情報処理学会論文誌 Vol. 52 No. 7 2268–2276 (July 2011)

推薦論文

Twitterを用いたコンテキストと入力文字列の相関関係分析

荒川 豊[†1] 田頭茂明[†1] 福田 晃[†1]

本研究の目的は，我々がこれまでに提案しているコンテキストアウェア日本語入力システムの実現に向けて，ユーザの位置と実際に入力された文字列との相関関係を明らかにすることである．本論文では，位置情報付き日本語データの中から，位置依存性の高いキーワードを抽出する手法を2つ提案する．データとしては，2009年12月から収集しているTwitter上のツイート約50万件を用いる．提案手法1では，あるキーワードを含むツイート群に対して，緯度と経度の標準偏差を求め，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．提案手法2では，複数の位置に依存しているキーワード（たとえば，チェーン展開している店舗名など）を高速に抽出するための手法として，探索を3階層（100 km の正方エリア，10 km の正方エリア，1 km の正方エリア）に分けて行うことにより，提案手法1では検出できない，全国に分散したキーワードがある確率以上で出現する1 km 正方エリアの高速な抽出を実現している．

Relationship Analysis between User Contexts and Input Word with Twitter

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

The objective of this study is to specify the relationship between user's context and really-used words for realizing the context-aware Japanese text input method editor. We propose two analytical methods for finding location-dependent words from a half million tweets including Japanese and geographical location, which have been collected since Dec. 2009. First method is to analyze the standard deviation of both latitude and longitude of all the tweets including a certain word. It is very simple way, but it cannot find out the keywords that depend on multiple locations. For example, tweets including famous department store's name has a large standard deviation, but they may depend on each location. Therefore, we propose three-tier breadth first search, where the searching area is divided into some square mesh, and we extract the area which includes tweets more than average of upper area. In addition, we re-divide the extracted areas into smaller areas. Our method can extract some locations for one keyword.

1. はじめに

これまで我々は，携帯端末における日本語の入力を改善する手法として，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日本語入力システム（IME：Input Method Editor）を提案している[1)–3)]．コンテキストアウェアIMEでは，初めて入力する地名やニッチなランドマーク名（たとえば，ビル名や交差点の名前，レストラン名）など，辞書データに登録されていないキーワードに関しては変換候補として提示できないという従来の問題点を解決するために，ユーザの位置情報に基づいて動的に辞書を更新する．この辞書により，駅の近くでは駅名が優先されたり，同じ「し」で始まる駅名でも現在位置により「新宿（しんじゅく）」,「品川（しながわ）」,「新橋（しんばし）」の順序が変わるといった入力支援機能が追加されるとともに,「九大伊都キャンパス」や「アクトシティ浜松」といった通常の辞書には登録されていない単語を変換候補として表示することを可能にしている．これまで，文献1) において概念を提示し，文献4) で応答速度の面から提案アーキテクチャを評価し，文献3) で総括している．

一方，その有効性を明らかにする手法の1つとして，近年爆発的に利用者が増大しているTwitter[5)] 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」といったキーワードが，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[6)]．しかしながら，これらのキーワードは，発見的に抽出したキーワードを地図上にマッピングして，視覚的にその偏りを示していただけであり，あるキーワードに対して位置依存性の有無を判定したり，あるキーワードが依存している位置を示すことはできていない．今回は，位置情報付き日本語データとしてTwitterを用いたが，今後Twitter以外にも位置

†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
本論文の内容は2010年3月の組み込み技術とネットワークに関するワークショップ2010にて報告され，モバイルコンピューティングとユビキタス通信研究会主査・幹事全員（全員）により情報処理学会論文誌ジャーナルへの掲載が推薦された論文である．

情報と文字列を含むデータが増加すると考えられ，そのようなデータの中から，キーワードがどの位置で頻繁に用いられているかを高速かつ的確に抽出することは，我々が提案しているコンテキストアウェア IME だけでなく，位置連携広告[7] の効果測定などさまざまな位置連携情報サービスにおいて，重要になってくると考えられる．

