Vol.2010-MBL-55 No.10
2010/9/2

[奨励講演]Twitter 分析に基づく位置依存文字列の抽出

荒川 豊[†1] 田頭茂明[†1] 福田 晃[†1]

本研究では，2009 年 12 月から 2010 年 6 月にかけて収集した位置情報付きツイート 50 万件の中から，位置依存性の高い文字列を抽出する手法を提案する．提案手法では，あるキーワードを含むツイート群に対して，緯度および経度の標準偏差をそれぞれ求め，ツイート群のばらつきの度合いから，そのキーワードの位置依存性を測る．しかし，この手法では，依存する位置が複数存在するキーワード（例えば，チェーン展開している有名店舗名など）を位置依存性の低い単語として判定してしまう．そこで，ある一定の割合以上のツイートを含むエリアを高速に抽出する二次元深さ優先探索を提案する．提案手法では，まず，エリアを 100 キロ四方のグリッドに分割し，それぞれのグリッド内のツイート含有率を計算する．次に，ツイート含有率がある閾値を超えたエリアを 10 キロ四方のグリッドに分割し，同様の判定を行い，最終的には 1 キロ四方のグリッドまで走査する．これらの分析により，1 つのキーワードに対して複数の位置依存性を抽出することが可能となる．

Extraction of Location Dependent Words from Twitter Logs

Yutaka Arakawa ,[†1] Shigeaki Tagashira [†1]
and Akira Fukuda[†1]

In this paper, we propose how to extract the location-dependent keywords from our database which includes 465254 tweets obtained from Dec. 2009 to June 2010. First, we analyze the standard deviation of latitude and longitude, which shows variation level. It is very simple way, but it can't find out the keywords which depend on several locations. For example, famous department stores distributed all over Japan have a large standard deviation, but they will depend on each location. Therefore, we propose two dimension breadth first search, where the searching area is divided into some square grid, and we extract the area which include tweets more than average. In addition, we re-divide the extracted areas into more small grids. Our method can extract some locations for one keywords.

1. はじめに

これまで我々は，携帯端末の入力を快適にする手法として，ユーザのコンテキストに応じて辞書を動的に変化させるコンテキストアウェア IME システムを提案している[1)–3)]．さらに，その有効性を明らかにする手法の一つとして，近年爆発的に利用者が増大している Twitter の位置情報付きツイートを分析することを提案し，「新宿」や「渋谷」といった文字列が，まさに「新宿駅」や「渋谷駅」周辺に偏って利用されていることを示している[4)]．しかしながら，これまでは地図上にプロットして視覚的にその偏りを示していただけであり，定量的に位置依存性を評価できていなかった．そこで，本研究では，その偏り具合を，ある文字列の位置依存性として定量化し，どのような単語がどの位置でよく利用されているかを明確にする．定量化する手法として，まず，緯度および経度の標準偏差を用いた手法を提案する．これにより，ある文字列を含むツイートがどの程度地理的なばらついているのかを数値として把握することが可能となる．しかしながら，電気屋や百貨店など複数の地域にランドマークとして存在するような文字列の場合，位置依存性があるにもかかわらず，標準偏差は大きな値をとってしまうと言う問題点がある．そこで，もう一つの手法として，二次元深さ優先探索を提案する．本手法は，探索エリアをグリッドに区切り，グリッド内に含まれる特定文字列を含んだツイート数が閾値を超えるか否かを判定する．もし閾値を超えるエリアが存在する場合，そのエリアを 10 分の 1 のグリッドで区切り，同様の判定を行う．これらの動作を 100km グリッドから 1km グリッドまで 3 階層で行い，1km グリッドの数を位置依存性の指標とする．本論文では，これまでに収集した位置情報付きツイート約 50 万件に対して，山手線の駅名や都道府県名など，数種類の文字列に関して，上記の分析を行い，それぞれの位置依存性を明らかにした．その結果，標準偏差の値からキーワードの位置依存性を定量化できること，また標準偏差が大きな値であっても二次元幅優先探索により位置依存性を抽出できることを明らかにした．一方，提案手法では，ツイートの地理的な偏りにより，いわゆる一票の格差ならぬ，1 ツイートの格差が生じてしまうため，改善の余地があることも明らかになった．以降では，第 2 章において，提案分析手法について説明し，第 3 章で分析結果を示す．最後に，第 4 章で本研究および今後の課題を総括する．

