ユーザの関心に応じた マイクロブログからの実世界観測情報の抽出

吉武　真人[a]　　新田　直子[b]　　馬場口　登[b]

大阪大学大学院工学研究科　〒565-0871 大阪府吹田市山田丘 2-1

a) *yoshitake@nanase.comm.eng.osaka-u.ac.jp*　b) {*naoko, babaguchi*}*@comm.eng.osaka-u.ac.jp*

概要　本研究では，世界中の利用者からリアルタイムの実世界観測情報が多く投稿されるマイクロブログから，ユーザの関心に応じた観測情報を抽出することを目的とする．ユーザの関心を表す単語としてクエリが与えられたとき，ユーザの関心に合致した観測情報は，クエリと意味的関連度の高い単語を，合致しない観測情報は，クエリと意味的関連度の低い単語を多く含むと考えられる．そこで，マイクロブログにて使用される単語間の意味的関連度を，現在までの投稿における単語の共起関係から逐次的に算出し，ユーザからクエリが与えられた時点で，各投稿に含まれる単語のクエリに対する関連度分布を抽出する．これを特徴量とした識別器により，任意のクエリに対し，適切な実世界観測情報の抽出を目指す．

キーワード　マイクロブログ，情報抽出，実世界観測情報，単語間関連度

1　はじめに

近年，人間が実世界を観測して得られた情報を観測時間や場所の情報とともに，マイクロブログや画像共有サイトをはじめとするソーシャルメディアで公開していることに着目し，ソーシャルメディア上の情報から実世界観測情報を獲得する研究が注目されている．人間は実世界の様々な場所に存在し，観測した情報の意味を解釈できるので，人間をセンサ（Citizen Sensor）[1] とみなして利用することにより，センサ設置のコストを抑えた上で多様な情報が獲得できる．マイクロブログの一つである Twitter では，主な投稿形式がツイートと呼ばれる 140 文字以下の短文であり，その手軽さによりリアルタイム性の高い様々な実世界観測情報が投稿されている．

Twitter を用いた実世界観測情報の抽出に関する既存研究では，観測対象の関連語を用いて観測対象に関連したツイートを抽出する手法が中心となっている．例えば，Sakaki ら [2] は，予め地震に関する単語を関連語として人手で設定することにより，地震に関するツイートを抽出し，震源地を推定した．また，土屋ら [3] は，予め準備した鉄道の運行トラブルに関するツイート集合から関連語を学習し，鉄道の運行トラブルを抽出した．

ユーザによって与えられたクエリを観測対象として，多様な観測対象の関連語を現在までのツイートから自動的に学習する手法も提案されている．Massoudi ら [4] や藤木ら [5] は，観測対象に関する特徴的な事象が発生した際に，その事象を表す単語とクエリの同一ツイート内での共起頻度が一時的に高くなると考え，クエリと短期的に共起する単語を関連語とした．この手法により，例えば，渋滞という観測対象に対して，渋滞が発生している場所を表す単語などが関連語となり，各地で発生している渋滞に関するツイートが抽出できると考えられる．

本研究では，ユーザの関心を表す単語であるクエリにより定められた観測対象に加え，例えば渋滞の要因となる事故や工事，通行規制など，ユーザからのクエリにより定められた観測対象に関連する対象の観測情報を同時に抽出することを考える．この場合，各対象に関する観測情報は独立していることが多いため，例えば，クエリとなる渋滞という単語に対して，事故の観測情報に含まれる単語の共起頻度が短期的に高くなる可能性は低い．しかし，観測対象同士が関連するため，渋滞と事故という単語対のように，対象を表す単語同士は，時間によらず，頻度は低いものの同一ツイート内に共起する可能性が高いと考えられる．そこで提案手法では，長期間に投稿されたツイート集合から，同一ツイートに断続的に共起する単語対に対して高く，共起しない単語対に対して低くなるような単語間の関連度を逐次的に算出する．クエリが与えられると，各ツイートに含まれる単語のクエリに対する関連度分布を特徴量として抽出し，この特徴量に基づく 2 クラス分類器を用いて，クエリに関連する対象の観測情報か否かを判定する．あらかじめ全ての単語間の関連度を算出しておくことにより，任意のクエリに対する関連度を示す特徴量抽出が可能となり，観測対象ごとに関連語や学習データを与える必要なく，関連観測情報が抽出できると期待される．

