単語間の関係性の経時変化を考慮したマイクロブログからの実世界観測情報の抽出

角谷　直人† 　新田　直子† 　馬場口　登†

† 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
E-mail: †sumiya@nanase.comm.eng.osaka-u.ac.jp, ††{naoko,babaguchi}@comm.eng.osaka-u.ac.jp

あらまし　マイクロブログにはユーザによるリアルタイムの実世界観測情報が多く投稿されるため，これらの自動抽出により実世界の状況が迅速に把握できる．ユーザから与えられるクエリにより観測対象が設定されるとき，クエリと関連性の高い単語を用いて，観測対象に関する観測情報が抽出可能と考えられる．提案手法では特に，単語間の関連性の経時変化に着目し，マイクロブログへの短期間の投稿から逐次的に学習した長期的・短期的という二つの側面の単語間関連性に基づき各投稿のクエリへの関連度を算出することにより，クエリで表される対象に加え，それに関連した対象に関する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，単語間関係

Real-world Observation Extraction from Microblog by Considering Temporal Change of Word Relationship

Naoto SUMIYA†, Naoko NITTA†, and Noboru BABAGUCHI†

† Graduate School of Engineering, Osaka University
2–1 Yamadaoka, Suita, Osaka, 565–0871 Japan
E-mail: †sumiya@nanase.comm.eng.osaka-u.ac.jp, ††{naoko,babaguchi}@comm.eng.osaka-u.ac.jp

Abstract　The current situation of the real world can be observed from a large number of messages posted on a microblog by people all over the world. Given a text query which represents the target to observe, the messages describing the target's current situation can be extracted by using related words. Especially, aiming at extracting the messages describing the current situations of not only the target represented by the query but also other related targets, this paper proposes a sequential approach for learning both the long-term and short-term relationships among words and a method for scoring each message based on the two types of relationships among words.

Key words　Microblog, Information Extraction, Real-world Observation, Word Relationship

1. はじめに

実世界の様々な時間，場所における状況は，一般に，実世界に設置されたセンサにより観測したデータの利用により捉えられる．例えば，道路で発生する渋滞や速度違反は，道路に設置した監視カメラやレーダを用いた，走行している車の台数やスピードの計測，定常状態から逸脱した異常状態の検知により検出される．また，地震は，全国各地に設置された地震計を用いた地面の揺れの大きさの計測により検出される．しかし，多様な実世界観測情報を多くの場所で検出するためには，観測したい情報に応じたセンサを各地に設置する必要があり，多大なコストが発生する．

そこで近年，人間が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を通して，各自が実世界を観測して得たさまざまな情報を，観測時間，場所と共に公開することに着目し，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．人間は実世界の多くの場所に存在し，五感を用いて感知したデータの意味を解釈する機能を持つため，このように人間をセンサ（Citizen Sensor）[1] とみなして利用することは，センサ設置のコストを抑えた上で，多様な情報を取得できるという利点がある．中でも，全世界で 2 億人以上のアクティブユーザを抱える Twitter [2] では，主な投稿形式がツイートと呼ばれる 140 文字以内の短いテキストであり，その手軽さから，テレビ番組やスポーツ観戦，災害情報，渋滞情報などさまざまな内容のリアルタイム性の高い実世界観測情報が大量に投稿されている．

Twitter を用いた実世界観測情報獲得に関する既存研究は，観測対象の関連語を用いて，観測情報を含むツイートを抽出する手法が主となっている．例えば，Sakaki ら [3] は，観測対象を地震に限定した上で，地震の観測時に用いられる単語を関連語として予め人手で設定し，関連語を含むツイートの追跡により震源地を推定した．また，土屋ら [6] は，観測対象を鉄道に限定した上で，予め準備した鉄道の運行トラブルの観測情報とな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情報を抽出した．これらの手法では，関連語や学習用データを人手で与える必要があるのに対し，ユーザが与えるクエリを観測対象と見なし，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が提案されている．Massoudi ら [4] や藤木ら [5] は，観測対象に何か特徴的な事象が発生した場合，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短期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，渋滞が発生した場所名が関連語となり，各地の渋滞の観測情報が抽出できると考えられる．

本研究では，これらの観測情報に加え，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，観測対象同士は関連しているため，渋滞と事故という単語のように，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，長期的，及び短期的という二つの側面での単語間の関連性に着目し，クエリに対する単語の長期的関連性に基づき観測対象を拡げ，クエリ，及びクエリと長期的関連性の高い単語に対する短期的関連性に基づき各観測対象に対する特徴的な事象を抽出する．この実現のため，本研究では，短期間のツイート集合から逐次的に単語間の長期的，短期的関連性を学習する方法，及び学習した単語間の関連性に基づき，各ツイートのクエリへの関連度を算出する方法を提案する．

2. 提 案 手 法

提案手法では，ユーザから与えられるクエリ q で表される観測対象を Q とし，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 及び Q に関連する対象 $Q'_n(n=1,\cdots)$ の観測情報を含むツイートを抽出することを目的とする．ただし観測対象 Q，Q'_n はそれぞれ単語 q，q'_n で表されるものとする．

