マイクロブログにおけるユーザの繋がりを利用した意見収集

河村 崇博† 浅野 泰仁‡ 吉川 正俊‡

†, ‡京都大学情報学研究科 社会情報学専攻 分散情報システム分野 〒606-8501 京都市左京区吉田本町

E-mail: †comssa@db.soc.i.kyoto-u.ac.jp, ‡{asano, yoshikawa}@i.kyoto-u.ac.jp

あらまし 高度情報化が進み携帯端末や家庭用計算機が普及した近年では，インターネット上に散らばる情報は爆発的に増加している．これら情報の扱いは難しく，ユーザは様々なソースからの情報を加味した上で意思決定を行うことが望ましい．近年ユーザ数が急増し，注目を集めているマイクロブログとして Twitter が存在する．Twitter では様々な背景を持ったユーザが多様な意見を発信しており，またユーザ間の関係性はフォローという形で表されている．本研究ではこのフォロー関係を利用することで，お互いにある程度関係性の薄いユーザの集合を求め，偏りのない情報ソースからの意見を収集する手法を提案する．また本研究はこのような意見群を用いて，意見の分布や新知識等を得ることで，ユーザの情報活用における意思決定の補助を目標とする．

キーワード ブログ・ソーシャルネットワーク，半構造・グラフデータマイニング

1. はじめに

近年では，家庭用計算機や携帯端末の普及により人々がインターネットに接続する機械が増加し続けており，また，SNS（ソーシャル・ネットワーキングシステム）やマイクロブログの流行により誰しもが容易にインターネット上に情報を発信することが可能となってきている．しかし，それらの多くは真偽が定かでない情報であり，各インターネットユーザには情報を適切に取捨選択する能力が求められる．あるトピックに対して情報の取捨選択を決定する方法としては，様々なソースからの情報を参照したうえで，総合的に自分が信頼できると考える情報を選ぶことが望ましい．また，それらの情報ソース間の関連性はできる限り薄いものであるとよい．情報ソースが偏ると，情報の指向も偏ると考えられ，結果的にユーザが偏った考えを選択することに繋がるかもしれない．逆に，全く異なる情報ソースを参照することで，その発信者の持つ背景に応じた新規性のある情報が手に入ると予想できる．

近年の代表的なマイクロブログサービスとして，Twitter がある．Twitter ではユーザは呟き（ツイート）として 140 文字以内の短文を投稿することができ，この手軽さ故に多くのユーザが日常の出来事や特定のトピックに対する自分の考え等を発信している．またユーザは自分が気に入った他のユーザを基本的に無許可でフォローすることができ，フォローを行うことで簡単にそのユーザの呟きの閲覧やそのユーザとのコミュニケーションをとることができる．この性質から Twitter では似た背景を持つユーザ同士がフォローという形で繋がる傾向がある．

本研究では，Twitter の持つ似たユーザが繋がりやすい性質を利用することで，お互いに関係性の薄いユーザの集合を求め，この集合から意見を抽出することで，情報ソースによるバイアスのかかっていない意見の収集方法を提案する．また，本研究では，この手法により集まった意見を利用して，情報の取捨選択におけるユーザの意思決定を補助することを目標とする．

本論文ではまず 2 節で情報収集に関連する研究の紹介を行う．そして 3 節で Twitter の特徴について詳しく触れ，4 節でどのように関連性の薄いユーザ同士の集合を求めるかを説明する．そして 5 節で実験計画を説明する．また，6 節で今後の展望について述べる．

2. 関連研究

ここでは本研究を行う上で参考としたシステムや研究について紹介をする．

インターネット上に散らばる情報を収集し分析する研究として NiCT が研究開発を行った WiSDOM[1]がある．WiSDOM ではあるトピックに対するニュース記事やブログ等を分析し，意見の種類や賛否の傾向を計算し，発信者のや意見の分布の表示を行う．同時に関連するキーワードや主要な文章なども提示する．あるトピックに対するキーワードや該当するブログ記事を収集するシステムでは他に株式会社きざしカンパニーの提供する kizasi.jp[2]がある．