そこで，本論文では，収集した位置情報付き日本語データの中から，位置依存性の高いキーワード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（以降，提案手法 1）と，3 階層幅優先探索を提案する（以降，提案手法 2）．提案手法 1 では，キーワードごとにツイートを抽出し，緯度および経度に関して，それぞれの標準偏差を求める．そして，それらの値がともにあるしいき値以下である場合は，そのキーワードの位置依存性が高いと判定する．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．また，この方式は，あるキーワードを含むツイート群に対して，標準偏差の計算を 2 回（緯度と経度）行うだけでよいため高速であるという特徴がある．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キーワードの場合，それぞれの地域においてミクロな位置依存性があったとしても，標準偏差は大きな値をとなるため，提案手法 1 では位置依存性を検出できないという問題点がある．そこで，このような複数の位置に依存したキーワードを抽出する手法として，3 階層幅優先探索を提案する（提案手法 2）．提案手法 2 では，高速な抽出を実現するために，探索エリアを 100 km の正方メッシュ（以降，100 km エリア），10 km の正方メッシュ（以降，10 km エリア），1 km の正方メッシュ（以降，1 km エリア）という 3 階層に分割し，100 km エリアから順に階層的に探索していく．まず，100 km エリアごとにそのキーワードを含むツイート数の割合（以降，出現率）を分析し，出現率が全国における出現率を超えるエリアを抽出する．そして，抽出されたエリアをより細かい 10 km エリアに区切り，再度．各 10 km エリアにおける出現率を分析し，上位層の出現率を上回るエリアを抽出する．ここで抽出された 10 km エリアを，さらに細かい 1 km エリアに区切り，同様の分析を行い，出現率の高い 1 km エリアを抽出する．最終的には，この 1 km エリアの有無により，そのキーワードの位置依存性の有無を判定する．

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，上述した 2 つの提案分析手法を用いて 80 個のキーワードに関して分析を行い，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 3 階層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．以降では，2 章において，Twitter を用いた位置情報を含む日本語データの収集手法について説明し，3 章において，2 つの提案分析手法について説明する．そして，4 章で分析結果を示す．5 章で関連研究を述べ，最後に，6 章において本研究および今後の課題を総括する．

2. Twitter を用いた位置情報を含む日本語データの収集手法

ここでは，本論文において分析対象のデータとなる Twitter（以降，ツイッター）について説明する．ツイッターとは，2006 年 7 月に Obvious 社が開始した，ユーザが 140 文字以内で「つぶやき（以降，ツイート）」を投稿することで，メールやメッセンジャよりも，ゆるいつながりを発生させるコミュニケーションサービスである．スマートフォンの普及が追い風となり，この 1 年でユーザが急増しており，2010 年 9 月の時点で利用者数 1 億 4,500 万人以上，1 日の新規ユーザ数 37 万人，1 日の平均ツイート数 9,000 万件と膨大であり，ツイートに割り当てられる固有の ID は，すでに 17 桁（1 京）に達している．ツイートは，基本的には誰からも閲覧できる状態（隠すことも可能）であり，統計では約 90%のツイートが公開状態である．また，2009 年 11 月に Geotagging API という，ツイートに対してつぶやいた場所の位置情報（緯度・経度）を付与できる API（Application Program Interface）がリリースされ，GPS を搭載した iPhone などの携帯端末からの位置情報付きのツイートが広まり，TweetMap[8] などの地図と連携したサービスや Foursquare[9] などの位置ゲームなど，ツイッターを用いた新たなサービス領域が生まれつつある．

我々は，このツイッターで公開されている位置情報付き日本語ツイートを収集することにより，これまで困難であった位置情報を含む日本語データを大量に入手できるのではないかと考え[6]，2009 年 12 月より**図 1** に示すツイート収集システムを稼働させて，これまでに約 200 万ツイートを収集している．このツイート収集システムの特徴は，Streaming API と Search API という 2 つの API を組み合わせることにより，位置情報付き日本語ツイートの収集効率を改善している点である．Streaming API は，リアルタイムなツイートを得ることができる API であるが，全ツイートからサンプリングされた約 10%のツイートだけである，そのため，Geotagging API リリース当時は，位置情報付きツイートそのものが少なく，Streaming API だけによる収集では，平均 39.9 ツイート/日（集計期間：2009 年 12 月 15 日〜2010 年 1 月 21 日）の収集にとどまっていた．そこで，我々は，位置情報を付与するか否かは，使用しているクライアントとプライバシに対する考え方という各ユーザの状況に依存すると考え，位置情報付きのツイートを発したユーザの過去のツイートと未来のツ

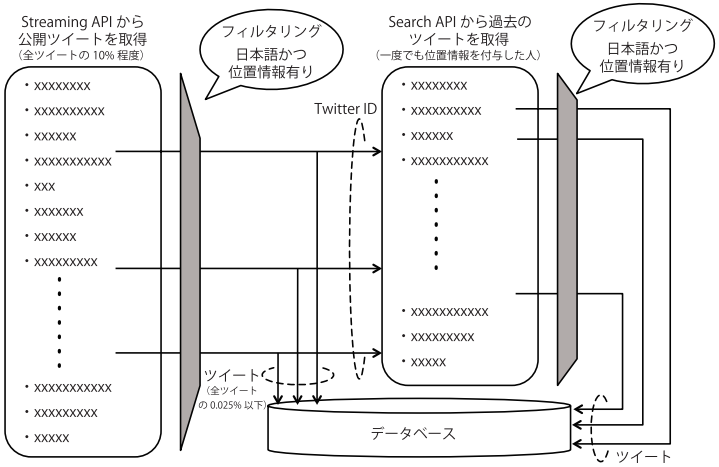

図 1　Streaming API と Search API を組み合わせたツイート収集システム
Fig. 1　Tweet collecting system with Streaming API and Search API.