†1 九州大学大学院システム情報科学研究院
Graduate School of Information Science and Electrical Engineering, Kyushu University

Vol.2010-MBL-55 No.10
2010/9/2

2. 位置依存文字列の抽出手法

位置依存文字列の抽出手法として，1）緯度経度の標準偏差による手法，2）2 次元幅優先探索手法を提案する．まず，緯度経度の標準偏差を用いた手法は，あるキーワードを含むツイート群に対して，緯度および経度の標準偏差をそれぞれ算出する．標準偏差の値は，ツイートの発信位置にばらつきが多い場合は大きくなり，ツイートの位置にばらつきが少ない場合は小さくなるため，この値からこのキーワードの位置依存性を測ることが可能となる．次に二次元幅優先探索について，図 1 に示す．この手法では，まず，あるキーワードを含むツイート群 $T^{keyword}$（ツイート数 $N^{keyword}$）を，その緯度と経度を元に 100km 単位の二次元メッシュ状の領域に分割する．このとき，各領域毎に，含まれるツイート数は，

$$N^{keyword}_{a,b,100} \quad (127 \leq a \leq 146, 26 \leq b \leq 46) \tag{1}$$

と表される．a と b は領域の左上の頂点の緯度．経度をそれぞれを示し，100 は辺の長さを表している．$N^{keyword}$ に対するツイート含有率は

$$P^{keyword}_{a,b,100} = N^{keyword}_{a,b,100} / N^{keyword} \tag{2}$$

と表すことができる．次に，$P^{keyword}_{a,b,100}$ がある閾値を超えている領域を抽出し，抽出された領域をより細かい 10km 単位の二次元メッシュ状の領域に分割し，1 つ上の上位層に含まれるツイート数 $N^{keyword}_{a,b,100}$ に対するツイート含有率を

$$P^{keyword}_{i,j,10} = N^{keyword}_{i,j,10} / N^{keyword}_{a,b,100} \quad (a \leq i \leq a + 100km, b - 100km \leq j \leq b) \tag{3}$$

と算出する．数式中には，わかりやすいように 100km と表記しているが，実際は，度 (10 進表記) (decimal degree:DD) に変換し，$100km = 0.9259266666667°$ を用いて計算を行っている．この中から，再度，$P^{keyword}_{i,j,10} > Threshold$ となる領域を抽出し，抽出された領域をより細かい 1km 単位の二次元のメッシュ状の領域に分割する．そして，1 つ上の上位層に含まれるツイート数 $N^{keyword}_{lat,lng,10}$ に対するツイート含有率を

$$P^{keyword}_{x,y,1} = N^{keyword}_{x,y1} / N^{keyword}_{i,j,10} \quad (i \leq x \leq i + 10km, j - 10km \leq y \leq j) \tag{4}$$

と算出する．最終的には，キーワードの位置依存性を，$P^{keyword}_{x,y,1} > Threshold$ となる領域の数で，を定量化する．この手法を用いることにより，あるキーワードが複数の位置に対して依存性を持ち，標準偏差が比較的大きな値になった場合にも，その位置を特定し，依存性を定量化できると考えている．キーワードの利用率が高い 1 キロ四方グリッドの検出は，単純に 1 キロ四方単位で全エリアを走査する方式も考えられるが，日本だけでも約 500 万エリアに分割されることになり，きわめて膨大な計算時間となる．一方，二次元幅優先探索は，100km 四方のエリアから順に絞り込んでいくことで，全エリアを探索する手法と比較して大幅な高速化を達成している．

