

短期持続型情報要求のための Twitter 上でのアドホックな フォローネットワーク構成手法

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
E-mail: 1ds09211s@google.com, 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、実世界での様々な事象に関してリアルタイムな情報が投稿されている。Twitter では、他のユーザをフォローすることによってその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、一般的にフォローによって構成される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。しかし、地域の祭りや音楽フェスティバルに参加したときなど、ユーザは参加している短期間だけ持続的に特定の事項に関する情報を取得したいという情報要求(短期持続型情報要求)を有する場合がある。本研究では、Twitter 上で適切なフォローネットワークを一時的に構成することによってユーザの短期持続型情報要求に応える手法を提案する。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

Automatic Construction of Ad-hoc Following Networks on Twitter for Getting Information about a Topic in Real Time

Shingo TAJIMA[†] and Taketoshi USHIAMA[‡]

[†]Department of Design, Kyushu University 4-9-1 Minami-ku, Fukuoka, 815-8540 Japan

[‡] Faculty of Design, Kyushu University 4-9-1 Shiobaru, Minami-ku, Fukuoka, 815-8540 Japan
E-mail: [†]1ds09211s@google.com, [‡]ushiama@design.kyushu-u.ac.jp

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 の中で利用者数が多く、投稿数が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。例えば、ユーザ A がユーザ B をフォローしたとき、ユーザ B の発信するツイートをユーザ A はリアルタイムに取得できる。ユーザは個別にフォローネットワークを構成できるため、自分の興味や関心に合わせてフォロー相手を選択してネットワークを構成することにより、自分にとって必要な情報を得ることができる。

Twitter では、リアルタイムな情報を簡潔なメッセージによって取得できる。また、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。こうしたことから、Twitter は、友人や知人との日常的なコミュニケーションの手段として利用する場合もあれば、気になるニュースや著名人の情報を得る手段等、様々な目的のために

として利用されている。

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行われない。つまり、Twitter においては、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 は次のような状況では有効的でないと考えられる。

- 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るが効率的でない。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られる。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。図1に通常のTwitterのフォローネットワークの概念図を示す。

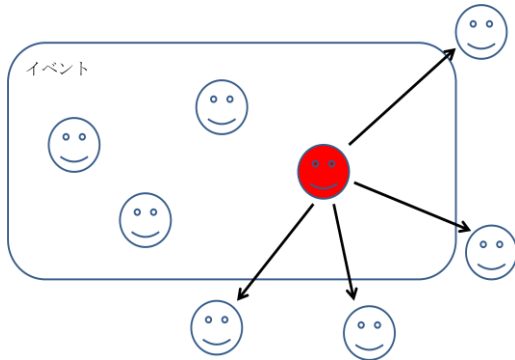


図1 通常のTwitterフォローネットワーク

本論文では上記の手法を実現するため、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である(図2)。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用な情報を提供する可能性があるTwitterユーザをレポートと呼ぶ。本手法では、適切なレポートに対して一時的なフォローネットワークを構成する。

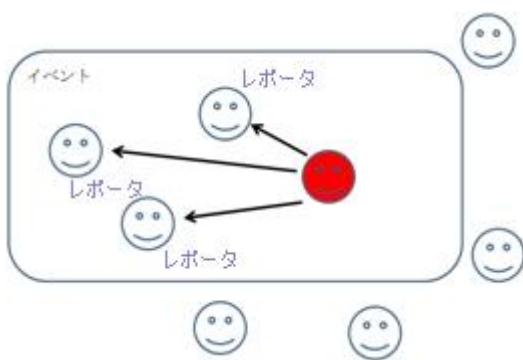


図2 本研究で提案するアドホックフォローネットワーク

図3に、提案システムの全体像を示す。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トを返すために、レポート候補者を収集する。
- (3) レポート候補者のツイートから、それぞれの候補者がレポートとしてふさわしいかを評価する。
- (4) 適切なレポートとして評価されたTwitterユーザを推薦する。
- (5) ユーザは、推薦されたレポート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得ることができる。