表 1　Streaming API から得られる 1 時間あたりのツイート数（計測期間：2010 年 10 月 6 日〜10 月 18 日）
Table 1　The number of tweet obtained from Streaming API per hour.

平均取得ツイート数	位置情報を含むツイート数（割合）	日本語かつ位置情報を含むツイート数（割合）
296,896.44	1,842.51 (0.62%)	75.45 (0.025%)

イート（定期巡回により取得）を Search API を用いて収集するシステムを構築している．これにより，収集効率は大幅に改善し，平均 4,392.1 ツイート/日（集計期間：2010 年 1 月 22 日〜2010 年 6 月 10 日）を収集することが可能となった．

表 1 に示すように，位置情報付与可能のクライアントが普及した現在（計測期間：2010 年 10 月 6 日〜10 月 18 日）でも，「日本語かつ位置情報が付与されたツイート」は，Streaming API で得られる 30 万ツイート/時のわずか 0.025%（75 ツイート/時，1,800 ツイート/日）であり，我々の収集方式（平均 3,420 ツイート/日，集計期間：2009 年 12 月 15 日〜2010 年 6 月 10 日）が有効であることが分かる．

3. 提案分析手法

収集した位置情報付き日本語データの中から，さまざまな文字列の位置依存性を明らかにすることを目的として，緯度と経度の標準偏差に基づく位置依存性分析手法と，3 階層幅優先探索による複数位置への依存性の検出手法を提案する．

3.1　緯度と経度の標準偏差に基づく位置依存性分析手法

提案手法 1 は，まず収集した位置情報付き日本語データから，位置依存性を調べたいキーワードを含むツイート群を抽出し，そのツイート群に関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．提案方式では，経験的にこの標準偏差の値がともに 1 以下である場合，そのキーワードに位置依存性があると判定する．

3.2　3 階層幅優先探索による複数の位置依存性の検出手法

提案手法 2 は，**図 2** に示すように，探索エリアを正方メッシュに区切り，エリアごとにこのキーワードを含むツイート数の割合（以降，出現率）を分析し，出現率があるしきい値を超えるエリアを次に探索エリアとして抽出する．次の階層に進むしきい値としては上位層（最上位層は，全エリア）の出現率を用いる．そして，抽出されたエリアをより細かい正方メッシュに区切り，再度それぞれエリアの出現率を分析する．これの分析を，100 km の正方エリア（100 km エリア），10 km の正方エリア（10 km エリア），1 km の正方エリア（1 km エリア）の 3 階層で行い，1 km エリアの有無により，キーワードの位置依存性の有無を判定する．

全エリアにおける収集したツイート数を N_{all}，あるキーワードを含むツイート数を $N_{all}^{keyword}$ とすると，このキーワードを含むツイートの出現率は，