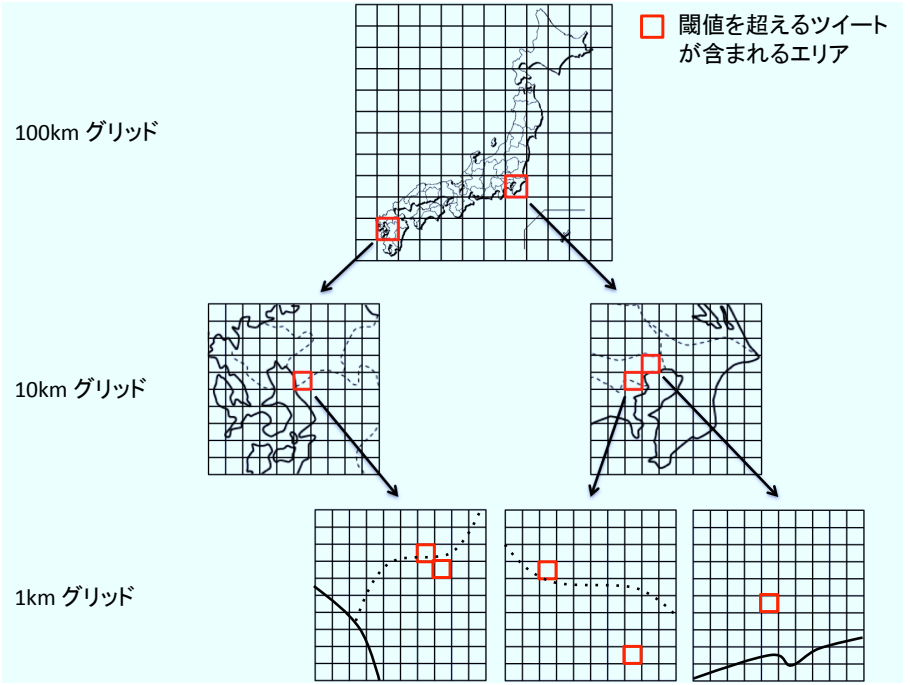

図 1 二次元幅優先探索

3. 分析結果

本研究で分析対象となるのは，2009 年 12 月 15 日から 2010 年 6 月 10 日までの間に収集した位置情報付き日本語ツイート 471275 件の内，北緯 26 度から 46 度，かつ東経 127 度から 146 度の範囲で発信された 465254 件である．

