短期持続型情報取得を目的とした Twitter 上での アドホックフォローネットワーク自動構成手法

田島 真悟† 牛尼 剛聡‡

†九州大学芸術工学部 〒815-8540 福岡県福岡市南区塩原 4-9-1
‡九州大学芸術工学部 〒815-8540 福岡県福岡市南区塩原 4-9-1
E-mail: †2ds13084s@s.kyushu-u.ac.jp, ‡ushiama@design.kyushu-u.ac.jp

あらまし Twitter は多くの利用者を有する代表的な SNS サービスのひとつであり、ユーザは実世界で起こっている様々な事象に関してリアルタイムな情報を投稿する。Twitter においてユーザは他のユーザをフォローすることで自らのフォローネットワークを構成し、フォローしたユーザが発するツイートをリアルタイムに取得できる。しかし、一般的に Twitter 上でのフォローネットワークは静的なものであり、目的に応じて変更されることはない。一方、ユーザが地域の祭りや音楽フェスティバルといった実世界イベントに参加した際、そのイベント内での情報を短期間だけ持続的に取得したいという情報要求を有するケースは多い。本研究ではこの問題を解決するため、Twitter 上で目的に応じたアドホックフォローネットワークを構成する手法を提案し、ユーザの短期持続型情報要求に応えることを目的とする。そして、プロトタイプシステムを利用した実験により提案手法の有効性を評価する。

キーワード Twitter、アドホックフォローネットワーク、情報共有

1. はじめに

近年、マイクロブログをはじめとする SNS が注目を集めている。多数の SNS の中で利用者数、投稿数ともに多いマイクロブログの代表例として Twitter がある。

Twitter では、ユーザは 140 文字以内でメッセージ(ツイート)を発信でき、そのツイートはフォローネットワークを通じて伝播される。特に、ユーザは自らのフォローネットワークを構成するためにフォロー対象のユーザからの承認を受ける必要がない。このため、自分の興味や関心に合わせてフォローする対象となるユーザを選択し、自分にとって必要な情報を得ることができる。

しかし、一般的に Twitter ユーザは友人、知人、家族、著名人をフォローすることが多く、構成したフォローネットワークを日常的に変化させることはない。つまり、Twitter において、多くのユーザは日常的に閉ざされたコミュニティ内でのみ情報共有を行なっていると考えられる。こうした特徴のため、Twitter は以下のような状況では有効的でない。

- 催し物やコンサート等の実世界イベントに参加したときに、イベントに関する情報を取得したい。
- スポーツの試合やテレビ番組を鑑賞しているときに、そのスポーツやテレビ番組に関しての情報を取得したい。
- 事故や災害に巻き込まれたときに、その事故や災害に関する情報を取得したい。

上記の種類の情報要求に関して共通する事項は、数時間から数日にかけて継続的な情報取得が要求されることである。すなわち、ハッシュタグやキーワードによる検索で上記の情報取得を継続的に実現する場合、検索要求を何度も繰り返し実行する必要があり効率的でない。従って、短期限定のイベント参加者内フォローネットワークを構成する手法が効率的であると考える。しかし、Twitter では同じ状況にあるユーザ間で動的にフォローネットワークを構成する機能を提供していない。そこで本研究ではこの問題を解決するため、ユーザの状況と目的に応じた一時的なフォローネットワークを自動的に構成する手法を提案する。本手法では、同じ状況を共有している Twitter ユーザを発見し、そのユーザを一時的にフォローすることによって、その状況に関する情報を Twitter からリアルタイムに取得するアプローチをとる。本稿では、このとき構成されるフォローネットワークをアドホックフォローネットワークと呼び、同じ状況を共有しており、その状況に関する有益な情報を提供する可能性がある Twitter ユーザをレポータと呼ぶ。図 1 にアドホックフォローネットワークの概念図を示す。