2　提案手法

提案手法では，ユーザの関心を表す単語としてクエリ q が与えられたとき，現在の直近の時区間において Twitter に投稿されたツイートから，q で表される観測対象，及びそれに関連する対象の観測情報を含むツイー

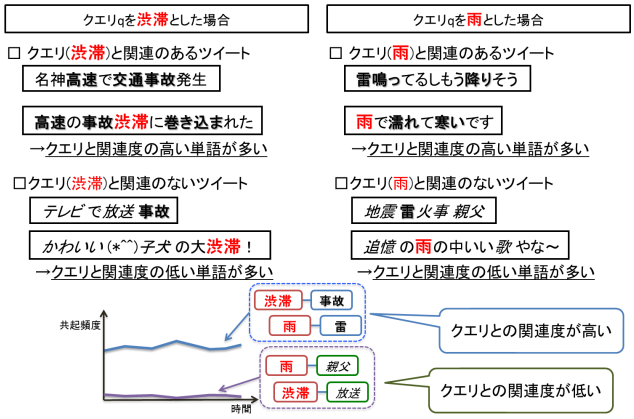

図 1 クエリに対する適合ツイート，非適合ツイートの例

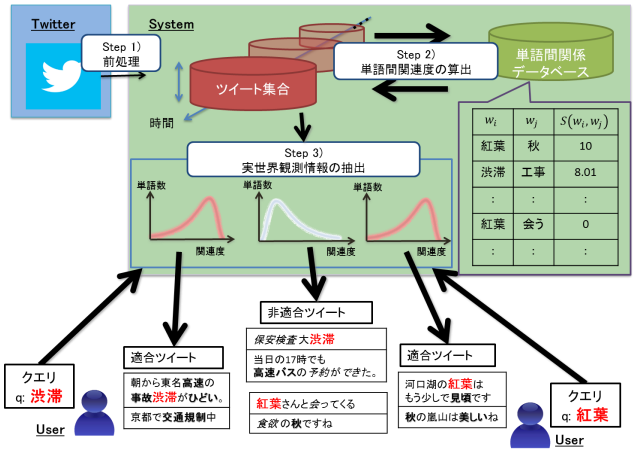

図 2 提案手法の概要

トを抽出することを目的とする．例えばクエリ q として渋滞が与えられたとき，q で表される観測対象は渋滞，それに関連する対象とは，渋滞の原因となる事故や，渋滞を回避する抜け道などが挙げられる．このようなユーザの関心に合致するツイートを適合ツイート，また，ユーザの関心に合致しないツイートを非適合ツイートと呼ぶ．例として，2 つの異なるクエリに対し，適合ツイートと，非適合ツイートを図 1 に示す．適合ツイートは，必ずしもクエリを含まないが，渋滞に対する事故や抜け道のように，一般にクエリから連想しうる単語を多く含むと考えられる．一方，非適合ツイートは，クエリやクエリから連想しうる単語を含む場合もあるが，クエリからは連想されない単語を多く含むと考えられる．また，クエリから連想しうる単語は，渋滞に対する事故などクエリと関連性が高く，時間によらず，頻度は低いものの同一ツイート内にクエリと共起する可能性が高い．一方，クエリから連想されない単語は，クエリとの関連性が低いため，時間によらずクエリと同一ツイート内に共起する可能性も低いと考えられる．

以上より，提案手法は，図 2 に示すように，以下のステップにより構成される．

Step1) ツイートの収集・前処理：

短い時区間ごとに Twitter からツイートを収集し，冗長ツイートや不要語などを除去する．

Step2) 単語間関連度の算出：

収集したツイートから共起単語対の抽出，及びその単語間関連度の算出を行い，単語対 (w_i, w_j) に対する単語間関連度 $S(w_i, w_j)$ を保持する単語間関係データベースを作成・更新する．