Vol.2010-MBL-55 No.10
2010/9/2

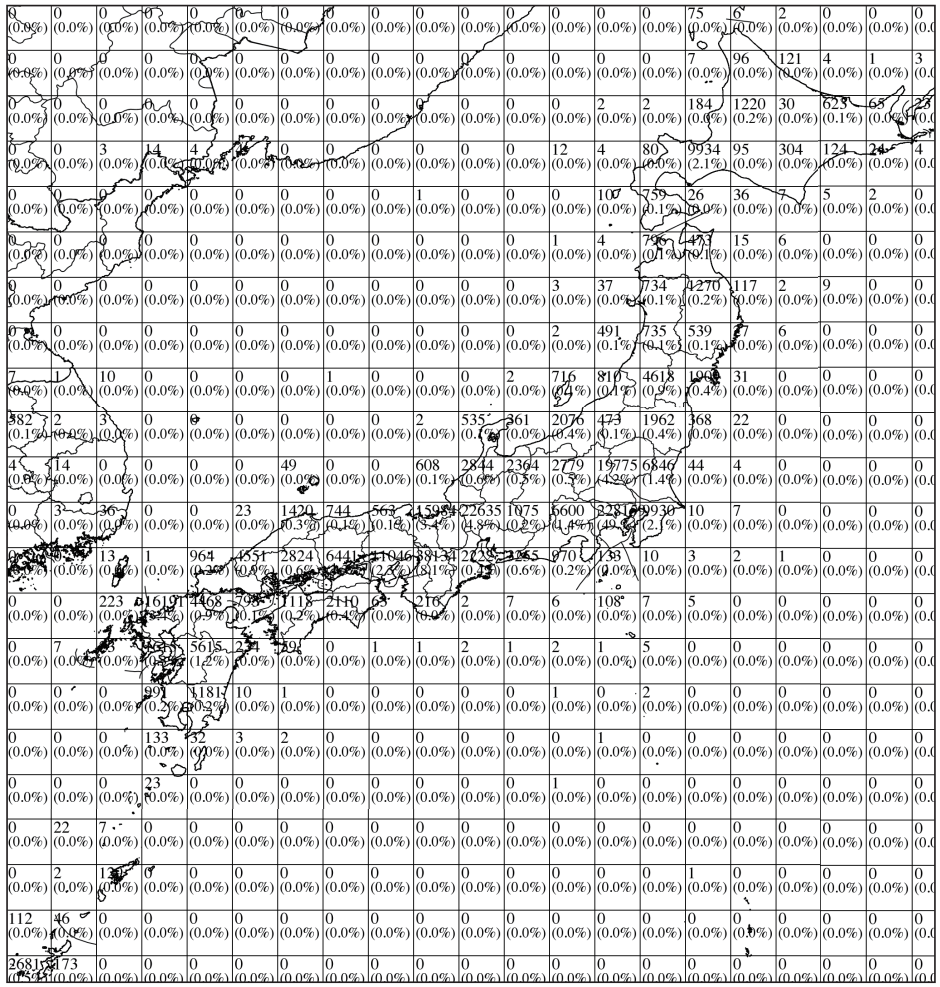

図 2 収集したツイートの地理的分布状況

図 2 に，465254 件のツイートの地理的分布状況を示す．この図は，対象となるエリアを 100km 四方のエリアに分割し，各エリアごとに含まれるツイート数および全ツイートに対する割合を地図上にマッピングしたものである．各エリア左上の数字がツイート数，および

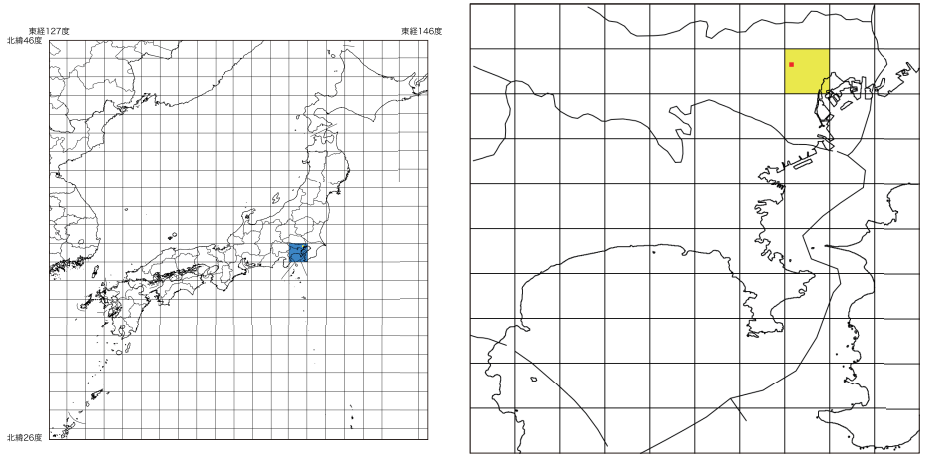

図 3 「新宿」を含むツイートの分布（閾値 15 %）

全ツイートに対する割合を示している．この図より，位置情報が付与されたツイートの約 50%は，東京から神奈川にかけた 100 キロ四方のエリアに集中しており，ツイッターの普及度合いは地域により大きく異なることがわかる．後述するが，このツイッターデバイドが分析に大きな影響を与える．