Twitter におけるユーザ分類の研究としては，眞野ら[1]がユーザの持つお気に入りツイートの情報を利用してクラスタリングする手法を提案している．この研究では 2 人のユーザが共通してお気に入り登録しているツイートの数をそのユーザ間の類似度として捉え，さらにお気に入りツイートからその 2 ユーザがどのようなワードで関係性を持っているかを示した．Hannon ら[2]はユーザの発信したツイートやフォロー情報からユーザ間の類似度を計算し，あるユーザと類似した他のユーザの推薦を行う手法を提案している．この研

[1] http://wisdom-nict.jp/
[2] http://kizasi.jp/

究ではツイート内容から推薦ユーザを決定する Contents-Base Filtering と，フォロー関係から推薦ユーザを決定する Collaborative Filtering を組み合わせた場合の性能比較を行なっており，2 つのフィルタリングを組み合わせたパターンが良い結果を出したことを示した．これらの研究は似た趣向を持つユーザを導出する手法であるが､本研究ではあえて Twitter におけるフォロー情報と SimRank を組み合わせる手法を採用した．何故ならば、お気に入り機能の場合、気になるツイートのブックマーク以外にも例えば自分に当てられたツイートへの感謝のような意味合いで使われることがあり，ユーザにより用途が異なるため類似度の計算としては弱いのではないかと考えたためである．また Hannon らの提案したツイート情報を利用する手法も、ユーザによってツイートの使い方に差異がありお気に入り機能を利用した場合と同様のことが考えられるためである．本研究ではフォロー情報であれば，どのユーザによっても利用方法に差異がないのではないかと予想した．

Wang[3]は Twitter における uniform なユーザのサンプリングの手法を提案している．これはランダムウォークを用いたサンプリングの応用となっている．しかし，本研究ではサンプリングされたユーザにバイアスがかかっていないことを保証したいため利用はしなかった。ただし，ユーザの繋がりを示すグラフを拡張したい場合には，計算量を抑えるために利用できる可能性はある．計算量の問題については 6 節で触れる．

3. Twitter

3-1. Twiter の仕組み

Twitter は近年ユーザ数が急増し，そのアカウント数は世界で総計 10 億を超えるマイクロブログである．Twitter ではユーザは 140 文字以内の短文を呟き (ツイート)という形で発信することができる．またユーザは他のユーザをフォローすることでそのユーザの呟きを自分のタイムライン上で閲覧することができる．

Twitter が他の SNS と大きく異なる点として，基本的にユーザのフォローに相手の許可を必要としないことが挙げられる．そのため，現実での交流の有無に関わらず，同じ興味を持つユーザ同士が繋がりやすいという特徴を持っている．また Twitter には鍵付きと呼ばれるプライベートアカウントが存在し，プライベートアカウントに対しては承認を得たユーザしかそのツイート，フォロー，フォローウィー等の情報を閲覧できない．

Twitter において，あるユーザの呟きを見る方法は大きく分けて 2 通りが考えられる．一つは上記のように

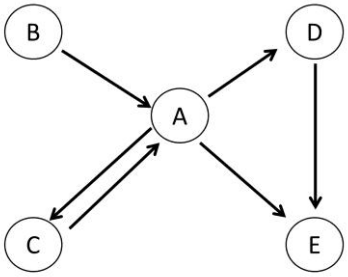

図 1

ツイートを閲覧したいユーザをフォローすることである．一般的に，これは自身の興味のあるユーザに対して行われ，日常的にタイムライン上でそれらのユーザのツイートを見ることとなる．この場合は必然的に偏った情報ソースからの意見が多くなると予想できる．もう一つが，呟き検索機能である．こちらは主にニューストピックやその時点で流行りとなっているワードに関する情報を収集したい時に用いられる．こちらは発言ユーザの背景に関わらず新しい順や，注目されている発言順にツイートを収集することができる．こちらの機能では，発言ユーザの背景を詳細に調べること無しには，情報ソースの偏りは判別できない．