$$P_{all}^{keyword} = N_{all}^{keyword}/N_{all} \tag{1}$$

と表すことができる．また，各正方エリアの左上頂点の座標（緯度：a，経度：b）とし，ある 100 km エリアに含まれるツイート数を，

$$N_{a,b,100} \quad (127 \leq a \leq 146, 26 \leq b \leq 46) \tag{2}$$

とし，ある 100 km エリアにおけるキーワードを含むツイート数を

$$N_{a,b,100}^{keyword} \quad (127 \leq a \leq 146, 26 \leq b \leq 46) \tag{3}$$

と表すと，このエリアでのキーワードを含むツイートの出現率は，

$$P_{a,b,100}^{keyword} = N_{a,b,100}^{keyword}/N_{a,b,100} \tag{4}$$

と表すことができる．

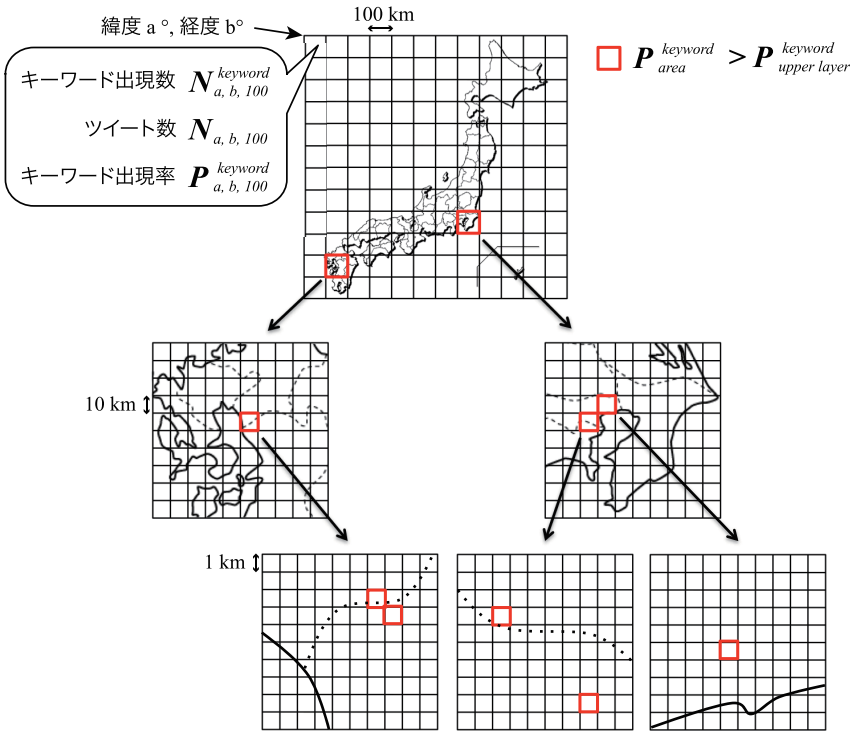

図 2　3 階層幅優先探索
Fig. 2　Three-tier breadth first search.

$P^{keyword}_{a,b,100}$ が $P^{keyword}$ を上回る場合は，全国の平均と比較して，このエリア（緯度：$a°$，経度：$b°$ を左上頂点とする 100 km エリア）は，このキーワードがよく利用されるエリアであることを表していることから，次の階層における探索対象エリアとする．この判定を左上の 100 km エリアから順にすべての 100 km エリアにおいて行い，次の階層で探索対象となる 100 km エリアを抽出する．

次の階層では，抽出された 100 km エリアをより細かい 10 km エリアに分割し，同様の判定を左上の 10 km エリアから順に行う．このとき，各 10 km エリアでのキーワードを含むツイートの出現率は，

$$P^{keyword}_{i,j,10} = N^{keyword}_{i,j,10}/N_{i,j,10} \quad (a \le i \le a+100\,\mathrm{km}, b-100\,\mathrm{km} \le j \le b) \tag{5}$$

と表すことができ，上位層での出現率 $P^{keyword}_{a,b,100}$ を上回る 10 km エリアを次の探索対象エリアと抽出する．数式中には，分かりやすいように 100 km と表記しているが，実際は，度（10 進表記）（decimal degree: DD）に変換し，100 km = 0.9259266666667° を用いて計算を行っている．

次の階層では，抽出された 10 km エリアをより細かい 1 km エリアに分割し，同様の判定を左上の 1 km エリアから順に行う．このとき，各 1 km エリアでのキーワードを含むツイートの出現率は，

$$P^{keyword}_{x,y,1} = N^{keyword}_{x,y,1}/N_{x,y,10} \quad (i \le x \le i+10\,\mathrm{km}, j-10\,\mathrm{km} \le y \le j) \tag{6}$$

と表すことができ，上位層での出現率 $P^{keyword}_{i,j,10}$ を上回る 1 km エリアが存在した場合，キーワードはその 1 km エリアにおいてよく利用されていることとなり，位置依存性があるといえる．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る場合にも，その位置（最小単位 1 km）を抽出することができる．キーワード出現率が高い 1 km エリアの検出は，単純に 1 km 正方メッシュに分割して全エリアを探索することで実現できるが，日本だけでも 4,432,320 エリアに分割されることになり，探索時間がきわめて膨大になる．一方，提案する 3 階層幅優先探索は，100 km エリアから階層的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．

4. 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 471,275 件のうち，北緯 26 度から 46 度，かつ東経 127 度から 146 度の範囲（以降，探索範囲）で発信された 465,254 件である．今回，位置依存性があると思われるキーワードとして，山手線の駅名やデパート名，家電量販店名，公園や施設名など，80 個のキーワードに対して，提案手法 1 および 2 を用いて分析した．

4.1　探索手法のは抽出速度（判定回数）に関して

まず，提案手法 2 の抽出速度に関する有効性を明らかにするため，あるキーワードが集中的に利用される 1 km エリアを発見するために必要な判定回数について分析する．判定とは，キーワードがその 1 km エリアに位置依存性があるかどうかの判定である．探索範囲を

表 2　判定回数が多くなる事例
Table 2　Results of keywords with much decision times.