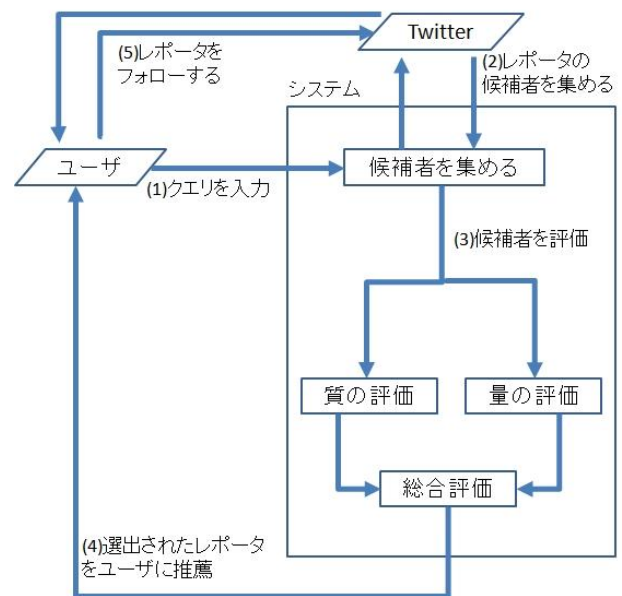


図3 システムの全体像

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。

一方、Twitter自体の特徴ではなく、Twitterの解析によってトレンドや世論、その他の社会的現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。彼らは、Twitterによる視聴率調査が従来の視聴率調査に対して優れている点について、

- コストがほとんどかからない(従来の視聴率調査方法ではテレビに専用の機器を接続するためにコストがかかる)。

- 視聴率だけでなく、そのテレビ番組に対する意見も同時に調査することができる。

と述べている。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(例「いまの地震すごかった!」)を意味し、negative はそうでないツイート(例「昨日の地震は本当に恐ろしかった」)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。

同様に、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 と negative 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。この場合における positive とは、実際にインフルエンザにかかっている人のツイート(例「今日は熱が 38 度ある…。インフルエンザは本当にきつい。」)を表し、negative ではそうでないツイート(例「インフルエンザが大流行しているようなので、気をつけましょう!」)を表す。彼らの制作したシステムによるインフルエンザ情報はインフルくん¹というウェブサイトで確認することができる。

また、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った。そして、得られたクラスタ内において、何らかの特殊な単語が頻出しているとき、クラスタ内でその単語に関するイベントが発生していると判断する手法を提案している。

3. 提案手法

本手法は、ユーザが入力したクエリに対してシステムがそれに関するツイートを発しているユーザを集めてレポート候補者とし、その中からレポートを選別して推薦する。ここで、レポート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

- 発してい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(レポータ)で選出を行う理由として以下が挙げられる。

- 実世界イベントに関するツイート 1 件 1 件に対して有益であるか否かの判断を行う場合、膨大な処理時間コストがかかるため現実的でない。一方で、有益なイベント関連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することができるため、少ない処理時間で済み、システムへの負担も小さい。
- イベント関連情報を継続的に発信しているユーザは、1、2 回しか発信していないユーザに比べ、そのイベント内容に関する詳細な情報を発信している場合が多い。例として、2013 年 1 月 24 日に行われた「ユニット祭り」と呼ばれる AKB48 のネット動画配信イベント中に「ユニット祭り」という単語でツイート検索を行うことで得られたユーザについて述べる。図 4 は、イベント関連情報を頻繁に発信していないユーザの例である。このユーザはツイートの中で、赤枠で囲まれたツイートが検索結果に現れた。しかし、そのツイート以降はこのユーザは対象とするイベント関連のツイートを発信していない。こうしたユーザは主に自分がそのイベントに参加していることを表すのみで、イベント内の詳細情報に関して言及していない場合が多い。次に、イベント関連情報を継続的、かつ頻繁に発信しているユーザの例を図 5 に示す。このユーザの例では、イベント内での様子を簡単な説明分と写真によって発信している。このように、イベント関連情報を頻繁に発信しているユーザはそのイベント内における詳細情報を述べている場合が多い。このことから、イベント関連情報を頻繁に発信しているユーザを選出することで、有益な情報を取得できる可能性が高いと考えられる。



図 4 イベントに関するツイートを 1 回だけ発信したユーザの例

¹ インフルくん: <http://mednlp.jp/influ>



図 5 イベントに関するツイートを継続的に発信しているユーザーの例

以下に、本手法におけるレポータ候補者の収集手法を述べ、候補者の中から上記の条件を満たすレポータを適切に選別する手法について述べる。

3.1 候補者の収集

本システムはユーザからキーワードを受け取ると、キーワード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、例えばキーワード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である。一方、キーワードが街で行われているイベントの名称であった場合、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キーワードをクエリとし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」というキーワードをシステムに与え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。このようにキーワードを直接含むツイートを発したユーザを求めることで、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点リードしてる。』というツイートは「オマーン」というキーワード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、ユーザが与えたキーワード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に、ユーザが与えたキーワードは直接含まないが、ユーザの与えたキーワードに深く関連するツイートを発見するため、関連ワードを利用する。例えば、「オマーン」というキーワード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ワードであると推定することができれば、関連ワード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 関連ワードの発見