ここでまず一つ目の問題として，Q'_n の観測情報の多くは，クエリとして与えられた単語 q を含まないことが挙げられる．例として，クエリとして渋滞が与えられたとき，渋滞と関連する事故の観測情報の多くが渋滞という単語を含まない．ただし Q と Q'_n は関連しているため，単語 q と q'_n は，頻度は高くないものの，時間によらず同一ツイート内に共起する可能性が高い．よって提案手法では，クエリ q に対して，長期的に同一ツイート内に共起する単語を q'_n として抽出することを考える．

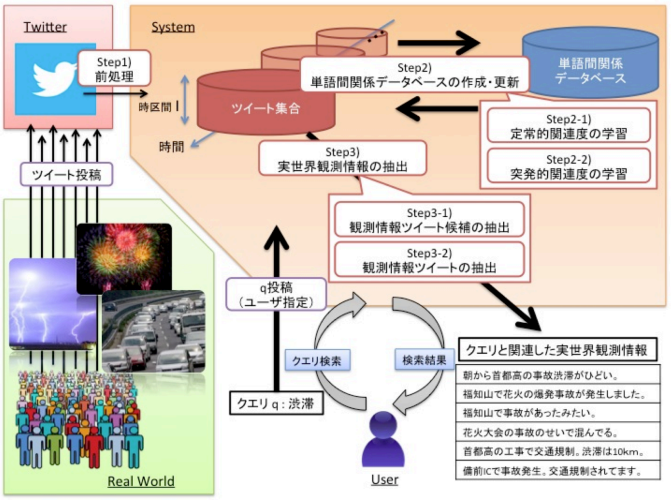

図 1　提案手法の概要

次に，単語 q 及び q'_n を用いたツイート抽出を行う場合，単語 q，q'_n を含むツイートは必ずしも Q，Q'_n の観測情報とは限らない，また，Q'_n の観測情報は必ずしも Q に関連しない，という二つの問題が考えられる．例として，渋滞，事故という単語を含むツイートは必ずしも渋滞，事故の観測情報とは限らず，また，事故の観測情報は，放送事故など，渋滞に関連しないものも多い．これらの問題に対しては，二つの解決策が考えられる．まず，これらのツイートは，事故に対する放送など，長期的関連性が q'_n とは高いが，q とは低い単語を含む可能性が高いと考えられる．また，渋滞や事故の発生した場所名など，単語 q や q'_n と共起頻度が短期的に高くなる単語を含むツイートは，Q や Q'_n との関連性が高く，逆に，q や q'_n と共起頻度の低い単語を含むツイートは，Q や Q'_n との関連性が低いと考えられる．

よって提案手法では，これら三つの問題に対応するため，単語間の長期的，短期的関連性を学習する．いずれの場合も，単語間の関係の経時変化を解析する必要があるため，短期間に投稿されたツイート集合から同一ツイート内の共起単語対を抽出し，現在までの学習結果を保持した単語間関係データベースの更新による逐次学習を行う．ユーザからクエリ q が与えられると，直近の時区間に投稿された各ツイートに対して，学習した単語間の長期的，短期的関連性に基づきクエリとの関連度を算出し，関連度の高いものをユーザに提示する．

提案手法は図 1 に示すように，以下のステップにより構成される．

Step1) 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) 単語間関係データベースの作成・更新：
収集したツイートから共起単語対の抽出，及び，その短期的関連度と長期的関連度の算出を行い，データベースを作成・更新する．

Step2-1) 短期的関連度の学習：
収集したツイートから抽出された共起単語対に対し，共起頻度の高い単語対に対して高く，低い単語対に対して低くなるよう設定した短期的関連度を算出する．

Step2-2) 長期的関連度の学習：
現在までに構築した単語間関係データベース，及び収集したツイートから抽出された共起単語対に対し，長期間にわたって共起する単語対に対して高く，まれにしか共起しない単語対に対して低くなるよう設定した長期的関連度を算出する．

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，単語間関係データベースに基づき，クエリ q 及び q に関連する観測対象 Q 及び $Q'_n(n=1,\cdots)$ の観測情報となるツイートを抽出する．

Step3-1) 観測情報ツイート候補の抽出：
直近の時区間に投稿された各ツイートに対し，単語間の長期的関連度に基づき Q との関連度を算出し，関連度の高いもののみを Q 及び Q'_n の観測情報となるツイートの候補として抽出する．

Step3-2) 観測情報ツイートの抽出：
抽出された各ツイート候補に対し，単語間の短期的関連度に基づき Q 及び Q'_n との関連度を算出し，関連度の高いもののみをユーザに提示する．

各ステップの詳細を次節以降で述べる．

2.1 前 処 理

時区間長 I ごとに Twitter に投稿された全ツイートを収集する．ただし，Twitter には，あるユーザが投稿したツイートを別のユーザがそのまま再投稿する際に用いられるリツイートという機能が存在する．リツイートを行うと，そのツイートがリツイートであることを示す「RT」という記号が，リツイート元のテキストデータに付与された状態で Twitter 上に投稿される．また，Twitter には同一のツイートを大量に投稿するスパムツイートが存在する．単語間の共起頻度はこれらのツイートによって大きく影響されるため，「RT」が付与されたリツイートや，重複するツイートを予め除去した上で，Twitter の Streaming API を用いて，ツイートを収集する．

また提案手法では，分析対象の単語を名詞のみに限定するため，収集したツイートに対して MeCab [7] による形態素解析を行い，名詞のみを抽出する．ただし，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名詞は不要な単語として除去する．

