検索　リアルタイム　ツイート

Technology Reports

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

Twitter社の提供するSNSは，世の中でリアルタイムに起きている出来事を伝えるメディアとして多くのユーザが利用している．ドコモでは，大量のツイートから日々の生活に役立つ情報（生活密着情報）を簡単に手に入れることができる各種検索サービス（電車運行，TV番組，ランドマーク）を提供している．本稿では，これら検索サービスの紹介，ならびに，サービスを支える検索技術について解説する．

サービス＆ソリューション開発部
鳥居 大祐　赤塚 隼
落合 桂一　角野 公亮

1. まえがき

Twitter[*1]は，リアルタイムに世の中の出来事が多数書き込まれ共有されるSNSとして定着しており，日々大量のツイート[*2]が共有され続けている．共有される内容はユーザの身の回りの情報など個人的なものも含まれる一方，日々の生活に役立つ公益性の高い情報（生活密着情報）も多く含まれる．しかし目的の情報を得るには，適切な検索キーワードをリアルタイムの状況に合わせて設定する必要があり，必ずしも簡単ではない．

そこでドコモでは，日々の生活で利便性の高いツイートを容易に閲覧できる各種専門検索サービスを提供することで，ユーザの利便性向上に努めている．今回開発したのはTwitterと親和性の高い電車運行，TV番組，ランドマーク[*3]に関連するツイートを専用に検索するサービスである．電車運行やランドマークに関するツイートは，現場からのレポートとして有用である．また，TV番組に関するツイートは，番組に対する視聴者のさまざまな感想や意見をリアルタイムに共有できTV視聴をより楽しくするものである．

これら各種検索の実現に重要なことは，各ドメイン（電車運行，TV番組，ランドマークなど）に関連のあるツイートを抽出し検索可能とすることであり，その際課題となるのは，各ドメインの関連ツイートを精度良く抽出することである．今回開発した手法ではドメインごとの辞書（例：電車路線辞書やランドマーク辞書）とツイート本文をマッチングさせて関連ツイートを抽出したうえで，ドメインごとにさらなる精度向上の工夫を凝らしている．

本稿では，各種検索サービスの紹介，ならびに，サービスを支える検索技術について解説する．

2. 電車運行関連ツイート検索

2.1 サービス概要

電車運行関連ツイート検索は，全国各鉄道路線に関する混雑状況や運行情報を含むツイート（以下，電車ツイート）を閲覧することができるサービスである．本検索システムでは，運転見合せや遅延といった電車固有のトラブルを自動で検知し，トラブルが発生していると思われる鉄道路線を注目度の高い鉄道路線として特定する．**図1**（a）に示すとおり，リアルタイム検索トップ画面のコーナー（「電車に関するツイート」と記載）には，地域ごとにボタンが

＊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．
＊3 **ランドマーク**：その土地の名所や象徴となるような建造物．