関連ワードを発見するために、ユーザが与えたキーワードを利用したクエリによる検索によって得られたツイート内における共起ワードを関連ワードの候補とする。例えば、「オマーン」という単語を含むツイートと共起語をその出現頻度でランキング化すると、図 6 に示す結果となった。

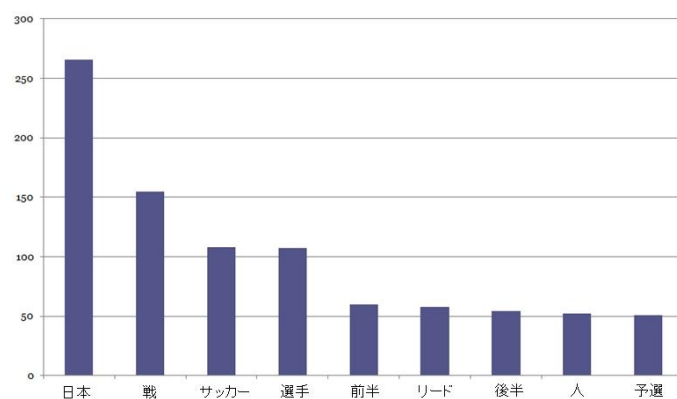


図 6
共起語ランキング(オマーンを含むツイート 1264 件中)

しかし、このランキング結果をそのまま関連ワードとして扱うのは不適切である。例えば、図 6 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。したがって、こうした語はキーワードである「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、共起語とキーワードの距離を利用する。距離としてはコサイン距離を用いる。コサイン距離は、以下の式(1)で表される。

$$\cos(X, Y) = \frac{|X \cap Y|}{\sqrt{|X| |Y|}} \quad \dots (1)$$

式(1)において、X, Y はそれぞれ図 7 に示すように、キーワード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 7 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。

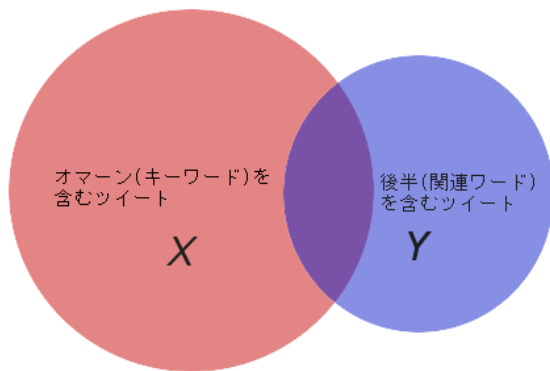


図 7 コサイン距離の概念図

しかし、Twitter 上で発信されているツイートから、それぞれの語を含んでいるツイート全てを抽出してその数を求めることは現実的に不可能である。そこで、ここでは以下に示す手順で擬似的にその集合の要素数の比を求める。

1. キーワードを含むツイート集合のうち、同時に共起ワードも含むツイートの数の割合を求める。

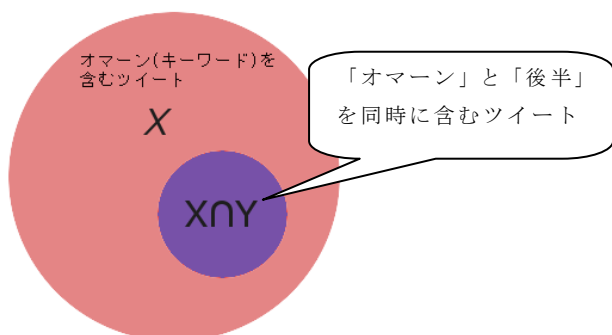


図 8 集合の比の求め方(1)

2. 関連ワードを含むツイート集合のうち、同時にキーワードも含まれるツイート数の割合を求める。

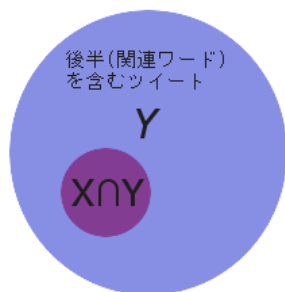


図 9 集合の比の求め方(2)

