Twitterにおけるイベントモニタリングのためのノイズ除去

伊川　洋平†　　村上　明子†

† 日本アイ・ビー・エム株式会社 東京基礎研究所　〒 135–8511 東京都江東区 5 丁目 6–52
E-mail: †{yikawa,akikom}@jp.ibm.com

あらまし　イベントに関する人々の声を収集するための情報源として Twitter が着目されている．Twitter からイベント関連のツイートを収集するには，事前に定義したイベント関連語を含むツイートを抽出するアプローチが一般的だが，収集したツイートには対象イベントと関係のないノイズツイートが多く含まれるという問題がある．このようなノイズツイートを除去するために，対象イベントと関係がないことを示唆するノイズキーワードを，ある時期に集中してイベント関連語について言及している “ホットユーザー” を利用して自動的に特定する手法を提案する．提案手法は，人手で作成した教師データや Wikipedia などの知識ベースを必要とせずにイベントとの関連性を判定できる点で新しい．また，イベント関連語で収集したツイートのみを用い，追加でデータを取得する必要がないことから，現実問題に対して適用可能な実用性の高い手法である．評価実験では,「東京モーターショー 2013」を対象イベントとして収集したツイートに対して，高精度でノイズツイートが除去できることを確認した．

キーワード　Twitter, イベントモニタリング, ノイズ除去

1. はじめに

特定のイベントに関する人々の声を収集するイベントモニタリングは，イベントに対する評判や要求を把握する上で，イベント主催者にとって重要である．従来のイベントモニタリングの手段として来場者アンケートが広く行われているが，実施コストが高く，大規模なイベントに対しては十分なサンプル数を確保するのが難しいという問題がある．そこで近年，イベントモニタリングのための情報源として Twitter が着目されている．Twitter は通常のブログと比較してリアルタイム性が高いという特徴を持つため，実際にその場所を訪れて体験をしている最中にツイートを発信する傾向が強い．そのため，人々がその場で感じたことがよりダイレクトにツイートに反映されることが期待され，人々の声を収集するための情報源として重要視されている．

本稿では,「東京モーターショー 2013」(注1) を題材として，Twitter を用いたイベントモニタリングについて検討を行う．モーターショーは世界の主要な都市で定期的に開催されている大規模なイベントである．東京モーターショー 2013 では 10 日間に渡って 178 の自動車関連会社が製品や技術を展示し，延べ 902,800 人が来場した．本イベントにおいては，調査員の面接による来場者アンケートが実施されたが，サンプル数は 860 人であり，来場者全体の 0.095% であった(注2)．面接による来場者アンケートは，回答内容の信頼性が高く有用な情報ではあるものの，サンプル数が極端に少ないため，有用な意見を取りこぼしている可能性が高い．そこで，より多くの人々の声を収集するために，Twitter が有用な情報源として期待されている．

Twitter からイベントに関連したツイートを収集するには，あらかじめイベント関連語を定義しておき，それらを含むツイートを抽出する．イベント関連語の候補としては，イベント名 (例: “モーターショー”) やイベント関連のハッシュタグ (例: “#motorshow”) が挙げられる．しかし，イベント関連ツイートが必ずイベント名やハッシュタグを含むとは限らないため，少数の語をイベント関連語としてツイートを収集すると，多くのイベント関連ツイートを収集できず取りこぼしがある可能性がある．以下，ツイートの例を示す．

#01 User A: 東京モーターショーに来ました
#02 User A: @friend おはよう！
#03 User A: ホンダブース混んでる...
#04 User A: XYZ123 のコンセプトカーがよかった

これらのツイートは全て同一のユーザー “User A” により発信されたもので，このうちモーターショーと関連のあるツイートは#01，#03，#04 である．ここで，イベント名とハッシュタグのみをイベント関連語としてツイートを収集すると，#03 と #04 を取りこぼしてしまう．よって，イベント関連ツイートを広く収集するためには，イベント関連語を充実させる必要がある．この例では，出展企業名 “ホンダ” や展示車名 “XYZ123” をイベント関連語に追加することで，#03 と#04 を収集することができる．

#03 と#04 を収集するための別のアプローチとして，イベント名やハッシュタグに言及したユーザーを特定してから，そのユーザーが発信したツイートを過去に遡って取得する方法が考えられる．しかし，Twitter のような膨大なデータ量のストリームデータに対して，過去に遡ってデータを取得するのはコストがかかる処理である．また，仮に収集したとしても，#02 のようなイベントと関連のないツイートが混入する可能性があり，これを除去するための手法を併せて検討する必要がある．よって，イベント関連ツイートを広く収集するには，イベント

(注1)：http://2013.tokyo-motorshow.com/
(注2)：http://2013.tokyo-motorshow.com/history/43_survey.html

関連語を充実させ，現在流れているツイートに対してイベント関連語を含んでいるものを抽出するアプローチが効率的である．

イベント関連語を充実させてツイートを収集するアプローチでは，イベントと関連のないツイートが多数収集されるという問題がある．出展企業名“ホンダ”をイベント関連語とした場合のノイズツイートの例を以下に示す．

#05 User X: ホンダ中古車が限定特価 http://...
#06 User Y: ホンダゴール !!!

