ソーシャルメディア上に投影された情報の偏在性及び遍在性の可視化

遠山 由自[†,a]　廣田 雅春[††,b]　石川 博[†††,c]　横山 昌平[†,d]

† 静岡大学大学院総合科学技術研究科　†† 大分工業高等専門学校 情報工学科

††† 首都大学東京システムデザイン学部情報通信システムコース

a) *gs15038@s.inf.shizuoka.ac.jp*　b) *m-hirota@oita-ct.ac.jp*

c) *ishikawa-hiroshi@sd.tmu.ac.jp*　d) *yokoyama@inf.shizuoka.ac.jp*

概要　Twitter の位置情報付きツイートにより，Twitter に投稿されている情報は地理的または時間的観点において，局所的または普遍的な分布を持っている．例えば，レストランに関するツイートでは，地理的観点において普遍的にツイートされるものとして全国チェーンのファミリーレストランがある．反対に局所的にツイートされるものとして，地域ごとに存在するご当地レストランがある．しかし，既存の検索システムでは，ある単語が持つこのような分布を把握することは困難である．そこで，本研究では，都道府県単位で Twitter の位置情報付きツイートを収集し，あるクエリとそのクエリが含まれている位置情報付きツイートの名詞との共起度を求めることにより，局所的単語と普遍的単語を識別し，地理的観点からそれらの分布の可視化を行うシステムを提案する．

キーワード　Twitter，ジオタグ，局所性

1　はじめに

近年，Twitter[1] ではユーザがツイートに位置情報を付与することができ，位置情報付きツイートの投稿が増加している．この位置情報付きのツイートにより，Twitter に投稿されている情報は地理的または時間的観点において，偏在性または遍在性を有する．

偏在性を有する情報とは，地理的観点においては地ビールのように特定の地域に集中してツイートされる情報を指し，また，時間的観点においては，特定の季節のみで流行する病気のようにツイートされる情報を指す．一方で，遍在性を有する情報とは，地理的観点においては大手のメーカーのビールのように地域を問わずツイートされる情報を指し，また，時間的観点においては，季節に関わらずツイートさている病気のように期間を問わずツイートされる情報を指す．本研究は，前者のような局所的な分布を持つツイート，後者のような普遍的な分布を持つツイートを効果的に可視化するシステムの構築を行った．

局所的な情報を抽出する研究として，長谷川ら [1]，三木ら [2] の研究が挙げられる．長谷川らは，ある地名が含まれているツイートを収集し，各地名との共起関係に基づき地域特徴語を抽出し，それを利用し Twitter からユーザの観光体験を抽出した．また，三木らは，位置情報付きツイートを用いて，エリアごとに地理的局所性の高いローカル語を決定し，それらを利用して位置情報が付与されていないツイートの発信位置の推定を行った．これらの研究は，局所的な情報である地域特徴語を抽出しているが，それらがどのように分布しているかは考慮していない．

そこで，本研究では，都道府県単位で Twitter の位置情報付きツイートを収集し，クエリとクエリが含まれている位置情報付きツイートの名詞との共起度を求めることにより，局所的単語と普遍的単語を識別し，地理的観点からそれらの分布の可視化を行うことを目標とする．本研究により，既存のクエリ検索では把握が困難である，地理的観点における局所的または普遍的な分布が把握でき，マーケティングや観光支援に繋がると考えられる．

2　予備実験

本実験では，Twitter から収集した位置情報付きツイートに含まれている単語が地理的観点において局所的または普遍的に分布することを検証する．今回の実験では図 1 に示している 7 種類のコンビニを人手で選択し，これらのコンビニの名前が含まれているツイートを地図上にマッピングし，その分布を確認した．データセットとして，2014 年 7 月 8 日から 2015 年 1 月 14 日の間に 555,337 人のユーザから収集した位置情報付きツイート 21,418,979 件を用いた．マッピングした結果，セブンイレブン，ファミリーマート，ローソンが様々な地域に普遍的に分布していることが分かった．一方，局所的に分布していたコンビニとして，図 2 の A に示している主に北海道に分布しているセイコーマートや，図 2 の B の新潟県や群馬県に局所的に存在しているセーブオンがあった．これらのことから，Twitter 上にある単語が地理的観点において局所的または普遍的に分布していることが分かり，今後 Twitter のデータを用いて，地理的観点から局所的単語または普遍的単語の抽出が可能である

[1] http://twitter.com/

図 1 各種コンビニのマーカー

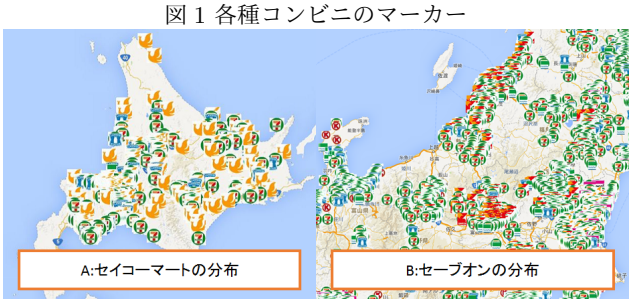

図 2 コンビニの分布

と考えられる．

3 提案システム

本システムでは，Twitter から位置情報付きツイートを収集し，クエリに対する局所的単語と普遍的単語を地理的観点から可視化する．システムの処理はクライアントサイドとサーバサイドの 2 つに分かれる．