例えば、上記の手順 1,2 による結果がそれぞれ 1/10 と 1/8 であったとき、その式は以下の式を得ることができる。

$$X : X \cap Y = 10 : 1$$

$$Y : X \cap Y = 8 : 1$$

$$\therefore X : Y = 10 : 8$$

上式から、今回の例におけるコサイン距離の計算式は以下のようになる。

$$\cos(X, Y) = \frac{1}{\sqrt{10 \cdot 8}} = \frac{1}{4\sqrt{5}}$$

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 11 件を図 10 に示す。縦軸はユーザがクエリとして与えたキーワード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、コサイン距離の値を、ユーザがクエリとして与えたキーワードと関連語に関する関連度とする。

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。

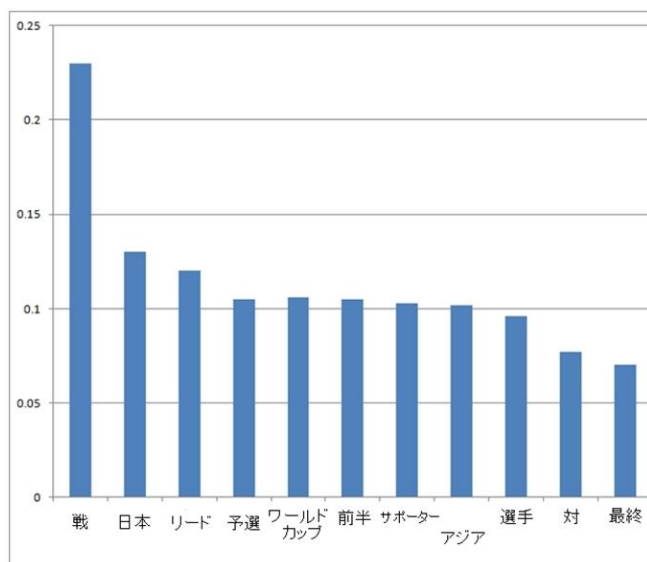


図 10 コサイン距離の計算結果

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) 発しているツイートの内容が、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言える。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

それぞれのレポータ候補者が、一定時間 L 以内に投稿したツイート件数を求め、量の評価の尺度とする。本研究では、レポータからリアルタイムに情報を得ることが必要であるため、 L を 5 時間とした。候補者が過去 5 時間以内に投稿したツ

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(2)にレポータ候補者 u の量の評価値 $av(\text{amount value})$ の定義を示す。変数 t はツイートを表し、変数 $T(u)$ は候補者 u が過去 L 以内に発信したツイートの集合を表す。

$$av(u) = \sum_{t \in T(u)} \phi(t) \dots(2)$$

$$\phi(t) = \begin{cases} 1 & \text{(ツイート内に含まれる関連語数が1以上の場合)} \\ 0 & \text{(ツイート内に関連語を含まない場合)} \end{cases} \dots(3)$$

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、基本的には文字数の少ないツイートよりも文字数の多いツイートの方が質の高いツイートであると判断する。しかし、単純に文字数を比較するだけでは、そのツイートがキーワードに対して情報量が高いとは判断できない。そこで、ツイートが含む関連語の有無を考慮する。2つのツイートを比較したとき、関連語をより多く含んでいるツイートの方が、キーワードに対する情報量が多いことが期待できる。従って、ツイートに含まれている関連語とキーワード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 11 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。そこで、このツイートの質の評価値を 0.35 とする。一方、例2のツイートでは、含まれる関連語は「試合」だけであり、その関連度は 0.04 であるため、このツイートの質の評価値は 0.04 となり、例1よりも質の低いツイートであると判断される。なお、キーワード自身の関連度は1で計算する。