まず，位置依存性が高いことが判明しているキーワードとして，文献[3] でも示した「新宿」を含むツイートの地理的分布を図 3 に示す．このとき，幅優先探索に用いる閾値は，15%としている．これは，上位のグリッドに含まれるツイートの 15%以上を含むエリアを次の探索エリアとすることを表す．閾値については，後述するが，閾値を変えることにより，抽出されるエリアが変化する．以降の図において，青の領域は設定した閾値を超えた 100km 四方のエリア，黄の領域は設定した閾値を超えた 10km 四方の領域，赤の領域は設定した閾値を超えた 1km 四方の領域である．左の図が日本全体を示しており，1 カ所だけ青のエリアがあることがわかる．それを拡大したものが右の図である．右図には，黄色のエリアがあり，その中に赤のエリアが 1 カ所だけ存在することがわかる．この図から,「新宿」というキーワードは，まさに新宿でよく利用されていることがわかる．

次に，位置依存性が低いと思われるキーワードとして,「なう」と「おはよ」を含むツイートの地理的分布を図 4 と図 5 にそれぞれ示す．このとき閾値は 5%とする．この図からは，「なう」や「おはよ」といったキーワードが首都圏でよく用いられるように見える．これは，

Vol.2010-MBL-55 No.10
2010/9/2

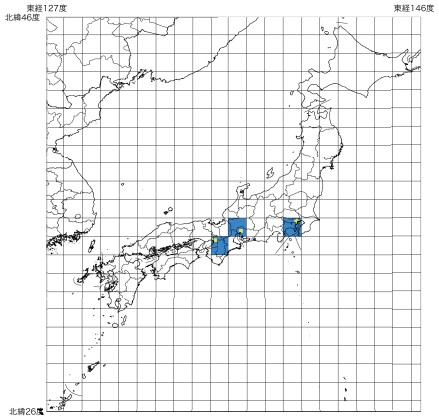

図 4 「なう」を含むツイートの分布（閾値 5 %）

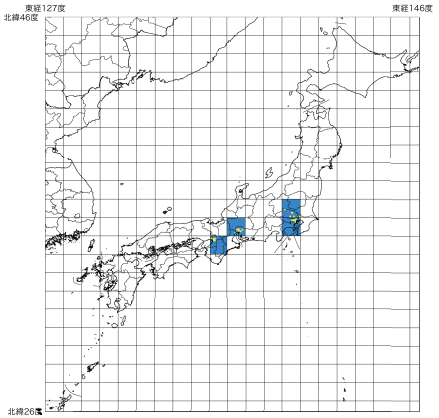

図 5 「おはよ」を含むツイートの分布（閾値 5 %）

100 キロ四方のエリアが抽出されるエリアは，そもそも総ツイート数が多いエリアであることから，ツイートデバイドの影響でこの地域で「なう」や「おはよ」といった単語がよく使われると誤判定しているためである．閾値を変化させた場合の結果は，表 1 を参照するとわかるが，1 キロ四方のエリアがすべて 0 カ所となる．これは，このような汎用的なキーワードは，全国的に分散しており，あり 1 カ所で極端に使われることがないことを意味している．

最後に，複数の位置に依存していると思われるキーワードとして，「ヨドバシ」を含むツイートの地理的分布を図 6 に示す．このとき，閾値は 5%としている．この図より，「ヨドバシ」というキーワードは，大まかに，福岡，大阪，東京で用いられており，それぞれを拡大すると，特に利用率が高い 1km 四方のエリアが複数存在することがわかる．具体的には，東京（右上）では秋葉原や新宿，福岡（左下）では天神，大阪（右下）では梅田，近辺において「ヨドバシ」というキーワードが利用されており，これはヨドバシカメラの実店舗の位置と近いことがわかる．また，表　1 をみると，「ヨドバシ」の緯度と経度に関する標準偏差は 1.872957738 と 2.715003459 となっており，バラツキが大きいこともわかる．これらの結果から，標準偏差を用いた手法では抽出できなかった，複数の位置に依存しているキーワードを，二次元幅優先探索により抽出できることがわかる．

表 1 は，上記分析を行った多種多様なキーワードの一例である．また，表 2 は，山手線の駅名に関してそぞれぞれ分析した結果である．表では，各キーワードに対して，件数，緯度

