マイクロブログを用いたリアルタイム地域情報の推薦

坂本　宏祐† Lim Jeongwoo† 新田　直子† 中村　和晃† 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: †{sakamoto,jeongwoolim}@nanase.comm.eng.osaka-u.ac.jp,

††{naoko,k-nakamura,babaguchi}@comm.eng.osaka-u.ac.jp

あらまし　近年モバイル端末を用いて，ユーザの現在の位置に応じた地域情報を推薦するサービスが実現されている．しかし，推薦される情報は必ずしも現在の状況に即したものではない．そこで本研究では，Twitter に代表されるマイクロブログには不特定多数のユーザから多くのリアルタイムな情報が投稿されることに着目し，マイクロブログを用いてリアルタイムな地域情報を推薦する手法を提案する．提案手法では，マイクロブログへの緯度経度付き投稿から，局所的に用いられる単語を，地域情報を表すローカル語として逐次抽出する．抽出された地域情報のトピック及びユーザの嗜好は，各ローカル語を含む投稿及びユーザの過去の投稿が表すと考え，これらの投稿に含まれる単語の意味的な類似性に基づき，ユーザの嗜好に応じた地域情報を推薦する．

キーワード　情報推薦，リアルタイム処理，マイクロブログ，地域情報

1. はじめに

近年，モバイル端末によりユーザの現在位置を取得し，それに応じて周辺の地域情報を提供する様々な位置情報サービスが存在する．地域情報とは，イベント，施設，特産品など特定の位置で観測される対象であり，例えば Google Maps [1] では，ユーザは任意のキーワードを入力することにより，キーワードに関連する周辺の地域情報を検索できる．この時，他ユーザによる地域情報に関する口コミや評価などが提示され，ユーザはそれらを参考に自身の嗜好に合致する情報を選択する．しかし，提示される情報は過去に投稿されたものが多く，必ずしも現在の状況に即したものではない．よって現状の位置情報サービスでは，突発的なイベントなどの地域情報の取得は困難である．

一方で，Twitter [7] などのマイクロブログに不特定多数のユーザからリアルタイムな情報が短いテキスト形式で投稿されることに着目し，特に投稿位置の緯度・経度を表すジオタグの付与された投稿を用いて，リアルタイムな地域情報を抽出する研究がなされている．例えば，Sasaki ら [5] は，地震や台風などの大規模災害を対象とし,「地震」や「揺れ」など特定の単語を含むジオタグ付き投稿の位置を追跡することにより，リアルタイムに震源地を推定した．また，Schulz ら [6] は交通事故などの小規模な事象を対象とし，単語に基づき投稿を判別する分類器を用いて新たに投稿された投稿から交通事故に関する投稿を抽出することにより，交通事故の発生現場などを特定した．

地震や交通事故など特定の対象に関連した地域情報を抽出する研究が行われている一方で，任意の対象に関する地域情報を検出する研究もある．Lee ら [8] は Twitter へのジオタグ付き投稿を用い，特定の空間領域における投稿数やユーザ数の急激な変化に基づき，夏祭りや花火大会など局所的に人が集中するようなイベントの検出を行った．また，Kamimura ら [2] は，Twitter に投稿されるジオタグ付き投稿に含まれる単語の空間的局所性をリアルタイムに解析することにより，最新の地域情報を表す単語を抽出した．

また，同じユーザからの複数の投稿からユーザの嗜好も抽出できる．例えば，Zhao ら [9] は Twitter の投稿から，購買に関するキーワードを用いて投稿の集合を抽出し，それらを構成する単語に基づいて購買傾向を判定する分類器を作成することにより，ユーザの購買傾向を抽出した．また Phelan ら [10] はユーザの Twitter への投稿履歴における単語と RSS から配信されるニュースを構成する単語の共通性に基づき，ユーザの興味のあるニュースをリアルタイムに推薦した．

このように，マイクロブログは複数ユーザによる投稿と地理空間の関係に基づきリアルタイムな地域情報を抽出できるだけでなく，複数の投稿に使用される単語の類似性に基づき，ユーザの嗜好なども抽出できる有用性の高い情報源である．そこで本研究では，マイクロブログの代表である Twitter を用い，まず Twitter への不特定多数からの投稿に対する逐次的に地理空間との関係を調べることにより，空間的局所性の高い単語を最新の地域情報を表すローカル語として抽出する．さらに，各ユーザの過去の複数の投稿，及び，ローカル語を含む複数ユーザからの投稿を収集し，これらの単語の意味的な類似性に基づき，ユーザの嗜好に応じたリアルタイムな地域情報を抽出する．これにより，より現在の状況に即した地域情報を提示する位置情報サービスが実現できる．

