位置情報付きツイートに基づく 地理的ユーザプロファイリング手法の提案

Proposal for Geographical User Profiling Method based on Geotagged Tweets

今井 規善 [1,a)] Noriyoshi Imai
奥 健太 [1,b)] Kenta Oku
服部 文夫 [1,c)] Fumio Hattori

概要：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も活用することで，日常的にツイートが発信されている地域，旅行時や出張時など特定の時期において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ことで，旅行などのユーザの非日常的な興味を抽出することが可能であると考える．

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][2]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*1や Yahoo!ロコ*2 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できる．

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タイ料理レストランを探していた．すると自動車で 10 分ほどのところに，1 件のタイ料理レストランが見つかったが，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
- 注目地点から多少離れた地域にあるタイ料理レストラン

前者の場合は，候補も多いと想定できるため，容易に代替案を提示することができる．一方で，後者の場合は，注目地点から離れすぎている場合には，アリスはそこにアクセスできず，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することができれば，その範囲内にあるタイ料理レストランを推薦することで，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*3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

[1] 立命館大学
Ritsumeikan University, 1-1-1 Nojihigashi, Kusatsu-city, Shiga 525-8577, Japan
a) is005088@ed.ritsumei.ac.jp
b) oku@fc.ritsumei.ac.jp
c) fhattori@is.ritsumei.ac.jp
*1 http://maps.google.co.jp/
*2 http://map.yahoo.co.jp/maps?
*3 https://twitter.com/

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も活用することで，日常的にツイートが発信されている地域，旅行時や出張時など特定の時期において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ことで，旅行などのユーザの非日常的な興味を抽出することが可能であると考える．

2. 関連研究

利用者の行動履歴やユーザ生成コンテンツなどから，その利用者の嗜好などをプロファイル化するという研究は多く行われている?．特に，近年はマイクロブログの一つである Twitter の普及に伴い，Web 上に膨大なユーザ生成コンテンツが蓄積されている．この Twitter により発信されたコンテンツを解析することで利用者の嗜好プロファイルを構築する研究がある [3].

Hannon ら [3] は，ユーザが発信したツイートから対象ユーザの嗜好を抽出し，ユーザプロファイリングを行っている．この研究では，対象ユーザのツイートの内容に基づきユーザの嗜好抽出を行う手法や，対象ユーザの Follower や Followee の関係に基づき嗜好抽出を行う手法を提案している．この研究に対し，本研究では嗜好抽出を行うのではなく，発信されたツイートの位置情報および時間情報に着目することで，ユーザが日常的・非日常的にどの地域において行動しているかを表す地理的なユーザプロファイルを構築することを目的としている．

藤坂ら [4] は，マイクロブログサービスの一つである Twitter において発信される位置情報付きツイートを情報源とし，そこから地域イベントを抽出する手法を提案している．特に，地域ごとの人々の通常活動状態との差異を考慮することで，通常とは異なる特別な活動が行われている地域を抽出することで地域イベントの抽出を行っている．藤坂らは，地域イベントの抽出を目的としていることから，重要な手掛かりとなる対象地域の時間的変化に着目している．

さらに，位置情報付きユーザ生成コンテンツから，多くの人々に関心をもたれている地点，つまり POI の抽出を行う研究がある．Crandall ら [5] は，flickr[*4] に投稿される位置情報付き写真データに対し，クラスタリング手法である Mean-shift 法を適用することで，多くの人に写真が撮られやすいようなランドマークを抽出し，地図上にそのランドマークの写真を提示するシステムを提案している．Zheng ら [6] は，独自に収集された GPS 軌跡データから，多くの人々が滞在する地点を抽出し，クラスタリングを行うことで，POI の抽出を行っている．

これらの研究が，多数のユーザから発信された位置情報付きユーザ生成コンテンツを集合知的に解析しているのに対し，本研究では，対象ユーザ一人に着目し，その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ている．

[*4] flickr: http://www.flickr.com/

表 1 ツイートテーブルの属性

属性	説明
id	ツイート ID
user_id	ユーザ ID
user_name	ユーザ名
text	ツイートの内容（テキスト）
year	ツイートが投稿された年
month	ツイートが投稿された月
week_of_year	ツイートが投稿された週
day_of_year	ツイートが投稿された年間の日
day_of_week	ツイートが投稿された週間の日
hour	ツイートが投稿された時刻
latitude	ツイートの緯度
longitude	ツイートの経度

3. 位置情報付きツイートの収集

マイクロブログサービスである Twitter が提供している streaming API[*5] を用いて位置情報付きツイートの収集を行い，ツイートテーブルに格納した．ツイートテーブルの属性を表 1 に示す．