Step3) 実世界観測情報の抽出：

ユーザからクエリ q が与えられたとき，算出した単語間の関連度に基づき，各ツイートに含まれる単語のクエリに対する関連度分布を特徴量として抽出する．この特徴量を用い，クエリに関連するか否かを判定する 2 クラス分類器を用いて，ユーザの関心に応じた観測情報を含むツイートを抽出する．

次節以降で，各ステップの詳細について述べる．

2.1 ツイートの収集・前処理

時区間 I ごとに Twitter に投稿されたツイートを収集する．ただし，あるユーザにより投稿されたツイートを別のユーザがそのまま再投稿したツイートであるリツイート，および同一の内容で大量に投稿されるスパムツイートは除去する．

また，収集したツイートに対して MeCab[6] による形態素解析を行い，一単語で意味を持つ単語の多い，名詞，動詞，形容詞のみを各ツイートから抽出する．活用形のある動詞と形容詞については，抽出時に原形に変換する．また，URL である「http://～」やユーザ名を表す「@～」をはじめとする英数字のみで構成される単語は，不要な単語として除去する．

2.2 単語間関連度の算出

収集したツイートを用いて，単語対 (w_i, w_j) に対する単語間関連度 $S(w_i, w_j)$ を保持する単語間関係データベースを作成・更新する．ただし，時区間長 I において，w_i と w_j の共起回数が 1 回の場合は，ノイズである可能性が高いため，共起回数 2 回以上の単語対のみを考慮する．また，「笑」のような極めて出現確率の高い一般的な単語は，関連性の無い単語とも頻繁に共起する可能性が高い．そこで，w_i と w_j の相互情報量 $B(w_i, w_j)$ を算出する．単語間の相互情報量 $B(w_i, w_j)$ が負の場合，w_i と w_j は相対的に共起せず，「笑」のような出現確率が高い一般的な単語を含む単語対に対しては，共起頻度が高くても相互情報量は低くなる．そのため，相互情報量 $B(w_i, w_j) \geq \beta$ を満たす単語対 (w_i, w_j) のみに対し，単

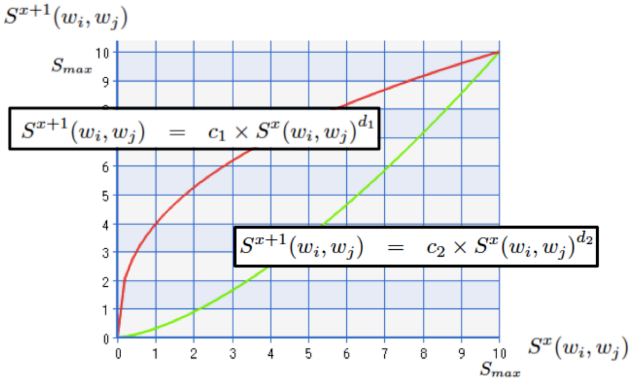

図 3 単語間関連度の更新

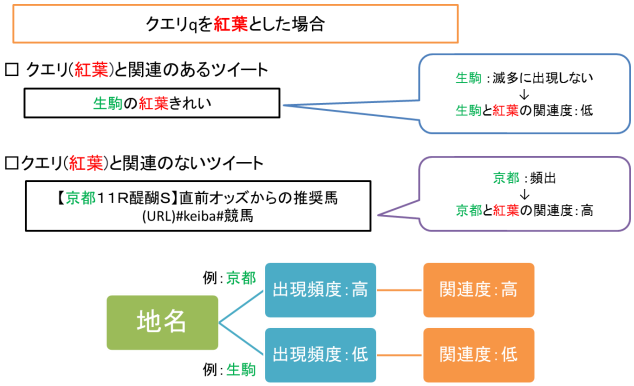

図 4 場所を表す単語による影響

語間の連続的な関連性を表す指標として，単語間関連度 $S(w_i, w_j)$ を以下のように算出する．

まず，(w_i, w_j) がデータベースに含まれていない場合，(w_i, w_j) を追加し，$S(w_i, w_j)$ の初期値として，$S(w_i, w_j) = 1$ と設定する．また，(w_i, w_j) がデータベースに含まれている場合，現在の単語間関連度を $S^x(w_i, w_j)$ とし，$S^{x+1}(w_i, w_j)$ を以下のように更新する．