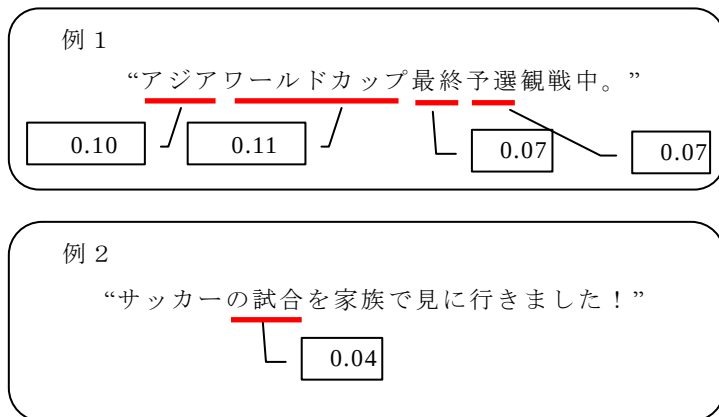


図 11 質の評価

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。キーワードに関係のな

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(3)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、 $qv(\text{quality value})$ を定義した。式(2)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。

$$qv(u) = \frac{\sum_{t \in T} \sum_{w \in W} \phi(t, w) \gamma(w)}{|T|} \dots(4)$$

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

$$\gamma(w) : \text{関連語 } w \text{ の関連度} \dots(5)$$

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、 $tv(\text{total value})$ を定義する。(1)および(2)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(4)に総合的な評価値の形式的な定義を示す。

$$tv(u) = av(u) \cdot qv(u) \dots(4)$$

3.4 ボットアカウントの除外

Twitter ではプログラムによる自動発言を行うアカウントが存在する。このようなアカウントはボットと呼ばれる。ボットの一例を図 12 に示す。本システムでのレポータ選出手法は、ユーザが発言したツイートに含まれる関連語を用いているため、こうしたボットアカウントがレポータとして選出される可能性がある。例えば、「ATARU」というクエリをシステムに入力した場合、図 12 のボットアカウントがレポータとして選出される可能性が高い。このボットアカウントが自動発言している文章全てに、「ATARU」という単語が含まれているからである。しかし、ボットアカウントはプログラムによって予め入力された文章を発言するだけであるため、実世界イベントに関するリアルタイムな情報を発信する可能性は低い。従って、こうしたボットアカウントはレポータから除外される必要がある。



図 12 ボットアカウントの例

ボットアカウントの一般的な特徴から、以下の3つの条件を設定し、これら全てを満たしている候補者はボットアカウントであると判断し、レポートから除外するようにした。

- 過去5時間以内に10件以上のツイートを発信している
- 全てのツイートにURLが含まれている
- 過去5時間以内のツイートの中で、他のユーザに対するリプライ(返信)ツイートを1件も含まない

4. 評価

4.1 実験方法と結果

本システムが推薦するレポートを評価するため、1月11日に大阪で行われた「十日戎」(通称:「えべっさん」)という祭りを対象とした実験を行った。

レポートの発信するツイートについて相対的な評価を行うため、「#えべっさん」のハッシュタグが付加されたツイートと比較した。また、関連語を用いた場合とそうでない場合の2パターンのレポートを用意し、関連ワードがレポート推薦に与える影響についても評価した。具体的な実験手順は以下の通りである。

- (1) 祭り当日の午後8時に、システムにキーワード「えべっさん」を入力し、関連語を用いた場合による推薦レポート、上位5名をフォローする。また関連語を用いなかった場合に推薦されたレポート上位5人に関しても同様にフォローする。
- (2) フォローを開始してから1時間後の時点で、(1)でフォローした関連ワードを用いた場合の推薦レポートグループによるツイートの最新20件、および関連ワードを用いなかった場合の推薦レポートグループによるツイートの最新20件をそれぞれの評価対象ツイートとする。
- (3) (2)と同じ時刻にて、ハッシュタグ「#えべっさん」で検索を行い、最新20件のツイートをハッシュタグを用いたツ

イートの評価対象とする。

- (4) 上記で得られた3パターンの20件のツイートに関して、被験者にアンケートを実施した。アンケートの内容は、「あなたが『えべっさん』に参加しているとして、Twitter上で『えべっさん』の情報をリアルタイムに得ようとしたとき、以下のツイートが出てきたら嬉しいかを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、イベント内において役立つであろうツイート』とした。また、それぞれのツイートがレポートによるものかハッシュタグによるものかを被験者には事前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20件の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、レポート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、 $p=0.02<0.05$ となり有意差があると判断された。

一方、関連語を考慮した場合によるレポート、そうでない場合によるレポートのツイートに対する評価値はほぼ同じ値を示しており、関連語がレポート推薦に与える影響が小さいことがわかった。この原因として、ツイートに含まれる単語とその関連度を利用して質を評価する際、キーワード自身の関連度を1とカウントしているのに対し、関連語の関連度は図10で示されるように、最大でも約0.23と比重が大変小さいことが一因として考えられる。したがって、ツイートの質を評価する際、関連ワードの関連度とキーワードの関連度を別の尺度で計算するといった考慮が必要である。