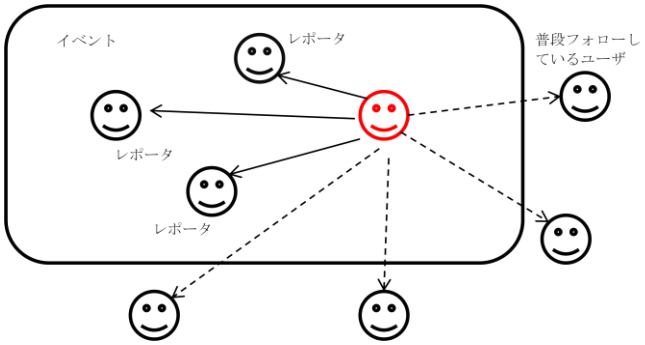

図 1 アドホックフォローネットワークの概念図

図 2 に、提案システムの処理の流れを示す。本システムは以下のように動作する。

(1) システムを利用するユーザがシステムに対してクエリを入力する。ここで一般的なクエリとしては、ユーザが参加

しているイベント名、鑑賞しているテレビ番組名などが考えられる。

(2) システムはそのクエリに対して適切なレポータを返すために、レポータ候補者を収集する。

(3) レポータ候補者のツイートから、それぞれの候補者がレポータとしてふさわしいかを評価する。

(4) 適切なレポータとして評価されたTwitterユーザを推薦する。

(5) ユーザは、推薦されたレポータを一時的にフォローすることにより、入力したクエリに関連する情報をレポータから得る。

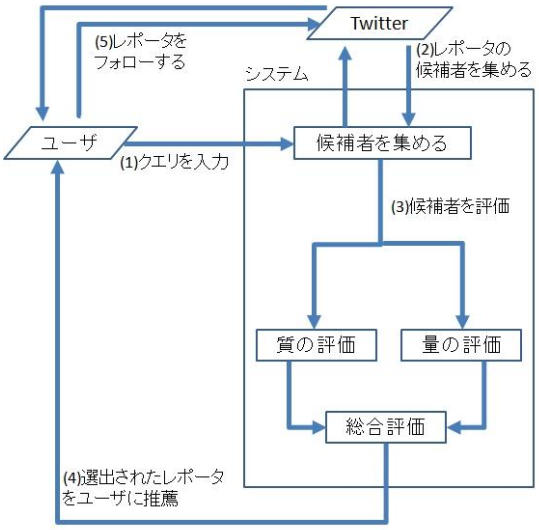

図 2 処理の流れ

2. 関連研究

Twitter は投稿数の多さ, 及び利用者の多さから数多くの研究が行われている。Bernardo ら[6]は、Twitter 上におけるフォロー・フォロワー数と実際の友人数との関連を調査する研究を行っている。また、Danah ら[7]は Twitter ユーザが利用するリツイート機能をその目的によって分類している。これらは、Twitter 自体の特徴を分析する研究である。

また、Twitter の解析によってトレンドや世論、その他の社会的事象や傾向を発見するという研究も活発に行われている。例えば、若宮ら[1]は、Twitter を解析することで、特定のテレビ番組を視聴していると考えられるユーザを抽出し、そのユーザ数からテレビ番組の視聴率を計算する手法を提案している。同様に、Jansen ら[2]は、ブランドに対する消費者の意見を Twitter から収集する手法を提案している。

さらに、Twitter はイベントや流行をいち早く発見する目的でも研究されている。例えば、榊ら[3]は、Twitter で流れているツイートを常にモニタリングすることで地震の発生をいち早く発見する手法を提案している。彼らは、「地震」「揺れた」といった単語を含むツイートを収集し、収集したツイートを support vector machine(SVM)[8] によって positive、negative の 2 種類に分類した。この場合における negative とは、実際にその瞬間に地震を体験した人のツイートを意味し、negative はそうでない人のツイートを意味する。positive に判定されたツイートの数から地震が実際に発生している確率を算出し、閾値により地震の発生を検出している。また、荒巻ら[4]は、Twitter をモニタリングすることでインフルエンザの流行を直ちに発見する手法を提案している。彼らも、前述した榊らの研究と同様、「インフルエンザ」という単語を含むツイートを SVM によって positive(実際にインフルエンザにかかっているユーザ)と negative(インフルエンザにはかかっていないユーザ)に分類することで実際のインフルエンザ患者数を割り出し、流行を検知している。同様に、杉谷ら[5]は、ツイートに付加されている位置情報を利用してローカルイベントを発見する手法を提案している。彼らは、ローカルイベントに関するツイートは、時間的にも位置的にも集中している場合が多いという事実に着目し、時間と場所の 2 つの要素でツイートのクラスタリングを行うことでイベントを発見するアプローチをとっている。

3. 提案手法

本手法では、ユーザが入力したクエリをもとに、システムがそれに関するツイートを発信しているユーザを集めてレポート候補者とし、その中からレポータを選別して推薦する。ここで、レポータとして推薦されるべきユーザは、以下のような特徴を持っていることが望ましい。

- ユーザが入力したクエリに関するツイートを継続的、頻繁に発していること
- 投稿するツイートの内容が有益であること

また、ツイート単位でなくユーザ単位で選出を行う理由として、処理時間コストの問題が挙げられる。つまり、実世界イベントに関するツイート全てに対して有益性の判断を行う場合、膨大な処理時間を必要とするため現実的でない。一方、有益なイベント情報を頻繁に発信すると期待できるレポータに対してアドホックフォローネットワークを構成する手法では、一度フォローネットワークを構成すれば以後はシステムの処理を要することなくレポータから有益な情報を取得できるため、少ない処理時間で済みシステムへの負担も小さい。