2.2 単語間関係データベースの作成・更新

収集したツイートを用いて，単語対 (w_i, w_j)，短期的関連度 $B(w_i, w_j)$，長期的関連度 $S(w_i, w_j)$ からなる単語間関係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．

2.2.1 短期的関連度の学習

収集したツイート中の共起単語対 (w_i, w_j) に対し，以下のように w_i と w_j の相互情報量を短期的関連度 $B(w_i, w_j)$ として算出する．相互情報量とは 2 つの確率変数の相互依存の尺度を表す量であり，単語間の相互情報量が高いほど関連が高いことを示す．

$$B(w_i, w_j) = \log \frac{P(w_i, w_j)}{P(w_i) \times P(w_j)} \tag{1}$$

$$P(w_i) = \frac{|\, T_i \,|}{N} \tag{2}$$

$$P(w_j) = \frac{|\, T_j \,|}{N} \tag{3}$$

$$P(w_i, w_j) = \frac{|\, T_i \cap T_j \,|}{N} \tag{4}$$

ただし，単語 w_i の出現確率を $P(w_i)$，単語 w_i，w_j の共起確率を $P(w_i, w_j)$，単語 w_i を含むツイート集合を T_i，T_i のツイート数を $|\, T_i \,|$，全ツイート数を N とする．

2.2.2 長期的関連度の学習

収集したツイート中の各共起単語対 (w_i, w_j) に対し，単語間の連続的な関連性を表す指標として，以下のように長期的関連度 $S(w_i, w_j)$ を算出する．ただし，単語間の相互情報量として算出した短期的関連度 $B(w_i, w_j)$ が負の場合，w_i と w_j は関連がない．また，「フォロー」や「笑」のような出現確率が高い一般的な単語を含む単語対の相互情報量は低くなるため，以下では短期的関連度 $B(w_i, w_j) \geqq \beta$ を満たす単語対のみを考慮する．

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，含まれている場合，現在の長期的関連度を $S^x(w_i, w_j)$ とし，$S^{x+1}(w_i, w_j)$ を以下のように更新する．

$$S^{x+1}(w_i, w_j) = c_1 \times S^x(w_i, w_j)^{d_1} \tag{5}$$

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．

$$S^{x+1}(w_i, w_j) = c_2 \times S^x(w_i, w_j)^{d_2} \tag{6}$$

ただし，$d_2 > 1$ とする．

式 (5)，(6) を図 2 に示す．$d_1 < 1$ とすることにより，式 (5) は連続して共起する単語対に対して，長期的関連度を上昇させる．特に，最初の数回で関連度が上昇し，連続する程に上昇の

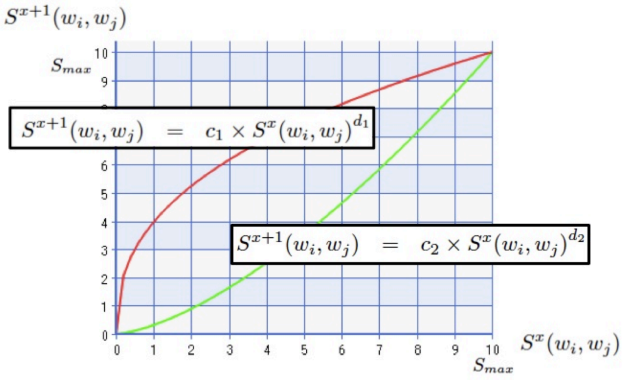

図 2 長期的関連度の更新

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 (6) は共起しない期間が連続する単語対に対して，長期的関連度を下降させる．特に，スコアが大きい場合は下降の度合いが大きく，連続するほどに下降の度合いが小さくなり 0 に収束する．式 (5) 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，単語対の長期的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．

$$c_1 = S_{max}^{(1-d_1)} \tag{7}$$

$$c_2 = S_{max}^{(1-d_2)} \tag{8}$$

最後に，長期的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，q で表される観測対象を Q とし，Q と関連のある観測対象を $Q'_n(n = 1, \cdots)$ としたとき，Q 及び Q'_n に関する観測情報であるツイート t を抽出する．まず，各ツイート t に対し，単語間の長期的関連度を用いて，観測対象 Q 及び Q'_n との関連度を算出し，関連度が高いものを観測情報ツイートの候補として抽出する．次に，各ツイート候補について，短期的関連度を用いて，観測対象 Q，Q'_n との関連度を算出し，関連度が高いものを観測情報ツイートとして抽出する．

2.3.1 観測情報ツイート候補の抽出

観測対象 Q と Q'_n は関連があるため，単語 q'_n は，クエリ q と日常的に同一ツイートに共起し，q との長期的関連度が高いと考えられる．よって，クエリ q 及び q と長期的関連度の高い単語 q'_n を含むツイートを，Q と Q'_n の観測情報ツイート候補として抽出する．ただし，単語 q'_n を含むツイートの中には，Q'_n の観測情報でないものや，Q と関連しない観測情報が含まれる．これらのツイートは，q'_n とは長期的関連度が高いが，q とは長期的関連度の低い単語を含む可能性が高い．

