1P-5

位置情報クラスタリングに基づく地理的ユーザプロファイリング手法

今井 規善†† 奥 健太† 服部 文夫†
† 立命館大学情報理工学部 †† 立命館大学大学院情報理工学研究科

1 はじめに

近年，地理情報検索および地理情報推薦が注目されている [1]．地理情報検索や地理情報推薦は特定のエリアの中から条件に合致したスポットを検索したり，推薦したりするものである．地理情報検索サービスとして，Google マップ* などが挙げられる．ユーザはこれらのサービスを用いることで，地図操作により注目している地域において興味のあるスポットを検索することが可能である． ここで，地理情報検索サービスを利用して，飲食店を探そうとしているアリスのシチュエーションを考える．アリスは居住地の近場で中華料理店を探しており，1 件の中華料理店が見つかったが，その店は休業日であった．このときにアリスに対して次のような飲食店を代替案として推薦できれば有用であると考えられる．

- 注目地点の近場にある別の種類の飲食店
- 注目地点から離れた地域にある中華料理店

前者の場合は，候補も多いと想定できるため，容易に代替案の提示が可能である．一方で，後者の場合は，注目地点から離れすぎている為，アクセスできず提示しても有用でないと考えられる．そこで，アリスが日常的に活動している範囲を推定し，その範囲内にある中華料理店を推薦することで，その推薦がアリスに受け入れられる可能性がある．日常的な活動範囲の推定として，単純に居住地からの距離に基づいて計算する方法が考えられるが，実際には，周辺の道路状況や利用可能交通機関，ユーザの交通手段などの要因があるため，このような要因を考慮に入れて推定することは容易でない．また，ユーザに自身の日常的な活動範囲を手動で入力してもらう方法も考えられるが，ユーザにとって負担となり現実的ではない．

そこで，本研究では，ユーザが日常的な出来事について気軽に発信することを可能にしている Twitter† に着目する．特に，GPS 付きスマートフォンなどの普及に伴い，個人が位置情報を添えてつぶやき（ツイート）を発信することが可能となっている．対象ユーザが日常的に発信している位置情報付きツイートを解析することで，そのユーザの地理的な日常行動範囲が推定できると考えている．本研究では，位置情報付きツイートからユーザの地理的な行動範囲をプロファイル化したものを，地理的ユーザプロファイルとよぶ．

本研究では，対象ユーザがこれまでに発信した位置情報付きツイートに基づき，対象ユーザの地理的ユーザプロファイルを構築する手法を提案する．特に，位置情報だけでなくツイートの投稿日時の情報を活用し，日常的にツイートが発信されている地域や旅行時や出張時など特定の時期のみツイートが発信されている地域の抽出を行う．これらの地域に着目することで，ユーザの日常行動範囲および非日常行動範囲を推定する手法を提案する．ユーザの日常行動範囲を推定することで，前述の例に挙げたような，ユーザの日常行動範囲に合わせた地理情報推薦が可能となる．また，非日常行動範囲におけるツイートを解析する事で，旅行などのユーザの非日常的な興味を抽出する事が可能であると考える．

2 関連研究

藤坂ら [2] は，位置情報付きツイートを用いて地域ごとに人々の通常の活動を定義し，分析対象のデータを比較し特別な活動が発生している地域を検知する．検知手法として，対象地域を決定し K-means 手法によりクラスタ化し，クラスタ内の位置情報付きツイートを時間軸別に分類する．その上で日常状態のデータを学習させ時間軸でツイートを分析した際，ツイートの数が著しく異なる時を異常状態とし，日常状態と異常状態を識別するものである．これに対し本研究では，1 人のユーザの位置情報付きツイートを階層的クラスタリングでクラスタ化する．その上で時間軸別に分類しツイートの集中度を分析し，地理的なユーザプロファイリングを行う事を目的としている．

3 提案手法

本章では対象ユーザがこれまでに発信した位置情報付きツイートに基づき，対象ユーザの地理的ユーザプロファイルを構築する手法について述べる．

Geographical User Profiling Method based on Location-based Clustering

†† Noriyoshi IMAI (is005088@ed.ritsumei.ac.jp)
† Kenta OKU (oku@fc.ritsumei.ac.jp)
† Fumio HATTORI (fhattori@is.ritsumei.ac.jp)
College of Information Science and Engineering, Ritsumeikan University (†)
Graduate School of Information Science and Engineering, Ritsumeikan University (††)

* http://maps.google.co.jp/
† https://twitter.com/

3.1 日常行動範囲と非日常行動範囲

ユーザが日常生活を営むために毎日や毎週定期的に訪れる地理的範囲の事を日常行動範囲と定義する．一方で，非日常的な活動を営む為特定の時期に短期的に訪れる地理的範囲の事を非日常行動範囲と定義する．