3.1 候補者の収集

本システムはユーザからクエリ(キーワード)を受け取ると、そのクエリに基づいてレポータ候補者を収集する。ここで対象となるレポータ候補者とは、例えばクエリがテレビ番組名であった場合は、そのテレビ番組の視聴者や公式アカウント等であり、クエリが街で行われているイベント名であった場合、イベントの参加者等がレポータ候補者として考えられる。

レポータ候補者を収集するために、まずユーザから与えられたクエリを用いてツイート検索を行い、そのツイートを発信している Twitter ユーザを候補者とみなす。例として、2012 年 11 月 4 日に行われた日本対オマーンのサッカーワールドカップの試合中に、「オマーン」というクエリをシステムに与えたしたときの事例を説明する。この例の場合、「オマーン」を

クエリとしてツイート検索を行った結果、『前半日本がオマーンに1点リード！』といったツイートが検索結果として抽出できる。このようにクエリを直接含むツイートを発したユーザを求めることで、レポータ候補者を収集できる。しかし、同じ状況を共有している(ここでは、同じサッカーの試合を見ている)Twitter ユーザは他にも存在すると考えられる。例えば、『ワールドカップの最終予選、日本が1点リード』というツイートは「オマーン」というクエリは含まれていないが、同じサッカーの試合について発しているツイート内容であると考えられる。このように、ユーザが与えたクエリを直接含むツイートを投稿していないユーザもレポータ候補者になる可能性がある。

上記の問題を解決するため、関連語を利用する。例えば、「オマーン」というクエリを含むツイート検索結果の内容から、「ワールドカップ」や「予選」といった単語が関連語であると推定することができれば、関連語でツイート検索を行うことによって、上のようなツイートも、とりこぼすことなく発見し、その発信者を候補者に追加できる。

3.2 関連語の発見

はじめに、関連語となりうる候補を収集する。本研究では、関連語を発見する手法としてトピック分類による方法とコサイン距離を利用した方法のそれぞれについて検証を試みた。;

3.2.1LDA を用いたトピック分類による手法

文書のトピックを分類する手法の一つとして、LDA(Latent Dirichlet Allo-cation)がある。LDA では、単語の潜在的トピックを推定することにより文書のトピック分類を行う。また、1つの文書が 1 つのトピックに属するのではなく、文書内に存在する各単語がそれぞれのトピックに確率的に属するとみなすのが特徴である。実際に LDA は小説の著者分類やニュース記事のトピック分類に利用されている。図 3 に LDA のモデル図を示す。図中に示されれるそれぞれの変数は、ディリクレ事前分布 Dir(α)および Dir(β)、トピック空間の多項分布 Multinomial(θd) 。単語空間の多項分布 Multinomial(ϕzi)、トピック数 T、文書数 D、各文書の単語数 N をそれぞれ表している。LDA の単語生成過程としては以下の通りとなる。

(1) 全トピック t においてディリクレ事前分布 Dir(β) から ϕt を抽出
(2) すべての文書 d においてもディリクレ事前分布 Dir(α) から θd を抽出
(3) 文書 d 内の i 番目の単語 wi において、抽出した文書 d の多項分布 Multinomial(θd)からトピック zi を抽出
(4) トピック zi の多項分布 Multinomial(ϕzi)から単語 wi を抽出

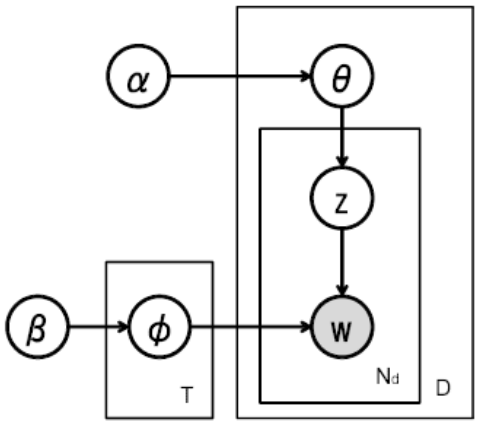

図 3 LDA のモデル

本研究では、LDA のパラメタを推定する計算手法としてギブスサンプリングを用いる。ギブスサンプリングでは、トピック j の確率分布はそれ以外のトピックの確率分布によって計算される。これを全トピックについて繰り返し計算を行うことにより、最適な ϕ と θ の値が推定される。ギブスサンプリングの定義式は以下で示される。

$$P(z_i = j | w_i = m, \mathbf{z}_{-i}, \mathbf{w}_{-i}) \propto \frac{C^{WT}_{mj} + \beta}{\sum_m C^{WT}_{mj} + V\beta} \frac{C^{DT}_{dj} + \alpha}{\sum_j C^{DT}_{dj} + T\alpha} \quad \cdots(1)$$