$$S^{x+1}(w_i, w_j) = c_1 \times S^x(w_i, w_j)^{d_1} \quad (1)$$

ただし，$d_1 < 1$ とする．最後に，収集したツイート中に共起しなかったデータベース中の単語対 (w_i, w_j) に対し，$S^{x+1}(w_i, w_j)$ を以下のように更新する．

$$S^{x+1}(w_i, w_j) = c_2 \times S^x(w_i, w_j)^{d_2} \quad (2)$$

ただし，$d_2 > 1$ とする．

式 (1)，(2) を図 3 に示す．$d_1 < 1$ とすることにより，式 (1) は連続して共起する単語対に対して，単語間関連度を上昇させる．関連度が高い程，上昇の度合いが小さくなり S_{max} に収束する．また，$d_2 > 1$ とすることにより，式 (2) は共起しない期間が連続する単語対に対して，単語間関連度を下降させる．関連度が低い程，下降の度合いが小さくなり 0 に収束する．また，式 (2) の傾きを，式 (1) の傾きより小さく設定することにより，共起する時区間が散発する場合も，単語対の単語間関連度を上昇させることができる．

また，c_1 と c_2 は，$S(w_i, w_j)$ の最大値 S_{max} により次式で決定される．

$$c_1 = S_{max}^{(1-d_1)} \quad (3)$$

$$c_2 = S_{max}^{(1-d_2)} \quad (4)$$

最後に，単語間関連度が初期値を下回る，すなわち，$S(w_i, w_j) < 1$ を満たす単語対 (w_i, w_j) をデータベースから削除する．

2.3 実世界観測情報の抽出

ユーザからクエリ q が与えられたとき，まず，適合ツイート候補の抽出を行う．あるツイートがクエリ q で表される対象の観測情報もしくは q に関連する対象の観測情報である場合，必ず q または q と関連する単語 w が含まれると考えられる．そこで，クエリ q もしくは，$S(q, w) \geq 1$ となる単語 w を含むツイートを適合ツイート候補として抽出する．

次に，各適合ツイート候補について，単語間関連度に基づき，以下の 3 つの特徴量を抽出する．

a) 関連度ヒストグラム

抽出すべき観測情報を含むツイートは，クエリ q もしくは q と単語間関連度が高い単語を多く含む可能性が高い．一方，抽出するべきでないツイートは，q と単語間関連度の低い単語を多く含む可能性が高い．よって，ツイートに含まれる各単語とクエリとの単語間関連度の分布を示す関連度ヒストグラムを特徴量として設定する．

b) クエリの出現回数

クエリ q の出現に関する情報は関連度ヒストグラムには含まれない．しかし，クエリが含まれているかどうかは，クエリ q で表される，もしくは q に関連した観測情報を含むツイートを抽出する上で，非常に重要な要素である．よって，ツイート内でのクエリの出現回数を特徴量として設定する．

c) 場所を表す単語の出現回数

場所を表す単語に対する単語間関連度は，実際の人間の感覚とは異なる数値になることが多い．例えば，都道府県名など，広い空間領域を表す単語は出現頻度が高く，多様な単語と共起するため，クエリと関連性の低い単語であってもクエリとの単語間関連度は高くなる．逆に，市町村名など，狭い空間領域を表す単語は出現頻度が極端に低く，クエリと関連性の高い単語であってもクエリとの単語間関連度は低くなる．図 4 に示す例では，出現頻度が高く，様々な単語と高い関連度を持つ京都という単語が，紅葉とは関連のない京都で開催されている競