これらのツイートは明らかにモーターショーとは関係のないツイートである．#05 は中古車販売の広告の例で，この場合“ホンダ”は自動車会社を意味しているが，このツイートはモーターショーとは関係がない．#06 はイベント関連語の語義の曖昧性が原因となって収集されるノイズツイートの例である．ここでは，“ホンダ”はサッカーの本田圭佑選手のことであり，得点を決めた瞬間に発信されたツイートである．このツイートもモーターショーとは関係がない．

また，イベントに関係あるかどうかがそのツイートのみからは判定できないツイートが存在する．以下に例を挙げる．

#07 User Z: ホンダのバイクがかっこよかった

この例では，“ホンダ”は自動車会社であると判断できるが，このツイートのみからは，モーターショーに関係があるかどうかの判断ができない．

以上の検討から，イベント関連語を用いて収集したツイートは，イベントとの関連性の観点で次の 3 種類に分類することができる．

- Relevant: ツイートは明らかにイベントと関連がある
- Irrelevant: ツイートは明らかにイベントと関連がない
- Unclear: ツイートがイベントと関連があるかどうかを判定できない

本研究は，Irrelevant のツイート，すなわちノイズツイートを，高精度に，低コストで特定することを目的とする．ここで“低コスト”とは，イベント関連語で取得したツイート以外のデータを新たに取得しないことを意味している．取得した膨大な数のツイート全てに対してコストのかかる処理を適用するよりも，最初に明らかにイベントと関係のないツイートを低コストで除去しておき，残ったツイートに対してより複雑な手法を適用する方が効率的である．また，取得したツイートに対して最初に適用するノイズツイート除去は，高精度であること，すなわち，Irrelevant と判定されたツイートの中にできるだけ Relevant や Unclear のツイートが含まれていないことが重要となる．

提案手法は，イベントに対して熱心にツイートしている“ホットユーザー”を利用してノイズツイートを特定する．ホットユーザーは，イベント関連語で収集したツイートをツイート発信者で集約し，何種類のイベント関連語に言及しているかをカウントすることで容易に発見することができる．例えば，#01，#03，#04 のツイートを発信した User A は，“東京モーターショー”，“ホンダ”，“XYZ123”の 3 種類のイベント関連語について言及しており，モーターショーにおけるホットユーザーである．このようなホットユーザーが発信したツイートは，イベントについて言及している可能性が高いことが期待される．この特徴を利用して，イベント関連語と共起しているキーワードに対して，イベントとの関連度合いを算出する．例えば，User A が発信した#03 の“ホンダ”と共起している“ブース”は，他のホットユーザーからも多数発信されており，モーターショーとの関連度合いが高いキーワードとなる．一方で，#06 のようなツイートは，サッカー中継を見ているユーザーによって多数発信されているが，モーターショーのホットユーザーからこのようなツイートが発信されることは稀である．従って，#06 の“ホンダ”と共起している“ゴール”は，モーターショーとの関連度合いが低いキーワード，すなわちノイズキーワードとして特定される．そして，ノイズキーワードを含むツイートがノイズツイートとして特定される．

本研究は，我々の知る限り，イベントモニタリングを目的としたノイズツイート除去に関する最初の研究である．提案手法は，イベント関連語で取得したツイートのみを用いてイベントと関連のないツイートを自動的に特定する．人手で作成した教師データを必要とせずにイベントとの関連性を判定できる点で新しく，また，Wikipedia など外部の知識ベースを必要としないため，結果が知識ベースの充実度合いに影響されることがない．評価実験では，「東京モーターショー 2013」を対象イベントとして収集したツイートに対して，高精度でノイズツイートが除去できることを確認した．

2. 提 案 手 法

本章では，対象イベントの関連語を用いて収集したツイートから，対象イベントについて言及していないノイズツイートを特定し除去するための手法について述べる．最初に，あらかじめ定義されたイベント関連語を含むツイートを収集する (2.1 節)．このようにして収集したツイートには，実際にはイベントについて言及していないノイズツイートが含まれる．そこで，イベントに対して熱心にツイートしている“ホットユーザー”を利用することで，イベントに関連したツイートを特定し，これを用いてノイズキーワードを特定する (2.2 節)．そして，このノイズキーワードを用いてツイートがノイズかどうかを判定する (2.3 節)．

2.1 ツイートの収集

イベントについて言及している可能性のあるツイートを広く収集するために，イベント関連語の集合 E を定義する．ここでイベント関連語とは，イベント関連ツイート中に一定の頻度で出現することが期待される語のことである．モーターショーの例では，イベント名やイベントのハッシュタグ，出展企業名，展示車名などがイベント関連語にあたる．提案手法では，イベント関連語を定義するための方法については特に定めない．対象イベントについて紹介した Web サイトやガイドブックを参考にして人手でイベント関連語を定義してもよいし，与えられたトピックに対する関連語を自動的に抽出するような手法を利用することも考えられる．何らかの方法によって定義されたイ

ベント関連語の集合を E とする．

続いて，イベント関連語の集合 E を用いてツイートを収集する．イベント関連語 $e \in E$ を 1 つ以上含むツイートを収集し，収集したツイート集合を TW_E とする．本稿では，ツイート t を，発信時刻 s_t，ユーザー u_t，メッセージテキスト m_t の三つ組とし，$t = (s_t, u_t, m_t)$ と記述する．収集したツイート集合 TW_E におけるユーザー集合を U，メッセージテキスト m_t に出現するイベント関連語の集合を $E(m_t)$ とする．ここで，$E(m_t) \subseteqq E$ であり，収集したツイートは少なくとも 1 つ以上のイベント関連語を含むことから，ツイート $t \in TW_E$ に対して $E(m_t) \neq \phi$ が成立する．また，メッセージテキスト m_t に出現するキーワードの集合を $W(m_t)$ とする．ここで，キーワードは m_t に出現する名詞（イベント関連語 $e \in E$ を含む）を想定しているが，提案手法においてキーワードの定義に制限はなく，名詞以外の語をキーワードとして利用することもできる．

