらは違う側面から投稿者を判定している．この二つの側面から評価することにより，どちらかに偏ることのない評価ができると考えている．二つの側面による指標を反映させるためにこれら二つの評価値を合わせる事を考える．

3.1.1 節と 3.1.2 節で算出した二つの数値を合わせて，重要度 I を算出する．具体的には 3.1.1 節で算出した (S, T) の組をある割合で組み合わせて，3.1.2 節の値を掛け合わせることで実現する．式は以下のようになる．

$$I = (s \cdot S + (1 - s) \cdot T) \cdot FF \qquad (0 \leqq s \leqq 1)$$

ここで s は S と T をどのような割合で組み合わせるかを決める値である．このようにして算出された指標 I を利用者に提示することによって，投稿者は投稿者の特徴や重要度を得ることができる．この数値の有用性を実験により明らかにする．

4. 評 価 実 験

4.1 実験の目的

我々は，フォロワー数というのは投稿者の人気を示すものであり，重要度を示す値ではないと考えている．そして投稿者の重要度は投稿したメッセージの RT 数とフォロワーのフォロワー数というもので推定できると考えている．そのため単にフォロワー数の多い順番で並べたランキングよりも，3. 章で提案した手法の方が重要度を示す指標になると考えている．この仮定が正しいことを示すために，取得したデータを基に 3.1.1 節で求めた (S, T)，3.1.2 節で求めた FF，3.2 節で求めた I と単純なフォロワー数のランキングを比較した．

4.2 実 験 手 順

本稿では Twitter を対象に実験データを作成した．Twitter ではメッセージにハッシュタグと呼ばれる「#」で始まるタグを付与することができる．そこで今回は 2010 年 1 月 1 日に投稿された,「#nhk」を含むメッセージを対象とした.「#nhk」というハッシュタグは，主に NHK の番組に対する実況を行っており，即時性が高く，一日平均数百件のメッセージが投稿される．また，NHK の番組は多岐にわたっており，ある特定の話題に限定されずに様々な投稿者がメッセージを投稿すると考えられる．特定の話題に限定されないコミュニティの方が，限定された閉鎖的なコミュニティよりも個人の興味によらない重要度というものが表れやすいと考えたため，我々はこのようなデータを選択した．その結果，収集したメッセージ数は 1163 件，それらのメッセージの投稿者 (RT しているメッセージがあった場合，RT 元の投稿者も含めた) は 383 人となった．実験として，まず投稿者一人あたり，どの程度メッセージを書いていたかを調査した．その後，その 383 人の投稿者に対して，それぞれ発信している情報が豊富かどうか，引用している情報が豊富かどうか，重要な投稿者といえるかどうかを人手で判断し，その結果を正解セットとした．そして 3. 章で提案した手法と，単にフォロワー数の多い順に並べた手法の精度，再現率を計算した．最後にそれぞれの手法における上位 5 人の投稿者名を調べ，異なっているかを見た．ここで精度 P と再現率 R は以下のように定義される．W を抽出結果中で適合している投稿者数，N を抽出結果の投稿者数，C を全投稿者の中で適合している投稿者数とすると

$$P = \frac{W}{N} \qquad R = \frac{W}{C}$$

で求めることができる．

4.3 実験結果と考察

まず予備実験として，投稿者が一日にどのくらいメッセージを投稿しているかを示したのが図 4 である．

この図からは，投稿者の約半数がメッセージを一件だけ投稿していることがわかる．つまり残りの半数は「#nhk」だけで一日にメッセージを複数投稿している．メッセージを用いて投稿者の重要度を算出する際に，メッセージの数は多い方が正確に推定できるといえるので，提案手法を用いるには投稿者が多数のメッセージを投稿していることが望まれる．そのため，約