表 1 実験結果

	レポート (関連語あり)	レポート (関連語なし)	ハッシュタグ
被験者 A	2.85	2.79	1.95
被験者 B	2.65	2.58	2.55
被験者 C	2.75	2.89	1.95
平均	2.75	2.75	2.15

4.2 考察

今回は評価実験の対象として、十日戎という参加者や規模の大きいイベントを選択することで好ましい実験結果を得ることができた。しかし、全てのイベントにおいて必ずしもこのような結果が得られるとは限らない。特に、小規模のイベントにおいては、関連するツイートを発信する人が少なく、有効なレポートを発見できる可能性は低い。今後の展望として、そうした小さなイベントに対しても評価実験を行うことで新たなアプローチを検討することを考えている。

また、現段階においてレポート候補者の質を評価する尺度としてツイート内に含まれる関連語とその関連度を利用しているが、この手法では不十分な点がいいくつか考えられる。例えば、推薦されるレポートのツイートが同じような単語を含んでばかりで内容が類似してしまうことがある。この問題に対する解決策として、ツイートのリツイートされた数、リプライ(返信)された数を質の評価基準の1つとすることや、候補者のフォロー・フォロワー関係を利用することでその候補者の信頼性を計ることを検討中である。

さらに、関連語の発見手法としてツイート内における共起語をもとに発見する手順を踏んでいるが、この場合、キーワードの略語や別名などは発見されにくい。なぜなら、同じツイート内にキーワードとその略語・別名を同時に含ませることは稀であるからである。しかし、このような略語や別名は重要な関連語の1つである。今後は、そうした共起語からは発見されづらい関連語も発見できる手法を検討していきたい。

5. 結論

本論文では、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、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。そこで、イベントに関して有益な情報を含むツイートを継続的かつ頻繁に発信しているユーザを発見してそのイベントの「レポート」とみなし、ユーザがレポートに対してフォローを行うことでイベントに関する有益な情報のみを継続的に取得できるアプローチを提案した。本研究でキーとなる点は、収集したレポート候補者の中から真のレポートを選別する基準である。今回は、各候補者が発信するイベント関連ツイートの「量」と「質」という2つの評価基準を設定し、評価を行った。2つの基準をもとにレポートが推薦され、ユーザは推薦されたレポートをフォローすることでイベント情報を得ることができる。

本システムの評価実験を行うため、1月11日に大阪で行われた「えべっさん」というお祭りを対象とした。実験では、システムが推薦したレポートが発信したツイート20件と、「#えべっさん」というハッシュタグによって検索を行ったツイート20件に関して被験者3人に評価してもらった。実験の結果、本システムのレポートが発信したツイートのほうが有益であるという結果を得ることができ、T検定によって評価値の平均に5%の有意差があると判断された。

本システムを制作するにあたって重要な課題は、いかに

TwitterAPI²へのリクエスト数を節約するかであった。本システムでは、ツイートの検索、および候補者が過去5時間以内に発信したツイートの調査においてTwitterAPIを利用しているが、TwitterAPIを利用できる回数は無制限でない。2012年1月24日時点で、TwitterAPIへのリクエスト制限は180回/15分である。また、過剰にTwitterAPIへのリクエストを行うことは処理時間の増長になりうる。そのため、本システムにおいては全候補者の優先度を設定し上位50件までの候補者に対してのみ評価を行うことや、関連ワードの発見も25単語までに限定することでAPIへのリクエスト数削減を行った。しかし、この制限は経験則に基づいて設定したものであり必ずしも適切でない。今後は、この制限範囲を変化させることによるレポート選出、および処理時間に与える影響を実験によって評価し、適切な制限範囲を検討する予定である。

参 考 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.jp/deim2011/>), 2011
- [2] Bernard J. Jansen, Mimi Zhang, Kate Sobel, Abdur Chowdury: Twitter Power - Tweets as Electronic Word 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, DICO2012(<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. System Sciences (HICSS), 2010 43rd Hawaii International Conference on.
- [8] T. Joachims. Text categorization with support vector machines. In Proc. ECML'98, pages 137-142, 1998.

² Twitter Developers: <https://dev.twitter.com/>