2.2 ホットユーザーを用いたノイズキーワードの特定

収集したツイートから対象イベントについて言及していないノイズツイートを特定するために，ノイズツイートであることを示唆するキーワードを特定する．本稿では，このようなキーワードをノイズキーワードと呼ぶ．ノイズキーワードは，イベント関連語 $e \in E$ それぞれに対して個別に規定される．

ノイズキーワードを特定するために，イベント関連語 $e \in E$ とキーワード w の組に対して，対象イベントとの関連度合いを表すイベント関連スコア $Rscore_w(e, w)$ を算出する．このスコアが低いキーワード w が，イベント関連語 e に対するノイズキーワードとなる．モーターショーの例では，$Rscore_w$(“ホンダ”, “バイク”) は高い値になり，$Rscore_w$(“ホンダ”, “ゴール”) は低い値になることが期待される．

提案手法では，$Rscore_w(e, w)$ を求めるために，対象イベントにおけるホットユーザーに着目する．ホットユーザーとは，ある期間において集中的にイベントに関連したツイートを発信しているユーザーのことである．ある期間においてホットユーザーであったユーザーが常にホットユーザーであるのではなく，集中的にイベントに関連したツイートを発信した期間においてのみホットユーザーとなる．提案手法は，ホットユーザーを利用してノイズキーワードを特定するために，次の仮定を用いる．

[仮定 1] ホットユーザーが発信したツイートは，ホットユーザーでないユーザーと比べて，イベントと関連している可能性が高い．

仮定 1 により，ホットユーザーによって発言されたツイートを手がかりとして $Rscore_w(e, w)$ を求めることができる．ここで，ホットユーザーでないユーザーが発信したツイートの中にも，イベントと関連したツイートが多数存在し得ることに注意する必要がある．従って，ホットユーザーでないユーザーが発信したツイートのみからノイズキーワードを特定することはできない．

以下，2.2.1 節において収集したツイート集合 TW_E におけるホットユーザーを特定する方法について述べる．続いて，2.2.2 節においてキーワードに対するイベント関連スコア $Rscore_w(e, w)$ を算出する方法について述べる．

2.2.1 ホットユーザーの特定

ホットユーザーを特定するために，各ユーザー，各期間毎にツイートを集約してツイート集合を構成する．最初に，期間の表記法について定義する．ツイートを収集した全期間 P を一定の時間幅で分割し，$P = \{P_1, P_2, \cdots, P_n\}$ とする．分割された各期間 P_i は，$P_i = [s_{i-1}, s_i)$，$s_i = s_{i-1} + \delta$ を満たすものとする．これは，全期間 P を一定の時間幅 δ で n 個の期間 $P_1, P_2, \cdots, P_n$ に等分することを意味する．続いて，ユーザー $u \in U$ が期間 $P_i \in P$ に発信したツイートの集合を $TW_E(P_i, u)$ とする．また，ツイート集合 $TW_E(P_i, u)$ において，メッセージテキスト中に出現したイベント関連語の集合を $E(P_i, u)$ とする．

期間 P_i におけるホットユーザーとは，期間 P_i において発信したツイート集合のメッセージテキスト中に，イベント関連語が θ 種類以上含まれているユーザーのことである．ホットユーザーの定義を次のように与える．

[定義 1] $|E(P_i, u)| \geqq \theta$ を満たすユーザー $u \in U$ を，期間 P_i におけるホットユーザーとする．ただし，θ は閾値パラメータである．

2.2.2 キーワードに対するイベント関連スコアの算出

ホットユーザーが発信したツイートはイベントに関連があるという仮定 1 に基づき，イベント関連語 e とキーワード w の組に対して，イベント関連スコア $Rscore_w(e, w)$ を算出する．$Rscore_w(e, w)$ は，ホットユーザーが発信したツイートにおいて e と w が共起する確率と，ホットユーザーでないユーザーが発信したツイートにおいて e と w が共起する確率の乖離度合いとして定義する．そして，この値が低ければ低いほど，ノイズキーワードである可能性が高くなる．

例えば，モーターショーのホットユーザーが発信したツイートにおける “ホンダ” と “バイク” の共起確率は，ホットユーザーでないユーザーの共起確率と比較して，極度に低くはならないことが予想される．なぜならば，モーターショーのホットユーザーは “ホンダ” と “バイク” を含むツイートを多数発信していることが期待されるためである．一方で，モーターショーのホットユーザーが発信したツイートにおける “ホンダ” と “ゴール” の共起確率は，ホットユーザーでないユーザーの共起確率と比較して，極端に低くなることが予想される．なぜならば，モーターショーのホットユーザーの大半はサッカーについてのツイートを発信していないと期待されるためである．