2. 提案手法

本研究は，ユーザ U の現在地 $g^U = (lat^U, lon^U)$ 及び嗜好に応じた地域情報 I^U の推薦を目的とする．より最新の地域情報を推薦するため，提案手法ではまず，多くのユーザが自身の周辺情報をツイートと呼ばれる短文で投稿する Twitter から，多様な地域情報をリアルタイムに抽出する．ここで地域情報は，施設やイベントなどある特定の位置で観測される対象であり，複数のユーザが同じ対象に関する観測情報を投稿する際，名称などその対象を表す同じ単語を用いる場合が多いと想定される．

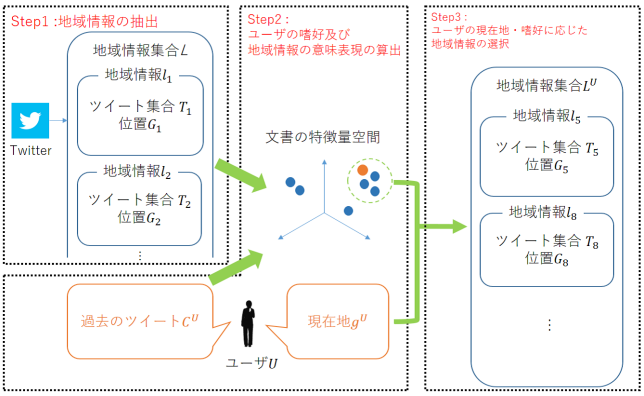

図 1　提案手法の概要

よって特定の位置でのみ複数のユーザによって使用される単語は地域情報を表す単語と考える．本研究では，このような地域情報を表す単語をローカル語と呼ぶ．

提案手法では，ジオタグ付きツイートが投稿されるごとに，ツイートに含まれる各単語の地理空間的局所性を確認することにより、常に最新の地域情報を表すローカル語を保持する地域情報データベースを構築する．次に，現在の地域情報データベースから，ユーザの現在地及び嗜好に応じて適切な地域情報を抽出する．まず，ユーザの現在地に近い位置の地域情報を表すローカル語を選択し，それぞれについて意味内容に基づくマッチングを行う．ここでは，ユーザの過去のツイート，及びローカル語を含む過去のツイートに用いられる単語の意味的な類似性を利用する．

以上を踏まえ，提案手法は，図 1 に示すように 3 つのステップにより構成される．

Step 1)　**地域情報の抽出**

ジオタグ付きツイート中に用いられる各単語の地理空間的局所性に基づき，各地の地域情報を表すローカル語 $l_k(k = 1, 2, \cdots)$ を抽出する．さらに，各ローカル語を含むジオタグ $g_{k,n} = (lat_{k,n}, lon_{k,n})$ の付与されたツイート $t_{k,n}$ を l_k により表される地域情報の観測位置，観測情報として収集する．これにより，地域情報集合 $L = \{l_k\}$，及び l_k に関する位置 $G_k = \{x_{k,n} | n = 1, \cdots, N_k\}$，ツイート集合 $T_k = \{t_{k,n} | n = 1, \cdots, N_k\}$ を含む地域情報データベースが構築される．

Step 2)　**ユーザの嗜好及び地域情報の意味表現の算出**

word2vec により学習される単語のベクトル表現は，各単語を意味的に表現できることが知られているため，以降では単語の意味表現と呼ぶ．Step1） により，ある一定時区間において収集したローカル語を含むツイート集合 $T = \{T_k | k = 1, 2, \cdots\}$ を用いて，word2vec により単語の意味表現を学習する．ユーザ U の過去のツイートに含まれる単語及び各地域情報 l_k のツイート集合 T_k に含まれる単語の意味表現に基づき，ユーザの嗜好 $\boldsymbol{f}^U$ 及び各地域情報の意味表現 $\boldsymbol{f}_k$ を算出する．

Step 3)　**ユーザの現在地・嗜好に応じた地域情報の抽出**

ユーザ U の現在地 g^U 付近の地域情報を，その意味表現 $\boldsymbol{f}_k$ とユーザの嗜好 $\boldsymbol{f}^U$ の類似度に基づきランク付けする．

次節以降で，各ステップの詳細について述べる．

2.1　地域情報の抽出

Twitter からまず，各地の地域情報を表すローカル語を抽出する．ローカル語は，ある特定の位置で観測される地域情報を表す単語であるため，空間的に限定された範囲で複数のユーザにより用いられると考えられる．よって，空間的局所性の高い単語をローカル語として抽出すればよいが，地域情報の人気度や変動性に基づき，局所性を観測できる時区間が異なる．よって，すべての単語に対して同じ時区間で局所性を判定するのではなく，単語ごとの出現履歴に応じて適切な時区間で局所性を判定することにより，各地の場所や特産品を表す語などの恒常的なローカル語と共に，イベントを表す語などの一時的なローカル語を抽出する．

