重力モデルとTF-IDFを用いたジオタグ付きTwitterデータからの観光地抽出と魅力の評価

前田 高志ニコラス[†,a]　吉田 光男[††,b]　鳥海 不二夫[†,c]　大橋 弘忠[†,d]

† 東京大学大学院工学系研究科　†† 豊橋技術科学大学

a) *maeda@crimson.q.t.u-tokyo.ac.jp*　b) *yoshida@cs.tut.ac.jp*　c) *tori@sys.t.u-tokyo.ac.jp*　d) *ohashi@sys.t.u-tokyo.ac.jp*

概要　本研究では，ジオタグ付き Twitter データから観光地抽出および各観光地の魅力推定の手法を構築する．人文地理学で用いられる重力モデルを改良し，Twitter データから得られた地域間の移動数と移動距離から目的地の魅力を推定する．さらに，各目的地について，その地点で投稿されたテキスト情報をもとに，TF-IDF 値の分布傾向から，そこが単に「便利な場所」であるか,「他にはない固有の魅力を持つ場所」であるかを峻別する．

キーワード　時空間データマイニング，位置情報，ジオタグ，Twitter，観光地推薦，重力モデル，TF-IDF

1　はじめに

近年，地方経済の疲弊や高齢化，震災後の復興に向けた動きの中で，地方都市の存立にとって，人の活発な地域間移動が重要であるという認識が広まってきた．中でも観光目的による人の流入は地方経済の重要な要素であり続けている．一方，観光庁などにより Twitter 位置情報を利用した観光情報抽出に注目が集まっており，Twitter 位置情報を用いた研究も増加している．しかしながら，現状はホットスポットの抽出や人気の経路推薦など，地理的に非常にミクロな対象に絞った物が多い．また，観光者の住所を考慮した研究はなされていない．

これまで，人の住んでいる場所と移動先の関係を考慮したものとして，人文地理学の領域では，人や物や情報の空間的フローを説明する空間的相互作用モデルの構築がなされてきた [2]．特に歴史が深いものはニュートンの万有引力を用いた重力モデルである．これは，2 都市間の流量が両都市の規模の積に比例し，距離に反比例するとしたモデルである．

空間情報学の領域では，Phithakkitnukoon ら [6] が，大規模な携帯電話の位置情報をもとにして，旅行者の行動を詳細に分析することに成功している．この研究では，旅行者の旅行頻度，移動距離，目的地，出発地，移動手段，現地での滞在時間の関係性を個人レベルで抽出し，その傾向の分析を行った．

一方，Twitter を利用した移動研究としては，観光ではなく日々の移動に関する研究ではあるが，若宮ら [5] が，位置情報付き Twitter データとパーソントリップ調査のデータをもとに，群衆の移動傾向を分析している．この研究では，地域間の移動について，移動距離・移動時間・移動量の 3 つの値をもとに各地域間の直感的な近接性を多次元尺度構成法により示すことに成功している．

これまでに述べたように，様々な先行研究はあるものの，Twitter 位置情報を利用した住所と観光地に関する広域な研究はなされていない．本研究は位置情付き Twitter データから得られる地域間移動情報から人の移動モデルを構築し,「各地の魅力」「距離が移動件数に与える影響」を算出する．また，Twitter のテキスト情報により各地域の魅力の原因がその地域固有の魅力に基づくものなのか，利便性に基づくものなのかを判定する．最終的にこの情報をもとに旅行者への，住所に応じた観光地推薦技術に活かす．

本研究は以下の手順によって前記の目的を達する．

1. まず，Twitter データから地域間移動数を抽出するため，各ユーザの居住地域と旅行・おでかけ目的の移動先の推定を行い，全ユーザについて集計を行う
2. 地域間移動数と目的地の魅力の間に成り立つモデルを立て，地域の持つ魅力を算出する
3. テキスト情報から,「他の地域にない魅力」を持つ場所を推定し，単に「便利な場所」と区別することで観光地を抽出する
4. 最終的に，居住地域に応じた観光地推薦を行う

2　各ユーザの居住地と移動先の分類

人の移動はもっぱら自宅や職場，学校といった特定の地点を日々往復する移動に占められている．Ester ら [3] の DBSCAN(Density Based Spatial Clustering Algorithm with Noise) を用いた，人の重要地点の探索方法が複数考案されている．本研究では，DBSCAN を Twitter 分析用に改良したアルゴリズムによって，各ユーザの居住地と移動先を特定する．