提案手法では，PMI（自己相互情報量）を用いてホットユーザーとそれ以外のユーザーにおけるイベント関連語 e とキーワード w の共起確率の乖離度合いを評価する．PMI は，2 つの事象 x と y が同時に起こる確率が，偶然の場合の確率と比較してどの程度起こりやすいかを測る指標であり，2 つの事象の関連度合いを測るのに用いられる．2 つの事象の関連性が高ければ正の値，関連性が低ければ負の値をとり，2 つの事象が独立の場合に 0 となる．事象 x，y についての PMI は，以下のように定義される．

$$PMI(x, y) = \log \frac{Pr(x, y)}{Pr(x)Pr(y)} \tag{1}$$

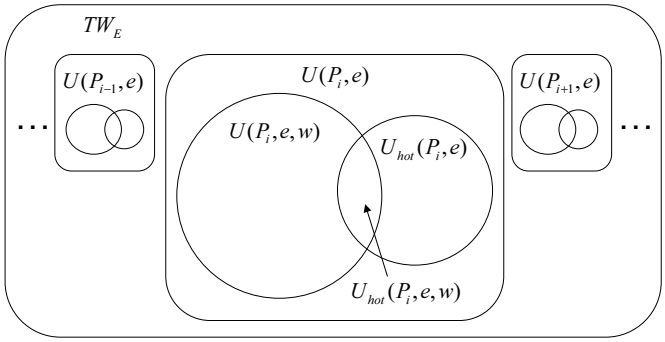

図 1　定義したユーザー集合の図示
Fig. 1　Illustration of defined user sets.

ここで，事象 x を “イベント関連語 e について言及したユーザーがキーワード w をツイートする事象”，事象 y を “イベント関連語 e について言及したユーザーがホットユーザーである事象” とすることで，PMI を用いてキーワード w と発信者がホットユーザーかどうかの関連性を測ることができる．PMI は 0 を境界として 2 つの事象に関連があるかどうかを弁別できることから，提案手法ではこれをイベント関連スコアとして採用する．

$$Rscore_w(e,w) = PMI(x,y) \tag{2}$$

以下，イベント関連スコア $Rscore_w(e,w)$ の算出方法を述べる．期間 P_i におけるホットユーザーの集合を $U_{hot}(P_i)$ とすると，$U_{hot}(P_i)$ は以下のように定義される．

$$U_{hot}(P_i) = \{u \in U \mid |E(P_i,u)| \geqq \theta\} \tag{3}$$

続いて，$PMI(x,y)$ を計算するために，4 つのユーザー集合を定義する．図 1 はこれらのユーザー集合をベン図で表したものである．期間 P_i において，イベント関連語 e を含むツイートを発信したユーザー集合を $U(P_i,e)$，そのうちのホットユーザーの集合を $U_{hot}(P_i,e)$ とする．これらのユーザー集合は以下のように記述される．

$$\begin{aligned} U(P_i,e) = \{u \in U \mid \exists t = (s_t,u,m_t),\\ s_t \in P_i \wedge e \in E(m_t)\} \end{aligned} \tag{4}$$

$$\begin{aligned} U_{hot}(P_i,e) = \{u \in U_{hot}(P_i) \mid \exists t = (s_t,u,m_t),\\ s_t \in P_i \wedge e \in E(m_t)\} \end{aligned} \tag{5}$$

期間 P_i において，イベント関連語 e とキーワード w を含むツイートを発信したユーザーの集合を $U(P_i,e,w)$，そのうちのホットユーザーの集合を $U_{hot}(P_i,e,w)$ とする．これらのユーザー集合は以下のように記述される．

$$\begin{aligned} U(P_i,e,w) = \{u \in U \mid \exists t = (s_t,u,m_t),\\ s_t \in P_i \wedge e \in E(m_t) \wedge w \in W(m_t)\} \end{aligned} \tag{6}$$

$$\begin{aligned} U_{hot}(P_i,e,w) = \{u \in U_{hot}(P_i) \mid \exists t = (s_t,u,m_t),\\ s_t \in P_i \wedge e \in E(m_t) \wedge w \in W(m_t)\} \end{aligned} \tag{7}$$

$PMI(x,y)$ を算出するための確率の値は，これらの 4 つのユーザー集合により次のように算出される．

$$Pr(x) = \frac{\sum_{P_i} |U(P_i,e,w)|}{\sum_{P_i} |U(P_i,e)|} \tag{8}$$

$$Pr(y) = \frac{\sum_{P_i} |U_{hot}(P_i,e)|}{\sum_{P_i} |U(P_i,e)|} \tag{9}$$

$$Pr(x,y) = \frac{\sum_{P_i} |U_{hot}(P_i,e,w)|}{\sum_{P_i} |U(P_i,e)|} \tag{10}$$

以上より，$Rscore_w(e,w)$ は次のように算出される．

$$\begin{aligned} &Rscore_w(e,w)\\ &= \log \frac{Pr(x,y)}{Pr(x)Pr(y)}\\ &= \log \frac{\sum_{P_i} |U_{hot}(P_i,e,w)| \times \sum_{P_i} |U(P_i,e)|}{\sum_{P_i} |U(P_i,e,w)| \times \sum_{P_i} |U_{hot}(P_i,e)|} \end{aligned} \tag{11}$$

このようにして算出された $Rscore_w(e,w)$ を評価することで，ノイズキーワードを特定することができる．PMI の性質から，$Rscore_w(e,w)$ が 0 よりも小さければ小さいほど，キーワード w がノイズキーワードである可能性は高くなる．

2.3　ノイズツイートの特定

ノイズツイートを特定するために，キーワード w のイベント関連スコア $Rscore_w(e,w)$ を用いて，ツイート t のイベント関連スコア $Rscore_t(e,t)$ を算出する．$Rscore_t(e,t)$ は，イベント関連語 e とツイート t の組に対して算出される．なぜならば，1 つのツイートに複数のイベント関連語を含む可能性があり，各イベント関連語 e に対して，$Rscore_w(e,w)$ を用いてノイズツイートかどうかを判定する必要があるためである．ツイート t のイベント関連スコア $Rscore_t(e,t)$ は，次の式により算出される．

$$Rscore_t(e,t) = \sum_{w \in W(m_t)} Rscore_w(e,w) \times Confidence(e,w) \tag{12}$$