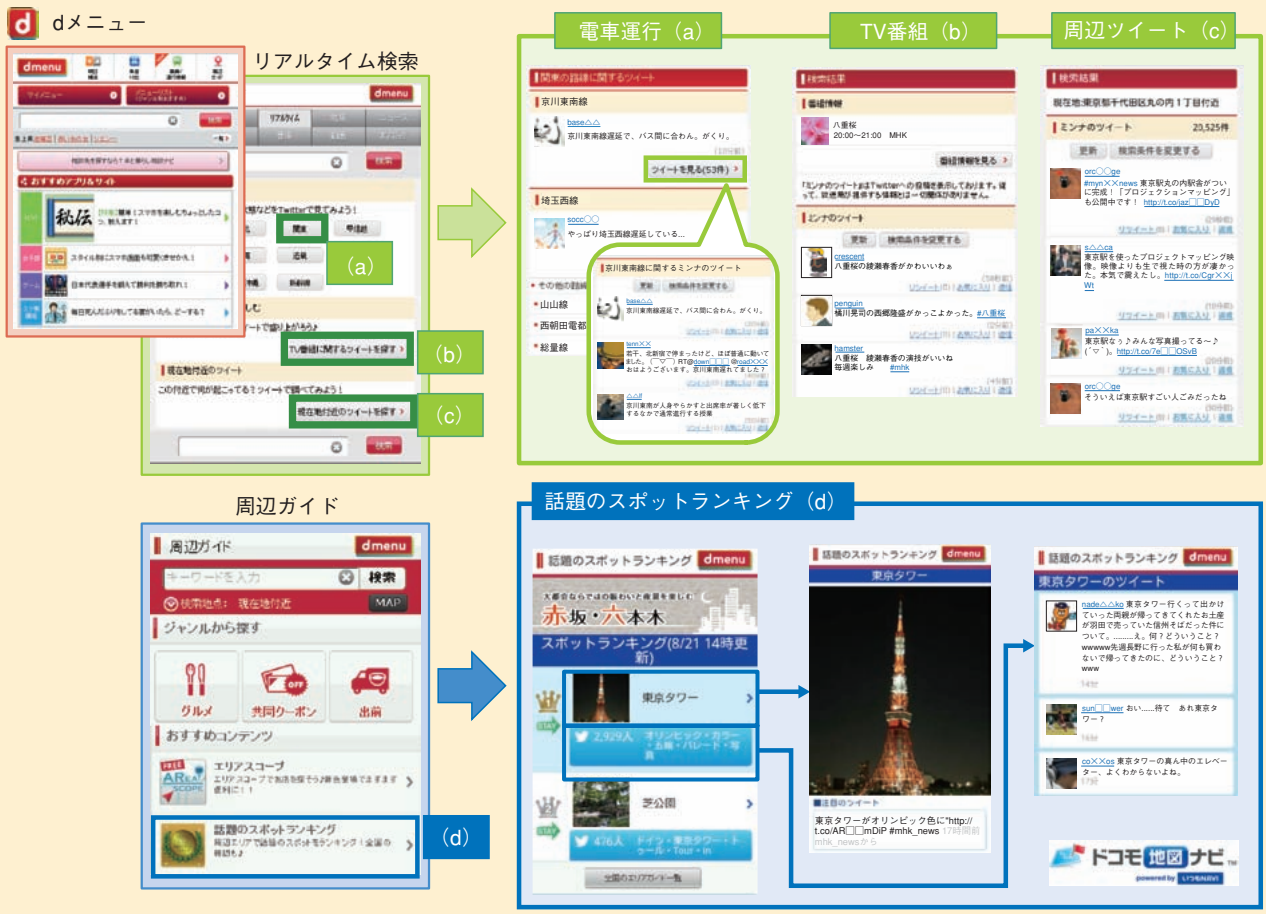

図1　生活密着情報を提供するリアルタイム検索のサービス概要

設置されている．ユーザは所望する地域を選択することで，該当地域の鉄道路線一覧を閲覧することが可能である．図の例では，「関東」のボタンをクリックすると関東の鉄道路線一覧ページに遷移している．このページではTwitter上で注目度が高い順序で鉄道路線名が表示されている．注目度が高い上位2つの鉄道路線（京川東南線，埼玉西線）については，関連する最新の電車ツイートを1つずつ表示している．鉄道路線の電車ツイート一覧を閲覧するには，該当の鉄道路線を選択し「ツイートを見る」ボタンをクリックすることで，指定された鉄道路線の電車ツイート一覧を閲覧することが可能である．

2.2 電車運行関連ツイート処理システム

図2に電車運行関連ツイート処理の概要図を示す．電車運行関連ツイート検索では，混雑状況および運行情報を含むツイートの数を鉄道路線ごとに求め，注目度の高い鉄道路線を抽出する必要がある．電車運行関連ツイート処理は，①ツイートと鉄道路線のマッピング，②混雑状況や運行情報に関連するツイートの抽出，③検索エンジンへのツイート登録，④注目路線判定の順に行われる．