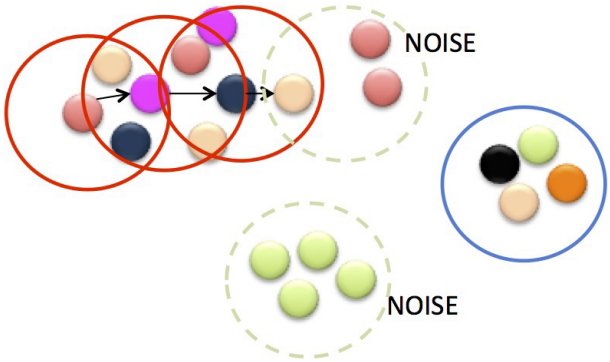

図 1　DBSCAN(異なる色は異なる日の Tweet であることを示す)

1. 一人のユーザに関し，期間内の全 Tweet の位置を抽出する
2. その中からどれか 1 点を取り出し，半径 4km 以内に 4 日分以上の Tweet がなければ，その点を Noise とみなす
3. 4 日分以上の Tweet があれば，それらを同じクラスタとみなす
4. クラスタ内の別の点が同様に半径 4km 以内に 4 日分以上の点を含むなら，それらの点も同じクラスタとする
5. クラスタを形成する点は高頻度滞在地とし，Noise となった点は低頻度滞在地とする

ここで，最も多くの日数の Tweet を持つクラスタの重心をそのユーザの居住地とし，そこから低頻度移動地を結んだものをそのユーザの地域間移動とする．4 日をしきい値としたのは，国土交通省観光庁の観光白書 [8] において，日本人の 1 回あたりの旅行の宿泊日数が 2.1 泊とあることから，4 日間同じ場所に滞在する旅行は稀であるという考えに基づいて設定した．距離に関するしきい値 4km に関しては暫定的な設定であり，期間のしきい値も含め，最適なしきい値の獲得及び検証は今後の課題とする．

3　出発地・到着地のクラスタリング

前述の手順で数多くの出発地と到着地が得られるが，空間的に近しいものをひとつにまとめたい．そこで Mean Shift Clustering[1] を用いる．Mean Shift Clustering とは以下のような，漸次的な手順によって近しい点同士をひとつにまとめる手法である．各ステップごとに各点が次に移る先の点を，自身を含めた近傍半径 r_n の円内に含まれるすべての点の重心とする．すべての点について，近傍半径 r_n の円内の点がそれよりさらに小さい収束半径 r_c の円内に収まれば，そこでこの処理を終了する．同じ収束半径に収まった点同士を同じクラスタとしてまとめ，各クラスタの重心をそのクラスタの代表点とする．

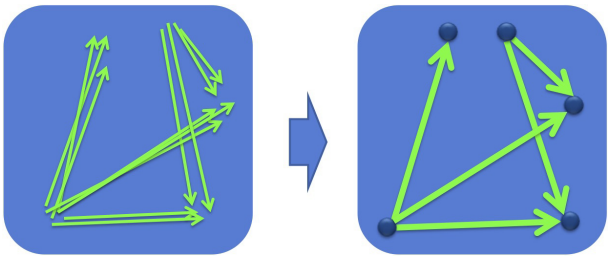

図 2 出発地・到着地のクラスタリング

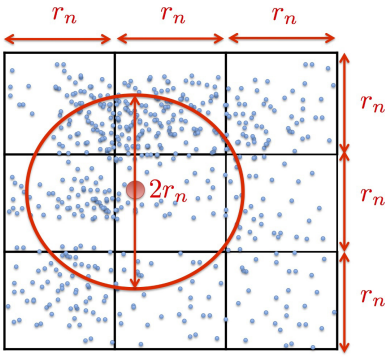

図 3 Mean Shift Clustering の効率化

本研究では，出発地のクラスタリングについては，r_n を $5km$，r_c を $1km$ とし，到着地のクラスタリングについては，r_n を $2km$，r_c を $1km$ とした．また，毎ステップですべての点同士の距離を求めると計算時間が膨大になるため，不要な計算を省く工夫をする．緯度・経度を等間隔で区切ったメッシュを張り，それらの各メッシュの縦・横の長さが r_n よりも長くなるようにする．このようにすれば，各点について，その点が含まれるメッシュと近傍の 8 メッシュ内の点以外は r_n 以上の距離にあるため，距離の計算をせずにすむ．これによって計算時間の短縮をはかる．なお，経度 1 秒あたりの長さは赤道から離れるほど短くなるため，本研究では日本最北端の宗谷岬の緯線における経度が r_n となる単位でメッシュを形成する．