以降，ユーザ u が発信したツイート集合を

$$T_u = \{t_{u1}, t_{u2}, \ldots\} \tag{1}$$

と表す．また，各ツイートの属性を指すときには，$t_{u1}.text$ のように，「ツイート . 属性名」の形式で表記する．

4. 地理的ユーザプロファイリング手法

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．

図 1 は，ある一人のユーザが一定期間発信してきた位置情報付きツイートを地図上にプロットしたものである．図のように，ユーザは特定の地域において集中してツイートを発信する傾向にあることがわかる．図 1 の例では，特に地域 A において大量のツイートが発信されており，それ以外では，地域 B，地域 C，地域 D においても集中してツイートが発信されていることがわかる．

さらに，ツイートが発信された時刻に着目すると，地域

[*5] https://dev.twitter.com/docs/streaming-apis

図 1　ユーザが発信した位置情報付きツイートの分布 (例)

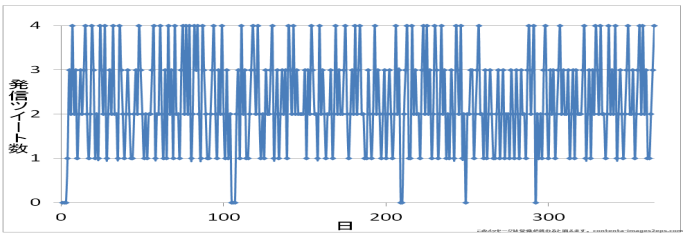

図 2　日常行動範囲における日ごとのユーザーの発信ツイート数

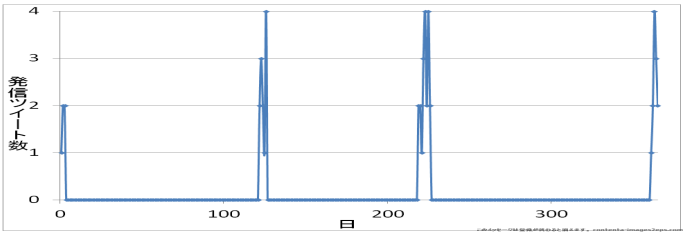

図 3　非日常行動範囲における日ごとのユーザーの発信ツイート数

A においては年間を通してほぼ毎日ツイートが発信されているのに対し，地域 B および地域 C においては，週一や月一など，ある一定の間隔をおいて発信されている．地域 D においては，特定の短い時期においてのみ発信されている．このことから，このユーザは，日常的には，ほぼ毎日，地域 A において行動していること，また週一や月一など定期的に地域 B および地域 C へ訪れていること，地域 D へは旅行など非日常的な行動として訪れていることがわかる．

以上のことより，対象ユーザが発信したツイートの位置情報および投稿日時に着目することで，対象ユーザの行動範囲を推定することができると考えている．本研究では，特に対象ユーザの行動範囲を以下の 2 種類に分類して考える．

(a)　日常行動範囲
(b)　非日常行動範囲

提案手法により構築される地理的ユーザプロファイルには，これら日常行動範囲および非日常行動範囲が含まれているものとする．以降，4.1 節では，日常行動範囲および非日常行動範囲の定義を示す．4.2 節では，地理的なユーザの興味領域を抽出する手法，4.3 節では，日常行動範囲および非日常行動範囲を抽出する手法について述べる．

4.1　日常的行動範囲と非日常的行動範囲

学校や職場への通学や通勤，スーパーへの日用品の買物など，ユーザが日常生活を営むために訪れる地理的範囲のことを日常行動範囲と定義する．日常行動範囲へは，ユーザは毎日や毎週といったようにほぼ定期的に訪れるということがいえる．

一方で，出張先や旅行先など非日常的な活動を営むために訪れる地理的範囲のことを非日常行動範囲と定義する．非日常行動範囲へは，ユーザは特定の時期に短期的に訪れるということがいえる．

頻繁にツイートを発信するユーザを考える．図 2 は，日常行動範囲における日毎のユーザの発信ツイート数を示したものである．図の横軸は時間軸を日単位（1 日から 365 日）で表しており，縦軸はツイートの発信頻度を示している．このように，頻繁にツイートを発信するユーザは，日常行動範囲においては，ほぼ毎日ツイートを発信していることがわかる．一方で，図 3 は，ある非日常行動範囲におけるツイートの発信頻度を示している．このように，非日常行動範囲においては，特定の時期に偏ってツイートを発信していることがわかる．