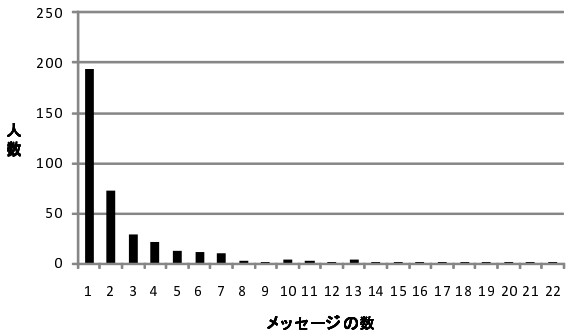

図 4　投稿者とメッセージ数の関係

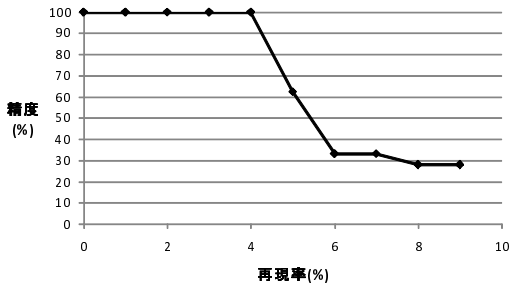

図 5　S 値の精度，再現率

半数が一日に複数メッセージを投稿しているという事は，提案手法がうまくいく可能性を示唆しているといえる．

4.3.1　引用に基づく手法の評価

次に，3.1.1 節で定義した，引用に基づく手法である S と T の精度と再現率を示す．3.1.1 節における定義式から，S は RT されているメッセージがない場合はいくらメッセージを送っていても同じ値 (V) になる．また T は RT しているメッセージがない場合に，いくらメッセージを送っていても同じ値 (0) になる．実験における条件として，メッセージ自体が本来持つ値である V は 1.0 とし，再帰的計算を 100 回行った．今回の実験で S が V より大きい値を持った投稿者 (メッセージを RT された投稿者) は全体の約 10%にあたる 40 人，T が 0 より大きい値を持った投稿者 (メッセージを RT した投稿者) は全体の約 12%にあたる 45 人だった．ランキングを行った後の S の上位 40 人，T の上位 45 人の精度と再現率を図 5，6 に示す．

まず S についてみてみる．383 人の中で，人手により発信する情報が豊富な投稿者と判断された投稿者数は 90 人だった．この投稿者達が正解セットとなる．実験結果から，S によってある程度発信する情報の豊富な投稿者を発見できているといえる．これはつまり，RT されているメッセージの内容は重要であるということを示せたといえる．上位にはフォロワー数の少ない投稿者もいるので，これまであまり人目につかなかった投稿者を見つけることができている．図 5 において上位 40 人とはこの日に RT されたメッセージを投稿した投稿者の数であり，これ以外の投稿者は全て同じ値になった．つまり，上位 40 人以外は同順位にランク付けされることとなった．

次に T についてみてみる．383 人の中で，人手により引用する情報が豊富な投稿者と判断された投稿者数は 97 人だった．T の方では，この手法によって値がついた (0 ではなかった) 投稿者は 45 人だった．図 6 を見てみると，こちらの方は再現率が増えても精度があまり下がらなかった．つまりこの日に RT をした投稿者は，半分近くが引用する情報が豊富な投稿者であるということである．このことから，引用する情報が豊富な投稿者は普段から引用する回数が多い投稿者であると考えられる．こちらも値がついた投稿者はまだ少ないので，より大規模な実験を行う必要があるが，良い結果が得られたと言える．

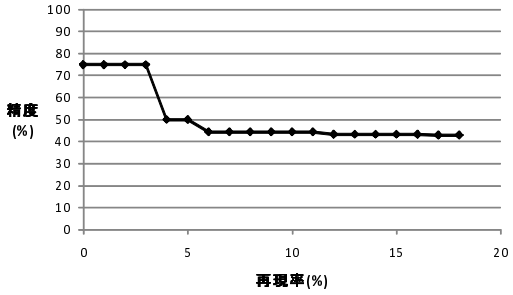

図 6　T 値の精度，再現率

そして今回の実験において最も高い値を出したメッセージは「試合終了。ガンバ 4-1 名古屋。ガンバ天皇杯連覇 #nhk #tennouhai」というものであった．新しいニュースをまだ知らない人にも伝えたいという思いから，このようなメッセージが RT されやすくなるのだと考えられる．このメッセージは投稿者が誰かは関係なく，新鮮であるという意味で価値があるので，我々の想定している重要なメッセージの一つであるといえる．またこれは，投稿者が Twitter に即時性を期待しているとも取ることができる．このようなメッセージを抽出できたことによって，提案手法の有用性が示せた．