キーワード	件数	提案手法 1（標準偏差）		提案手法 2（エリア数）			判定回数
		緯度	経度	100 km	10 km	1 km	
今日	27,332	1.777582278	3.161078953	27	138	173	19,294
なう	34,703	1.780606803	2.954744238	48	191	150	27,797

1 km 正方メッシュに分割して，エリアごとに判定を行った場合，判定回数はエリア数と同じ 4,432,320 回となる．一方，階層化して探索を行う提案手法 2 では，100 km エリア（462 個）に関しては全エリアで判定を行うものの，10 km エリア，1 km エリアに関してはその上位層において一定の割合以上のツイートが出現するエリアに絞り込まれるため，判定回数を削減することができる．今回分析した位置依存性の高いと思われる 80 個のキーワードの判定回数は，平均 986 回（最低 462 回，最大 2,741 回）であり，全探索に対して約 4,500 分の 1 に抑制できている．

ただし，今回の分析対象には含まれていないが「今日」や「なう」といった明らかに位置依存性の低いキーワードを分析した場合，判定回数は大幅に増大する（**表 2** を参照）．これは，幅優先探索の枝刈り効果が低くなるため，探索すべきエリアが増えるからである．今後の課題となるが，判定回数に上限値を設けるか，最初に抽出される 100 km エリアの数から早期に位置依存性の低いキーワードと判定する仕組みが必要と考えている．

4.2　全国に 1 カ所しかない施設名の分析

まず，提案手法 1 および提案手法 2 の基本的な有効性を明らかにするため，確実に抽出すべきキーワードとして，全国に 1 カ所しかない施設名に関して分析を行った．その結果を**表 3** に示す．まず，提案手法 1 に関しては，緯度および経度の標準偏差はきわめて小さな値となっており，ある位置での出現率がきわめて高いことを示している．提案手法 2 に関しては，1 km エリアをそれぞれ 1 カ所抽出していることが分かる．抽出された位置を示したものを**図 3** に示す．黄が抽出された 10 km エリアを表し，赤が抽出された 1 km エリアを示す．この図より，的確に依存している位置を抽出していることがが分かる．

4.3　視覚的に依存性が判明していたキーワードに関する分析

次に，位置依存性が高いことが視覚的に判明しているキーワードとして，文献 3) でも示した「新宿」「渋谷」といった駅名がある．そこで，山手線の駅名の中で，「し」で始まる，「新宿」「渋谷」「品川」「新橋」に関して分析した．分析結果を**表 4** に示す．いずれのキーワードも，緯度と経度の標準偏差が 1 以下であり，位置依存性が認められる．これは，提案

表 3　全国に 1 カ所しかない施設名の分析結果
Table 3　Result of the name of unique facilities.

キーワード	件数	提案手法 1（標準偏差）		提案手法 2（エリア数）			判定回数
		緯度	経度	100 km	10 km	1 km	
イクスピアリ	16	0.04972532	0.12422364	1	1	1	693
ラーメン博物館	13	0.036384111	0.023934838	1	1	1	704
東京体育館	11	0.008292145	0.10497707	1	1	1	683
湘南海岸公園	11	0.00028748	0.033721379	1	1	1	693

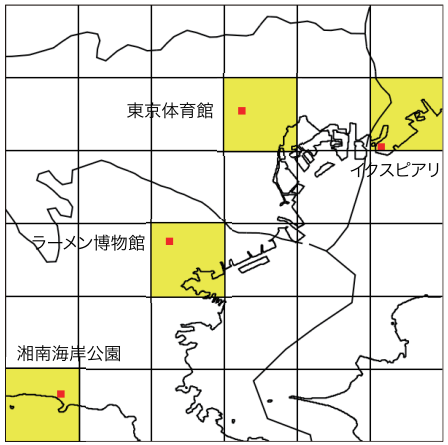

図 3　提案手法 2 で抽出した全国に 1 カ所しかない施設名の地理的分布
Fig. 3　Geographical distribution of the name of unique facilities.

表 4　山手線の「し」で始まる駅名に関する分析結果
Table 4　Results of stations which name start with “shi”.