具体的には，ツイートを収集している地理空間全体を，各エリアのツイート数がほぼ均等となるよう J 個のエリア $A = \{a_j | j = 1, \cdots, J\}$ に分割する．ローカル語は地名や特産品，イベントなど各地で観測される対象を表す名称が多く，主に名詞から構成されると考えられるため，ジオタグ付きツイートが投稿される度，形態素解析により品詞のタグ付けを行い，名詞のみを抽出する．ここで，例えば “Michigan” という単語が州の名であるのに対し “Michigan Stadium” は施設名であるように，ローカル語を複合名詞として抽出することにより示す場所や意味がより限定されると考えられるため，複合名詞を抽出する．

ジオタグ付きツイートが投稿されるごとに，ツイートに含まれる Z 個の名詞 $u_z\,(z = 1, \cdots, Z)$ の出現履歴を更新する．u_z の局所性は，テキストマイニングの分野で，コーパス中のドキュメントごとに固有の単語を抽出するために用いられる TFIDF 法を応用し，以下の式により算出される．

$$tfidf_{u_z}^{max} = h_{u_z}^{max} \cdot idf_{u_z} \tag{1}$$

$$idf_{u_z} = \log \frac{J}{|A_{u_z}|}, where\ A_{u_z} = \{a_j \mid h_{u_z,j} \neq 0\} \tag{2}$$

ただし，$h_{u_z,j}\,(1, \cdots, J)$ はエリア j における単語 u_z の出現頻度，$h_{u_z}^{max} = \max_j h_{u_z,j}$ は全エリアの中の u_z の最大出現頻度，$|A_{u_z}|$ は u_z が出現したエリア数，J は全エリア数を表す．また，出現頻度はユーザにつき各エリアで 1 回のみカウントする．

u_z が特定のエリアに頻繁に出現したときに $tfidf_{u_z}^{max}$ は高くなるため，$tfidf_{u_z}^{max} \geqq R$ を満たしたとき u_z をローカル語として抽出する．さらに，イベントを表す語のような一時的に空間的局所性を持つローカル語は，それ以外の時区間ではどのエリアでも使用され得る．よって，長期間の出現履歴を保持していた場合，一時的な空間的局所性を検知することが困難となる．そこで $tfidf_{u_z}^{max} < r$ を満たしたとき，u_z の出現履歴を一度削除し，新たに収集し始めることにより，一時的な空間的局所性を正しく検知できるような時区間での出現履歴を保持するよう対処する．また，$tfidf_{u_z}^{max} \geqq R$ を満たして u_z をローカル語 l_k として抽出するときも，u_z のこれまでの出現履歴を削除し，新たに収集を開始する．これにより，一度ローカル語として抽出された u_z が新たな出現履歴において $tfidf_{u_z}^{max} < r$ を満た

表 1　ボットによるツイート

アカウント ID	緯度	経度	ツイート
191092262	25.77508716	-80.2732666	News AIDS : New mouse model technology could speed the search for an AIDS vaccine URL
186478529	25.77508716	-80.2732666	News AIDS : New mouse model technology could speed the search for an AIDS vaccine: Researchers at Boston URL
186478529	25.77508716	-80.2732666	News AIDS : Researchers harness antibody evolution on the path to an AIDS vaccine: A vaccine needs to elicit those URL

したとき，過去の一時的なローカル語であると判断し，ローカル語集合から削除する．

ここで，ボットと呼ばれる機械により自動的に行われるツイートが問題になり得る．ボットの中には，表 1 に示すように異なるアカウントを用いて，近接した位置から非常に類似したツイートを投稿するものがある．よって，このようなツイートに含まれる単語の空間的局所性が高くなるが，このような単語は地域情報を表すものとしては適切ではないことが多い．よって，u_z の出現履歴に u_z を含むジオタグ付きツイートを記録しておき，$tfidf_{u_z}^{max} \geqq R$，もしくは $tfidf_{u_z}^{max} < r$ を満たすとき，出現履歴に含まれるツイート間の単語の類似度を Jaccard 係数により算出する．Th_e 以上の類似度を持つすべてのツイートをボットによるツイートとして削除した後，再度 TFIDF 値に基づくローカル語判定を行う．

以上の処理により，ローカル語 l_k を抽出すると共に，ローカル語と判定された際，出現履歴に記録されたジオタグ付きツイートを保存することにより，l_k に関する位置 $G_k = \{x_{k,n} | n = 1, \cdots, N_k\}$，ツイート集合 $T_k = \{t_{k,n} | n = 1, \cdots, N_k\}$ を含む地域情報データベースが構築される．

2.2　ユーザの嗜好及び地域情報の意味表現の算出

ユーザの嗜好に応じた地域情報を推薦するため，ユーザの過去のツイート，及びローカル語を含む過去のツイートに用いられる単語の意味的な類似性を利用する．ここで，Twitter への投稿には特有な表現が多く用いられるため，単語間の意味的な類似性は Twitter への投稿自身から学習することが望ましい．特に，地域情報などに用いられる単語間の意味関係を学習するため，前節の手法により，ある一定区間において抽出したローカル語を含むツイート集合を用いる．