馬に関連するツイートにおいて，関連度の高い単語として関連度ヒストグラムに含まれる．逆に，生駒という単語は紅葉の場所を表しているにも関わらず，出現頻度が極めて低く，紅葉と共起していないために，関連度の低い単語として関連度ヒストグラムに含まれる．このように，場所を表す単語は誤分類の原因となり得るが，実世界観測情報は特定の場所で観測された情報であり，観測場所の情報を含むことが望ましい．よって，場所を表す単語がツイートにどれだけ含まれているかを特徴量として利用することにより，実世界観測情報か否かの判定精度の向上が期待される．そこで，MeCab により地域を表す固有名詞に分類された単語を場所を表す単語とし，このような単語のクエリとの単語間関連度は関連度ヒストグラムに含めず，ツイート内での出現回数を特徴量として設定する．

これらの 3 つの特徴量を並べた特徴量ベクトルに基づき，予め用意しておいた学習データを用いて，サポートベクターマシン（SVM）により分類器を生成し，適合ツイート候補を分類する．分類の信頼度の高いツイートから順にユーザに提示する．

3 実験

2013/10/25〜2013/11/27 に，日本語を用いて投稿されたツイートを収集し，そのうちのべ 32 日間のツイート 21,134,159 件を実験に用いた．また，I を 24 時間とし，各パラメータは $S_{max} = 10$，$\beta = 2.0$，$d_1 = 0.4$，$d_2 = 1.5$ とした．収集したツイート集合より単語間関連度を学習し，単語間関係データベースを作成・更新した．また，2013/11/27 のツイート集合から，渋滞，遅延，地震，紅葉，雨という 5 つのクエリに対して，それぞれ適合ツイート例 227 個，非適合ツイート例 354 個を人手で抽出し，これらを学習データとして，実世界観測情報抽出に用いる分類器を学習した．

11/7，11/16 において，それぞれクエリ q を遅延，警報，津波，試合，災害として観測情報の抽出を行った後，抽出したツイートのうち上位 10，30，50 件のツイートに対して，それぞれのクエリに対する適合ツイートとして適切であるか評価した．分類対象が多く，再現率を算出するのが困難なため，抽出結果は，以下の式で定義される適合率 P_N，平均適合率 AP_N により評価する．適合率は，抽出したツイートのうち適合ツイートの割合を，平均適合率は，適合ツイートが抽出したツイートの上位に存在するかを表す指標である．

$$P_N = \frac{R}{N} \tag{5}$$

$$AP_N = \frac{1}{R}\sum_{k=1}^{N}(P_k \times rel(k)) \tag{6}$$

ただし，R は上位 N 件中の適合ツイート数であり，$rel(k)$ は上位 k 件目のツイートが適合ツイートなら 1，適合ツイートでないなら 0 とする．

表 1 評価実験結果

q	日時	P_{10}	P_{30}	P_{50}	AP_{10}	AP_{30}	AP_{50}
遅延	11/7	**1.00**	**0.87**	**0.86**	**1.00**	**0.98**	**0.94**
遅延	11/16	**1.00**	**0.93**	**0.84**	**1.00**	**0.97**	**0.94**
警報	11/7	0.60	$P_{24} = 0.63$		**0.87**	$AP_{24} =$ **0.75**	
警報	11/16	*0.00*	*0.03*	*0.02*	*0.00*	*0.04*	*0.04*
津波	11/7	**0.80**	0.57	*0.36*	**0.86**	**0.78**	**0.76**
津波	11/16	0.60	**0.77**	**0.78**	0.50	0.67	**0.70**
災害	11/7	*0.30*	*0.13*	*0.22*	*0.32*	*0.30*	*0.22*
災害	11/16	**0.70**	**0.83**	**0.78**	**0.79**	**0.81**	**0.81**
試合	11/7	*0.20*	*0.27*	*0.38*	*0.18*	*0.25*	*0.31*
試合	11/16	*0.30*	0.50	0.56	*0.30*	*0.43*	*0.48*