3-2. Twitter からのグラフ作成

twitter におけるフォロー関係は一方向もしくは双方向であり，ユーザの繋がりは，図 1 のようにユーザをノード，フォロー関係を辺とする有向グラフとして表すことができる．Twitter ではユーザがフォローしているユーザのことを「フォローウィー」，逆にあるユーザをフォローしているユーザを「フォロワー」と呼ぶ．図 1 の A を例にとると，B，C が A のフォロワー，C，D，E が A のフォローウィーとなる．

一般的に，Twitter 上では，ユーザは自分の興味に合うユーザをフォローするため，同じユーザからフォローされているユーザ同士は似ているユーザ同士である可能性が高い．図を例にとると，同じユーザ A からフォローされているため，ユーザ C，D，E は似ているユーザである可能性が高い．逆に誰からもフォローをされていないユーザ B は，他のユーザと比べて異なる背景を持つユーザと考えられる．また当然のことであるが，ユーザ A，C のように相互にフォローしているユーザの組み合わせは実際にユーザ同士の交流を行なっていたりと親密度の高い≒非常に似ているユーザと考えることができる．

Twitter ではこのフォロー関係を上手く利用することで，あるユーザと確かに関係のないユーザを導出できる．このようにして関係の薄いユーザを次々にピックアップすることで，偏りのない情報ソース群が得られると考えることができる．

4. 意見収集のプロセス

4-1. 類似度計算

ユーザ同士の関連性は類似度で表される．有向グラフにおけるノード同士の類似度の計算方法としてはSimRank[4]があり，Twitter のフォロー機能が持つ性質に大凡合致している．SimRank は類似したノード同士からリンクされているノード同士は類似しているという考えに基づいた計算方法であり，2 ユーザ a，b 間の SimRank の値 $R(a, b)$は式 (1)，(2)で表される．

$$R_0(a,b) = \begin{cases} 0 & (if\ a \neq b) \\ 1 & (if\ a = b) \end{cases} \qquad (1)$$

$$R_{k+1}(a,b) = \frac{C}{|I(a)||I(b)|} \sum_{i=1}^{|I(a)|} \sum_{j=1}^{|I(b)|} R_k\left(I_i(a), I_j(b)\right) \qquad (2)$$

一般的なグラフでは $I(a)$は点 a にリンクされている辺の集合を表すが，Twitter においては，$I(a)$，$I(b)$はそれぞれ，ユーザ a，b のフォローウィー集合であり，$I_i(a)$，$I_j(b)$ はその要素となる．また C は定数である(C=0.8 等が用いられる)．式 (1)，(2) に表されるように類似度は再帰的に計算され，反復の回数は k で表される．これと同様にしてフォロワー集合を計算に用いることもできる．Jeh ら[4]の論文によると，SimRank は n=4 程度で，一定の値に収束する傾向があることが示されている．

4-2. 意見収集の手法

このセクションでは実際の意見収集のプロセスを説明する．このプロセスのフローチャートを図 2 に示した．

意見収集では，まず基準となるユーザを決め，基準ユーザ集合の要素とする．基準ユーザは今後，意見収集する際のユーザの関係性の比較対象となる．即ち，基準ユーザと関係性の薄いユーザの意見を収集することとなる．その後，そのユーザを中心としたフォローグラフを生成する．次に SimRank を用いてユーザ間の類似度を計算する．ここまでが前処理である．

次に何らかのクエリが与えられた場合，まずクエリワードを含むツイートを発信しているユーザを全てフォローグラフの中からピックアップし，その中から基準ユーザとの類似度が閾値以下のユーザをランダムに選択する．そして，そのユーザを基準ユーザ集合に追加し，得られたツイートを保存する．そして今度は基準ユーザ集合の全てのユーザとの類似度が閾値以下のユーザを選択し，同様の処理を繰り返していく．このプロセスを，クエリワードを含むツイートを発信しておりかつ閾値を下回るユーザが見つからなくなるまで，もしくは規定数の意見が集まるまで繰り返す．