以上のことから，各行動範囲においてツイートの投稿日時の分布を調べることで，その行動範囲が日常行動範囲であるのか，非日常行動範囲であるのかを推定することができると考える．次節以降，まずユーザの発信ツイート集合からユーザの興味領域を抽出し，その興味領域が日常行動範囲であるか非日常行動範囲であるかを推定する手法について述べる．

4.2　位置情報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．

(a)　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．
(b)　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．
(c)　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．
(d)　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．

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

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．

4.3 時間的特徴に基づく日常・非日常行動範囲の抽出

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．ここでは，各日時におけるツイートの投稿数をベクトルの要素とした多次元特徴ベクトルにより，クラスタの時間的特徴を表現する．具体的には，クラスタ C_i について，次のベクトルを定義する．

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

$$\boldsymbol{F}^w(C_i) = (c^w(C_i, 1), c^w(C_i, 2), \ldots, c^w(C_i, 53)) \quad (4)$$

$$\boldsymbol{F}^m(C_i) = (c^m(C_i, 1), c^m(C_i, 2), \ldots, c^m(C_i, 12)) \quad (5)$$

ここで，$\boldsymbol{F}^d(C_i)$ 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．同様に，$\boldsymbol{F}^w(C_i)$ は，1 年を 53 週として扱ったときの各週のツイート投稿数をベクトルの要素とした，53 次元特徴ベクトルである．1 月 1 日が 1 週目，12 月 31 日が 53 週目に相当する．このとき，$c^w(C_i, t)$ は，クラスタ C_i において，t 週目にユーザが投稿したツイート数である．さらに，$\boldsymbol{F}^h(C_i)$ は，1 年を 12 カ月として扱ったときの各月のツイート投稿数をベクトルの要素とした，12 次元特徴ベクトルである．このとき，$c^m(C_i, t)$ は，クラスタ C_i において，t 月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および非日常行動範囲を推定する．ただし，日常行動範囲およ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 (3), (4), (5) をそれぞれ次式のように正規化した特徴ベクトルを用いる．

$$\boldsymbol{F}^{d*}(C_i) = (c^{d*}(C_i, 1), c^{d*}(C_i, 2), \ldots, c^{d*}(C_i, 365)) \quad (6)$$

$$\boldsymbol{F}^{w*}(C_i) = (c^{w*}(C_i, 1), c^{w*}(C_i, 2), \ldots, c^{w*}(C_i, 53)) \quad (7)$$

$$\boldsymbol{F}^{m*}(C_i) = (c^{m*}(C_i, 1), c^{m*}(C_i, 2), \ldots, c^{m*}(C_i, 12)) \quad (8)$$

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，少なくとも一つのツイートが投稿されている場合は 1 が，そうでない場合は 0 が与えられる．

このとき，$\boldsymbol{F}^{d*}(C_i)$ 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．同様に，$\boldsymbol{F}^{w*}(C_i)$ または $\boldsymbol{F}^{m*}(C_i)$ 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，ユーザはほぼ毎週または毎月ツイートを発信していることを表す．そこで，いずれかの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．

5. おわりに

本研究では位置情報付きツイートから，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した．今後，第 4 章で述べた手法において，適切な閾値 Δ および θ の設定方法について検討し，実際の位置情報付きツイートを用いた評価実験を行う．

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)「コンテキスト限定価値を考慮した情報推薦方式」（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] K. W.-ting Leung, D. L. Lee, and W.-chien Lee, “CLR : A Collaborative Location Recommendation Framework based on Co-Clustering Categories and Subject Descriptors,” in SIGIR ’11:Proceedings of the 34th international ACM SIGIR conference on Research and development in Information, 2011, pp. 305–314.

[2] M. Ye, P. Yin, W. C. Lee, and D. L. Lee, “Exploiting geographical influence for collaborative point-of-interest recommendation,” in SIGIR ’11:Proceedings of the 34th international ACM SIGIR conference on Research and development in Information, 2011, pp. 325–334.

[3] J. Hannon, M. Bennett, and B. Smyth, “Recommending twitter users to follow using content and collaborative filtering approaches,” in Proceedings of the fourth ACM conference on Recommender systems, 2010, pp. 199–206.

[4] 藤坂達也, 李龍, 角谷和俊, “Twitter ユーザの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,” in WebDB Forum 2010, 2010.

[5] D. Crandall, L. Backstrom, D. Huttenlocher, and J. Kleinberg, “Mapping the World ’s Photos,” in WWW ’09:Proceedings of the 18th international conference on World wide web, 2009.

[6] V. W. Zheng, Y. Zheng, X. Xie, and Q. Yang, “Collaborative Location and Activity Recommendations with GPS History Data,” in Proceedings of the 19th international conference on World wide web, 2010, pp. 1029–1038.