ツイート解析　地名判定技術　位置情報

Technology Reports

位置に関連するツイート解析技術とその応用

Twitter*1は，リアルタイムに情報が共有されるSNSであり，それを解析することで，実世界で起こるさまざまなイベントを検出したり，注目スポットを抽出するなど，世の中の動向を知ることができる．本稿では，ツイートを解析し位置に関連付ける技術を紹介する．また，位置に関連付けたツイートの活用事例として，地図上へのリアルタイムなツイートマッピング，紅葉など特定のキーワードに対する盛り上がり地域の検出，通信エリア品質の分析への応用について紹介する．

サービス＆ソリューション開発部† 落合 桂一（おちあい けいいち）
鳥居 大祐（とりい だいすけ）
先進技術研究所 菊地 悠（きくち はるか）
山田 渉（やまだ わたる）

1. まえがき

Twitterは，リアルタイムにユーザの投稿が共有されるソーシャルメディアとして広くユーザに利用されており，それを解析することで，注目されているイベントやスポットを検出したり，話題になっているニュースなど，今起きている世の中の動向を知ることができる．例えば，ドコモが提供するdメニュー リアルタイム検索[1]の「周辺ツイート検索」や周辺ガイドの「話題のスポットランキング」[2]では観光地に関連するツイート*2をリアルタイムで収集することで，周辺の話題スポットに関する情報を即座に把握することができる[3]．このような位置情報サービスを提供するにあたり重要なのは，地名やスポットに関する情報を“リアルタイム”かつ“正確”に収集できることである．そこでドコモではそれを可能とするツイート解析・位置関連付け技術を開発した．

本稿では，ツイートを位置に関連付ける技術について解説する．また，位置に関連付けたツイートの活用事例として，地図上へのリアルタイムなツイートマッピング，紅葉など場所と関連する特定のキーワードに対する盛り上がり地域の検出方法，携帯電話のつながりやすさなどの通信エリア品質の分析への応用について解説する．

2. ツイート・位置情報関連付方法

位置に関連するツイートを集める方法は3種類存在する．

①ジオタグ付きツイート

ジオタグと呼ばれる緯度経度の情報を付加して投稿されたツイートを利用する．

②テキスト解析による地名判定

ツイート本文をテキスト解析して地名を抽出し，地名辞書を用いて位置と関連付ける．

③位置に関連付けたアカウントの利用

ツイートを投稿するユーザのプロフィールを基に位置に関連付けることができる．特に観光スポットや施設にそれぞれ公式アカウントがある場合，アカウントを位置に関連付けることができる．

以下にその具体的な方法を述べる．なお，前述の3種類のうち①は，すでに位置情報と関連づけられているため，本稿では②と③について解説する．

† 現在，サービスイノベーション部

*1 Twitter：Twitterという名称やロゴ，Twitterバードは，アメリカ合衆国また他国々におけるTwitter, Inc.の登録商標．
*2 ツイート：Twitter社の提供するマイクロブログサービスにおける記事のこと．

2.1 テキスト解析による地名判定

前述の3種類の中では，ジオタグ付きツイートや位置に関連付けたアカウントのツイートを利用するよりも，テキスト解析を行って位置と関連づける方式のほうが抽出可能なツイート数が多い．このため情報抽出の観点では，②の方式が有効である．テキスト解析によってツイートと位置を関連付ける場合，ツイート文中に含まれる地名が，同名で異なる場所を示す名称や，人名などの地名以外の意味で使われることを考慮し曖昧性を除去する必要がある[4]．そこで同名の異なる場所に関する曖昧性解消に共起*3語を用いた地名判定技術を，また地名以外の意味で使われた場合の曖昧性解消には地名候補の前後の単語を見て地名かどうか判別することができるCRF（Conditional Random Fields）*4による地名判定技術を用いる．