図 3 に例を示す．渋滞という単語は，事故や交通，高速などと長期間にわたって共起する．つまり，クエリ q を渋滞としたとき，q'_n となる事故や交通，高速などを含むツイートは，渋滞と関連度が高いと考え，これらを抽出する．しかし，例えば事故を含むツイートは必ずしも事故の観測情報とは限らず，また渋滞と関連しない観測情報である場合もある．このようなツイートは，図 3 に示すように，渋滞と長期的関連度が低く，事故と長期的関連度が高い放送や遅延などの単語を含む可能性が高い．

そこで，ツイート t に含まれる単語 w の長期的関連度を用いて，ツイート t の観測対象 Q への関連度 $Score_Q(t)$ を以下のように算出する．

$$Score_Q(t) = \sum_{\substack{S(q,q') \geqq \gamma \\ q' \in t}} S(q, q') - \sum_{\substack{S(q',w) \geqq \gamma \\ S(q,q') \geqq \gamma \\ q', w \in t \\ w \neq q'}} S(q', w) \tag{9}$$

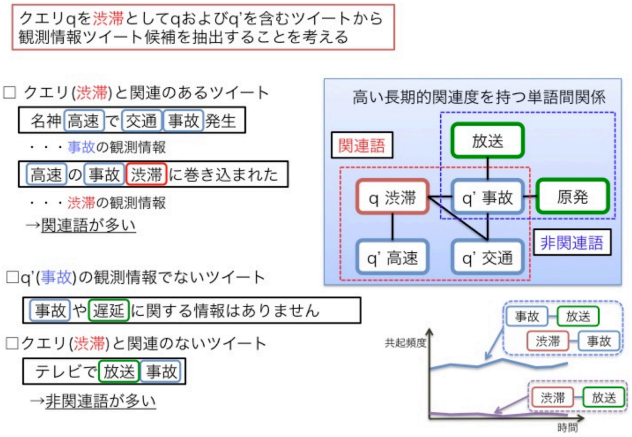

図 3　観測情報ツイート候補の抽出例

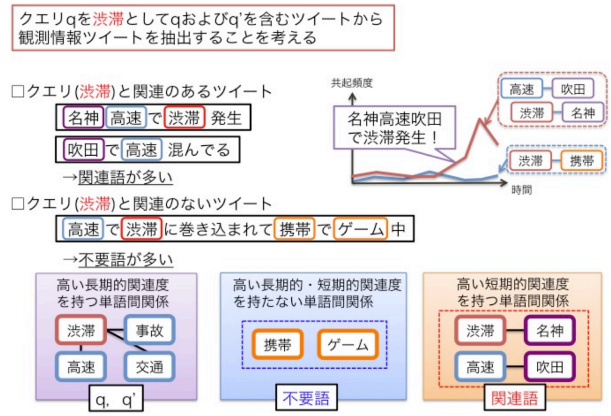

図 4　観測情報ツイートの抽出例

第一項では，クエリ q に対する長期的関連度が高い単語 q' を q の関連語とみなし，$S(q, q')$ を加算する．第二項では，q' に対する長期的関連度が高く，かつクエリ q に対する長期的関連度が低い単語を q の非関連語 w とみなし，$S(q', w)$ を減算する．

以上の処理を行った結果，$Score_Q(t) < \gamma$ となったツイートは，関連語より非関連語を多く含むため，クエリ q と関連が低いと考えられる．よって $Score_Q(t) \geqq \gamma$ を満たすツイート t を，観測対象ツイート候補とする．

2.3.2 観測情報ツイートの抽出

観測対象 Q，Q'_n に関する特徴的な事象の観測情報となるツイートは，単語 q，q'_n と短期的に共起頻度が高くなる単語を含む可能性が高いと考えられる．

例えば，図 4 に示すように，渋滞が発生した際，その観測情報を複数のユーザが投稿することにより，発生地を表す東名や京都という単語と渋滞の短期的関連度が高くなる．一方，携帯やゲームといった単語は，偶然同一ツイート内で，渋滞と共起したと考えられる．

そこで，ツイート t に含まれる長期的関連度が低い単語 w の短期的関連度に基づき，ツイート t のクエリ q 及び q'_n への関連度 $\hat{Score}_{Q,Q'_n}(t)$ を以下のように設定する．

表 1　q を渋滞としたときの q'_n

11/13			
事故	道路	交通	高速

11/20			
事故	道路	交通	高速

11/27			
事故	道路	交通	高速

表 2　q を清水寺としたときの q'_n

11/13
-

11/20
京都

11/27	
京都	紅葉

表 3　q を遅延としたときの q'_n

11/13							
列車	本線	発生	電車	伝言板	点検	鉄道	提供
遅れ	人身	証明	事故	関東	影響	運行	

11/20							
列車	本線	発生	電車	伝言板	点検	鉄道	提供
遅れ	人身	証明	事故	関東	影響	運行	運休

11/27							
列車	本線	発生	電車	伝言板	点検	鉄道	提供
遅れ	人身	証明	京王	関東	運行	運休	

$$\hat{Score}_{Q,Q'_n}(t) = Score_Q(t) + max_{\substack{w\in t \\ w\in R' \\ q_m\in R}} B(q_m, w) - \alpha S \tag{10}$$