ここで，$Confidence(e,w)$ は高頻度語に対して高い重み付けを行うための関数である．この重み付け関数は，低頻度語がイベント関連スコアに与える影響を低減し，高頻度語を重視することで，より信頼性の高いスコアを得ることを目的としている．本稿では，e と w を両方含むツイートを発信したユーザー数の対数を重み付け関数として採用している．

$$Confidence(e,w) = \log \sum_{P_i} |U(P_i,e,w)| \tag{13}$$

最終的に，$Rscore_t(e,t)$ を評価することでノイズツイートを特定する．

［定義 2］　ノイズツイートの集合 TW_{Noise} を次のように定義する．

$$\begin{aligned} TW_{Noise} = \{t \in TW_E \mid \exists t = (s_t,u_t,m_t),\\ e \in E(m_t) \wedge Rscore_t(e,t) < \tau\} \end{aligned} \tag{14}$$

ここで，τ は閾値パラメータである．

パラメータ τ については，$Rscore_w(e,w)$ が PMI をベースにしていることから，$Rscore_t(e,t)$ が 0 よりも小さいツイートがノイズツイートであることが期待される．従って，$\tau = 0$ が適切な設定であると予想されるが，パラメータ τ の調整については，4.3 節で議論することにする．

3. データセット

本稿では，「東京モーターショー 2013」を対象イベントとしたノイズツイート除去の評価実験を行った．対象イベントの開催場所は東京ビッグサイト，開催期間は 2013 年 11 月 22 日から 12 月 1 日までの 10 日間であった．対象イベントに関連するツイートを収集するために，東京モーターショーについて紹介したガイドブックや Web サイトを参考にして，人手でイベント関連語リストを作成した．イベント名やイベント関連のハッシュタグ，出展企業名，ブランド名，製品名など，147 のイベント関連語を選定した．日本語表記と英語表記を対象とし，“日産” と “ニッサン” のような表記ゆれについても考慮してイベント関連語リストを作成した．

作成したイベント関連語を含むツイートを Twitter Search API(注3) を用いて収集した．収集したツイートのうち，イベントモニタリングの観点で興味の対象外である以下の 4 種類のツイートを除外した．

- リツイート
- URL を含むツイート
- ボットアカウントからのツイート
- 日本語以外のツイート

URL を含むツイートを除外した理由は，ニュース記事など他の Web サイトの情報を再発信しているだけのものが多いためである．しかしながら，写真共有サービスや個人ブログなど特定のドメインから発信された URL を含むツイートは，ユーザーの意見や感情を含んでおり，イベントモニタリングにおいて有用である可能性がある．そこで，そのようなサービスの URL のホワイトリストを作成し，このリスト中に含まれる URL を含むツイートは除外しないようにした．また，ツイート収集期間において，2 回以上同じメッセージを投稿したユーザーをボットアカウントとみなして除外した．

イベント関連語リストを用いて収集したツイートから上記 4 種類のツイートを除外し，残った 596,611 ツイートを実験対象のツイート集合 TW_E とした．また，各ツイートのメッセージテキスト m_t に対して形態素解析を行い，名詞を抽出して $W(m_t)$ とした．

4. 評 価 実 験

提案手法の有効性を検証するために，3. 章で述べた東京モーターショー 2013 のデータセットを用いて，次の 3 つの項目について評価実験を行った．

（1） ホットユーザーの妥当性

ホットユーザーに関する仮定 1 の妥当性を検証するために，言及したイベント関連語の異なり語数が多いユーザーが発信したツイートは，異なり語数が少ないユーザーと比べて，イベントについて言及している可能性がより高くなることを示す．

（2） キーワードに対するイベント関連スコアの妥当性

2.2.2 節で述べた方法で算出された，キーワードに対するイベント関連スコアの妥当性を検証する．

（3） ノイズツイート特定の精度

2.3 節で述べた方法でノイズツイートを特定し，精度について評価を行う．また，パラメータ τ の設定について考察する．

本実験では，データセットにおいて出現頻度が高い上位 10 の自動車会社，スバル，日産，トヨタ，ホンダ，三菱，スズキ，ベンツ，ポルシェ，マツダ，ジャガーを対象として評価を行った．

4.1 ホットユーザーの妥当性

言及したイベント関連語の異なり語数が多いユーザーが発信したツイートは，異なり語数が少ないユーザーと比べて，イベントについて言及している可能性がより高くなることを確認する．これにより，提案手法が前提としている，ホットユーザーの特徴に関する仮定 1 の妥当性が示されることになる．

期間 P_i において，言及したイベント関連語の異なり語数が 1 から 5 のユーザーを，ランダムサンプリングによりそれぞれ 100 ユーザー選択する．これらのユーザーが発信したツイートが対象イベントに言及しているかどうかを，各ユーザーが発信したツイートの内容を元に人手で判定する．

表 1 に結果を示す．1 列目は，期間 P_i において，あるユーザーが発信したツイートに含まれるイベント関連語の異なり語数を表している．2 列目は，全期間 P における延べユーザー数で，例えばユーザー u が期間 P_3 と P_5 においてイベント関連語について言及している場合は 2 回カウントされることになる．3 列目は，2 列目でカウントしたユーザーが対象イベントについて言及している割合で，ランダムに 100 ユーザーをサンプリングして人手で判定した結果である．

表 1 から，対象イベントに言及しているユーザーの割合は，言及しているイベント関連語の異なり語数が多いほど高くなる傾向が見て取れる．この結果は，ホットユーザーに関する仮定 1 が妥当であることを示している．また，この結果に基づき，ホットユーザーを特定するための閾値パラメータを $\theta = 3$ と設定して以後の実験に用いる．

表 1 イベント関連語の異なり語数とイベントに言及したユーザーの割合の関係
Table 1 The unique number of event-related words and the ratio of users who mentioned the event.