表 1 に結果を示す．クエリ q を警報としたときの 11/7 のデータに関しては，適合ツイートが 24 件しか抽出されなかったため，上位 10，24 件に対して適合率，平均適合率を評価した．

クエリ q が遅延の場合，適合率，平均適合率が共に高い結果が得られた．11/7 において，クエリ q を遅延としたとき，正しく抽出された適合ツイートと除外された非適合ツイートの一部を表 2 に示す．ここで，クエリもしくはクエリとの単語間関連度 $S(q, w) \geq 1.0$ を満たす単語 w を太字で，クエリとの単語間関連度を持たない単語を下線で，ローカル語に分類された単語を二重下線で示している．1-1，1-4 のように，クエリ q を含む適合ツイートが多くの関連語により正しく抽出された．また，1-2，1-3 のように，クエリ q を含まないが，クエリ q に関連する観測情報も，単語間関連度に基づき正しく抽出された．クエリ q が遅延の場合，各鉄道会社などがそれぞれの形式に従って投稿しているツイートが多く存在する．このように，多くの観測情報が投稿される観測対象がクエリである場合，高い適合率で適合ツイートを抽出できると考えられる．また，1-3，1-4 のような一般のユーザによって投稿された適合ツイートも，下位ではあるが正しく抽出された．一方，1-5，1-6 のように，クエリ q もしくは $S(q, w) \geq 1.0$ を満たす単語 w を含む非適合ツイートは，クエリとの関連度を持たない単語を多く含むため，正しく除外された．

次に，11/7，11/16 においてクエリ q を警報としたとき，抽出された適合ツイートの例を表 3，4 に示す．同様の日程において，クエリ q を津波としたときの適合ツイート例を表 5，6 に，クエリ q を災害としたときの適合ツイート例を表 7，8 に，クエリ q を試合としたときの適合ツイート例を表 9，10 に示す．

クエリ q を警報とした場合，11/7 は台風の影響により，関東や東北などで雨風が非常に強かったため，2-1，

表 2 11/7 における q を遅延としたときの適合ツイート，非適合ツイート例

ID	順位	ツイート本文
1-1	1	unko_kanto 東葉高速線【列車遅延】JR 中央 総武 線 (各 停) 内で車両点検を 行っ た影響で、一部列車に遅れが 出 ています。(11/07 09:30) #駅の伝言板 #栃木 県運行速報
1-2	21	11/07 17:15 #京成 線 #Kanto 16:43 頃、都営浅草線内で発生 し た人身事故の影響で、一部列車に遅れや運休が 出 ています。(17:09) Y378 #TrainDelay
1-3	36	都営 浅草 で事故か？その影響？
1-4	37	埼京線、濃霧の影響で遅延…… この時間に？ 明日の朝濃霧で 高崎 線・宇都宮 線が遅延 し ても おかしくはないけど、大丈夫 だろう。
1-5	非適合	最近この時間に 変え たけど遅延か腹痛発生 し たら ギリギリ だよなー
1-6	非適合	東北方面 の 夜行 列車はスノーパル 2355 と 尾瀬夜行 が あり ますし

表 3 11/7 における q を警報としたときの適合ツイート例

ID	順位	ツイート本文
2-1	4	風強いな 思 たら暴風警報出とんか
2-2	6	@(ID) 今、初めて 知っ たんだけど、秋田 県暴風警報出てるよ (◇;)
2-3	19	北日本 と 北陸 強風などに注意 - NHK (URL)

表 4 11/16 における q を警報としたときの適合ツイート例

ID	順位	ツイート本文
3-1	1	ぼくの今日の運勢です 恋愛運 ★★★★★ 金運 ★★☆☆☆ 健康運 ★★★★★ 仕事運 ★★★★☆ 棚ボタ警報発令。美味しいラッキーがいろいろ落ちてきますぜ。お楽しみに。 ラッキーアイテム 『ワイン』→ (URL)
3-2	5	おれの今日の運勢です恋愛運 ★★★★★ 金運★★★☆☆ 健康運 ★★☆☆☆ 仕事運 ★★☆☆☆ お つかれ顔。 (URL)
3-3	26	【警報・注意報情報】 16 日 06 時現在 福岡 地方 警報・注意報発令中