4　モデル化と指標値の定義

4.1　概要

本研究では，2 地点間の移動件数は下記 4 つの値によって決まると考え．到着地の魅力を推定するために，これらの指標値が互い持つ関係性をモデル化する．

- 出発地の放出力

  出発地の放出力が高ければ，その地点からの移動件数と移動距離が大きくなる．

- 到着地の魅力

到着地の魅力が高ければ，その地点への移動件数と移動距離が大きくなる．

- 移動コスト

  ２地点間の移動コストが高ければ，その区間の移動件数が減る．本研究では，移動距離をコストの指標値として用いる．

- 到着地の競合

  出発地の周りに多くの魅力ある到着地があれば，それぞれの到着地への移動件数が分散する．

4.2 目的地選択のモデル

出発地点 s に存在するユーザが，数ある目的地から目的地 e を選択する確率 $P(s \to e|s)$ を以下の式で表す．

$$P(s \to e|s) = \frac{A_e}{D_{se}^{\alpha}} \Big/ E_s \tag{1}$$

- A_e：目的地 e の持つ絶対的な魅力（未知変数）
- D_{se}: 出発地 s と到着地 e の間の距離（既知変数）
- α: 距離が目的地の魅力に与える影響を決定する係数（未知変数）
- E_s: 出発地 s の周囲の魅力の総和（未知変数）

  $E_s = \sum_k \frac{A_k}{D_{sk}^{\alpha}}$ によって与える．

出発地点 s に存在するユーザが，数ある目的地から目的地 e を 選択する確率 P は，実データによる観測値を用いると下記のように表すことができる．

$$P_{obs}(s \to e|s) = \frac{T_{s \to e}}{\sum_k T_{s \to k}} \tag{2}$$

- $T_{s \to e}$：データで得られた，出発地 s から目的地 e へ移動した移動件数（既知変数）

4.3 重回帰分析による指標値の算出方法

式 (1) と式 (2) を等号で表し，両辺の対数をとり，重回帰分析が可能なように式変形を行う．

$$\log\left(\frac{T_{s \to e}}{\sum_k T_{s \to k}}\right) = \sum_i x_i \cdot \log A_i - \alpha \cdot \log D_{se} - \sum_i y_i \cdot \log E_i + c \tag{3}$$

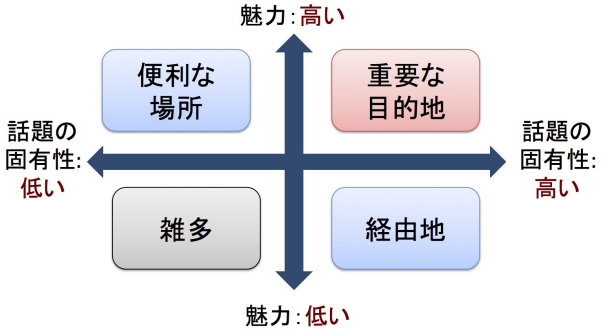

図 4 地域の魅力と固有性

ここで，左辺は被説明変数であり，データから求めることができる．説明変数は x_i，D_{se}，y_i であり，係数は $\log A_i$，α，$\log E_i$ である．また，c は定数項である．説明変数の D_{sd} は緯度経度から GRS80 楕円体をもとにした計算によって求める．また，x_i は入力データの目的地が e であるとき，$i = e$ である x_i を 1 とし，それ以外を 0 とする．同様に y_i は入力データの出発地が s であるとき，$i = s$ である y_i を 1 とし，それ以外を 0 とする．

Mean Shift Clustering をもとに出発地・到着地をクラスタリングしたのちに，各 2 地点間の移動件数を算出し，式 (3) に代入して，重回帰分析により，各目的地の魅力（A_e），距離が魅力に与える係数（α），各出発地の周囲の魅力（E_s）を求める．