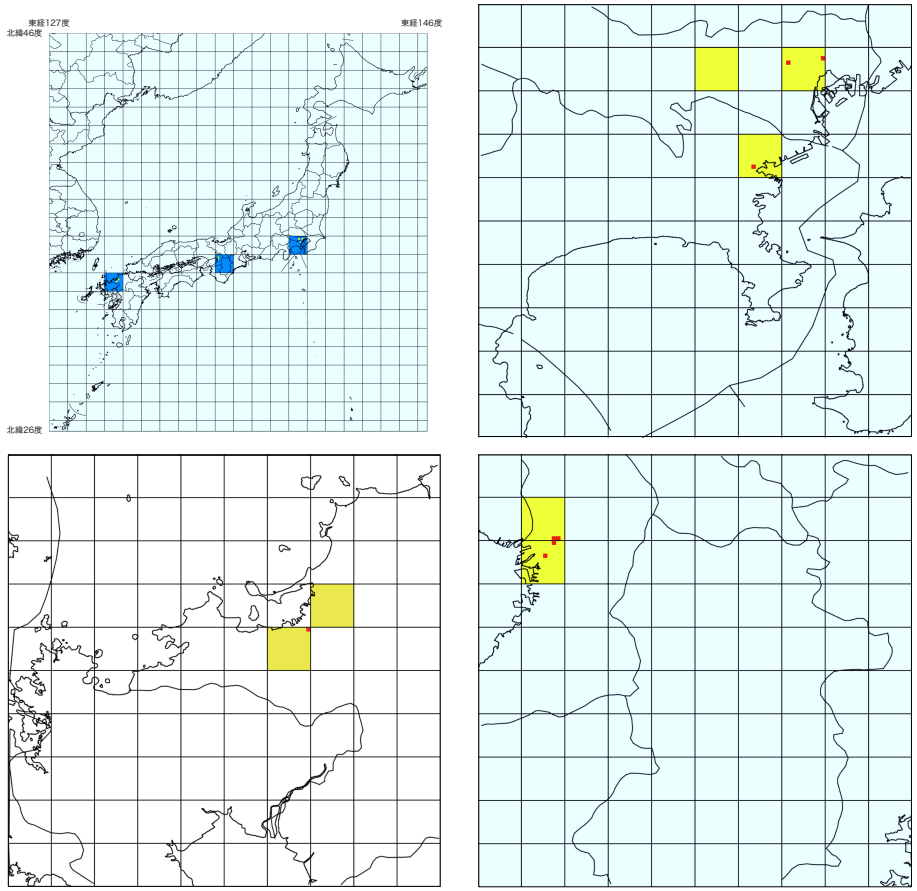

図 6 「ヨドバシ」を含むツイートの分布（閾値 5 %）

に対する標準偏差，経度に対する標準偏差，閾値を 5%とした場合の結果，閾値を 10%とした場合の結果，閾値を 15%とした場合の結果を示している．さらに，見やすいように，標準偏差の値が緯度，経度ともに 1 以下の場合（標準偏差だけで位置依存性が判定可能なエリア）や，標準偏差が共に 1 以下の場合に閾値 15%を超える 1km エリアがあるか否か（標準偏差による判定の正当性を確認）や，標準偏差は 1 以上であるが閾値 15%を超える 1km エ