課題としては以下のようなものが挙げられる．

- より大規模な実験
- 計算式をより根拠あるものにしていく
- 時間を考慮したメッセージの評価

計算式に関しては，正規化というものを単にメッセージ数やフォロワー数で割ることによって実現しているので，RT 数とメッセージ数や RT 数とフォロワー数の関係などを調べることによって，より正確な計算式になると考えられる．また，メッセージの評価としては，今の所同じメッセージを RT したメッセージは全て同じ値にしている．投稿されてすぐ RT したメッセージと，時間が経ってから RT したメッセージは重要度が変わると考えられる．また，RT したメッセージをさらに RT したメッセージは二次情報といえ，これも重要度が下がると考えられるのでその辺りも考慮していく必要がある．

4.3.2　選択関係に基づく手法と選択数の多さを用いた手法の比較

次に，3.1.2 節で定義した，選択関係に基づく手法である FF の精度と再現率を調べた．その結果を図 7 に示す．

正解セットとして，人手により重要であると判断された投稿者は，383 人中 103 人であった．図 7 を見ると，精度がある程度の高さを保っていることから，ある程度重要な投稿者を発見

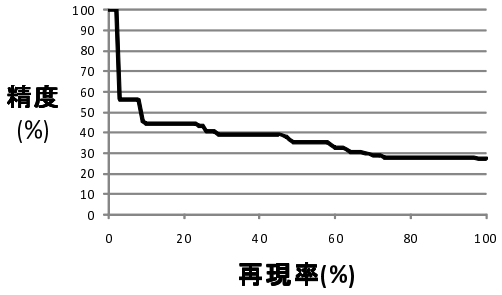

図 7　FF 値の精度，再現率

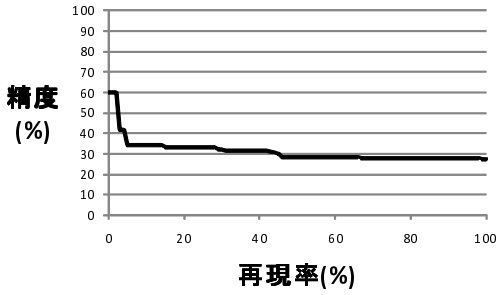

図 8　フォロワー数の精度，再現率

できているといえる．また，FF のランキングの上位 5 人のフォロワー数を調べてみたところ，一番少ない投稿者で 47 人であり一番多い投稿者で 706 人だった．このことから，FF の算出方法においてフォロワー数を用いているが，フォロワー数の多さと FF の高さにはあまり関係がないといえる．よって，ランキングの上位にいる投稿者がますます上位の立場を堅固にするということもない．そしてフォロワー数が少なくても重要な投稿者と認識される可能性が十分あるので，Twitter を始めたばかりの人にも従来より比較的簡単にフォロワーがつくことがある．また上位の投稿者のメッセージを見たところ，ある特定の話題に偏ったりすることもなく，かといって一般的な話題だけではなく個人的な内容のメッセージも多数見られた．つまり個人の興味によらない指標となっているといえる．そのため FF を使うことによって，フォロワー数と同じように，違うコミュニティに属している 2 人の投稿者を容易に比較することができる．比較のために，同じ正解セットに対して，単純にフォロワー数の多い順に並べたランキングの精度と再現率を調べた．その結果を図 8 に示す．

フォロワー数によるランキングと比較しても良い結果が得られたため，FF は指標として十分使える可能性があると考えられる．

課題として，フォロワー数の非常に多い投稿者への対応が考えられる．FF によるランキングの上位の投稿者のフォロワーを見てみたところ，多くにフォロワー数が 20 万人を超すような投稿者がみられた．FF は，フォロワーのフォロワー数の平均のようなものであるので，一人そのような投稿者がいると FF の値が上がる．つまり，FF はフォロワー数の非常に多い投稿者にフォローしてもらっているかどうかという指標になってしまっている．また，フォローをしてもらった投稿者にフォロー し返す「フォロー返し」というものも多く見られる．そのため FF が高い数値を示していても，投稿者の投稿するメッセージが重要であるわけではなく，フォロワー数の多い人にフォロー返しをしてもらっただけという可能性がある．それらの問題を考えるには，投稿者が誰をフォローしているのかも考慮する必要があると考えられる．