ギブスサンプリングの結果求められる ϕ と θ の値は以下で示される。

$$\phi_{mj} = \frac{C^{WT}_{mj} + \beta}{\sum_m C^{WT}_{mj} + V\beta} \quad \cdots(2)$$

$$\theta_{dj} = \frac{C^{DT}_{dj} + \alpha}{\sum_j C^{DT}_{dj} + T\alpha} \quad \cdots(3)$$

C^WTmj は単語 m がトピック j に割り当てられた回数、C^DTdj は文書 d がトピック j に割り当てられた回数、V は全単語数、T は全トピック数をそれぞれ表している。

本研究では、1 ツイートを 1 文書として扱い、上記のギブスサンプリングの計算を行うことでツイートのトピック分類を行う実験を試みたが、ほとんどの場合適切なトピック分類を実現することができなかった。その理由として、Twitter の場合 1 ツイートが 140 字以内という制限があることから、1 文書に含まれる単語数がニュース記事や小説の文書における単語数と比べて圧倒的に少ないからであると考えられる。LDA のギブスサンプリングでは、同一トピックでは共通の単語を含みやすいという事実を前提としていることから、1 文書における単語数が少ないとトピック推定も困難となる。さらに、LDA のギブスサンプリングは処理コストが大きいというデメリットも存在

する。本研究のシステムでは、Twitter の状況に応じたリアルタイム性の高いレスポンスが要求される。そのため、処理時間の大きいアルゴリズムを用いることは適切ではない。

3.3.2 コサイン距離を用いた手法

次に、クエリ単語とのコサイン距離を利用した関連語発見手法について述べる。本手法ではまずコサイン距離の計算を行う対象とするべき「関連語の候補単語」を収集する。関連語の候補単語として、クエリを含むツイートを検索して得られた検索結果のツイート内における共起語を利用する。「オマーン」をクエリとしてツイートを検索した際、検索結果のツイート内における共起語をその出現数順でランキング化すると、図 **4** に示す結果となった。

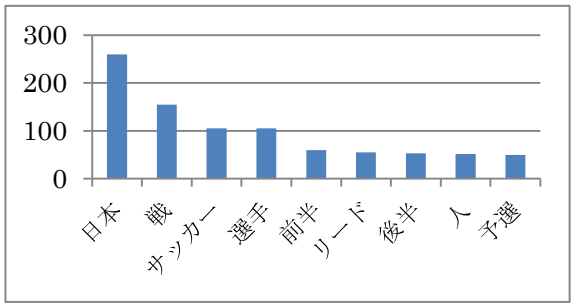

図 4 共起語(オマーンを含むツイート 1264 件中)

図 **4** で示されるような共起語の出現数を、その共起語とクエリとの関連度であると解釈するには不適切な場合がある。例えば、図 4 の中で、「後半」という語と「人」という語は、同程度の頻度で出現している。しかし、「人」という語は、その試合に関する話題以外でも数多く使われていると考えられる。従って、こうした語はクエリである「オマーン」に対する関連度は低いとみなすべきである。この問題を解決するために、共起語とクエリの単語のコサイン距離を利用する。コサイン距離は式(4)で示される。

$$\cos(X,Y)=\frac{|X\cap Y|}{\sqrt{|X||Y|}} \quad \ldots(4)$$

式(4)において、X,Y はそれぞれ図 5 に示すように、クエリの単語を含むツイート集合 X、および共起語の1つを含むツイート集合 Y を表す。図 5 では、共起語の1つの例として「後半」という語を用いている。

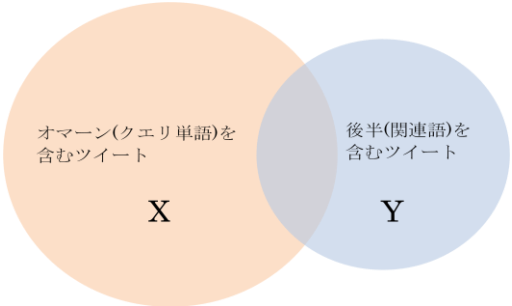

図 5 コサイン距離の概念図

コサイン距離の計算をそれぞれの共起語に対して適用した結果のうち、上位 9 件を図 6 に示す。縦軸はユーザがクエリとして与えた単語とのコサイン距離を表す。前述した「人」のような、一般語に対する値は下がり、その代わりにサッカーの試合における特有の語が高い値となっている。本研究では、このコサイン距離の値を、ユーザがクエリとして与えた単語とそれぞれの関連語との関連度とする。