表 5 11/7 における q を津波としたときの適合ツイート例

ID	順位	ツイート本文
4-1	1	【地震情報】[22:18 頃] ▼震源：福島 県 浜 通り（N37.1゜ E140.7゜）▼ 深 さ：約 10km ▼ 規模：M3.9 ▼最大震度：2 ▼津波の心配 なし （気象庁 (URL) ） #earthquake
4-2	22	@(ID) 福島 が震源でした。 最近、地震が多いのです。

表 6 11/16 における q を津波としたときの適合ツイート例

ID	順位	ツイート本文
5-1	5	FNN 16 日午前 3 時 58 分、鹿児島 県 十島 村で震度 3 の地震 津波の心配 なし (URL)
5-2	21	【地震速報】千葉 県 北西 部でマグニチュード 5.8、最大震度 4 の地震発生！←先週「17 日～19 日に巨大地震発生の 恐れ」って あっ たんだが … (URL)
5-3	37	地震！千葉 で M5.5！最大震度 4 だそうです！

表 7 11/7 における q を災害としたときの適合ツイート例

ID	順位	ツイート本文
6-1	3	悩ム…非常用防災トイレ『シートイレ』 ５０回分 災害・断水時でも安心簡易トイレ。 #AmazonJP #アマゾン ＝＝＞ (URL)
6-2	6	トイレトイレ！ トイレ！ トイレえええええ
6-3	8	【非常時・緊急避難 用品】 #8: トイレ非常袋 10 回分入り KM-012 (URL) #地震 #amazon

表 8 11/16 における q を災害としたときの適合ツイート例

ID	順位	ツイート本文
7-1	4	伊豆大島土砂災害から 1 カ月 35 人 死亡4 人 安否不明。伊豆大島で起きた大 規模 な土砂災害から 16 日で 1 カ月が たち ました。35 人が 死亡 し、今も 4 人が 行方不明 のままです。被害 現場 では、朝から 住民 らが 花 を 手向け、冥福 を 祈り(Д;)
7-2	32	【地震情報】 16 日 20 時 44 分頃 ○ 震度3 東京 都２３区
7-3	38	フィリピン 台風、死者 は４４６０人に - TBS News (URL) #東南アジア

2-2，2-3 のような適合ツイートが正しく抽出された．しかし，11/16 においては 3-3 のような正しく抽出された適合ツイートも存在するが，3-1，3-2 のような誤抽出が多く見られた．これは，3-1，3-2 のような非常に内容の似ているツイートが多数存在し，その中に 3-1 のような警報を含むツイートも多く存在したため，単語間関連度

表 9 11/7 における q を試合としたときの適合ツイート例

ID	順位	ツイート本文
8-1	4	最後まで … 希望 をすてちゃいかん あきらめたらそこで試合終了だよ by安西 監督
8-2	6	野球 日本 代表「侍 ジャパン」、強化試合に向け 台湾 に出発　(URL) #FNN
8-3	19	@(ID) 諦めたらそこで試合終了ですよ

表 10 11/16 における q を試合としたときの適合ツイート例

ID	順位	ツイート本文
9-1	15	【動画】［国際 親善試合］オランダ 代表 2-2 日本 代表（2013/11/16） #日本 代表 (URL)
9-2	18	第 92 回 全国 高校サッカー選手権大会　大阪 決勝 履正 社 1-1東海大仰 星　延長戦へ。
9-3	26	高校の 部 第二試合　試合終了　関東 第一高 3-8 沖縄尚学 高　試合中盤から 終盤 は、試合の 流れ は 沖縄 に。二番手投手の 久保 くんが 投打 に活躍！ (URL)