前処理として，$t_{k,n}$ から，地域情報 l_k の意味内容に関係が小さいと考えられる単語として，“http(s)://...” という形式の URL，ユーザ名を表す “@” で始まる単語を除去する．さらに，一般的な語の関係性を学習するため，ツイートの最後に付与されることの多い,“ @ yankee stadium ”のように,“ @ ” の後ろにスペースを入れた上で単語が続く表現も地名が記述されたものとして削除する．また，l_k 自身も削除した上で word2vec [3], [4] を適用する．word2vec は，学習用コーパスとして大量の文書集合が与えられたとき，各単語の周辺単語を推定するような 2 層からなるニューラルネットワークを用いて，同じコンテキストで用いられる単語対ほど近くなるような各単語のベクトルを学習する．ここでは，各ローカル語と共に用いられる単語対が近くなるよう，T_k を一つの文書とみなし，全てのローカル語に対する文書集合 $T = \{T_k | k = 1, 2, \cdots\}$ を学習用コーパスとして word2vec に与える．

次に，学習した単語の意味表現を用いて，ユーザの嗜好，及び各ローカル語で表される地域情報の意味表現を算出する．ユーザの嗜好はユーザの過去のツイートに含まれる単語集合，地域情報 l_k のトピックはローカル語を含むツイート集合 T_k に含まれる単語集合を用いて算出可能と考えられる．これらの単語集合の中には “baseball” のように限られた地域情報に関するツイートに出現する地域情報のトピックに関連が大きいと考えられる単語と，“go” のように多くの地域情報に関するツイートに出現し，地域情報のトピックとは関連が小さいと考えられる単語がある．そこで，同様に TFIDF 法を用いて T_k 中の単語 $u_{k,m}$ に対し，l_k に対する重要度 $w_{k,m}$ を算出する．l_k の意味表現 $\boldsymbol{f}_k$ は以下の式のように，重要度の高い，つまり l_k に特有な単語ほど，意味表現を強く反映させて算出する．

$$\boldsymbol{f}_k = \sum_{m=1}^{M_k} w'_{k,m} \cdot \boldsymbol{v}_{k,m} \tag{3}$$

$$w'_{k,m} = \frac{w_{k,m}}{\sum_{m=1}^{M_k} w_{k,m}} \tag{4}$$

$$w_{k,m} = tf_{u_{k,m}} \times log \frac{|T|}{|\{T_k | u_{k,m} \in T_k\}|} \tag{5}$$

ただし，$u_{k,m}$ はすべての $t_{k,n} \in T_k$ から上述の通り不要な語を除去した後に残った M_k 個の単語であり，$T_k = \{u_{k,m} | m = 1, \cdots, M_k, w_{k,m} \in t_{k,n}\}$，$\boldsymbol{v}_{k,m}$ は $u_{k,m}$ の意味表現，$tf_{u_{k,m}}$ は T_k 中の $u_{k,m}$ の出現頻度である．

ユーザ U の嗜好も同様に，複数のユーザの過去のツイートを収集し，各ユーザのツイート中の各単語に対して重要度を算出した上で，意味表現 $\boldsymbol{f}^U$ を算出する．

2.3　ユーザの現在地・嗜好に応じた地域情報の抽出

位置情報サービスでは一般的にユーザの現在地付近の地域情報を提示する．そこで，ユーザの現在地と空間的に近い地域情報を推薦するために，ユーザ U の現在地の緯度・経度 g^U を含むエリアで出現したローカル語 l_k を選択する．

ユーザ U の嗜好を表す意味内容表現 $\boldsymbol{f}^U$ に対し，意味内容表現 $\boldsymbol{f}_k$ とのコサイン類似度の高いものから順番にローカル語 l_k をユーザの嗜好に応じた地域情報として提示する．

3.　評 価 実 験

Twitter の Streaming API を用いてアメリカ本土を緯度が 24 度から 49 度，経度が-125 度から-66 度の範囲と設定し 2016 年 9 月 8 日から 2016 年 10 月 9 日のうちの 30 日間に投稿され

表 2　“kelly shorts stadium” を含むツイート

ツイート本文
fired up for that cmich football win @ Kelly/Shorts Stadium URL
Babe showing the boys the football facility. ?? #CMU #cmuhomecoming2016 @ Kelly/Shorts Stadium URL
It’s a beautiful day for a football game! @ Kelly/Shorts Stadium URL

表 3　“johnson space center” を含むツイート

ツイート本文
Getting spacey #joeandgraciedoamerica2016 @ NASA - Johnson Space Center in Houston, TX URL
Sooo now I want to be an astronaut. #NASA #jetpack ? @ NASA - Johnson Space Center in Houston URL
Cockpit of the Space Shuttle @ NASA - Johnson Space Center in Houston URL