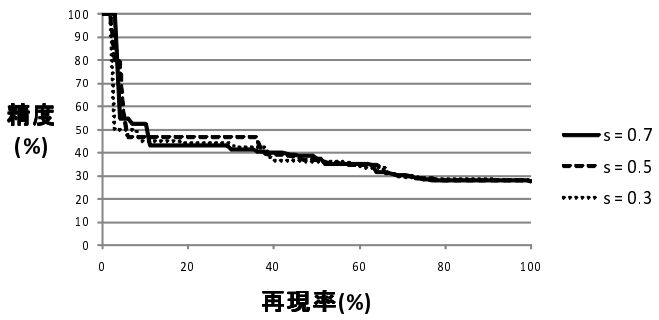

図 9　I の精度，再現率

S	T	FF	フォロワー数
tirashiori	hakkinton	kuya_00	KATOKICHIcoltd
wakakit0	bottonbenjo	magurohonsha	burarimachi
ESQ_JPN	bgyfromosaka	yanagi_moon	gopochan
tub0yaki_shiro	asante8	UmiSola	hashtagsjp
kim_take_mac	miyaby	hajime0130	mikeexpo

$I(s=0.3)$	$I(s=0.5)$	$I(s=0.7)$
kuya_00	kuya_00	kuya_00
hakkinton	hakkinton	UmiSola
nkeisuke	nkeisuke	hajime0130
bgyfromosaka	UmiSola	Mukunokiy
miyaby	hajime0130	Otecchi

表 1　それぞれの手法で抽出した上位 5 人の投稿者の ID

4.3.3　混合手法の評価

次に，3.2 節で定義した，混合手法である I を用いて，$s = 0.3, 0.5, 0.7$ の三通りについて精度と再現率を調べた．

図 9 を見ると，この三通りにあまり違いはないことがわかる．. これは発信する情報が豊富な投稿者の方が重要なのか，引用する情報が豊富な投稿者の方が重要なのかは判断できないということだといえる．ただ，この三つ全てにおいて，フォロワー数によるランキングを上回っていたので，提案手法の有効性を示せたと考えている．課題としては，I の算出方法の改善が挙げられる．S や T と FF を比較したときに，FF の方が大きな値になっており，I の値が FF の値に大きく左右されてしまった．また，s もどれが一番適切かは決まっていない．そのため，何らかの方法でこれらの値を正規化して計算すると，よりよい結果を導く可能性がある．また，どちらかに偏っている投稿者の方が重要であるのならば，S と T を合わせずに二つ組のまま用いる方がよいかもしれない．

最後に，全ての手法においてランキングを行った際の，それぞれの手法の上位 5 人の投稿者 ID を表 1 に示す．この表 1 を

見ると，I によるランキングの上位が FF によるランキングの上位と似ていることがわかる．ここから I が FF に大きく左右されていることが見て取れる．また，フォロワー数の多さで並べた順とは違う結果になっている．つまり，フォロワー数の多さでは見つけることのできない投稿者を発見することに成功している．その点で，これらの指標の新たな可能性を示せている．

全体を通して，これらの提案手法が従来にはない新たな指標となる可能性を秘めていることがわかった．改良の余地はあるものの，投稿者の妥当な重要度推定が十分可能であるということを示せたと考えられる．

5. おわりに

本稿では，マイクロブログにおいて重要な投稿者を発見するために，二つの側面を用いて重要度を推定する手法を提案した．一つ目は重要度を推定する投稿者本人のメッセージの引用回数を基にする手法である．ここでは発信する情報が豊富かどうかと，引用する情報が豊富かどうかという二つの指標を相互的に定義し，再帰的に計算することによって重要度を算出した．二つ目は，投稿者の選択関係を表すネットワークを用いて，投稿者の選択関係から投稿者の重要度を算出する手法である．具体的には投稿者を選択している投稿者の数を分母に，さらにそれらの投稿者を選択している投稿者の数を分子にした値を重要度とした．