イベント関連語の異なり語数	全期間 P における延べユーザー数	対象イベントに言及したユーザーの割合
1	127,437	0.16
2	16,786	0.45
3	5,565	0.59
4	2,465	0.73
5	1,272	0.66

(注3)：https://dev.twitter.com/docs/api

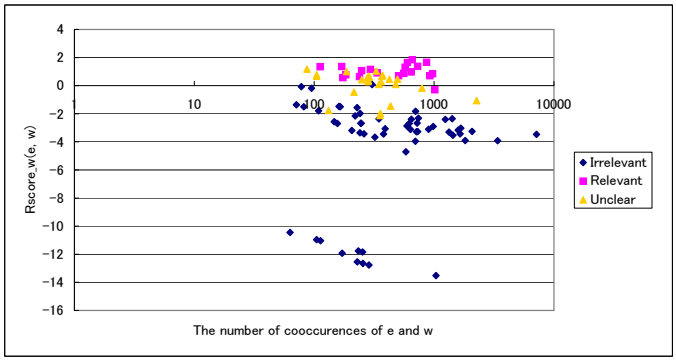

図 2　各クラスの $Rscore_w(e,w)$ の分布
Fig. 2　Distributions of $Rscore_w(e,w)$ for each class.

4.2　キーワードに対するイベント関連スコアの妥当性

対象とする自動車会社 10 社それぞれについて，ツイート内で会社名と共起している頻度が高い上位 10 のキーワードを抽出する．抽出された合計 100 の会社名 e とキーワード w の組に対して，人手により以下の 3 つのクラスに分類した．

- Relevant: e と w の組は，モーターショー関連の話題であることを強く示唆している．
- Irrelevant: e と w の組は，モーターショー関連の話題ではないことを強く示唆している．
- Unclear: e と w の組からは，モーターショー関連の話題かどうかを判断できない．

例として，“ホンダ” と “ブース”，“三菱” と “コンパニオン” が Relevant，“日産” と “スタジアム”，“三菱” と “口座” が Irrelevant，“スズキ” と “バイク”，“ホンダ” と “エンジン” が Unclear に分類された．

図 2 に，e と w の組に対するイベント関連スコア $Rscore_w(e,w)$ の分布を示す．図において，横軸は e と w の共起数，縦軸は $Rscore_w(e,w)$ の値である．この結果から，キーワードに対するイベント関連スコア $Rscore_w(e,w)$ により，3 つのクラスが弁別されている様子が確認できる．また，Relevant のキーワードは $Rscore_w(e,w) > 0$ の領域に，Irrelevant のキーワードは $Rscore_w(e,w) < 0$ の領域に，そして，Unclear のキーワードは $Rscore_w(e,w) = 0$ 付近に分布している様子が見て取れる．また，特に Irrelevant キーワードについて，頻度が高くなるほどイベント関連スコアがより低く算出される傾向が見て取れる．以上の結果から，キーワードに対するイベント関連スコアは妥当に算出されており，このスコアを用いることでノイズツイートを高精度に特定できることが期待される．

4.3　ノイズツイート特定の精度

ノイズツイート特定の精度を評価するために，評価データを作成した．対象の自動車会社 10 社それぞれに対して，会社名を含むツイートをランダムに 100 ツイート抽出し，人手により以下の 3 つのクラスに分類した．

- Relevant: ツイート t は明らかにモーターショー関連のツイートである．
- Irrelevant: ツイート t は明らかにモーターショー関連ではないツイートである．
- Unclear: ツイート t からは，モーターショー関連のツイートかどうかを判定できない．

各ツイートに対して人手により 3 つのクラスへの分類を行った．その際に，添付された画像や動画，ユーザープロフィール等，ツイートのメッセージテキスト以外の情報は参照せずに分類を行った．評価データにおける各クラスごとのツイート数を表 2 に示す．

表 2　評価データにおける各クラスごとツイート数
Table 2　Distributions of three classes in evaluation data.

会社名	Relevant	Unclear	Irrelevant
スバル	8	18	74
日産	1	12	87
トヨタ	6	15	79
ホンダ	17	35	48
三菱	19	13	68
スズキ	7	36	57
ベンツ	4	6	90
ポルシェ	3	5	92
マツダ	9	28	63
ジャガー	1	4	95
Total	75	172	753

ここで，ノイズツイートを特定するためのパラメータ τ について検討する．2.3 節では，PMI の性質から $\tau = 0$ が適切な設定であるとしたが，この妥当性について確認する．図 3 は τ の値を変化させることにより，ノイズツイート特定の精度がどのように推移するかを示している．この図から，$\tau = 0$ を境にノイズツイート特定の精度が急激に減少している様子が見て取れる．