ただし，R は q 及び q'_n の集合，R' は $Score_Q(t)$ 算出時に用いなかった単語の集合であり，S は $B(w, q_m) < \delta$ を満たすような観測対象と関連しない単語 w と，共起回数が 1 回の単語 w の数とする．つまり，2.3.1 節で算出された関連語のいずれかに対し短期的関連度が高い単語 q_m が t に含まれる場合，該当する関連語が表す観測対象の特徴的な事象を表す単語とみなし，さらに関連度を加算する．一方で，短期的関連度が低い単語が t に含まれる場合，これを不要語とみなし，関連度を減算する．

最後に，算出された $\hat{Score}_{Q,Q'_n}(t) \geqq \gamma$ を満たすツイートをユーザに提示する．

3. 評 価 実 験

提案手法の有効性を以下の二つの観点から検証する．

(1)　長期的関連度と短期的関連度の有用性

(2)　単語間関係データベースの更新の必要性

Twitter の StreamingAPI を用いて，2013/10/25〜11/27 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 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 $S_{max} = 10$，$d_1 = 0.4$，$d_2 = 1.5$，$\gamma = 9.0$，$\beta = 4.0$，$\alpha = 4.0$，$\delta = 1.0$ とした．

3.1　長期的関連度と短期的関連度の有用性

まず，長期的関連度に基づく関連語がどのように変化するのかを確認するために，11/13，20，27 の一週間ごとの日付において，クエリ q との長期的関連度 $S(q, q'_n) \geqq \gamma = 9.0$ を満たす，つまり長期的関連度に基づき q の関連語と判定された単語 q'_n を表 1，2，3 にそれぞれ示す．また表 4 に，クエリ q を含むツイート数，q もしくは q'_n を含むツイート数，抽出した観測情報ツイート候補数を示す．表 1，2，3 において下線を引いた単語は，3 日間で共通して抽出された単語 q'_n である．q を変化させたとき，それぞれ渋滞，遅延，清水寺から一般的に連想しうる単語が，13 日，20 日，27 日の間で，大きく変化することなく抽出されていることが分かる．q が遅延の場合，q'_n に遅延の発生場所である路線名や都市名が確認できる．これは長期に渡り何度も遅延がその場所で観測されることを示している．一方で，q が渋滞の場合，q'_n には渋滞の発生場所である地名などは確認できず，関連語に変化がない．これは，渋滞が特定の場所で長期にわたって観測される対象ではないことを示している．q が清水寺の場合，13 日においては q'_n に該当する単語はなかったが，20 日に京都，さらに 27 日に紅葉が q'_n となった．これは，紅葉の時期に近づくにつれて，紅葉の名所として知られる清水寺を訪れる観光客が増加し，紅葉に関する実世界観測情報が多く投稿されたことに起因すると考えられる．このことは，表 4 で示すように，清水寺を含むツイート数が増加していることからも読み取れる．

表 4　11/13〜27 における q ごとの観測情報ツイート候補数の比較

検索日時	q	q を含むツイート数	q，q'_n を含むツイート数	観測情報ツイート候補数
11/13	渋滞	37	725	198
11/20	渋滞	41	764	286
11/27	渋滞	38	789	214
11/13	遅延	54	3,211	1,095
11/20	遅延	66	3,299	967
11/27	遅延	84	3,460	1,197
11/13	清水寺	7	7	7
11/20	清水寺	14	530	140
11/27	清水寺	19	803	256

表 4 より，各クエリ q を含むツイート数は，その日に収集された全ツイート数の 0.01 %程度となっている．q'_n を含むツイートも合わせると，大幅にツイート数が増加する．これは q'_n に一般的な単語が多く含まれるためである．例えば，クエリ q が遅延の場合，q'_n には電車や関東など，多くの人が日常的に使用する単語が含まれ，それらの単語を含むツイート数は膨大になる．これらのツイートには，クエリと関連する観測情報ツイートも含まれているが，9 割以上はクエリと無関係なツイートである．

次に，クエリ q に応じた観測情報が抽出可能かを確認するため，同じ検索日時において，異なるクエリ q を与え，観測情報ツイートを抽出した．ただし，クエリ q で表される観測対象に特徴的な事象が発生した日を検索日時に設定した．11/23 において，クエリ q を渋滞，遅延としたとき，正しく抽出された観測情報ツイートと除外されたツイートの一部を表 5，6 に示す．ここで，$Score_Q(t)$，$\hat{Score}_{Q,Q'_n}(t)$ の算出時に加算された単語を太字で，減算された単語を下線で示している．また，rank は各ツイートの順位である．