クライアントサイドでは，ユーザはシステムに対して「ビール」や「ラーメン」など任意のテキストを入力する．この入力に応じてサーバサイドから生成された局所的単語の分布が地図上に可視化される．

サーバサイドでは，クエリが発行されたら，クエリを含むツイートの名詞を取得する．Jaccard 係数を用いて，クエリ q と取得した名詞 n との共起度 $coop(q,n)$ を求める．

$$coop(q,n) = \frac{T_{q\cap n}}{T_{q\cup n}} = \frac{T_{q\cap n}}{T_q + T_n - T_{q\cap n}} \quad (1)$$

$T_{q\cap n}$ は q と n を共に含むツイートの個数，$T_{q\cup n}$ は q と n のどちらか一方のみを含むツイートの個数，T_q，T_n はそれぞれ q，n を含むツイートの個数を表している．クエリとの共起度が高く，投稿された都道府県が一定数以下である局所的単語を求め，その結果をクライアントサイドに返す．

4 実験

本実験では，キーワードに対して都道府県ごとに共起する単語の中に局所的または普遍的な単語が含まれるか検証を行った．まず，都道府県ごとにあるキーワードが含まれているツイートを取得し，そこから名詞を抽出する．次に，抽出した名詞とキーワードとの共起度を算出し，共起度が高い順に共起単語を並べた．データセットは予備実験と同様である．今回はキーワードとして，「ホームセンター」,「日本酒」と共起する名詞を調べた．

表 1 に都道府県ごとの「ホームセンター」と共起度が高い上位 3 単語の一例を示す．表 1 より,「コーナン」や「コメリ」が様々な都道府県で共起度が高い名詞の上位 3 つに存在していた．このことから,「コーナン」や「コメリ」は全国的に分布していることが分かる．一方，佐賀に「ユートク」，山梨に「オーツル」が共起度が高い名詞の上位 3 つ含まれていたが，その他の都道府県には

表 1 「ホームセンター」と共起度が高い上位 3 単語

都道府県	1 位	2 位	3 位
茨城	コーナン	コメリ	ロイヤル
神奈川	カーマ	コーナン	くろがね
新潟	ムサシ	コメリ	ひらせ
山梨	くろがね	オーツル	センター
三重	カーマ	コーナン	コメリ
大阪	コーナン	ダイキ	ロイヤル
兵庫	ムサシ	コーナン	コメリ
和歌山	ムサシ	コーナン	コメリ
香川	コーナン	NUboard	コメリ
佐賀	ユートク	センター	ホーム
長崎	コーナン	コメリ	センター

含まれていなかった．よって,「ユートク」と「オーツル」は特定の都道府県のみに存在する局所的なホームセンターであることが分かる．また,「日本酒」では，地域ごとの地酒の名前である「一白」,「獺祭」,「鍋島」が「日本酒」と共起度が高い名詞として抽出されていた．これらのことから，あるキーワードに対して共起している名詞の中に，様々な都道府県に存在する普遍的な単語と，特定の都道府県のみに局所的に存在する単語が存在することが分かった．

5 おわりに

本研究では，都道府県単位で Twitter の位置情報付きツイートを収集し，ユーザが入力したクエリと，そのクエリが含まれている位置情報付きツイートの名詞との共起度を求めることにより，局所的単語と普遍的単語の分布を地理的観点において，可視化することを提案した．実験では，位置情報付きツイートを用いて，クエリに対して共起している名詞の中に，局所的単語と普遍的単語が存在することを確認した．

今後は，提案システムの実装を行い，抽出した局所的単語と普遍的単語を地図上に可視化する．また，地ビールの分布など正解のあるデータとの比較を行い，本システムの精度を評価する予定である．

謝辞

本研究の一部は JSPS 科研費 15K00421 および首都大学東京傾斜的研究費（全学分）学長裁量枠戦略的研究プロジェクト戦略的研究支援枠「ソーシャルビッグデータの分析・応用のための学術基盤の研究」によるものです．この場を借りて，深く御礼申し上げます．

参考文献

[1] 長谷川　馨亮，馬　強，吉川　正俊：Twitter からの地域特徴語辞書の構築とその観光情報検索への応用，第 6 回データ工学と情報マネジメントに関するフォーラム，2014.

[2] 三木　翔平，新田　直子，馬場口　登：単語の地理的局所性の経時変化を考慮したツイートの発信位置推定，第 6 回データ工学と情報マネジメントに関するフォーラム，2014.