続いて，$\tau = 0$ の周辺における，ノイズツイート特定の精度を調査するために，$\tau = 0$，$\tau = \epsilon$ における精度を比較する．ここで ϵ は，限りなく 0 に近い正の値である．定義 2 によれば，$\tau = 0$ は $Rscore_t(e,t) < 0$ のツイートを，$\tau = \epsilon$ は $Rscore_t(e,t) \leqq 0$ のツイートをそれぞれノイズとみなすことを意味する．図 4 は，各会社ごとの $\tau = 0$ と $\tau = \epsilon$ におけるノイズツイート特定の精度を差異を示している．この結果から，ほとんどの会社において $\tau = \epsilon$ よりも $\tau = 0$ の方が精度が高いことが分かる．

表 3 に，$\tau = 0$ における精度，再現率，F 値を示す．この結果から，提案手法は $\tau = 0$ において高い精度を達成し，F 値も実用的な値になっていることが分かる．ここから少しでも τ の値を増やせば，図 4 に示したように精度は急激に減少する．結論として，提案手法において $\tau = 0$ が適切なパラメータであり，提案手法の適用にあたっては τ を調整する必要がないことが期待される．

5.　関 連 研 究

我々の知る限りでは，本研究はイベントモニタリングを目的としたノイズツイートの特定という問題に取り組んだ最初の研究である．また，ホットユーザーの概念を用いてノイズツイートを特定するアプローチは既存研究において見当たらない．本

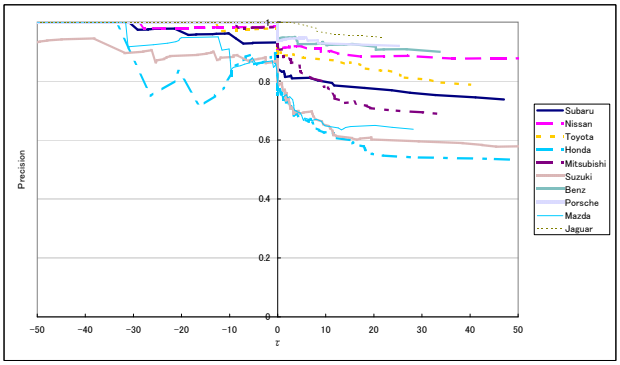

図 3　ノイズツイート特定における τ と精度の関係

Fig. 3　Relationships between τ and the precision in noise tweet filtering.

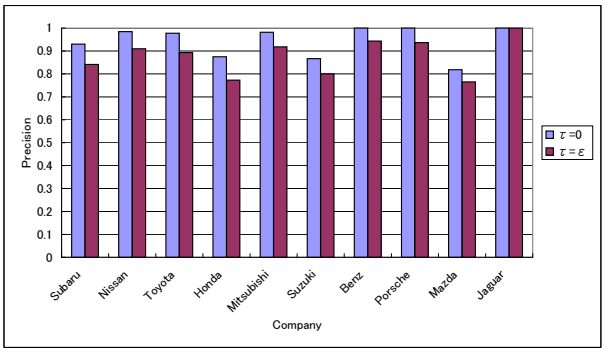

図 4　各会社における $\tau = 0$ と $\tau = \epsilon$ での精度の比較

Fig. 4　Comparison of precisions between $\tau = 0$ and $\tau = \epsilon$ for each company.

表 3　$\tau = 0$ におけるノイズツイート特定のパフォーマンス

Table 3　Performances of noise tweet filtering when $\tau = 0$.

会社名	Precision	Recall	F-measure
スバル	0.93	0.73	0.82
日産	0.98	0.74	0.85
トヨタ	0.98	0.56	0.72
ホンダ	0.88	0.53	0.66
三菱	0.98	0.81	0.89
スズキ	0.87	0.68	0.76
ベンツ	1.00	0.46	0.63
ポルシェ	1.00	0.54	0.70
マツダ	0.82	0.57	0.67
ジャガー	1.00	0.85	0.92
Total	0.94	0.65	0.77

章では，ノイズツイートを特定するために適用可能な関連研究を示し，本研究との差異を示す．

ノイズツイートを特定する問題は，複数の意味を持つ語に対して意味を一意に決定する，語義の曖昧性解消に関する研究 [9] と関連している．なぜなら，複数の意味を持つイベント関連語に対して，曖昧性を解消することでイベントと関連がないことが明らかになり，ノイズツイートであることが判定できる場合があるためである．ただし，例えば “ホンダ” が自動車会社であったとしてもモーターショーと関連がない場合（中古車販売の広告ツイートなど）もあるため，語義の曖昧性解消だけでは十分ではないことに注意する必要がある．

語義の曖昧性解消の研究は長い歴史があるが，多くの研究は，新聞記事のような文法規則に沿って書かれた，十分な長さのあるテキストを対象としている．一方で，近年の研究では，ツイートへの適用を想定し，話し言葉で書かれた短いテキストを対象とした研究が行われている．[1] は教師あり学習のアプローチを用いて，また，[11] [13] は Wikipedia を用いてツイート中の語に対して語義の曖昧性解消を行っている．

また，ツイート中の語に対して知識ベースにおけるエントリに関連付けを行う，Entity Linking に関する研究 [2] [3] [4] [5] [6] が近年注目されている．これらの研究はいずれも Wikipedia を前提としたアプローチであり，Wikipedia が持つカテゴリ情報などを利用してエントリとの関連付けを行う．

以上で述べたアプローチは，いずれも本研究におけるノイズツイートを特定するために適用することが可能である．しかし，人手により作成された教師データや知識ベースを必要とすることから，提案手法のアプローチとは異なる．