①のツイートと鉄道路線のマッピングは，サーバ側に鉄道路線名とその揺らぎを登録した鉄道路線辞書を用いて，ツイートに路線名が含まれるか部分検索を行うことで実現する．路線名が含まれないツイートは以降②③④の処理を行わない．②の混雑状況や運行情報に関連するツイートの抽出は，あらかじめ混雑状況

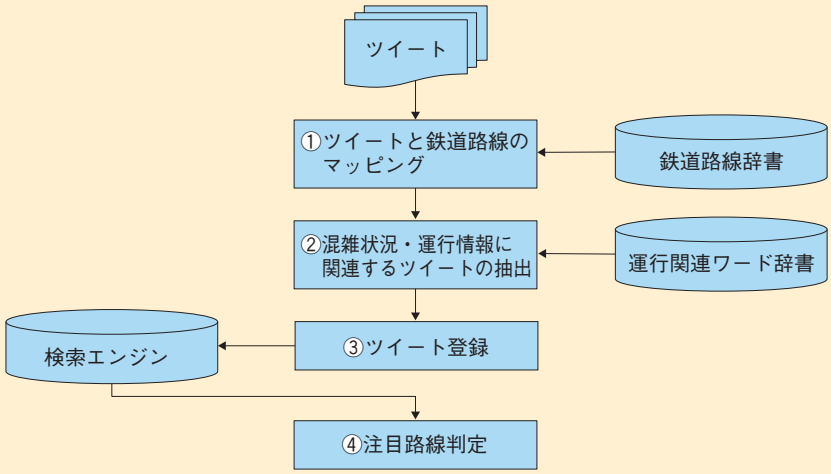

図2　電車運行関連ツイート処理

および運行情報に関連する単語を運行関連ワード辞書に登録しておき，ツイートが辞書に登録された単語が含まれるか部分検索を行うことで実現する．③のツイート登録ではツイートを検索エンジンに登録する．①と②で混雑状況や運行情報を含まないツイートに関しては検索エンジンに登録しない．④の注目路線判定では①と②で絞りこまれたツイートの数を直近一定期間内で集計する．注目度の値がしきい値以上の場合は注目度の高い鉄道路線として画面上に表示する．①②③の処理は数秒で完結するため，ツイートからリアルタイムな電車の混雑状況および運行情報の確認ができる．④の処理は数分おきにバッチ処理で行われる．

3. TV番組関連ツイート検索

3.1 サービス概要

TV番組関連ツイート検索サービスは，放送中のTV番組の内容に対する感想やコメントのツイートを閲覧できるサービスである（図1（b））．TV番組に対して言及しているツイートをリアルタイムに推定し，番組ごとにまとめて表示することにより，ソーシャルビューイング[*4]が実現され，ユーザはTV視聴をより楽しむことができる．

本サービスのトップ画面には，全国ネットで放送中の番組を表示し，ツイート検索画面への遷移が容易になるよう工夫を行った．また，放送地域を選択することで，ローカル番組に関するツイート検索を可能としている．さらに，番組へのツイートの紐付けは，放送中の番組だけでなく1週間後までの放送予定番組に対して行われており，放送日時を指定することでそれらの番組に対するツイートを閲覧することが可能である．

現在放送中の番組に関連するツイートの推定は，放送局のハッシュタグ[*5]に加えて，リアルタイムに抽出した番組特有のハッシュタグやワード（以下，特徴語）を利用することで実現している．また，ツイート検索画面において番組に対して言及していると推定されたツイートは，投稿時間順だけでなく，ツイートに対する注目度を指標とした話題順での並び替えも可能である．

3.2 TV番組関連ツイートのリアルタイム抽出処理

図3に，TV番組に関連するツイートのリアルタイム抽出処理の構成図を示す．放送中のTV番組に関するツイート検索は，番組内容の移り変わりを考慮する必要がある．そのため，本処理では，番組関連ツイート抽出に使用する番組特徴語と番組ハッシュタグを抽出する「(A) 番組特徴語・ハッシュタグ抽出」と，現在放送中のTV番組と推定されたツイートを紐付けて検索エンジンに登録する「(B) 番組関連ツイート抽出」が並行して行われる．