表 1 キーワードに関する分析結果

キーワード	件数	標準偏差（緯度）	標準偏差（経度）	閾値 5 %			閾値 10 %			閾値 15 %			標準偏差が共に 1 以下	標準偏差が共に 1 以下の場合に閾値 15 % を超える 1km エリアがあるか否か	標準偏差は 1 以上だが 閾値 15 %を超える 1km エリア
				100	10	1	100	10	1	100	10	1			
なう	34703	1.780606804	2.954744238	3	14	2	1	1	0	1	1	0			
今日	27332	1.777582278	3.161078953	3	15	2	1	1	0	1	1	0			
おはよ	12870	1.841242083	3.652545806	4	28	3	1	1	0	1	1	0			
ラーメン	2934	1.845647549	3.047124398	4	18	10	1	1	0	1	1	0			
カレー	2621	2.003874009	2.828997743	3	11	4	1	2	0	1	1	0			
横浜	2440	0.540794598	0.995894859	1	4	2	1	1	1	1	1	1	○	○	
うどん	1286	1.394118804	3.33749605	4	19	6	1	1	0	1	1	0			
ヨドバシ	693	1.872957738	2.715003459	3	7	8	2	4	4	2	3	3			■
松屋	364	1.051085624	2.00895423	3	12	11	1	1	0	1	1	0			
阪急	359	0.471899878	0.789153208	3	16	26	2	9	7	2	7	4	○	○	
西武	342	0.527610551	0.956369306	2	13	35	1	4	1	1	2	1	○	○	
ビックカメラ	338	1.455766352	2.597212698	3	9	13	1	2	3	1	1	1			■
神宮	320	1.079864254	1.915854377	3	8	10	1	1	3	1	1	3			■
藤沢	298	0.115262604	0.109895935	1	2	3	1	2	1	1	2	1	○	○	
ディズニー	236	0.961334883	1.808024695	1	4	6	1	4	2	1	2	1			■
ららぽーと	175	0.262700693	0.868447204	3	12	15	2	6	5	1	2	2	○	○	
箱根	150	0.343275572	0.791063259	2	10	16	2	6	6	2	5	2	○	○	
東武	149	0.712048762	0.621171878	2	16	16	2	10	5	2	8	4	○	○	
うなぎ	140	0.688492886	2.438501601	4	27	38	3	10	3	1	1	0			
IKEA	124	0.562286917	2.028247209	5	24	36	4	10	7	2	4	3			■
熱海	112	0.498460099	0.772737061	2	11	23	1	1	3	1	1	2	○	○	
高島屋	108	0.365236665	1.68770487	4	13	21	2	3	4	2	3	3			■
伊勢丹	100	0.549851468	1.671129512	3	11	18	2	5	2	1	2	1			■

リアがある（標準偏差による判定では抽出できないエリア）などに印をつけている．

表 1 をみると,「なう」「今日」「おはよ」「ラーメン」「カレー」といった汎用的なキーワードは件数も多く，標準偏差も大きく，閾値 10%では 1km エリアが存在しないことがわかる．こうしたキーワードは位置依存性がないといえる．これらの単語は，二次元幅優先探索を行っても依存が検出されない．一方,「横浜」「熱海」「箱根」といった地名は標準偏差が小さく位置依存性があるといえる．また,「ヨドバシ」「IKEA」「伊勢丹」といった複数の店舗を持つ店の名前は，依存する位置が複数存在するため，標準偏差は大きくなるが，二次元幅優先探索によりその依存性が抽出できていることがわかる．

表 2 から，山手線の駅名は，比較的標準偏差が小さく，位置依存性が高いことがわかる．「東京」は駅名以外に多数用いられていたり,「田町」は「有田町」「添田町」など様々な地名に含まれている用語でもあることから，標準偏差が比較的大きい値になったと考えられる．

4. まとめと今後の課題

本研究では，収集した位置情報付きツイート 50 万件の中から，位置依存性の高い文字列を抽出する手法として，緯度および経度の標準偏差を用いた手法と，ある一定の割合以上のツイートを含むエリアを高速に抽出する二次元深さ優先探索を提案した．提案手法を用いることにより，標準偏差の値，あるいはある一定割合以上のツイートを含む 1km 四方エリアの存在により，そのキーワードの位置依存性を定量化することが可能となる．しかしなが