⑴共起語による地名判定技術

共起語による地名判定技術とは，各地名はその近隣の地名や，その地域特有の語と共起しやすいという仮定に基づき，目的の地名かそれ以外の語かの判定を行う技術である．

例えば，京都にある「円山公園」は北海道にある「円山公園」よりも「京都」という地名と共起しやすい．また地名の「松島」は，「松尾芭蕉」の所縁の地であるため，人名の「松島」に比べて，「松尾芭蕉」という語と共起しやすい．このような共起しやすい語を本稿では共起語と呼ぶ．

共起語による地名判定のフローを**図1**に示す．あらかじめ曖昧性のある地名には，各地名の共起語を紐づけて共起語DBに登録しておく．そして，地名DBと照らし合わせることで，ツイート群から地名を含むツイートを抽出した後，地名に曖昧性があるか判定し，ある場合には，登録した共起語DBと照らし合わせてさらにその地名の共起語を含むツイートだけを抽出する．このようにして曖昧性のある地名と，ツイートとの正確な関連付けを実現している．

⑵CRFを用いた地名判定技術

前述の共起語による地名判定技術では，地名に加えて共起語を含むツイートだけを抽出していた．この手法は高い精度で地名とツイートの関連付けを行うことが可能であるが，共起語を含まないツイートを関連付けることができない．しかし，共起語を含まないツイートでも，地名とそれ以外の語では文章中での語の使われ方を見ることで判別することができる．

例えば，

「松島でご飯を食べた」

「松島さんはご飯を食べた」

という文章中にはどちらも共起語はない．しかし「松島」の付近の「～で」や「～さん」といった特徴を見ることで，それぞれの「松島」が地名か人名かを判別することができる．このような特徴のことを自然言語処理分野では素性（そせい）と呼ぶ．

地名やそれ以外の語に関する素性を分類器*5に学習させることで，文章中の語が地名かそれ以外かを自動で判定させることができる．本手法では分類器としてCRFを用いている．

素性抽出とCRFの学習のフローを**図2**に示す．CRFによる分類ではまず，ツイート群より抽出された地名を含むツイートに対して，人手で地名や人名といった正解値をつけていく．次に正解値がつけられたツイートから，素性を抽出し，分類性能が高くなるようにCRFを学習させていく．

CRFによる地名判定のフローを**図3**に示す．ツイート群より抽出された地名を含むツイートのうち，地

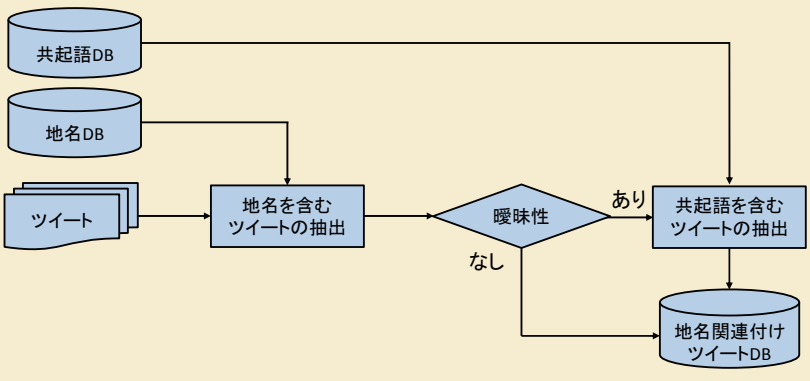

図1　共起語による地名判定のフロー

＊3 **共起**：ある単語とある単語が1つの文章中に出現すること．

＊4 **CRF**：条件付確率場．入力された要素が連なったもの（系列）に対して，その特徴量に基づいてあらかじめ定められたラベルを付与する手法の一種．ここでは文章中の地名の曖昧性の解消のために，文章を構成するそれぞれの語について，周辺の語を参照して地名またはそれ以外かのラベルを付与する．

＊5 **分類器**：入力を，その特徴量に基づいてあらかじめ定められた分類先のいずれかに分類する装置．

名とそれ以外の語で曖昧性があるか判定し，曖昧性があるものに対して，図2のステップで学習させたCRFで，地名かそれ以外の語かの判別を行い，地名とツイートの関連付けを行う．このようにして共起語がない場合でも地名とそれ以外の語の曖昧性を解消し，正確な地名とツイートの関連付けを実現している．