を逐次的に算出していくにつれ，このようなツイートに含まれる単語と警報との関連度が高くなったことが原因であると考えられる．

クエリ q を津波とした場合，4-1，4-2，5-1，5-2，5-3 のように津波の原因となる地震の情報や，それに伴う津波の情報を含むツイートが正しく抽出された．表 1 より，11/7 でツイートの数を増やすに従って適合率が低下しているのに対し，11/16 では適合率が低下していないことがわかる．これは，11/16 に人口の多い関東圏で比較的大きな地震が発生したため，これに関するツイートが増加し，下位においても 5-3 のように地震に関するツイートが得られたためであると考えられる．

クエリ q を災害とした場合，11/7 の適合率が低い原因として，6-1，6-3 のような災害対策用品の宣伝をしているツイートが多く存在するため，トイレなどの単語と災害との関連度が高くなり，6-2 のようなツイートが誤抽出された．しかし，11/16 は関東圏で比較的大きな地震が発生したことや，フィリピンでの台風被害に関するニュースなど，災害に関するツイートが多く投稿されたため，津波の場合と同様に高い適合率を示した．また，7-1，7-2，7-3 のように土砂災害や地震，台風といった様々な災害に関するツイートが正しく抽出された．

最後に，クエリ q を試合とした場合，11/7 においては，8-2 のような野球の国際試合に関するツイートが正しく抽出されたが，8-1，8-3 のような漫画のセリフを使ったツイートが多く抽出されたため，適合率が低くなった．一方，11/16 は，土曜日で高校生の野球やサッカーの試合や，サッカー日本代表の親善試合があったため，9-1，9-2，9-3 のようなツイートが正しく抽出され，11/7 に比べて高い適合率が得られた．

4　まとめ

本研究では，ユーザの関心を表す単語としてクエリが与えられたとき，ユーザの関心に応じた実世界観測情報を抽出する手法を提案した．提案手法では，Twitter への投稿から算出した単語間関連度に基づき，ツイートに含まれる単語のクエリに対する関連度の分布を抽出する．この分布を用いて適合ツイートと非適合ツイートに分類することによって，クエリを限定することなく，ユーザの関心に応じた実世界観測情報を含むツイートを抽出する．2013 年の 32 日間に投稿されたツイートに対し，提案手法によりクエリと関連するツイートを抽出し，抽出結果が適切であるかを主観評価で確認した．抽出精度はクエリによって大きく変わるものの，クエリとして与えられた観測対象に関する観測情報が Twitter に多く投稿される場合には，高い適合率で抽出された．また，クエリに関連した事象が実世界で発生すると適合率が高くなり，発生しなければ適合率は低くなった．

問題点として，内容が非常に似ているツイートが大量に投稿された場合，関連性がないと考えられる単語対に関しても単語間関連度が高くなることがある．その影響により，一部のクエリにおいて適合率が著しく低下した．そのため，今後の課題として，このようなツイートをスパムツイートに含めて除去する必要があると考えられる．

参考文献

[1] Sheth, A.: Citizen Sensing, Social Signals, and Enriching Human Experience, IEEE Internet Computing, Vol. 13, No. 4, pp. 87-92, 2009.

[2] Sakaki, T., Okazaki, M. and Matsuo, Y.: Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors, Proc. WWW, pp. 851-860, 2010.

[3] 土屋圭，豊田正史，喜連川優 : マイクロブログを用いた鉄道の運行トラブル状況抽出に関する一検討，情処研報 IFAT，Vol. 111, No. 31, pp. 1-6, 2013.

[4] Massoudi, K., Tsagkias, M., Dijke, M. D., et al.: Incorporating Query Expansion and Quality Indicators in Searching Microblog Posts, Proc. ECIR, pp. 362-367, 2011.

[5] 藤木紫乃，上田高徳，山名早人 : 経時的な関連語句の変化を考慮したクエリ拡張による Twitter からの情報抽出手法，DEIM forum，C9-5, 2013.

[6] MeCab Japanese morphological analyzer, https://code.google.com/p/mecab