(A) の番組特徴語・ハッシュタグ抽出処理の流れを以下に述べる．まず，①TV番組メタデータデータベース（DB）から現在放送中のTV番組を一意に特定するID（以下，番組ID）とTV局を特定し，ツイート抽出対象の番組とする．また，②解析用ツイートDBから過去一定時間分の全ツイートを抽出して解析対象ツイート群とする．次に，③対象番組を放送しているTV局のハッシュタグが付与されたツイートを解析対象ツイート群から抽出し，番組ツイート群とする．このとき，対象の番組と在京キー局で放送されている番

＊4　**ソーシャルビューイング**：ソーシャルメディアを通じて，複数のユーザが同一のイベントや放送コンテンツの視聴体験を共有すること．

＊5　**ハッシュタグ**：ツイートを投稿する際に「#」記号で始まる単語を付与することで，同じ話題のツイートを他のユーザが見つけやすくする機能（(例) #jishin，#地震）．

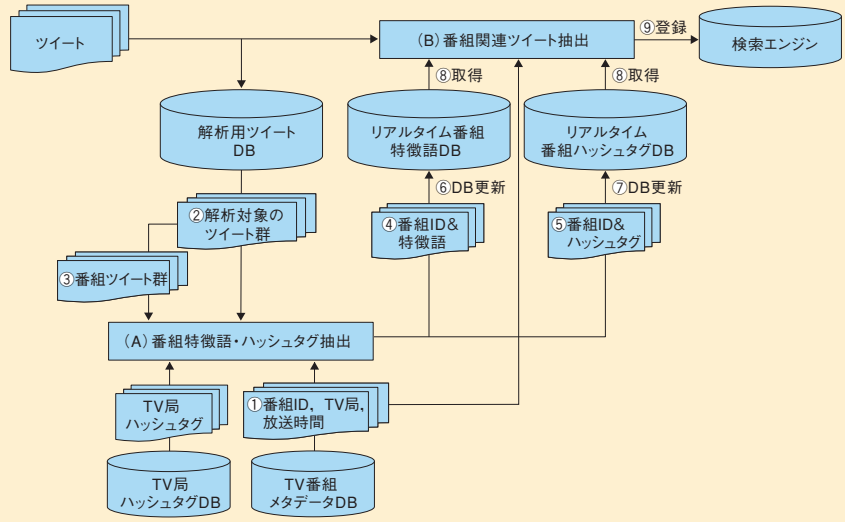

図3　TV番組関連ツイートのリアルタイム抽出処理

組が同一である場合，対象の在京キー局のハッシュタグも番組ツイート群の抽出に使用する．次に，④番組ツイート群を特徴付けるワードを抽出し，番組IDを紐付けて番組特徴語とする．さらに，⑤解析対象ツイート群に含まれるすべてのハッシュタグに対し，各ハッシュタグが付与されたツイート群と番組ツイート群の類似度を計算し，類似度が高いハッシュタグに番組IDを紐付けて番組ハッシュタグとする．最後に，⑥リアルタイム番組特徴語DBと⑦リアルタイム番組ハッシュタグDBとの更新を行う．これらの処理を一定時間ごとに行うことで，リアルタイム番組特徴語DBとリアルタイム番組ハッシュタグDBは常に最新の状態に保たれる．

(B) の番組関連ツイート抽出処理では以下のように処理を行う．まず，⑧ツイート抽出対象番組の番組IDに紐付けられた番組ハッシュタグと番組特徴語をそれぞれリアルタイム番組特徴語DB，リアルタイム番組ハッシュタグDBから取得する．そして，リアルタイムに取得しているツイート本文中のワード，ハッシュタグとマッチングを行うことで，対象のTV番組について言及していると推定されるツイートを抽出し，⑨検索エンジンへ登録を行う．