た 6,655,763 件のジオタグ付きツイートを収集した．

これらを用い，まず 2.1 節の手法を用いてローカル語の抽出し，その結果について考察した後，抽出されたローカル語を含むツイートを投稿したユーザに対して，提案手法により適切なローカル語を推薦できるか実験により検証する．

3.1　ローカル語の抽出

アメリカ本土を各エリアのツイート数が均等となるよう $J = 256$ エリアに分割し，各ツイートに対して，Brill’s Tagger [12] による品詞のタグ付け，TermExtract [11] による複合語の抽出を行った．得られた全ての名詞に対して，$R = 16.63$, $r = 6.97$ として逐次的にローカル語を抽出した結果，30 日間のすべてのツイートを処理したときの地域情報データベースには 41,938 語のローカル語が存在した．ただしパラメータは, [2] の結果を踏まえて設定した．表 2 ～ 7 に，ローカル語とそれを含むツイートの例を示す．ただしツイート本文中の “URL” は “http(s)://...” という形式の URL である．表 2 ～ 4 から “kelly shorts stadium”，“johnson space center”，また “louis international airport” のようにサッカー関連，宇宙関連，空港といった明示的なトピックを持ち，特定の場所を表すローカル語が抽出できていることが分かる．また，表 5 から，テキサス州という広い空間範囲を表す “tx” というローカル語は，映画，スポーツ，空港のように様々なトピックを併せ持つことが分かる．さらに，表 6，7 から，“oklahoma state fair” や “international space orchestra” のように，祭事や音楽団によるコンサートなど，短期間に行われるイベントなどを表すローカル語も抽出されていることが分かる．このように空間的に狭い範囲を表し特定のトピックをもつものから，空間的に広い範囲を表し様々なトピックが含まれるもの，さらに時間的に狭い範囲を表すものまで，多種多様なローカル語が取得できていることが分かる．以降はこれらの抽出したローカル語を用いて実験を行う．

3.2　地域情報の推薦結果の評価

地域情報の推薦手法の評価を行うため，3.1 節で取得したローカル語の内，まず初めの 26 日間のツイート集合を学習用コーパスとして word2vec により単語の意味表現を学習した．

表 4　“louis international airport” を含むツイート

ツイート本文
Leaving St. Louis. @ Delta Terminal - Lambert-Saint Louis International Airport URL
Last leg of the trip. 2 more days then home for the weekend! (@ Lambert-St. Louis International Airport - @flystl) URL
Leaving on a jet plane...! (@ Lambert-St. Louis International Airport - @flystl in Saint Louis, MO) URL

表 5　“tx” を含むツイート

ツイート本文
Watching a scary moovie! (@ San Carlos, Texas in San Carlos, TX) URL
Early for soccer practice (for once) (@ Preston Meadow Park in Plano, TX) URL
Early morning flights! (@ Big Spring Executive Airport (KBPG) in Big Spring, TX) URL
I’m at Carmike 16 Movie Theater for When the Bough Breaks in El Paso, TX URL

表 6　“oklahoma state fair” を含むツイート

ツイート本文
Love me some fair food with my hubby!! #fairfood #statefair #okstatefair2016 @ Oklahoma State Fair URL
My first day of several working State Fair stories. @ Oklahoma State Fair URL
Today at the Fair with @VSLeadership was great!! #joinvsl @ Oklahoma State Fair URL

表 7　“international space orchestra” を含むツイート

ツイート本文
with the international space orchestra before their support set #SRhollywoodbowl #sigurroslive··· URL
The International Space Orchestra. Yes, music is rocket science. #hollywoodbowl @ Hollywood Bowl URL

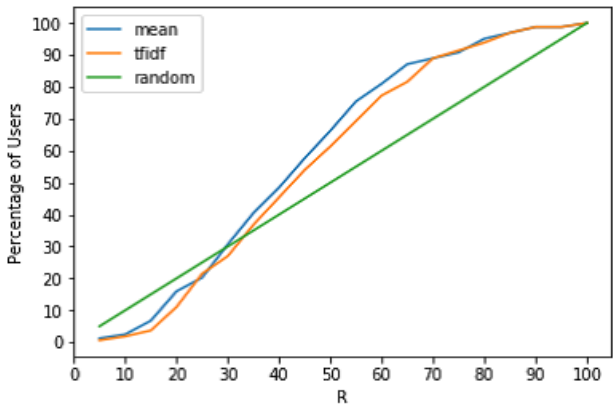

図 2　上位 R%に選択されたユーザの割合

ここで，本研究ではユーザの嗜好に合ったトピックをもつローカル語を推薦することを目的としているため，様々なトピックを併せ持つローカル語は推薦に適していないと考えられ

表 8 “mean”，“tfidf” の両方において上位に推薦されたローカル語 “bbva compass stadium”

R_m	R_t	S_m	S_t	ローカル語のツイート	ユーザの過去のツイート
1.6%	0.8%	0.844	0.860	Got to see a Dynamo win with the bffs!! @ BBVA Compass Stadium URL.	NWSL Regular Season: September 25, 2016: Houston Dash vs Seattle Reign at BBVA... URL.
				Gonna see this game good! Arriba los Tigres!! @ BBVA Compass Stadium URL.	At the @HoustonDynamo game with @paulwall-baby #ForeverOrange URL.