コサイン距離を用いた関連語発見手法では、LDA を用いたトピック分類による手法と比べて処理コストも少なく、かつ正確な関連語を発見しやすい。そこで本研究では関連語発見手法として、このコサイン距離を用いた手法を採用する。

関連語を抽出した後、3.1 で述べたように関連語によるツイートの検索を行い、それらのツイートを発信しているユーザも候補者に追加する。

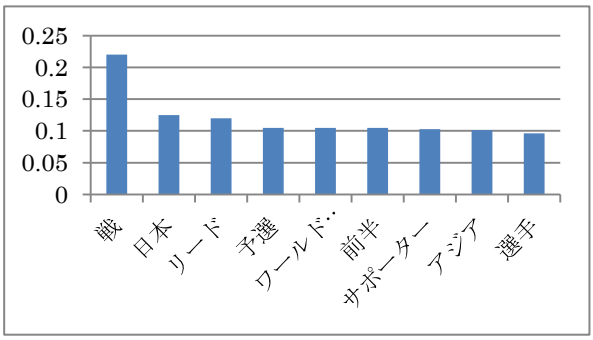

図 6 コサイン距離の計算結果

3. 3 レポータの選出

次に、得られたレポータ候補者の中からレポータを選出する。レポータとして求められる条件は、前述したように、

(1) その状況に関するツイートを頻繁に発する見込みのあるユーザ

(2) 投稿したツイートの内容が、高い情報量を持ち有益であるユーザ

である。つまり、(1) はそのユーザが発する状況関連ツイートの「量」的な側面を表し、(2) はそのユーザが発する状況関連ツイートの「質」的な側面を表しているといえる。以下に「量」および「質」を評価する方法について述べる

3.3.1 量の評価

本研究では、量の評価手法として 2 つの指標を検討した。1 つめの指標は、それぞれのレポータ候補者がクエリ入力時点から L 時間前以降に投稿した関連ツイートの件数を求め、量の評価の尺度とする方法である。本研究では L=5 時間と設定した。候補者が過去 5 時間以内に投稿したツイートの中で、3.2 で求めた関連語が 1 語でも含まれていれば、それを関連ツイートとみなす。式(5)に本手法における定義式 av(amount value)を示す。変数 t はツイートを表し、変数 T(u)は候補者 u が過去 L 時間以内に発信したツイートの集合を表す。

$$av(u) = \sum_{t \in T(u)} \phi(t) \quad \ldots(5)$$

$$\phi(t) = \begin{cases} 1 (\text{ツイート内に含まれる関連語数が1以上の場合}) \\ 0 (\text{ツイート内に関連語を含まない場合}) \end{cases} \quad \ldots(6)$$

2 つめの量の評価手法は、各候補者が過去にツイートを投稿した時刻を考慮した方法である。例えば、過去 5 時間以内に投稿したツイート数が同じ 10 件であるユーザ A とユーザ B がいたとき、ユーザ A は 4 時間前に最後のツイートを投稿しており、ユーザ B は直近 15 分以内のツイート数が多かった場合、現時刻の直後にツイートを投稿する確率が高いのはユーザ B のほうであると考えられる。従って、過去 L 時間以内のツイート数だけでなく、各ツイートの投稿時刻を考慮した計算手法を提案する。式(7)に本手法の定義式を示す。e^{-λ h(t)}は、ツイート t を投稿した時刻がより過去であるほど小さくなっていき、直近であるほど大きくなる。これにより、より直近にツイートを投稿しているユーザに対して高い評価値がつくこととなる。

$$av(u) = \sum_{t \in T(u)} \phi(t) \cdot e^{-\lambda \cdot h(t)} \quad \ldots(7)$$

$h(t)$＝現時刻とツイートtを投稿した時刻の差(単位は時)

3.3.2 質の評価

レポータ候補者が発信するそれぞれのツイートの質を評価する基準について述べる。

本研究では、質の高いツイートを情報量の多いツイートと解釈するため、ツイートが含む関連語の数を考慮する。関連語をより多く含んでいるツイートほど、クエリの話題に対する情報量が多いと考えられるためである。従って、ツイートに含まれている関連語とその関連度を利用して、個々のツイートの質の評価の尺度とする。

例えば、図 7 に示す例1のツイートの場合、含まれる関連語は「アジア」、「ワールドカップ」、「最終」、「予選」の 4 つであり、その関連度の合計値は 0.35 である。従って、このツイートの質の評価値は 0.35 と計算される。