4. ランドマーク関連ツイート検索

4.1 サービス概要

ランドマーク関連ツイート検索として「周辺ツイート検索」と「話題のスポットランキング」の2種類のサービスを提供している．

「周辺ツイート検索」では，ユーザの現在地に基づいて，周辺の駅や施設に言及しているツイートを検索できる検索サービスである（図1 (c)）．図中に示す「現在地付近のツイートを探す」というリンクを選択することで，ユーザの現在地付近のツイートを検索できる．本サービスではTwitterの特性であるリアルタイム性を活かして，自分の周辺でまさに今起きているホットな情報を検索できる．また，検索結果から次に述べる話題のスポットランキングへ遷移することもできる．

「話題のスポットランキング」は，Twitterで話題になっている観光スポットのエリア別ランキングを提供するサービスである（図1 (d)）．ユーザの現在位置に対応するエリアやユーザが選択したエリアに関する観光スポットランキングを閲覧できる．ランキングページから遷移した施設詳細画面では，該当施設についてのツイートの中から選ばれた注目のツイート，スポットに関連するワード，住所や営業時間などの基本情報を閲覧できる．また，地図アプリやご当地ガイドなどドコモ地図ナビ™＊6サービスのアプリとの連携，該当施設についてのツイート詳細閲覧，施設名称と関連ワードを利用した検索を行うことができる．

4.2 周辺ツイート検索システム

周辺ツイート検索を実現するためには，ツイートに含まれる場所を抽出する必要がある．しかしながら，場所を表すワード抽出の際，駅や施設の名称の中には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名称を考慮する必要がある．例えば，同名で異なる場所を示す例として，円山公園があ

＊6　**ドコモ地図ナビ™**：「ドコモ地図ナビ」およびそのサービスロゴは㈱NTTドコモの商標．

る．円山公園は札幌市と京都市の2カ所に存在する．また，地名以外の意味で使われる名称の例としては，宮城県にある日本三景の1つである松島が挙げられる．場所の特定にはこのような曖昧性を解消する必要があり，以下に述べるツイート登録処理にて対応している．

図4に周辺ツイート検索システムにおける検索エンジンへのツイート登録処理の流れを示す．まず，駅や施設の名称，位置のデータを施設データとしてあらかじめシステム内に保持しておく．検索エンジンには，ツイートに駅や施設の名称があるか判定し，ツイートと位置の関連付けを行ったうえで登録する．上記で述べた曖昧性の解消には，既存研究[1]と同様の処理を行っている．具体的には，同一文書中では曖昧性のある地名は近隣の地名と共起[*7]しやすい（例：京都の「円山公園」と「京都市」は同一ツイートに含まれやすい）という仮定に基づいて判定している．このような近隣の地名は各施設データに紐付ける形で保存している．ツイートと位置の関連付けの処理は，システムがツイートデータを受信した際にリアルタイムに行われる．

次に，検索システムの構成を**図5**に説明する．ユーザが検索する際は，まず移動端末からユーザの位置情報（緯度・経度）をサーバに通知する．次に，ユーザの位置情報を基準点とし，事前に定めた検索半径内にある近隣施設のツイート検索を行う．この段階でツイートが所定の数に満たない場合は，検索半径を拡大して再度検索を行う．これにより検索結果に一定量のツイートを確保できる．しかしながら，半径拡大が頻繁に発生すると検索のレスポンスが遅くなるという問題がある．そこで，本システムではあらかじめ各施設の緯度・経度を検索基準点として一定数のツイートが確保できる検索半径を保存しておく．そして，ユーザの緯度経度を受け取ったサーバでは，近隣施設データ取得半径判定機能で，ユーザから最も近い施設を検索し，保存しておいた検索半径を利用して実際に検索エンジンに問合せを行う．これにより，場所によるツイート数によらず高速に検索でき，かつ，ツイート数を確保した検索が行える．