表 2 山手線の駅名に関する分析結果

キーワード	件数	標準偏差（緯度）	標準偏差（経度）	閾値 5 %			閾値 10 %			閾値 15 %			標準偏差が共に 1 以下	標準偏差が共に 1 以下の場合に閾値 15 % を超える 1km エリアがあるか否か	標準偏差は 1 以上だが 閾値 15 %を超える 1km エリア
				100	10	1	100	10	1	100	10	1			
東京	6653	1.118969213	1.870019472	1	4	1	1	1	1	1	1	0			
新宿	3372	0.307684917	0.530890082	1	1	2	1	1	1	1	1	1	○	○	
渋谷	2755	0.362913771	0.58612702	1	2	2	1	1	1	1	1	1	○	○	
池袋	1249	0.282709556	0.470970645	1	2	4	1	1	3	1	1	2	○	○	
品川	1181	0.329082904	0.55298274	1	2	3	1	2	2	1	1	1	○	○	
上野	842	0.553581874	1.015299087	1	3	3	1	2	3	1	2	2			■
秋葉原	826	0.33386486	0.794864398	1	2	1	1	1	1	1	1	1	○	○	
恵比寿	753	0.330393149	0.338590489	1	1	3	1	1	2	1	1	1	○	○	
新橋	646	0.34456659	0.609530609	1	1	4	1	1	2	1	1	2	○	○	
目黒	518	0.402998747	0.621846226	1	3	4	1	2	2	1	1	2	○	○	
神田	498	0.600563657	1.646472784	1	1	4	1	1	2	1	1	2			■
田町	490	1.1752865	2.599571518	1	1	3	1	1	2	1	1	2			■
代々木	481	0.375809284	0.493880335	1	2	5	1	2	3	1	2	3	○	○	
有楽町	453	0.15011584	0.615522763	1	2	2	1	1	1	1	1	1	○	○	
原宿	347	0.311440385	0.510431216	1	3	2	1	1	2	1	1	2	○	○	
大崎	287	0.956519757	1.095373452	1	2	2	1	2	2	1	1	1			■
五反田	284	0.559527399	0.726720418	1	2	2	1	1	2	1	1	2	○	○	
浜松町	170	0.622381038	0.206610268	1	1	2	1	1	1	1	1	1	○	○	
高田馬場	158	0.496130954	1.018177569	1	1	4	1	1	3	1	1	2			■
大塚	128	0.608677752	1.987208567	1	3	3	1	3	3	1	3	1			■
駒込	127	0.036614886	0.017057673	1	2	3	1	1	3	1	1	2	○	○	
日暮里	126	0.187998782	0.83024648	1	3	2	1	2	2	1	2	1	○	○	
巣鴨	103	0.067719622	0.280127845	1	2	2	1	2	1	1	1	1	○	○	
御徒町	101	0.025655458	0.03199448	1	2	3	1	2	1	1	1	1	○	○	

ら，ツイートの絶対数が地域によって異なるツイートデバイドの問題により，件数は少ないが位置に依存しているキーワードを検出できていないことも明らかになった．

謝辞 本処理系の開発, 及び検証は, 日本電信電話株式会社 NTT サービスインテグレーション基盤研究所と国立情報学研究所の提供する研究設備, 回線を利用した共同研究の一環として実施している．ここに記して謝意を示す．

参 考 文 献

1) 末松慎司，荒川　豊，田頭茂明，福田　晃：ネットワークを用いたコンテキストアウェア日本語入力支援システムの提案，信学技報，NS2009-136, Vol.109, No.326, pp.89–94 (2009).
2) 荒川　豊，末松慎司，田頭茂明，山口雄輔，田中裕大，福田　晃：[技術展示] ネットワーク連携コンテキストアウェア日本語入力支援システムの実装，信学技報，MoMuC2009-58, Vol.109, No.380, pp.31–34 (2010).
3) 荒川　豊，末松慎司，田頭茂明，福田　晃：コンテキストアウェア IME システムの提案と実装，情報処理学会 マルチメディア, 分散, 協調とモバイル (DICOMO2010) シンポジウム，No.4D-1, pp.914–922 (2010).
4) 荒川　豊，田頭茂明，福田　晃：Twitter におけるコンテキストと単語の相関関係分析，情報処理学会研究報告, SLDM/EMB/MBL/UBI 合同研究発表会「組込み技術とネットワークに関するワークショップ ETNET2010」，Vol.2010-MBL-53, No.50, pp. 1–7 (2010).