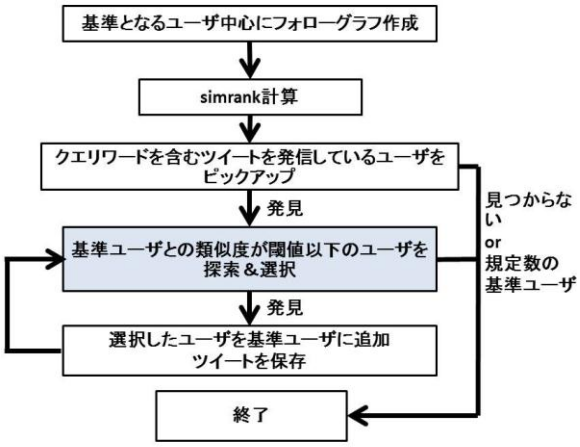

図 2

尚，プライベートアカウントについては，ツイート，フォロー，フォローウィーを収集できないため一見意見収集に利用出来ないように見えるが，他のユーザからのフォロー情報は利用できる．よって，これらはフォローグラフ生成と，SimRank 計算のみに利用し，閾値以下のユーザ探索以降のプロセスでは無視することとする．

このプロセスにより情報ソースに偏りのない意見集合が求められると予想される．

5. 実験計画

今回の実験のために，ある Twitter ユーザ（学生）を中心に 2 ホップ先までのフォローウィー (合計 843501 ユーザ)を収集した（2012/12/21～2012/12/25）．また，各ユーザについて呟きを 200 件ずつ取得しておく(2013/01/08～)．この集合に対して，前セクションで提案した手法を用いて，閾値や基準ユーザの規定数を変化させた時にそれぞれどのような結果が得られるかを調べる．同時に，意見収集結果を比較するための対象として，逆に似通ったユーザ(≒類似度の高いユーザ)の集合からも意見を収集する．こちらはユーザ探索の際に，逆にある閾値以上のユーザを集めるようにすることで実現できる．

手法の評価としては，2 つの方法で得られた意見集合がどのように異なるかの考察，また実際に新規性のある意見や知識が得られるかで判断をする．

6. 今後の展望

前セクションで立てた計画に沿って実験を行い，結果を考察することが第一の課題となる．

また，今回提案した手法は計算量が非常に多い点で問題がある．例えば，フォローグラフを全体的に拡大することを考える．ユーザの平均フォローウィー数をnとし，mホップ先まで拡大する場合，その計算量は$O(n^m)$となる．SimRankの場合，計算の反復回数をK，$|I(a)||I(b)|$の平均をd_2とした時，その時間計算量は$O(Kn^{2m}d_2)$となる．このため，無闇にグラフを拡大することは望ましくない．計算量の増加を最小限にしてフォローグラフを拡大する手法の考案が望まれる．現在検討している方法としてはグラフの部分的な拡大である．あるユーザのフォローウィーを収集しフォローグラフを拡大することを考えた時，意見収集の効率を最大化するユーザを発見することで，そのユーザ中心のグラフ拡大により問題の改善を図ることができる．

参　考　文　献

[1] 眞野裕也，青山俊弘，“ミニブログユーザの記事指向を用いたクラスタ発見”，Journal of JACT, Vol.15, No.3, pp.43-46 (2010)

[2] John Hannon，Mike Bennett，Barry Smyth，“Recommending Twitter Users to Follow Using Content and Collaborative Filtering Approaches”, Proceedings of the fourth ACM conference on Recommender systems, pp.199-206 (2010)

[3] Tianyi Wang, Yang Chen, Zengbin Zhang, Peng Sun, Beixing Deng, Xing Li, “Understanding GraphSampling Algorithms for Social Network Analysis”, proceedings of the ICDCSWorkshops, 2011, pages 123-128

[4] Glen Jeh, Jennifer Wisdom, “SimRank: A MEasure of Structuring-Context Similarity”, proceeding of the KDD, 2002, pages 538-543

[5] Minas Gjoka, Maciej Kurant, Carter T. Butts, Athina MarkOpoulou, “Walking in Facebook:A Case Study of Unbiased Sampling of OSNs”, proceeding of theINFOCOM, 2010, pages 2498-2506