1-4，1-5，1-6，2-1，2-4 のように，クエリ q を含む観測情報

表 5　11/23 における q を渋滞としたときの観測情報ツイート

ID	rank	$Score_Q(t)$	$\hat{Score}_{Q,Q'_n}(t)$	ツイート本文
1-1	1	29.99	36.28	京葉道路で事故か。オートバイの高速事故ってほぼ命にかかわるから恐い
1-2	2	19.99	27.49	京葉道路で追突事故 5 人死傷 http://t.co/M0kCgTUYto ニュース
1-3	3	19.99	27.49	【N】京葉道路で追突事故 5 人死傷 http://t.co/SqZhHTA9NO
1-4	7	20	27.19	うげー東名下りで海老名から 30km も事故渋滞してるこれ解消されるまで待機
1-5	8	19.99	27.18	海老名 SA なう。ここまでの東名高速は大渋滞ちう。
1-6	9	19.99	27.18	東名高速渋滞ひどいですよ
1-7	圏外	9.98	-18.01	交通費も全て会社持ちの研修！研修内容 は、地獄 巡りと全国の 支社 の方との 交流www バタバタ になるかなぁとは思ったけど、研修 参加 希望 しちゃった〜楽しんできまーす
1-8	圏外	9.99	-14.00	道路に写し出された 自分 の 影 が、コート のせいもあって ハンコック に見えるわ。わら わが美しいからじゃ！って 遊び やって仕事から帰ってる
1-9	圏外	-29.94		11/2314:18 ＃東京メトロ副都心線＃ Kanto13:44 頃、東急東横線内で 発生 した 人身 事故の影響で、一部 列車 に 運転 変更が出ています。(14:16)
1-10	圏外	-39.79		運行 情報に関するお知らせです。16 時 43 分現在■東京メトロ副都心線・一部列車 遅延 原因：人身事故 発生 時刻：13 時 44 分頃 発生 場所：東急東横線反町駅
1-11	圏外	-49.86		unkokanto 東急東横線【列車遅延】13:44 頃、反町駅で 発生 した 人身 事故の影響で、現在も 列車 に 遅れ が出ています。(11/2315:45) ＃駅の伝言板＃東京都 運行 速報
1-12	圏外	-49.76		unkokanto 東急東横線【平常 運転】反町駅で 発生 した 人身 事故の影響で、列車 に 遅れ が出ていましたが、16:45 現在、ほぼ平常通り 運転 しています。(11/2316:45) ＃駅の伝言板＃茨城県 運行 速報

表 6　11/23 における q を遅延としたときの観測情報ツイート

ID	rank	$Score_Q(t)$	$\hat{Score}_{Q,Q'_n}(t)$	ツイート本文
2-1	1	89.93	97.35	unkokanto 東急東横線【列車遅延】13:44 頃、反町駅で発生した人身事故の影響で、現在も列車に遅れが出ています。(11/2315:45) ＃駅の伝言板＃東京都運行速報
2-2	2	79.43	85.22	[関東]2013/11/2310:00 ＃京王高尾線 05:45 頃、京王線内で発生した人身事故の影響で、現在も一部列車に遅れや運休が出ています。
2-3	4	60.26	67.67	unkokanto 東急東横線【平常 運転】反町駅で発生した人身事故の影響で、列車に遅れが出ていましたが、16:45 現在、ほぼ 平常 通り運転しています。(11/2316:45) ＃駅の伝言板＃ 茨城 県運行速報
2-4	11	49.94	57.35	11/2314:18 ＃東京メトロ副都心線＃ Kanto13:44 頃、東急東横線内で発生した人身事故の影響で、一部列車に運転変更が出ています。(14:16)
2-5	12	49.94	51.33	運行 情報 に関する お知らせ です。16 時 43 分現在■東京メトロ副都心線・一部列車遅延原因：人身事故発生時刻：13 時 44 分頃発生 場所：東急東横線反町駅
2-6	89	19.94	25.67	京王線人身事故で各停しかこない。。バイト間に合うかな (´ω ‘) 小田急にすればよかった（泣）
2-7	90	19.94	25.67	クソ京王線は朝の 5 時から人身事故なんですね
2-8	圏外	9.99	-10.00	マック の ポテト。。塩 がまったくついてなくて味がしなかった 3 つとも。。まじないわー車の点検迫ってたでクレームも言い損ねた いらいら
2-9	圏外	9.99	-12.38	プラズマ という物質は無い。個体 や 気体 などと 一緒 で『状態』の名称だ。一般 的に プラズマ は 気体 で 放電 する事で生まれ、電磁波 の 交差 でも発生する。
2-10	圏外	-9.97		まもなく 上野 駅から最終列車成田空港行が 発車 します
2-11	圏外	-9.99		辻堂 通過 列車のためご 注意 下さい
2-12	圏外	-68.69		京とれいん bot;阪急梅田 駅 3 号線から 13:52 発 快速 特急・京都河原町行き (130 列車)、まもなく 発車 いたします。停車駅 は十三・淡路・桂・烏丸です。京都河原町 終点 には 14:35 に 到着 いたします。

ツイートが多くの関連語により正しく抽出された．例えば渋滞が発生した際には，1-4，1-5，1-6 のように短期的関連度に基づき学習された東名や海老名などの関連語により，渋滞の観測情報が正しく抽出されたことが分かる．

また，1-1，1-2，1-3，2-2，2-3，2-5 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，長期的関連度に基づき学習された関連語により正しく抽出された．クエリ q が渋滞の場合の結果を着目すると，11/23 において京葉道路でオートバイの追突事故が発生した影響で，事故と京葉道路という単語の短期的関連度が上昇し，1-1，1-2，1-3 のようなクエリ q を含まないツイートに対し $\hat{Score}_{Q,Q'_n}(t)$ が上昇した．クエリ q が遅延の場合も同様に，京王線と東急東横線で発生した人身事故の影響で，事故と東急東横線や反町という単語の短期的関連度が上昇していることも確認できる．