キーワード	件数	提案手法 1（標準偏差）		提案手法 2（エリア数）			判定回数
		緯度	経度	100 km	10 km	1 km	
新宿	3,372	0.307684917	0.530890082	1	1	13	683
渋谷	2,755	0.362913771	0.58612702	1	1	6	683
品川	1,181	0.329082904	0.55298274	1	1	6	693
新橋	646	0.34456659	0.609530609	1	1	5	683

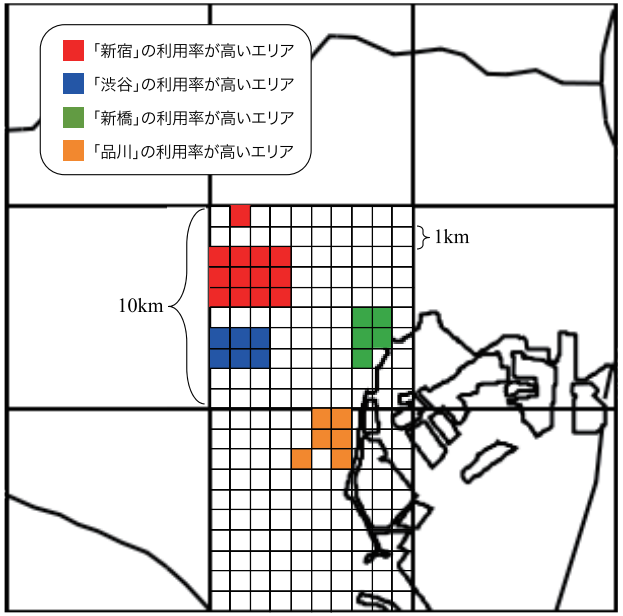

図 4　提案手法 2 で検出した 1 km エリアの地理的分布
Fig. 4　Geographical distribution of 1 km area extracted by our proposal 2.

表 5　全国に展開する家電量販店名に関する分析結果
Table 5　Results of electrical department chain located all over Japan.

キーワード	件数	提案手法 1（標準偏差）		提案手法 2（エリア数）			判定回数
		緯度	経度	100 km	10 km	1 km	
ヨドバシ	693	1.872957738	2.715003459	5	10	14	2,148
ビックカメラ	338	1.455766352	2.597212698	7	13	14	2,741

手法 2 においても 1 km エリアが抽出されていることからも分かる．また，1 km エリア単位で見た場合に，複数の位置に依存性があることも分かる．その様子を地図上にマッピングしたものが**図 4** である．この図より，各キーワードは隣接した複数の 1 km エリアで出現率が高いこと，各キーワードの依存位置に明確な違いがあること，またその位置が実際の山手線の駅近辺であることが明らかになった．

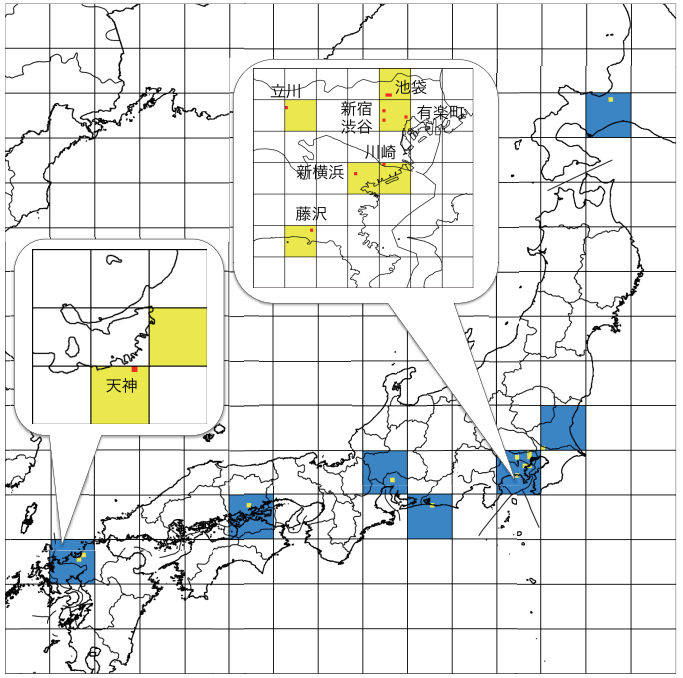

図 5　提案手法 2 で検出した「ビックカメラ」の利用頻度が高いエリアの地理的分布
Fig. 5　Geographical distribution of 1 km areas where the word “biccamera” is frequently used.