2.2 位置に関連付けたアカウントの利用

観光スポットや施設にそれぞれ公式アカウントがある場合がある．これらの公式アカウントでは，その施設や場所に関する有益なツイートが発信されていることが多い．例えば，ショッピングモールの公式アカウントではセールやフェアの情報があり，市役所では地域の催しの情報などがある．そこで施設や場所の公式アカウントの調査を行い，それらの関連付けを行っている．それら公式アカウントのツイートを利用することで，より多くの有益なツイートを抽出できる．

3. 地図上へのツイート表示と活用事例

地図上へのツイート表示を行うためには，各ツイートに対して表示位置となる緯度経度を割り当てる必要がある．表示位置はツイートごとに決定するものと，ツイートをつぶやいたアカウントにより決定する方法を併用する（**表1**）．

位置に関連するツイートの活用においては，以下に示す2つの観点が存在する．

・リアルタイムに状況把握が必要かどうか
・特定のトピックに関するツイートの抽出が必要かどうか

3.1 地図上でのツイートのリアルタイム表示

地図上でのツイートリアルタイム表示では，ある特定の場所に関連して今まさにつぶやかれていることを，地図上で閲覧できる．特定トピックへの限定は行わず，どのようなことが話題になっているか，場所ごとに確認することができる．ツイートの背景色は位置情報の付与方法に応じ，テキスト解析を用いた場合は青色，公式アカウントの場合は橙色，ジオタグ付きツイートの場合は緑色としている（**図4**）．

位置に関連付け可能なツイートの数は，地域によって件数に大きなば

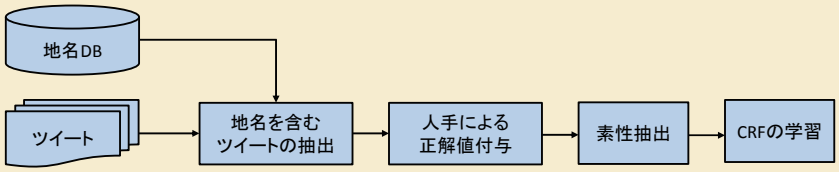

図2　素性抽出とCRFの学習のフロー

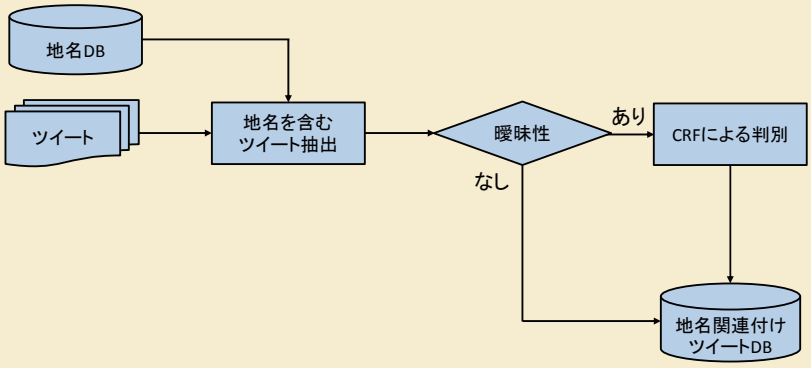

図3　CRFによる地名判定のフロー

表1　地図上におけるツイートの表示位置

ツイート種別	地図上の表示位置
ジオタグ付きツイート	ツイート自身に付与された緯度経度
テキスト解析したツイート	テキスト解析で抽出した地名の緯度経度
公式アカウント（市区町村のアカウントなど）によるツイート	公式アカウント（店舗・役所など）の所在地

らつきがある．また地図の縮尺や画面サイズにより画面上にツイートが表示しきれない場合や，表示できるツイートが一件も存在しない場合もある．

このため，表示エリア内で位置と関連付けられたツイートが多数存在する場合には，ユーザにとって有用性が高いと考えられるツイートを優先的に表示する．具体的には，リツイート*6数が多いツイートを最初に表示し，同一エリアを一定時間表示し続けた場合は優先度が低いツイートに置き換える．