4.3 話題のスポットランキングのツイート処理

図6に話題のスポットランキングにおけるランドマーク関連ツイート処理を示す．最初の施設情報・ツイート関連付けでは4.2節の周辺ツイート解析システムと同様に，同名地名や地名以外の意味で使われる単語の曖昧性解消を行っている．ここでも施設情報とツイートの関連付けはリアルタイムに行っている．この段階で施設に対してツイートが関連付けられており，ここで施設別にツイート数をカウントする．各施設はご当地ガイドアプリで定めているエリアに分類されており，エリアごとにツイート数の降順にソートすることでエリア別ランキングを生成する．その後，エリア別ランキング上位の

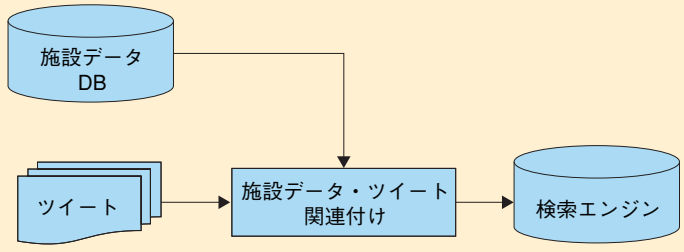

図4　周辺ツイート検索におけるツイート登録処理

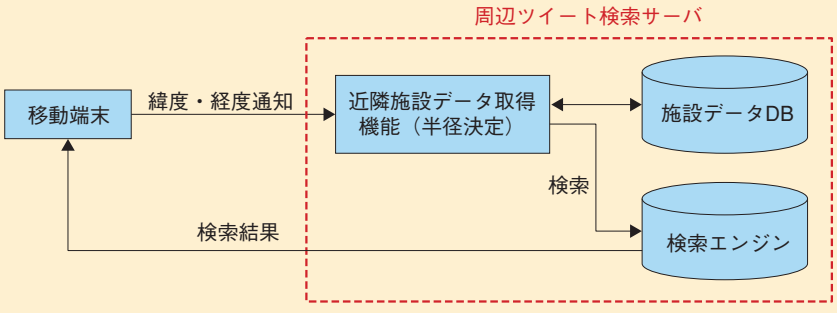

図5　周辺ツイート検索のシステム構成

＊7　**共起**：ある単語とある単語が1つの文章に同時に出現すること．

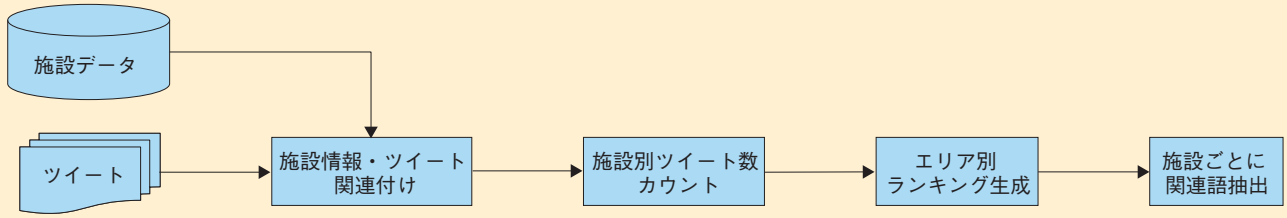

図6　話題のスポットランキングにおけるツイート処理

施設に対して，各施設に関連付けられたツイートから施設名と単語の共起度，単語自体の珍しさを指標として関連語を抽出する．エリア別ランキングの生成および関連語の抽出は一定時間ごとに蓄積されたツイートデータを対象に処理を行う．

5. あとがき

本稿では，日々の生活に役立つツイート（電車運行，TV番組，ランドマーク）を検索できるリアルタイム検索サービスの紹介，ならびに，サービスを支える検索技術について解説した．今回の開発では，各サービスのドメインに合わせて関連ツイートを精度良く抽出可能な検索技術や，目的の情報が探しやすい検索インタフェースに注力し，単なるキーワード検索以上の利便性提供を行った．

今後は関連ツイート抽出技術の精度向上を行うとともに，リアルタイムデータを活用して旬の情報をプッシュで届けるといった，さらに利便性の高いサービスを提供できるよう研究開発を行っていく．

文　献

[1] E. Amitay, N. Har'El, R. Sivan and 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.