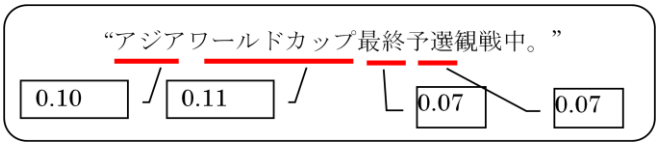

図 7 質の評価

この質の評価を、量の評価と同様にそれぞれの候補者が過去 5 時間以内に発信したツイート全てに対して行い、ツイート 1 件当たりの質の評価の平均値をとる。クエリの話題に関係のないツイートや情報量の少ないツイートまでも頻繁につぶやいているユーザは、量の評価では高く評価される可能性が高いが、質の評価においては 1 件あたりの平均値が下がることにより著しく低く評価されるため、推薦されにくくすることができる。

式(8)に質の評価値の計算式を示す。質の評価値を表す関数として、qv(quality value)を定義した。式(5)と同様に、変数 u は各候補者、変数 t はツイートを表し、変数 T は各候補者が過去 L 時間以内に発信したツイートの集合を表す。変数 w は関連語を表し、変数 W は 4.2 章で求めた関連語の集合を表す。

$$qv(u) = \frac{\sum_{t \in T} \sum_{w \in W} \varphi(t,w)\gamma(w)}{|T|} \quad \ldots(8)$$

$$\varphi(t,w) = \begin{cases} 1 (\text{ツイート}t\text{に関連語}w\text{が含まれる場合}) \\ 0 (\text{ツイート}t\text{に関連語}w\text{が含まれない場合}) \end{cases}$$

$\gamma(w)$：関連語wの関連度 $\quad \ldots(9)$

3.3.3 総合的な評価

レポータ候補者に対して計算した量の評価値と質の評価値の積を、そのレポータ候補者に対する総合的な評価値とする。総合的な評価値を計算する関数として、tv(total value)を定義する。式(5)および式(8)の評価基準により、関連ツイートを過去 L 時間以内に頻繁につぶやいているユーザ、かつそれらのツイート1件あたりの情報量が高いユーザほど、上位にランキングされる。式(10)に総合的な評価値の形式的な定義を示す。

$$tv(u) = av(u) \cdot qv(u) \quad \ldots(10)$$

4. 評価

4.1 レポータとハッシュタグ検索の比較

4.1.1 実験方法と結果

本システムが推薦するレポータについて定量的に評価するため、2013 年 5 月 26 日に行われたソフトバンク対ヤクルトのプロ野球の試合を対象とした実験を行った。システムに与えるクエリとしてソフトバンクホークスのハッシュタグである「#sbhawks」を用いた。本実験における量の評価手法は、ツイートの投稿時刻を考慮しない式(5)の計算式を用いた。具体的な実験手順を以下に示す。

(1) 試合当日の午後 8 時 55 分に、システムに「#sbhawks」をクエリとして入力し、推薦レポータ、上位 5 名をフォローする。

(2) フォローを開始してから 30 分間で(1)の推薦レポータ 5 名が発信したツイート全 41 件を推薦レポータグループの評価対象ツイートとする。

(3) (2)と同じ時間範囲内において、ハッシュタグ「#sbhawks」が付加されたツイート全 157 件を(2)のツイート数と揃えるめに等間隔で間引いたツイート41件を比較評価のベースラインとなるツイート集合とする。

(4) 上記で得られた計 82 件のツイートに関して、被験者 5 名にアンケートを実施した。アンケートの内容は、「あなたがこの試合の状況をリアルタイムに知りたい場面を想定し、それぞれのツイートを読んだときにそれが試合状況の情報を十分伝えているかを 5 段階で評価してください」というものである。評価基準は、1 が『不必要な情報、出てこないほうが良い』、3 が『どちらでもない。出てきても出てこなくも構わない』、5 が『有益なツイート、試合状況に関する情報を十分に伝えているツイート』とした。また、それぞれのツイートが(2)のレポータによるものか(3)のハッシュタグ検索から得られたものかは被験者に知らせないようにした。

アンケートによる評価実験の結果を表 1 に示す。各項目の値は、それぞれの被験者が全ツイートに対して評価した値の平均値を示している。ここで示されるように、レポータが発信したツイートへの評価値はハッシュタグ検索で得られたツイートへの評価値を上回っている。評価値に対して T 検定を行ったところ、p=2.24*10^(-12)<0.05 となり有意差があると判断された。

表 1 得られたツイートに対する評価

	レポータ	ハッシュタグ
被験者 A	3.02	2.25
被験者 B	3.34	2.28
被験者 C	3.17	2.5
被験者 D	3.18	2.33
被験者 E	3.25	1.72
平均	3.19	2.22