表 9 “mean”，“tfidf” の両方において下位に推薦されたローカル語 “amc magic johnson”

R_m	R_t	S_m	S_t	ローカル語のツイート	ユーザの過去のツイート
87%	84%	0.596	0.568	Having a movie day. Viewing #magnificentseven. @ AMC Magic Johnson Harlem 9 URL.	I’m at John F. Kennedy International Airport in Queens, NY, NY URL
				Enjoying a movie with my little man... My kids are my motivation, my everything. (@ AMC Magic Johnson Harlem 9) URL	I’m at Adolfo Surez Madrid-Barajas Airport - @aena in Madrid URL

表 10 “mean” で上位に推薦されたローカル語 “thewit”

R_m	R_t	S_m	S_t	ローカル語のツイート	ユーザの過去のツイート
34%	51%	0.806	0.713	Love a Thursday date night?? @ ROOF on theWit URL.	“I wonder if @MissyNev know who they playin today??” #GoTigers
				Hanging with my family. We got a babysitter. Woop Woop @ ROOF on theWit URL	“I’m still nice for record”

表 11 “tfidf” で上位に推薦されたローカル語 “foodie”

R_m	R_t	S_m	S_t	ローカル語のツイート	ユーザの過去のツイート
39%	22%	0.844	0.847	Sunset Station Casino hosted Foodie Fest. Food, beer, wine, rides and games @ Las Vagas URL	Finally got to start this movie last night! #kingsglaivefinalfantasyxv #finalfantasyxv… URL
				Las Vegas last day here at Foodie Fest..... #LasVegas Day 3:Come get the must try items!!!?The… URL	Dance Party! URL

る．そこで，各ローカル語に対し，ローカル語を含むツイート集合の中に大きな重要度を持つ単語が存在すれば，そのローカル語は明示的なトピックを持つと仮定し，27 日目から 30 日目に抽出されたローカル語の内、そのツイート集合が 20 単語以上で構成され，かつ (4) 式で表される重要度が 0.1 を越える単語を持つローカル語を選択したところ，合計で 3,653 個のローカル語が選択された．次に，Twitter 上からランダムに選択した投稿者 1,000 人の過去のツイートを 1 人当たり最大 3,000 件取得し，選択したローカル語の内，当日にローカル語を含むツイートを投稿したユーザ 178 人からローカル語を含むツイートを投稿する直前までの過去のツイートを同様に最大 3,000 件抽出し，学習した word2vec を用いて，2.2 節の手法に基づき各ユーザの嗜好を表す意味表現を算出した．その上で，各ユーザに対し，ローカル語を含むツイートを投稿したときのユーザの位置を含むエリアを位置情報として持つローカル語を抽出し，ユーザがツイートを行う前日までに行われた各ローカル語を含むツイート集合から同様にローカル語のトピックを表す意味表現を算出し，ユーザの嗜好を表す意味表現との類似度が高い順に推薦を行った．ただしローカル語が示す位置情報は複数エリアに含まれ得るため，最も頻出したエリアのみをローカル語が示すエリアとした．

図 2 にユーザが実際に投稿したローカル語が上位 R%に推薦されたユーザの割合を示す．図中の “tfidf” は式 (5) を用いて算出した $w'_{k,m}$ を用いてローカル語に対する意味表現を算出した場合であり，“mean” は文書の意味表現を各構成単語の意味表現の平均とした場合，すなわち，ローカル語を含むツイート集合中の各単語の出現頻度 $tf_{u_{k,m}}$ を式 (3) において $w'_{k,m}$ の代わりとして用いて意味表現を算出した場合である．また “random” はユーザに無作為に地域情報を推薦した場合である．図より，ユーザの嗜好とローカル語の意味表現の類似性を考慮することにより，無作為に推薦した場合と比較すると，ユーザが実際に投稿したローカル語が正しく上位に推薦される傾向にあることが分かる．しかし，単語への重み付けによる違いはほとんど見られなかった．一つの要因として，ユーザの嗜好を表す意味表現を算出する際に無作為に選択した 1,000 人の投稿者の過去のツイートにトピックの偏りがあったため，ユーザのツ