引用に基づく手法は投稿者自身を評価基準にしているが，選択関係に基づく手法は周りからの評価を評価基準にしている．この二つの視点からの手法を合わせることによって，投稿者を多面的に評価できると考えた．

そして実験では引用に基づく手法と選択関係に基づく手法と，二つを合わせた手法の三種類の提案手法の有用性を明らかにした．引用に基づく手法の結果は，精度が 20%〜50%となっていた．比較的高い値となった要因として，引用回数を用いた点が挙げられると考えている．引用は他の投稿者からの評価と捉えられるので，引用が多いメッセージは他の投稿者からの評価が高いため，重要度が高いと言えたと考えられる．今後の課題としては，マイクロブログというのは普通のブログに比べて即時性が高く，時間が経つと価値が失われるメッセージが多いと考えられる．よって時間を考慮したモデルを考えることが精度を上げるうえで重要であると考えられる．

選択関係に基づく手法の結果は，精度が 40%〜50%となっていた．この要因として，多数の投稿者に選択されている投稿者は元々多数の投稿者とつながっているので，選択する投稿者を吟味して決定するという可能性がある．今後の課題としては，非常に多い投稿者に選択されている投稿者によって，ランキングの精度が悪くなっている可能性があるので，選択した投稿者数の対数を取るなどして，より妥当な計算式にしていく必要がある．

二つを合わせた手法の結果は，精度が 30%〜50%となっていた．この要因としては，引用に基づく手法と選択関係に基づく手法という違った側面から定義した重要度を合わせたことにより，多面的な判断が可能になったことが挙げられる．ただ選択関係に基づく手法によって算出された重要度が高く反映される結果になったため，選択関係に基づく手法と結果が大きく変わらなかった．今後は合わせる際に正規化する事を考えていくべきである．

まとめとして，提案した手法は選択数の多さを用いた手法を上回り，新たな指標としての可能性が示せた．今後の課題として，より大規模な実験をすることによって信頼性の高い結果を得ることと，マイクロブログの特性をさらに分析して重要度の算出に反映させていくことを考えている．

謝辞 本研究の一部は，文部科学省科学研究費補助金 (課題番号 20300036, 20500104, 21013026, 20700101) によります．ここに記して謝意を表します．

文　献

[1] K. Fujimura, T. Inoue, and M. Sugisaki. The eigenrumor algorithm for ranking blogs. In *WWW Workshop on the Weblogging Ecosystem*, 2005.

[2] A. Kritikopoulos, M. Sideri, and I. Varlamis. BlogRank: ranking weblogs based on connectivity and similarity features. In *Proceedings of the 2nd international workshop on Advanced architectures and algorithms for internet delivery and applications*, p. 8. ACM, 2006.

[3] 中島伸介, 舘村純一, 原良憲, 田中克己, 植村俊亮. 重要な blogger 発見を目的とした blog スレッド解析手法. 知能と情報, Vol. 19, No. 2, pp. 156–166, 2007.

[4] N. Agarwal, H. Liu, L. Tang, and P.S. Yu. Identifying the influential bloggers in a community. In *Proceedings of the international conference on Web search and web data mining*, pp. 207–218. ACM, 2008.

[5] Daniel M. Romero Bernardo A. Huberman and Fang Wu. Social networks that matter: Twitter under the microscope. *First Monday*, Vol. 14, No. 1-5, January 2009.

[6] A. Java, X. Song, T. Finin, and B. Tseng. Why we twitter: understanding microblogging usage and communities. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, pp. 56–65. ACM, 2007.

[7] Owen Phelan, Kevin McCarthy, and Barry Smyth. Using twitter to recommend real-time topical news. In *RecSys*, pp. 385–388. ACM, 2009.

[8] 岩木祐輔, アダムヤトフト, 田中克己. マイクロブログにおける有用な記事の発見支援. *The First Forum on Data Engineering and Information Management (DEIM)*, pp. A6–6, 2009.

[9] 桑原雄, 稲垣陽一, 草野奉章, 中島伸介, 張建偉. マイクロブログを対象としたユーザ特性分析に基づく類似ユーザの発見および推薦方式. 情報処理学会データベースシステム研究発表会, Vol. 149, No. 18, pp. 2B–2, 2009.