時間軸と特徴語によるツイートマッピングシステム

1085041　手塚　悠太

（指導教員　速水 治夫 教授）

1. はじめに

近年，爆発的に普及しているTwitterは特にリアルタイム性に優れ，投稿文字数の少なさから携帯電話やスマートフォンといったモバイル端末からの投稿も多く，モバイル端末からの投稿はその利用特性から地理位置情報が付加されたツイートもされる．そこで，これらのツイートを集積・地図上へ展開する事で地域毎の事象の変遷を記録するのに役立つ．

しかし，集められた地理位置情報が付加されたツイートを単純に地図上に展開しても煩雑となるばかりで実用性を得ることは出来ず，リアルタイムに優れる反面過去のツイートが二次利用されにくい傾向にある．

また，実際にGPS測位情報が付加されたツイートは全体のツイートと比較して非常に少数で有り，特定の場所に関連したツイートにGPS測位情報がないために地図上へマッピングできないといった問題もある．

そこで本論文ではツイート本文を解析することによってそのツイートのGPS測位情報を取得してマッピング可能化すると共に，時間軸を取り入れて期間指定によるツイートの抽出を行い，さらにツイートの文章をそのカテゴリやグループに自動的に分類することで新たな視点からツイートを検索可能にするシステムを提案する．

2. 関連研究

2.1 実世界センサーとしてのTwitterの可能性[1]

この研究はツイートから花粉症に関する情報を収集し，花粉症に関連するツイートの量を都道府県別に集計・色分けして可視化するシステムである．ただし，ツイート自身を可視化するのではなく，花粉症に関するツイートの全体量から花粉症患者数や花粉飛散地域の分布表記を行うシステムである．地理位置情報の取得はGPS測位情報の他，ユーザの現在位置設定から行われる．この現在位置設定はユーザが自由に決定・入力することができる項目で有り，必ずしも正確な位置情報が入力されているとは限らない．

3. 提案システム（サーバ）

本サーバシステムは主に Apache，PHP，MySQL で構成されている．主な機能を以下に示す．

3.1 ツイートデータの取得フェーズ

本システムで利用するためのツイートデータは Twitter API を介して取得を行う．

まず，利用者は自身のツイートを取得するため，Twitter に本システムを登録する．登録を行うと本システムからユーザの各種データにアクセス可能になるため，ユーザのツイート，フレンド一覧，フレンドのツイートを取得する．

データ取得後のツイートについては Twitter Streaming API を介してその都度取得処理を行う．

3.2 形態素解析フェーズ

このフェーズでは取得したツイートデータを形態素解析 API へポストすることでツイートの本文を形態素解析して結果を受け取る．実際の形態素解析部については汎用性の面から本システムとは独立して稼働する Web API として実装を行った．

なお，形態素解析を行う際の辞書データとして Wikipedia キーワード，はてなキーワード，日本郵政の郵便番号データの地名を特徴語として辞書化し，形態素解析システムに組み込んでいる．

3.3 ジオコーディングフェーズ

GPS 測位情報の無いツイートをマッピングするためツイート内に存在する特徴語からジオコーディング処理を行う．ジオコーディング処理は大規模なデータベースが予め必要となるため，今回は Yahoo! Open Local Platform の API を用いてジオコーディングを行う．

4. 提案システム（クライアント）

4.1 特徴語リスト表示

特徴語リスト表示部では，マップ範囲のツイートを解析した結果を表示している．

特徴語の右にチェックボックスを用意しており，true，false によってマップにマッピングされているツイートのフィルタリングを行う．

4.2 マップ表示

マップの表示には Google Map の拡張 API である Android Maps を利用しており，その上にサーバから取得したツイートデータのマッピングを行う．

GPS 測位情報を持つツイートは赤いマーカーで描画され，位置情報が付加された特徴語を持つツイートは緑色のマーカーで描画される．

5. 評価実験

本システムの有効性を調査するため，評価実験を行った．実験は，ツイートマッピング可能化性能，特徴語自身の GPS 測位情報取得可能率，ツイートマッピングにかかる処理時間の評価を行った．その一部を下記に示す．

5.1 GPS 測位情報付加率の調査

本システムによるマッピング可能率の調査のため，ツイートへの GPS 測位情報を調査した，その結果を以下の表 1 に示す．

なお，データベース A は私個人の ID をルートとして 2 ホップ分のユーザのツイートを収集したもので，データベース B は 1 ホップ分のユーザのツイートの中で GPS 付加情報があるツイートや特徴語が含まれているツイートのみを収集したデータベースである．

結果を見ると，全データで 0.62%，実験用データでは 0.84%のツイートが GPS 付加情報を有している事になる．

表 1　GPS 測位情報付加率

DB	全体数	GPS 付加ツイート数	付加率
A	26,029,068	161,125	0.62%
B	436,068	3,650	0.84%

5.2 本システムによるマッピング可能化率

本システムによりツイートデータを解析後のマッピング可能化率を調査した．その結果を以下の表 2 に示す．

表 2　マッピング可能化率

DB	全体数	可能化ツイート数	可能化率
B	436,068	361,521	82.9%

5.3 特徴語のジオコーディング成功率

本システムにより特徴語として選定された語句のジオコーディング成功率を調査した．その結果を以下の表 3 に示す．

表 3　ジオコーディング成功率

DB	全体特徴語数	成功数	成功率
B	81,135	37,503	46.2%

5.4 マッピング可能化ツイートの精度

本システムによりマッピング可能化されたツイートのマッピング精度を調査するため，マッピング可能化済みツイートを 100 件ランダムで取得し，本文で示す場所と解析結果の GPS 測位情報の比較を行った．その結果を下記の表 4 に示す．

表 4　マッピング可能化ツイートの精度

精度	詳細	ツイート数
A	本文中で示した地域へマッピング	37
B	同一地名別地域へのマッピング	8
C	人物名と地名を誤認	22
D	他の地名，固有名詞と誤認	27
E	その他・評価不能	6
合計		100

6. 考察と今後の課題

これらの評価実験により，ツイート本文の解析を行って特徴語を抽出し，それらに基づいてツイートマッピングを行う本システムでは，従来の GPS 測位情報を用いてツイートをマッピングする場合よりもより多くのツイートをマッピング可能となったことを確認出来たが，同じく評価実験の結果から課題も浮き彫りとなって見えてきた．

現状，解析結果のツイートデータをマッピングする場合，37%がおおよそ適切な位置にマッピングされるが，残りのツイートは不適切な位置にマッピングされてしまう．不適切な位置にマッピングされたツイートは視認性や情報整理の妨げとなり，特徴語での自動フィルタリング処理の意義を失ってしまう．

今後の課題として，UI の改善，同意語・同義語を考慮した分類，特徴語判定精度の向上等，より正確な位置へマッピングできるよう取り組んでいく．

参考文献

[1] 高橋哲朗(富士通研)・野田雄也(ニフティ)，実世界のセンサーとしての Twitter の可能性，信学技報，vol. 110，no. 400，NLC2010-38，pp. 43-48，2011 年 1 月．