次に、システムが行なう質的評価の手法が適切であるかについて検討する。図 8 に、システムがレポータにつけた質的評価値と、被験者がレポータにつけた質的評価値(候補者のツイートの評価の平均値)を比較した図を示す。

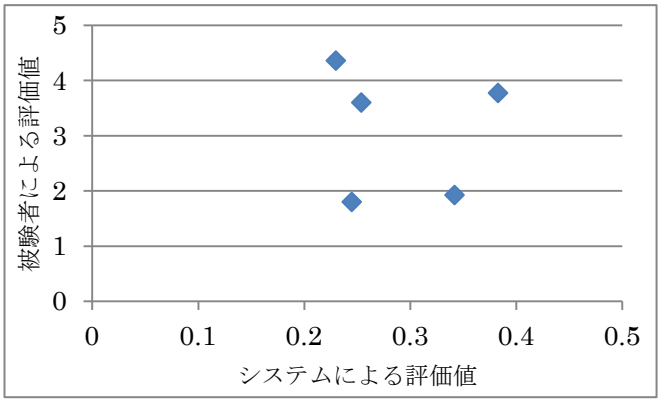

図 8 質的評価の比較

この図において、正の相関がとれれば理想であるが、実際の相関係数は-0.11 であり、システムはレポータのツイートの質を正しく評価できていないと考えられる。つまり、ツイートの質を評価する方法としてツイートに含まれる関連語の数を計るだけでは不十分であると言える。そこで、ツイートの文字数と、被験者がつけた評価値に相関があるか調べた。その結果を図 **9** に示す。このグラフでは、相関係数が 0.81 となっている。つまり、ツイートの文字数が多いものほど、被験者は質の高いツイートであると評価する可能性が高いことを示す。従って、今後は本システムにおける質の評価基準としてツイート文字数も考慮することを検討したい。

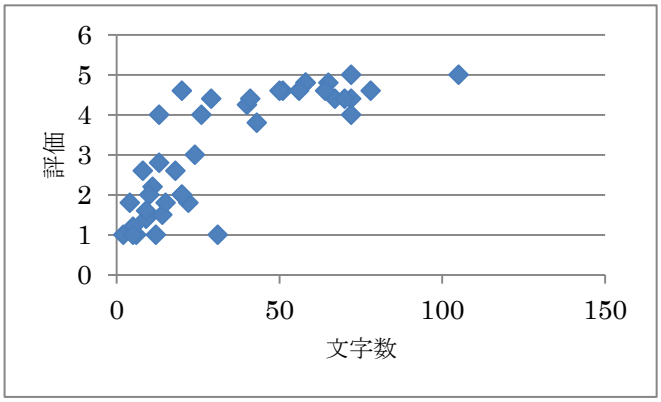

図 9 ツイート文字数と評価の比較

4.2 量の評価手法の比較

次に、量の評価手法についての比較実験を行った。対象としたイベントは2014年2月に行われたオリンピック競技「ハーフパイプ」、及び同月に起きた阪急京都の運行トラブルである。実験を行った時間帯はそれぞれのイベントが行われている(起こっている)最中とし、システムにあたえたクエリは「ハーフパイプ」、及び「阪急京都」である。実験の主な手順は 4.1 と同じである。本実験の目的は量の評価手法の改善であり、すなわちより多くのツイートをする確率の高いユーザを推定することである。従って、フォロー後に発言したツイート数のみを比較検討した。**図 10** にツイート投稿時刻を考慮しない量の評価値(式(5))とそのレポータが直後 15 分間で発言したツイート数を比較したグラフを示す。このグラフにおいて、相関係数は 0.36 であった。次に、**図 11** にツイート投稿時刻を考慮した量の評価値(式(7))とそのレポータが直後 15 分間で発言したツイート数を示す。このグラフの場合だと、相関係数は 0.57 となり、**図 10** の相関係数よりも大きく向上した結果となった。このことより、量の評価においてツイート投稿時刻を考慮した計算式(7)の方が、直後にツイートをする確率の高いユーザを推定できると言える。