4.4　複数の位置に依存性があると思われるキーワードに関する分析

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，全国に展開する家電量販店名である「ヨドバシ」と「ビックカメラ」に関して分析を行った．分析結果を**表 5** に示す．提案手法 1 を用いた場合，緯度および経度の標準偏差は 1 よりも大きくなっており，位置依存性はないと判断される．しかしながら，提案手法 2 を用いた場合は 1 km エリアをそれぞれ 14 カ所抽出しており，複数の位置に依存性があることが分かる．それを検証するために,「ビックカメラ」に関して，1 km エリアの地理的な分布を示したものが**図 5** である．この図から，福岡，岡山，名古屋，浜松，札幌といったビックカメラの地方の各店舗周辺を抽出できていることが分かる．特に関東エリアは，渋谷，新宿，池袋，有楽町，立川，新横浜，藤沢，川崎など，細かい 1 km エリアを的確に抽出していることが分かる．これらの結

果から，標準偏差を用いた提案手法 1 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，3 階層幅優先探索により抽出できる可能性を示した．

5. 関 連 研 究

今回の我々の提案は，ツイッターに関して，a) ツイートの蓄積，b) ツイートの分析，c) 地図情報との連携，という 3 つの側面があるため，それぞれに関して，いくつかの関連研究を示す．

ツイートの蓄積に関しては，自身のつぶやきを Blog 形式で保存する Twilog[10] や複数人の発言を編集して記事として保存できる Togetter[11] などが有名である．しかしながら，これらは自身のつぶやきに限定されているうえ，位置情報はいっさい考慮されていない．一方，米国議会図書館は，全米デジタル情報基盤整備・保存プログラム（NDIIPP）に基づいて 2006 年以降の全ツイートを保存していることが近年明らかになったが，そのデータを利用することは不可能である．

ツイッターに関する全体的な分析は，Krishnamurthy らが，2008 年にツイッターのユーザ特性に関して調査を行っている[12]．また，田中らは，文献 13) において，リンクを含むツイートに絞った分析を行い，bot と人間との区別などを検討している．白木らは，文献 14) において，モバイル端末上でのコンテキストウェアなアプリケーション推薦を実現するために，ツイッターにおける特徴的な言い回し「なう」に着目し,「なう」を含む発言から状況を推定する手法を検討している．さらに，文献 15) では，感情認識のためのコーパスに用いることが検討されている．しかしながら，我々のように，位置情報と入力されたキーワードの関係性を明らかにしている研究はこれまでにない．

また，学術的な研究以外の分析サービスとしては，TwiTraq[16] や TweetStats[17] など 100 種類を超える多くのサービスがあるが，いずれも自分のつぶやきに対するアクセス解析である．ツイッター全体に対する分析サービスとしては，ツイッター社自身が国別のアクセスランキングやさまざまな統計データを提示しているが，細かな 1 km 単位での統計情報などは提供されていない．ほかにも，dBuzz[18] や Twipple[19] などよく利用されているキーワードを分析するサービスがあるが，いずれも時間軸に対して，最近よく利用されているキーワードを抽出しているだけであり，位置は考慮されていない．

ツイッターと地図情報の連携としては，Streaming API を観察し位置情報付きのツイートだけをツイッターのつぶやきを地図上にマッピングした TweetMap[8] が有名である．これは，Streaming API から得たツイートの中で位置情報が付与されていたツイートだけをリアルタイムに地図上に表示しているだけであり，過去のツイートを蓄積されているわけでないので，分析のデータとして用いることはできない．また，近年のツイッタークライアントには，位置情報を指定すると周辺のツイートを表示する機能が搭載されているが，これに関しても Search API から結果を表示しているだけであり，過去データの閲覧や分析はできない．

6. お わ り に

本論文では，コンテキストの中でも特に位置に焦点を絞り，入力されたキーワードとその位置の相関関係を明らかにすることを目的として，ツイッター上の日本語かつ位置情報が付与されたツイートを収集と分析を行った．収集したデータの中から，位置依存性の高いキーワードを抽出する手法として，緯度と経度の標準偏差を用いた分析手法と 3 階層幅優先探索という 2 つの手法を提案した．2009 年 12 月 10 日から 2010 年 6 月 10 日の 140 日間にかけて収集した 471,275 ツイートに関して分析を行い，提案手法 1 を用いることにより，全国に 1 カ所しかない施設名や有名な駅名などは的確に抽出できることを明らかにした．また，提案手法 2 を用いることにより，全探索と比較して大幅に探索時間を削減しつつ，複数の位置に依存した家電量販店名などもその位置を的確に抽出できることを明らかにした．また，提案手法 2 を用いることにより，提案手法 1 よりも詳細に 1 km 正方メッシュ単位での依存位置を探索することが可能になった．

今後の課題として，位置依存性が低いキーワードにおいて提案手法 2 の判定回数が増加する問題の改善と，膨大な時間がかかるため今回は行っていないが実際に全探索を行った場合の検出精度の分析を行う予定である．また，キーワードを軸とした分析だけでなく，各 1 km エリアでの K-top キーワードの抽出手法なども検討するとともに，他の空間分析手法との比較も行う予定である．