3.2 位置情報ベースのクラスタリングに基づく興味領域の抽出

ユーザが発信したツイート集合を位置情報に基づきクラスタリングすることで，ユーザの興味領域を抽出する．本研究では，クラスタリング手法として，階層的クラスタリングを用いる．階層的クラスタリングとは，距離の近いデータ同士を統合することで集合を作成し，さらにその集合同士を統合し，階層的に集合を作り出す手法である．

階層的クラスタリングに基づく，ユーザの興味領域の抽出手順を下記に示す．

1. ユーザ u がこれまでに発信したツイート集合を，$T_u = \{t_{u1}, t_{u2}, \ldots, t_{un}\}$ とする．
2. 初期状態として，1 個のツイートだけを含む n 個のクラスタを作る．
3. 2 個のクラスタ C_i および C_j の間の距離 $d(C_i, C_j)$ を算出し，最も距離の近い 2 個のクラスタを統合する．
4. この統合を，すべてのツイートが 1 個のクラスタに統合されるまで繰り返す．

ここで，2 個のクラスタ間の距離 $d(C_i, C_j)$ は次式により算出される．

$$d(C_i, C_j) = \sqrt{(w_{ix} - w_{jx})^2 + (w_{iy} - w_{jy})^2} \quad (1)$$

ここで，(w_{ix}, w_{iy}) は，クラスタ C_i の重心座標（経度，緯度）を表し，クラスタ C_i に含まれるツイート集合の経度・緯度から算出される．

階層的クラスタリングを実行することで，階層的に K 個のクラスタが作成される．このとき，クラスタ C_i に含まれるツイート集合の重心に対する標準偏差が閾値 Δ 以下となるようなクラスタ C_i を，ユーザの興味領域を表すクラスタとして定義する．

3.3 時間的特徴に基づく行動範囲の抽出

クラスタ C_i に含まれるツイート集合の投稿日時に基づき，クラスタの時間的特徴化を行う．クラスタ C_i については次のベクトルを定義する．

$$\vec{F}^d(C_i) = (c^d(C_i,1), c^d(C_i,2), \ldots, c^d(C_i,365)) \quad (2)$$

ここで，$\vec{F}^d(C_i)$ は，1 年を 365 日として扱ったときの各日のツイート投稿数をベクトルの要素とした，365 次元特徴ベクトルである．このとき，$c^d(C_i, t)$ は，クラスタ C_i において，t 日目にユーザが投稿したツイート数であ る．提案手法では，クラスタ C_i の時間的特徴に基づき，ユーザの日常行動範囲及び非日常行動範囲を推定する．ただし，日常行動範囲及び非日常行動範囲を推定するためには，各日時におけるツイート投稿数よりも，その日時において一つでもツイートが投稿されたか否かの情報で十分である．そこで，式 (2) を次式のように正規化した特徴ベクトルを用いる．

$$\vec{F}^{d*}(C_i) = (c^{d*}(C_i,1), c^{d*}(C_i,2), \ldots, c^{d*}(C_i,365)) \quad (3)$$

ここで，$c^{d*}(C_i, t), c^{w*}(C_i, t), c^{m*}(C_i, t)$ には，それぞれ，各時刻 t において，ツイートが 1 つ以上の場合は 1 が，それ以外は 0 が与えられる．このとき，$\vec{F}^{d*}(C_i)$ において 1 となる要素が多いとき，クラスタ C_i において，ユーザはほぼ毎日ツイートを発信していることを表す．また特徴ベクトルにおいて，1 となる要素の割合が閾値 θ 以上となる場合，クラスタ C_i は日常行動範囲であると定義し，そうでない場合は，非日常行動範囲であると定義する．

4 定性分析

提案手法の妥当性を検証するために，実データに基づく定性的評価を行った．対象ユーザは京都府京都市中京区を出身地とするユーザで京都市周辺のツイートの投稿割合が 365 日の内 261 日投稿されており，0.72 と高い値を示したため日常行動範囲と推定した．一方大阪市周辺では，ツイートの投稿割合が 365 日の内 19 日投稿されており，0.05 と低い値であった為非日常行動範囲と推定した．

5 おわりに

本研究ではユーザの地理的ユーザプロファイルを構築することを目的とし，対象ユーザがこれまでに発信した位置情報付きツイートから日常行動範囲と非日常行動範囲の推定を行った．また，実データに基づく定性的評価を行った．今後は，ユーザの日常行動範囲および非日常行動範囲を識別するための適切な閾値の設定方法などについて検討する．

謝辞

本研究の一部は，文部科学省科学研究費補助金若手研究（B）（研究代表者：奥健太，課題番号 23700132）による．ここに記して謝意を表します．

参考文献

[1] 奥健太，西崎剛司，服部文夫: 地域限定性スコアに基づく位置情報付きコンテンツからの地域限定語句の抽出, 情報処理学会論文誌：データベース, Vol5，No.3（TOD55），pp.97-116，2012.

[2] 藤坂達也，李龍，角谷和俊：集合経験知を用いた地域的ノーマル状態に基づく地域イベントの発見, WebDB Forum 2011, 2G-1-1, 2011.