本研究では，ツイート単体からイベントに関連しているかどうかを判定できない場合は Unclear としていたが，Twitter のような短いテキストからコンテキスト情報を取得するために，ユーザープロフィールを活用する研究 [7] [8] [10] [12] も行われている．本研究では Unclear としていたツイートが，これらのアプローチを適用することで Relevant か Irrelevant かを判定できる可能性がある．[7] は，Wikipedia を用いてツイートのユーザープロフィールを分析するための初期検討を行った．[10] は，Wikipedia を用いてユーザーの興味を分析するためのグラフベースの枠組みを提案した．[8] は，Wikipedia の編集履歴をユーザープロフィールに用いて，語義の曖昧性解消を行う手法を提案した．[12] は，ユーザープロフィールを用いて，ツイートをニュース，イベント，オピニオンのクラスに分類した．これらの手法は，ユーザープロフィールや知識ベースを利用する手法だが，本研究で提案した手法により事前にノイズツイートを除去しておくことで，処理コストを削減することができる．

6.　結論と今後の課題

本論文では，イベントモニタリングを目的としてノイズツイートを特定する手法を提案した．提案手法は，人手によって作成された教師データや知識ベースを必要としない点で新しく，実用的である．提案手法は，ホットユーザーという概念を用いてイベントと関連がないことを示唆するノイズキーワードを特定し，最終的にノイズツイートを特定する．また，評価実験により提案手法がノイズツイートを高精度に特定できることを示した．

今後の課題としては，提案手法をより大規模なイベントに適用することが挙げられる．このようなイベントでは，イベント関連語を人手で定義するコストが増大する恐れがあるため，これを支援するための手法が重要になる．また，国際的なイベントでは，複数の言語に対して言語横断的にイベント関連のツイートを収集し，ノイズツイートを除去する手法が必要である．

本論文では，イベントモニタリングを目的としてノイズツ

イートの特定を行ったが，提案手法の適用先はイベントモニタリングに限定されるものではない．分析トピックに対して関連語を定義することができ，ホットユーザーが存在するような分析トピックであれば提案手法を適用することができる．例えば，都市のモニタリングは提案手法の有望な適用先のひとつである．自治体関係者は自分たちが管轄する都市が Twitter 上でどのように言及されているかを知ることで，都市のプロモーションやブランド戦略に活用したり，風評や突発的な事件，事故を素早く把握し対応することができる．このように，提案手法をイベントモニタリング以外の目的で利用したときの有効性の検証も今後の課題である．

文　献

[1] A. Davis, A. Veloso, A.S. da Silva, W. Meira Jr., and A.H. F. Laender. “Named Entity Disambiguation in Streaming Data.” Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1. Association for Computational Linguistics, 2012.

[2] P. Ferragina, and U. Scaiella. “TAGME: On-the-fly Annotation of Short Text Fragments (by Wikipedia Entities).” Proceedings of the 19th ACM international conference on Information and knowledge management. ACM, 2010.

[3] A. Gattani, D.S. Lamba, N. Garera, M. Tiwari, X. Chai, S. Das, S. Subramaniam, A. Rajaraman, V. Harinarayan, and A. Doan. “Entity Extraction, Linking, Classification, and Tagging for Social Media: A Wikipedia-based Approach.” Proceedings of the VLDB Endowment 6.11, pp.1126-1137, 2013.

[4] Y. Genc, Y. Sakamoto, and J.V. Nickerson. “Discovering Context: Classifying tweets through a semantic transform based on Wikipedia.” Foundations of Augmented Cognition. Directing the Future of Adaptive Systems. Springer Berlin Heidelberg, pp.484-492, 2011.

[5] C. Li, J. Weng, Q. He, Y. Yao, A. Datta, A. Sun, and B. Lee. “TwiNER: Named Entity Recognition in Targeted Twitter Stream.” Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval. ACM, 2012.

[6] E. Meij, W. Weerkamp, and M. Rijke. “Adding Semantics to Microblog Posts.” Proceedings of the fifth ACM international conference on Web search and data mining. ACM, 2012.

[7] M. Michelson, and S.A. Macskassy. “Discovering Users’ Topics of Interest on Twitter: A First Look.” Proceedings of the fourth workshop on Analytics for noisy unstructured text data. ACM, 2010.

[8] E.L. Murnane, B. Haslhofer, and C. Lagoze. “RESLVE: Leveraging User Interest to Improve Entity Disambiguation on Short Text.” Proceedings of the 22nd international conference on World Wide Web companion. International World Wide Web Conferences Steering Committee, 2013.

[9] R. Navigli. “Word Sense Disambiguation: A Survey.” ACM Computing Surveys (CSUR) 41.2, 2009.

[10] W. Shen, J. Wang, P. Luo, and M. Wang. “Linking Named Entities in Tweets with Knowledge Base via User Interest Modeling.” Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2013.

[11] D. Spina, J. Gonzalo, and E. Amigo. “Discovering Filter Keywords for Company Name Disambiguation in Twitter.” Expert Systems with Applications 40.12, pp.4986-5003, 2013.

[12] B. Sriram, D. Fuhry, E. Demir, H. Ferhatosmanoglu, M. emirbas. “Short Text Classification in Twitter to Improve Information Filtering.” Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval. ACM, 2010.

[13] S.R. Yerva, Z. Miklos, and K. Aberer. “Entity-based Classification of Twitter Messages.” International Journal of Computer Science & Applications 9.EPFL-ARTICLE-174746, pp.88-115, 2013.