表 12　“bbva compass stadium” における $tf_{u_{k,m}}$，$w_{k,m}$

ローカル語				投稿者			
“mean”		“tfidf”		“mean”		“tfidf”	
単語	$tf_{u_{k,m}}$	単語	$w_{k,m}$	単語	$tf_{u_{k,m}}$	単語	$w_{k,m}$
game	0.071	dynamo	0.073	houston	0.118	dynamo	0.097
see	0.054	game	0.054	vs	0.083	vs	0.082
dynamo	0.036	see	0.035	dash	0.047	images	0.066
year	0.036	year	0.030	posted	0.046	photos	0.053
youve	0.018	battery	0.028	facebook	0.046	facebook	0.052

表 13　“amc magic johnson” における $tf_{u_{k,m}}$，$w_{k,m}$

ローカル語				投稿者			
“mean”		“tfidf”		“mean”		“tfidf”	
単語	$tf_{u_{k,m}}$	単語	$w_{k,m}$	単語	$tf_{u_{k,m}}$	単語	$w_{k,m}$
premiere	0.182	premiere	0.202	ny	0.021	lol	0.007
movie	0.091	marvel	0.119	new	0.020	knicks	0.006
day	0.091	netflix	0.107	york	0.016	litt	0.006
viewing	0.091	cage	0.104	get	0.009	get	0.005
luke	0.091	viewing	0.104	lol	0.009	st	0.005

表 14　“thewit” における $tf_{u_{k,m}}$，$w_{k,m}$

ローカル語				投稿者			
“mean”		“tfidf”		“mean”		“tfidf”	
単語	$tf_{u_{k,m}}$	単語	$w_{k,m}$	単語	$tf_{u_{k,m}}$	単語	$w_{k,m}$
woop	0.091	woop	0.146	chicago	0.025	radisson	0.018
love	0.045	irvine	0.081	tonight	0.015	wit	0.014
thursday	0.045	babysitter	0.075	wit	0.012	blu	0.012
date	0.045	betta	0.074	work	0.012	gotham	0.011
night	0.045	robert	0.054	magnificent	0.011	lsd	0.010

表 15　“foodie” における $tf_{u_{k,m}}$，$w_{k,m}$

ローカル語				投稿者			
“mean”		“tfidf”		“mean”		“tfidf”	
単語	$tf_{u_{k,m}}$	単語	$w_{k,m}$	単語	$tf_{u_{k,m}}$	単語	$w_{k,m}$
fest	0.138	fest	0.125	te	0.012	noah	0.014
food	0.069	sponsoring	0.052	day	0.010	stoked	0.009
squad	0.034	food	0.051	work	0.010	fantasy	0.007
sponsoring	0.034	scratch	0.046	haha	0.008	work	0.006
dj	0.034	strip	0.046	good	0.007	time	0.006

イートを構成する単語の TFIDF 値を適切に算出できなかったことが考えられる．

表 8，9 に，“tfidf” と “mean” の両者においてユーザが実際に投稿したローカル語が上位に推薦されたローカル語 “bbva compass stadium” 及び下位に推薦されたローカル語 “amc magic johnson”，表 10 に “tfidf” よりも “mean” において上位に推薦されたローカル語 “thewit”，表 11 に “mean” よりも “tfidf” において上位に推薦されたローカル語 “foodie” のそれぞれの R，類似度 S，ローカル語を含むツイートと投稿者のツイートを示す．ただし R_m，R_t はそれぞれ “mean”，“tfidf” における R を表し，S_m，S_t もそれぞれ “mean”，“tfidf” における S を表している．また，表 12，13，14，15 にそれぞれのローカル語における $tf_{u_{k,m}}$，$w_{k,m}$ の上位 5 個を示す．

“bbva compass stadium” は “HoustonDynamo” というサッカーチームの本拠地であるスタジアムを表すローカル語である．表 8，12 から，ローカル語を含むツイートには “dynamo” や “game” といったサッカーチームに関連する単語が頻出している．投稿者は “HoustonDynamo” のファンであると想定される人物であり，“bbva compass stadium” および “HoustonDynamo” に関するツイートを多く投稿している．これによりローカル語のトピックを表す意味内容と投稿者の嗜好を表す意味内容の類似度が大きくなり，“mean” 及び “tfidf” の両方において正しく推薦されたと考えられる．

“amc magic johnson” は映画館を示すローカル語である．表 9 のローカル語のツイートを見ると映画関連のトピックが出現していることが分かる．一方，投稿者は旅行が趣味であり，様々な場所において写真とともにその場所を表す単語とともにツイートを投稿している．これにより，様々な場所がもつトピックにより投稿者の嗜好が表され，特定のトピックを表す “amc magic johnson” というローカル語との類似度が下がったと考えられる．