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：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) 末松慎司，荒川　豊，田頭茂明，福田　晃：ネットワーク連携コンテキスアウェア IME の高速化手法，電子情報通信学会総合大会，No.B-15-18 (2010).
5) Twitter. http://twitter.com/
6) 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).
7) 株式会社シリウステクノロジー：adLocal. http://twitter.com/
8) TweetMap. http://tweetmap.info/
9) Foursquare. http://foursquare.com/
10) Twilog. http://twilog.org/
11) 株式会社トゥギャッター：Togetter. http://togetter.com/
12) Krishnamurthy, B., Gill, P. and Arlitt, M.: A few chirps about twitter, *Proc. 1st Workshop on Online Social Networks*, pp.19–24, ACM (2008).
13) 田中淳史，田島敬史：twitter のツイートに関する分類手法の提案，第 2 回データ工学と情報マネジメントに関するフォーラム，No.A5-1 (2010).
14) 白木敦夫，矢野幹樹，酒井佑太，小澤俊介，杉木健二，松原茂樹，河口信夫：モバイルアプリケーション推薦のための Twitter 発言者の状況の推定，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.1G-5, pp.251–257 (2010).
15) Pak, A. and Paroubek, P.: Twitter as a corpus for sentiment analysis and opinion mining, *Proc. LREC 2010* (2010).
16) 株式会社ユーザローカル：TwiTraq. http://twitraq.userlocal.jp/
17) TweetStats. http://tweetstats.com/
18) 電通：電通バズリサーチ dBuzz. http://www.dbuzz.jp/
19) Twipple. http://twipple.jp/

(平成 22 年 10 月 25 日受付)
(平成 23 年 4 月 8 日採録)

推　薦　文

本論文では，ツイッターの位置情報付き「つぶやき」から位置依存性の高いキーワードを抽出する方法として，a) あるキーワードを含むつぶやき群に対して，緯度と経度の標準偏差を求めて，位置依存性を検出する方法と，b) 複数の位置に依存しているキーワードを高速に抽出するため探索を 3 階層に分けて行う方法を提案し，実際の位置情報付きつぶやきデータを用いて，上記方法の可能性を示した．位置情報を含むツイートの分析には新規性があり，他の研究にも応用可能で，関連分野への寄与が大きい．よって，ここに研究会推薦論文として推薦する．

（モバイルコンピューティングとユビキタス通信研究会主査　竹下　敦）

荒川　豊（正会員）

1977 年生．2001 年慶應義塾大学理工学部情報工学科卒業．2003 年同大学大学院修士課程修了．2004 年同大学 COE 研究員．2006 年同大学大学院博士課程修了．博士（工学）．2006 年同大学院特別研究助手．2007 年同大学院特別研究助教．2009 年 3 月より九州大学大学院システム情報科学研究院助教．2010 年 4 月より同大学システム LSI 研究センター助教（兼務）．主として，ネットワークアプリケーション，ネットワークプロトコル，トラヒック制御に関する研究に従事．APCC 2008 Best Paper Award（2008 年），情報処理学会 MBL 研究会優秀論文賞（2009 年），DICOMO 優秀論文賞および優秀プレゼンテーション賞（2010 年），Mashup Award 6 GeoHack 賞および沖電気工業賞（2010 年），情報処理学会山下記念研究賞（2011 年），第 3 回フクオカ Ruby 大賞奨励賞（2011 年）．IEEE，電子情報通信学会各会員．

田頭　茂明（正会員）

1996 年龍谷大学理工学部卒業．1998 年奈良先端科学技術大学院大学情報科学研究科博士前期課程修了．2000 年同大学情報科学研究科博士後期課程修了．博士（工学）．2000 年広島大学工学部助手．2007 年同大学大学院工学研究科助教．同年九州大学高等研究院特別准教授，同大学大学院システム情報科学研究院特任准教授．モバイル・ユビキタスコンピューティング，システムソフトウェアの研究に従事．情報処理学会山下記念研究賞（2009 年），電子情報通信学会通信ソサイエティ活動功労賞受賞（2009 年）．IEEE，電子情報通信学会各会員．

福田　　晃（フェロー）

1977 年九州大学工学部情報工学科卒業．1979 年同大学大学院工学研究科修士課程情報工学専攻修了．同年日本電信電話公社（現 NTT）武蔵野電気通信研究所入所．1983 年九州大学助手．1989 年同大学助教授．1994 年奈良先端科学技術大学院大学教授．2001 年九州大学大学院システム情報科学研究院教授，2008 年九州大学システム LSI 研究センター長（兼任），現在に至る．工学博士．組み込みソフトウェア，ユビキタスコンピューティングに関する研究に従事．情報処理学会研究賞（1990 年），Best Author 賞（1993 年）等を受賞．情報処理学会フェロー，電子情報通信学会，ACM，IEEE Computer Society，日本 OR 学会各会員．