一方，表示エリア内に位置と関連付けられたツイートが存在しない場合には，画面外のツイートおよび方向を指し示す矢印を表示する（**図5**赤枠部分）．また表示された矢印をクリックすることで，画面外のツイートが関連付けられた位置へ地図をスクロールすることができる．これによりユーザは地図の縮尺を変更したり，画面をスクロールさせてツイートを探したりする手間が省け，効率的にツイートを確認することができる．

3.2　エリアの盛り上がり検知

エリアを考慮した分析は位置関連分析の中でも特徴的なものの1つである．単一スポットのみでなく，エリアに含まれる複数のスポットを考慮した分析を行うことで，エリア全体の特徴を捉えることができる．例えば，エリアに含まれるスポットの紅葉に関連するツイート数を指標化することで，そのエリアの紅葉の見ごろを推定することができる．また，**図6**に示すように，過去のツイートから計算した指標によりエリアを見ごろ順に並べ替えると紅葉前線を再現することができる．

エリア全体の特徴を捉えるにあたっては，エリア内に含まれる複数のスポットによってエリア全体の指標とし，エリア内の単一スポットのみによる指標とならないようにする．例として，あるエリアに含まれる3つのスポットに関するツイートの発

図4　地図上におけるツイートの表示位置

図5　画面外ツイートへのナビゲーション

*6 **リツイート**：他者が投稿したツイートを，内容に変更を加えることなく再投稿したツイートのこと．また個々のツイートについて再投稿された回数のことをリツイート数と呼ぶ．

生パターンを考える．**図7**(a)では1つのスポットから3つのツイートが，図7(b)では3つのスポットから1つずつツイートが発生している（例えば，紅葉に関するツイートを想定）．盛り上がり指標を単純にエリア内に発生している全ツイート数とすると，両者ともに3となり変わらない．しかしながら，前者においては何らかのイベントがあったなどスポットにおける特有の事象のみを捉えている場合も想定されるので，すべてのスポットにおいてツイートが発生している後者の方がエリア全体に起きている事象（例えば，エリアにおける紅葉の見ごろ）を捉えるのには適していると考えられる．このような多様性の観点を取り入れた指標としては，ツイートが発生しているユニークスポット数（1スポットで複数のツイートが発生していても1つとカウント）などがある．

多様性の観点に加え，エリア全体でのツイート発生数などの量的観点を考慮することで，エリアの盛り上がり度を定義できる．図7(b)の例では，各スポットから1つずつツイートが発生しているよりも5つずつ発生している方が盛り上がり度が高いと考えられる．期間ごとにエリアにおける多様性と発生量の観点を組み合わせた指標を計算することで，そのエリアにて最も盛り上がった期間（最も紅葉が見ごろであった日）を推定することができる．

このようなエリアを考慮した分析は紅葉だけでなく，桜への適用や台風，ゲリラ豪雨，竜巻，地震などの災害への応用も期待される．

3.3　通信エリア品質の分析

本節では，位置に関連付けたツイートを活用した，通信エリア品質分析への応用について解説する．データ処理フローを**図8**に示す．位置に関連付けたツイートから「つながる」「つながらない」などの通信サービス関連語と，通信会社名を含むツイートを抽出することで，各場所についての通信品質について言及しているツイートを抽出することができる．

抽出した結果を地図上に表示するシステムの画面イメージを**図9**に示す．図9(a)および(b)は表示されている地図の中で通信品質についてのツイートが抽出された地名のリストお