“tfidf” よりも “mean” において上位に推薦されたローカル

語である “thewit” は “Roof on theWit” というバーの名前を表すローカル語である．表 14 から，ローカル語のツイートと投稿者の過去のツイートはともに目立ったトピックを持たないと考えられる．一方，表 14 から，$tf_{u_{k,m}}$ よりも $w_{k,m}$ において，“woop” という単語が大きな値を持っていることが分かる．このように，“tfidf” においては，トピックを表すうえであまり重要でないと考えられるが，特定のローカル語においてのみ頻出するために大きな重要度をもつ単語がローカル語の意味内容に大きく影響を与えている．それに対して “mean” においては，それらの重要でないと考えられる単語の重要度が低くなっているため，“tfidf” よりも “mean” においてランキングが上昇したと考えられる．

“mean” よりも “tfidf” において上位に推薦されたローカル語である “foodie” はラスベガスで行われた祭事である，“Great American Foodie Fest ” を表している．表 11 から，“foodie” というローカル語は祭事関連のトピックをもっていることが分かる．また，投稿者はゲームを趣味としており，過去のツイートには娯楽関係の単語が多く出現している．表 15 から，“foodie” というローカル語は $tf_{u_{k,m}}$，$w_{k,m}$ ともに祭事関連の単語が大きな値を持っていることが分かるが，一方で投稿者のツイートでは $tf_{u_{k,m}}$ において，“work” という娯楽のトピックを持つとは考えにくい単語が大きな値を持っており，それに対して $w_{k,m}$ では “fantasy” というような娯楽関連の単語が大きな値を持っている．これにより “mean” より “tfidf” においてローカル語のトピックを表す意味内容と投稿者の嗜好を表す意味内容の類似度が向上したと考えられる．

以上の考察から，提案手法を用いることにより，投稿者のツイート及びローカル語のツイートを考慮した推薦ができていることが分かる．図 2 において，“random” より低くなっている部分が存在する要因としては，過去のツイートにおいて明確なトピックが存在しない投稿者に対して推薦を行ったため，明示的なトピックを持つローカル語を選択できなかったためであると考えられる．

4. ま と め

本研究では，マイクロブログからユーザの嗜好及び地域情報をリアルタイムに抽出し，両者の意味的な類似性に基づいて，ユーザの嗜好・現在地に応じた地域情報をリアルタイムに推薦する手法を提案した．アメリカ本土において抽出した緯度・経度付きツイートを用いて，ローカル語とローカル語を含むツイートの抽出し，単語の意味表現を学習することにより，ユーザの過去のツイート及び地域情報を構成する単語の意味表現から両者を特徴づけ，類似性に基づき地域情報を推薦することで，無作為に推薦するよりも正しくユーザの嗜好に応じた地域情報の推薦が可能となった．一方で，TFIDF 法を用いた場合でも精度の向上はみられなかった．今後の課題として，明示的なトピックをもつ単語の意味内容を強く反映させる手法の検討が挙げられる．

5. 謝　　辞

本研究の一部は，科学研究費補助金（基盤（C）26330137，基盤（S）16H06302）の助成を受けたものである

文　　献

[1] “Google Map,” https://maps.google.com
[2] T. Kamimura, N. Nitta, K. Nakmura, and N. Babaguchi, “On-line Geospatial Term Extraction from Streaming Geotagged Tweets,” Proc. International Conference on Multimedia Big Data, pp.322–329, 2017.
[3] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient Estimation of Word Representations in Vector Space,” Proc. International Conference on Learning Representations, 12 pages, 2013.
[4] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, “Distributed Representations of Words and Phrases and their Compositionality,” Proc. International Conference on Neural Information Processing Systems, pp.3111-3119, 2013.
[5] T. Sasaki, M. Okazaki, and Y. Matsuo, “Earthquake Shackes Twitter Users: Real-time Event Detection by Social Sensor,” Proc. International Conference on World Wide Web, pp.851-860, 2010.
[6] A. Schulz, P. Ristoski, and H. Paulheim, “I See a Car Crash: Real-time Detection of Small Scale Incidents in Microblog,” Proc. Extended Semantic Web Conference, pp.22-33, 2013.
[7] “Twitter,” https://www.Twitter.com.
[8] R. Lee, S. Wakamiya, and K. Sumiya, “Discovery of Unusual Regional Social Activities Using Geo-tagged Microblogs,” World Wide Web, Vol.14, No.4, pp.321–349, 2011.
[9] W. X. Zhao, Y. Guo, Y. He, H. Jiang, Y. Wu, and X. Li, “We Know What You Want to Buy: A Demographic-based System for Product Recommendation On Microblogs,” Proc. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp.1935-1944, 2014.
[10] O. Phelan, K. McCarthy, and B. Smyth, “Using Twitter to Recommend Real-Time Topical News,” Proc. ACM Conference on Recommender Systems, pp.385-388, 2009.
[11] “Term Extract,” http://gensen.dl.itc.u-tokyo.ac.jp.
[12] E. Brill, “A Simple Rule-based Part of Speech Tagger,” Proc. Workshop on Speech and Natural Language, pp.112-116, 1991.