表 7　11/18 における q を清水寺としたときの観測情報ツイート

ID	rank	$Score_Q(t)$	$\hat{Score}_{Q,Q'_n}(t)$	ツイート本文
3-1	1	19.37	28.44	京都行きたい?。ライトアップされた清水寺見たい?。
3-2	2	10	19.06	清水寺行ってきやしたー?ライトアップ綺麗やた?http://t.co/9bwfEYoeXO
3-3	3	10	19.06	清水寺のライトアップきてもーたがな (*j) めっちゃきれい !!!
3-4	6	10	15.98	あの清水寺でも今日はそんなに混んでなかった。やはり観光するなら平日がオススメ！
3-5	8	9.37	15.79	京都橘と藤枝東が 2 回戦で。全国高校サッカー選手権大会の組み合わせ抽選。
3-6	9	9.37	15.79	藤枝東 vs 京都橘って (´ω ‘) 好カードすぎるだろ。一回戦で消えて欲しくないカード
3-7	10	9.37	15.79	藤枝東と京都橘楽しみすぎる
3-8	11	9.37	15.79	全国高校選手権組み合わせが決定!初戦から昨年準 V 京都橘 vs 藤枝東!!— ゲキサカ [講談社]http://t.co/JrZlMggrgO@gekisaka さんから
3-9	12	9.37	15.79	うぉぅ、なんてこった！初戦が藤枝東とは（＾＿＾;）RT @ SDr1207 : 一回戦京都橘か。がんばれまじがんばれ藤枝東！
3-10	13	9.37	15.79	一回戦から京都橘 VS 藤枝東みれるとかやばい !! めっちゃおもろなりそうやん !!!

表 8　11/23 における q を清水寺としたときの観測情報ツイート

ID	rank	$Score_Q(t)$	$\hat{Score}_{Q,Q'_n}(t)$	ツイート本文
4-1	1	19.31	26.26	昨日は紅葉狩りをしに京都へ行ってきました！昼間の紅葉もライトアップされた紅葉も綺麗でした！http://t.co/uj7w6b64Ek
4-2	2	19.65	24.74	なにそのオシャレ RT@smokycatdeq : 清水寺から八坂神社まで、紅葉を見ながらツイートをしまくる。そういう 仕事。
4-3	3	19.31	20.87	浄瑠璃 寺の紅葉まっさかり@京都府木津川市 http://t.co/ECezQCXpwN
4-4	4	19.31	20.46	山科の 毘沙門 さん　その４　＃京都　＃紅葉 http://t.co/Y5wJ9Lcv2V
4-5	5	19.65	19.65	@FutaH06 京都駅に売ってたよ清水寺と京都駅でほとんどのお土産買えるからまとめて買うとよい
4-6	6	19.65	19.65	やっぱり清水寺の紅葉はきれい！http://t.co/DkJXWnclHT
4-7	11	19.31	19.31	@miwa0124 京都は紅葉真っ盛りです !!
4-8	21	9.65	16.79	銀閣寺！紅葉綺麗 !! http://t.co/blb07T6ozi
4-9	32	9.65	15.33	@okeihannet 京都・東福寺行きたい !!!
4-10	33	9.65	15.32	今日は紅葉日和やね@東福寺通天橋 http://t.co/p2OQWBdDlw

一方，1-7，1-8，1-9，1-10，1-11，1-12，2-8，2-9，2-10，2-11，2-12 のように，クエリ q もしくは q'_n を含む観測情報ではないツイートは，非関連語により正しく除外された．1-7，1-8，2-8，2-9 は短期的関連度に基づく非関連語を多く含むため，正しく除外された．また，1-9，1-10，1-11，1-12，2-10，2-11，2-12 は長期的関連度に基づく非関連語を多く含むため，正しく除外された．例えば，1-9，1-10，1-11，1-12 は，渋滞とは関連の低い電車の事故に関連するツイートであり，渋滞と関連度の高い事故という単語を含む一方で，事故と関連の高い人身や遅延といった単語を含むため除外された．これらのツイートは，クエリ q が遅延の場合には，遅延と人身という単語の長期的関連度が高いため，2-1，2-3，2-4，2-5 のように正しく抽出されており，渋滞と遅延は，どちらも q'_n として事故を含むが，ツイートに含まれるそれ以外の長期的関連度，短期的関連度に基づき決定されるクエリ q との関連語，非関連語により，クエリ q に応じた観測情報ツイートが抽出されることが確認できる．

以上の結果より，長期的関連度および短期的関連度に基づき決定されるクエリ q の関連語，非関連語により，ツイートのクエリ q に対する関連度 $Score_Q(t)$，$\hat{Score}_{Q,Q'_n}(t)$ が適切に上昇，下降することが分かる．よって，クエリ q に応じた実世界観測情報の抽出における，長期的関連度および短期的関連度の有用性が示された．

3.2　単語間関係データベースの更新の有用性

ユーザがクエリを与えたときの状況に応じた観測情報が抽出可能かを確認するため，同じクエリ q に対し，異なる検索日時において観測情報ツイートを抽出した．11/18 と 11/23 において，クエリ q を清水寺としたとき，正しく抽出された観測情報ツイートと除外されたツイートの一部を表 7，8 に示す．