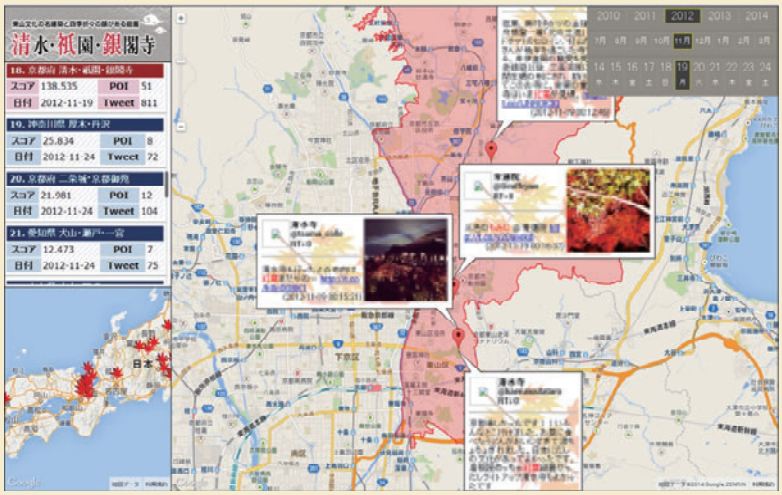

図6　エリア盛上り検知技術を応用した紅葉前線の再現

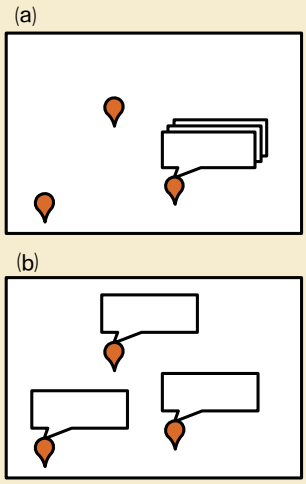

図7　エリアにおけるツイート発生パターン

よびジオタグとテキスト解析で抽出されたツイート数を表示している．図9(c)は通信品質についてのツイートが抽出された場所を地図上で示している．図9(d)は選択した地名に対する通信品質に関するツイートをタイムライン表示している．また，ツイートの内容が「つながる」などのポジティブな場合はテキストの背景をピンクにし，「つながらない」などネガティブな内容の場合は背景を水色にする．そうすることで，選択した地名に対してのユーザのコメントを視覚的に把握することができる．

このように，通信品質に関するツイートを地図上に表示することで，通信エリア品質改善業務の支援を行える情報を提供できる．

4. あとがき

本稿では，ツイートを位置に関連付ける方法と，位置に関連付けたツイートの活用事例として地図上へのリアルタイムなツイートマッピング，紅葉など場所と関連する特定のキーワードに対する盛り上がり地域の検出方法，携帯電話のつながりやすさなどの通信エリア品質の分析への応用について解説した．

今後は，実サービスへの展開や他の位置情報関連データのとの組合せによる解析を行っていく．

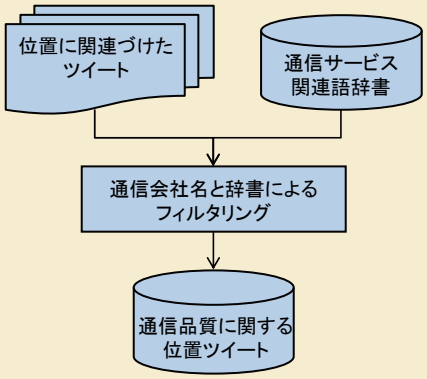

図8　通信エリア品質ツイートの処理フロー

図9　通信エリア品質の確認画面

文　献

[1] NTTドコモ："dメニュー リアルタイム検索．"
http://realtime.search.smt.docomo.ne.jp/
[2] NTTドコモ："話題のスポットランキング ドコモ地図ナビ．"
http://s.dmapnavi.jp/kanko/ranking/ranking_top.php?geo=&val=&linkfrom=T009
[3] 鳥居，ほか："生活密着情報を提供するリアルタイム検索サービスの開発，"
本誌，Vol.20，No.4，pp.12-17，Jan. 2013.
[4] E. Amitay, N. Har'El, R. Sivan, A. Soffer: "Web-a-Where: Geotagging Web Content," SIGIR'04 Proc. of the 27th annual international ACM SIGIR conference on Research and development in information retrieval, pp.273-280, 2004.