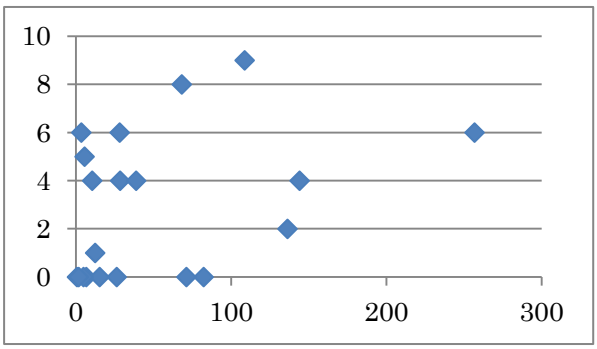

図 10 ツイート時刻を考慮しない量の評価値とレポータのツイート数

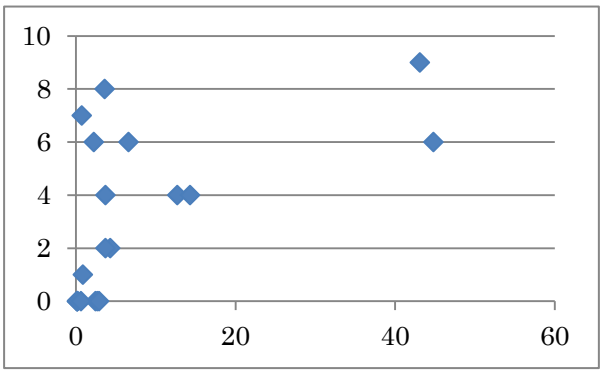

図 11 ツイート時刻を考慮する量の評価値とレポータのツイート数

4.3 考察

今回の実験で、レポータが発言するツイートとハッシュタグ検索で得られるツイートをそれぞれ被験者 5 名に評価してもらった結果、提案手法である前者の方が有益なツイートを多く含むという結果を得ることができた。一方で、システムがレポータを選定する基準の 1 つである質の評価については十分適切に行われていないことがわかった。これに対する改善策として、候補者のツイートの文字数やリツイート数、リプライ数を新たな尺度に加えることを検討している。さらに、ツイート自体の評価だけでなく、候補者自身のプロファイルであるフォロー数やフォロワー数も評価基準として利用できると考えられる。これらの要素を新たに加えることにより、さらなるレポータ選別精度向上を検討したい。

5. 結論

本研究では、Twitter を利用して実世界イベントに関する情報をリアルタイムに取得したいという要求を満たすため、イベントに関して有益な情報を頻繁に発信しているユーザをレポータとして発見することで、アドホックなフォローネットワークを自動的に構成するシステムを制作した。従来手法として、ハッシュタグを用いることで実世界イベント情報をリアルタイムに取得する方法が存在する。しかし、ハッシュタグの付加はツイート発信者の任意であるため、ハッシュタグが付加されたツイートの情報量・有益性が保証されていないといったことや、継続的な情報取得を行う場合、検索要求を何度も繰り返す必要があるといった欠点が挙げられる。評価実験では、このハッシュタグ検索で得られるツイートと推薦レポータ 5 名の発言するツイート、それぞれの有益性を被験者に評価してもらった結果、提案手法である後者のツイートの方が有益なツイートを多く含むという結果が得られた。また、T 検定によって評価の平均値に有意差があることを示すことができた。

文　献

[1] 若宮翔子, 李龍, 角谷和俊 : Twitter-based TV Audience Behavior Estimation for Better TV Ratings, DEIM Forum (http://db-event.jpn.org/deim2011/), 2011

[2] Bernard J. Jansen, Mimi Zhang, Kate Sobel, Abdur Chowdury : Twitter Power - Tweets as ElectronicWord of Mouth, Journal of the American Society for Information Science and Technology, Vol.60, Pages 2169-2188, 2009

[3] Takeshi Sakaki, Makoto Okazaki, Yutaka Matsuo : Earthquake Shakes Twitter Users - Real-time Event Detection by Social Sensors, WWW '10 Proc. of the 19th international conference on WWW, Pages 851-860, 2010

[4] 荒牧英治, 増川佐知子, 森田瑞樹 : Twitter Catches The Flu - Detecting Influenza Epidemics using Twitter, EMNLP '11 Proc. of the Conference on Empirical Methods in Natural Language Processing, Pages 1568-1576, 2011

[5] 杉谷 卓哉 : Twitter における投稿メッセージの時空間的局所性の解析によるローカルイベント検出手法, DICOMO2012(http://www.dicomo.org/), Pages 1704 - 1711, 2012

[6] Huberman, Bernardo A., Romero, Daniel M. and Wu, Fang, Social Networks that Matter: Twitter Under the Microscope (December 5, 2008). Available at SSRN: http://ssrn.com/abstract=1313405 or http://dx.doi.org/10.2139/ssrn.1313405

[7] Danah boyd, Scott Golder, Gilad Lotan. Tweet, Tweet, Retweet: Conversational Aspects of Retweeting on Twitter. 43rd Hawaii International Conference on System Sciences (HICSS), 2010.

[8] T. Joachims. Text categorization with support vector machines. In Proc. ECML'98, pages 137–142, 1998.