18 日の時点で q'_n は京都という単語のみであったが，23 日の時点では京都と紅葉という二つの単語が q'_n として抽出された．18 日には，3-5，3-6，3-7，3-8，3-9，3-10 のように，その日に発表された京都の高校サッカー選手権の観測情報ツイートが多く抽出された．これらは清水寺の関連観測情報としては不適切と考えられるが，清水寺と長期的関連度の強い関連語として京都が抽出されたため，京都における特徴的な事象に関する観測情報として抽出された．しかし，23 日には，紅葉のシーズンを迎えた京都に関するツイートが増加したことに伴い，清水寺と紅葉の長期的関連度が高くなり，4-1，4-2，4-3，4-4，4-6，4-7，4-8，4-10 のように，清水寺周辺の紅葉の観測情報ツイートが多く抽出された．このように，クエリを与えたときの状況

表 9　評価実験結果

クエリ q	日時	P@10	P@30	P@50	AP@10	AP@30	AP@50
渋滞	11/16	0.60	*0.45*	*0.40*	0.67	0.59	0.51
渋滞	11/23	0.70	0.50	0.56	**0.93**	0.84	0.62
遅延	11/19	**0.90**	0.70	0.74	**0.99**	**0.95**	0.79
遅延	11/26	**1.00**	**1.00**	**0.98**	**1.00**	**1.00**	**0.99**
清水寺	11/18	*0.30*	*0.17*	*0.16*	0.56	0.41	*0.34*
清水寺	11/23	0.60	0.57	0.56	**0.90**	**0.70**	0.65

に応じた関連語を抽出するためには，単語間関係データベースを更新する必要であることが分かる．

3.3　抽出ツイートの評価

提案手法により抽出したツイートが，クエリ q と関連する観測情報ツイートとして適切であるかを主観評価で確認した．異なる日付，異なるクエリ q を用いて観測情報の抽出を行った後，三人の被験者に，抽出したツイートのうち $\hat{Score}_{Q,Q'_n}(t)$ の上位 10，30，50 件のツイートに対して，クエリ q の観測情報ツイートとして適切であるかを判断させ，2 人以上の被験者が適切と判断したツイートを正解ツイートとした．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正解ツイートの割合を，平均適合率は，正解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．

$$P@N = \frac{R}{N} \tag{11}$$

$$AP@N = \frac{1}{N}\sum_{k=1}^{N}\left(P@k \times rel\left(k\right)\right) \tag{12}$$

ただし，R は上位 N 件中の正解ツイート数であり，$rel\left(k\right)$ は上位 k 件目のツイートが正解ツイートなら 1，正解ツイートでないなら 0 とする．

表 9 に結果を示す．クエリ q が遅延の場合，Twitter に投稿される観測情報が多いため，抽出するツイート数を増やしても適合率，平均適合率が共に高い結果が得られた．一方，クエリ q が渋滞の場合，主な観測者が運転手であるため，Twitter に投稿される観測情報が少なく，抽出するツイートの数を増やすごとに適合率に低下した．しかしこの場合でも，平均適合率は高く，少ない観測情報ツイートが上位に抽出されていることが分かる．また，クエリ q が清水寺の場合，18 日に比べて，23 日では適合率，平均適合率が共に高い結果となった．これは，18 日が月曜日なのに対し，23 日は紅葉のシーズンを迎えた土曜日であり，多くの観光客が，清水寺とその周辺において紅葉に関する観測情報を Twitter に投稿したためと考えられる．

このように，抽出精度はクエリ q に依存するものの，多くの観測情報が Twitter に投稿される観測対象がクエリ q である場合，高い適合率で観測情報の抽出が可能であり，投稿される観測情報が少ないクエリ q である場合でも，抽出したツイートの上位に多くの観測情報が抽出された．

4. ま と め

本研究では，Twitter に投稿されるツイートから，ユーザが与えたクエリと関連のある実世界観測情報を抽出するため，実世界の状況に応じて経時変化する長期的・短期的関連性という二つの側面を持つ単語間関係を，短期間に投稿されたツイート集合から逐次的に学習し，各ツイートに対し，現在の状況に応じたクエリとの関連度を算出する手法を提案した．2013 年の 32 日間に投稿されたツイートに対し，異なるクエリを用いて提案手法を適応し，長期的関連性により，ユーザから与えられたクエリと関連する対象にまで観測対象が拡大され，クエリを含まない関連情報が抽出されること，および，短期的関連性により，観測対象に特徴的な事象が発生した際，それらの観測情報が抽出されることを確認した．加えて，異なる時間において同一のクエリを用いて提案手法を適用し，単語間関連性の更新により，現在の状況に応じた実世界観測情報が抽出されることを確認した．今後の課題として，関連度算出に用いるパラメータの適切な設定方法が挙げられる．

文　献

[1] Amit Sheth, “Citizen Sensing, Social Signals, and Enriching Human Experience”, IEEE Internet Computing, Volume 38,Issue 4, pp 87-92, 2009.

[2] “Twitter”, http://twitter.com/

[3] Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo, “Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors”, Nroceedings of the 19th International Conference on World Wide Web(WWW2010), pp. 851-860, 2010.

[4] K.Massoudi, M.Tsagkias, M.D.Dijke and W.Weerkamp ,“Incorporating query expansion and quality indicators in searching microblog posts”, In Nro. of the 33rd European Conference on Advances in Information Retrieval(ECIR), 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 2013 C9-5,2013.

[6] 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”, 情報学基礎研究会報告 2013-IFAT-111(31), PP. 1-6, 2013.

[7] “mecab Japanese morphological analyzer”, https://code.google.com/p/mecab/