5 テキスト情報による到着地の固有性判定

5.1 地域の魅力と話題の固有性

移動モデルによって各地の魅力を求めることが可能であるが，その魅力がその土地の持つ固有性に起因するのか，あるいは単純に利便性があるからだけなのかを判断する必要がある．ここでは，Twitter のテキスト情報を用いて，ユーザたちの Tweet からその土地の固有性に着目した投稿が多いのか，あるいは，ほかの多くの地域でも話される話題が多く占めているのかに注目することで，固有性の判定を行う．これにより，図 4 の第 1 象限のように魅力が高く話題の固有性も高い場所は，その土地の持つ固有の魅力に人が惹かれて訪れていることを示す．魅力が高く話題の固有性が低い場所は，大規模店舗などの利便性の高い場所を示すと考えられる．魅力が低く，話題の固有性が高い場所は経由地として一時滞在する場所を示すと考えられる．

5.2 テキスト情報に注目した固有性の算出

土地の固有性を評価するために，Twitter 投稿記事のテキスト情報を用いる．各到着地において投稿された Tweet のうち，各ユーザの低頻度移動地のものを集め，それをまとめてひとつの文書とする．文書内での出現回数が多い単語が，他の文書でほとんど出現しなければ，

その地域は他にはない魅力を持っていると考えられる．例えば，富士山の周辺では「富士山」「御来光」「〜合目」という単語が多く，これは他の場所でほとんど出現しない．逆に「ご飯」「テレビ」「サッカー」といったありふれた単語が主要となる地域では，固有の魅力をあまり持たないと考えられる．

関連研究として，ジオタグの付与されていない Twitter 投稿記事のテキスト情報から投稿時の地理的位置を推定する Cheng ら [4] の研究がある．これらの研究では単語の地理的な局所性に注目して投稿時の地理的位置を推定している．三木ら [7] は，単語の地理的局所性を算出するために，ジオタグのついた投稿記事から各場所ごとの単語の TF-IDF を算出している．

本稿は地理的局所性の高いローカル語が投稿される割合が多い場所とそうでない場所を判定することで，その土地の固有性を評価する．ここでは，三木ら [7] と同様に，キーワードの重要度を表現する TF-IDF を指標に用いる．これは主に文書を特徴づける単語に高い値を割り振るものであり，文書の要約や同ジャンルの文書のクラスタリングに用いられる．TF-IDF は TF(Term Frequency: 単語の出現頻度) と IDF(Inverse Document Frequency: 逆文書頻度) の積によって求める．

$$tfidf_{w_i,d} = tf_{w_i,d} \cdot idf_{w_i,d} \tag{4}$$

$$tf_{w_i,d} = \frac{N_{w_i,d}}{\sum_k N_{w_k,d}} \tag{5}$$

$$idf_{w_i,d} = log\frac{|D|}{|d : d \ni w_i|} \tag{6}$$

ここで，$N_{w_i,d}$ は文書 d に含まれる単語 w_i の出現回数を示す．$|D|$ は全文書数を示す．

地域の固有性が高い場所では，ユーザが他の地域にない，その地域固有の単語を多く発し，地域の固有性が低い場所ではその逆となると考えられる．このため，前者の地域の文書（その地域の全 tweet のテキスト情報の結合）では，一部の単語の TF-IDF が極端に高くなり，その他大部分の単語との落差が大きくなる．逆に地域の固有性が低い場所では，文書内の TF-IDF の分布はよりなだらかなものとなる．このためここでは，各文書の TF-IDF が高いものから上位 10 %のものの総和をその地域の固有性と定義づける．

6 実データによる計算結果

6.1 有効ユーザ数・決定係数

2014 年 4 月から 2015 年 3 月にかけて，1ヶ月ごとに計算を行った．表 1 は各月の有効ユーザ数と式 (3) の重回帰分析の決定係数を表す．ここで有効ユーザ数とは DBSCAN によって，高頻度滞在地と低頻度滞在地の両方を持つユーザの数を示す．

表 1 有効ユーザ数と決定係数

	有効ユーザ数	決定係数
2014/4	81115	0.754117611
2014/5	79870	0.754309567
2014/6	86167	0.779786745
2014/7	93809	0.774493922
2014/8	107418	0.728441143
2014/9	95723	0.737483792
2014/10	85012	0.753256551
2014/11	83743	0.742581565
2014/12	106951	0.742476591
2015/1	105444	0.743801537
2015/2	99846	0.761829237
2015/3	124954	0.737153887

表 2 距離が移動件数に与える影響

	距離の累乗係数
2014/4	0.930512480
2014/5	0.945892090
2014/6	0.915936481
2014/7	0.907708275
2014/8	0.864301811
2014/9	0.940809110
2014/10	0.916424284
2014/11	0.947433676
2014/12	0.912618666
2015/1	0.854242589
2015/2	0.853498501
2015/3	0.894034621

6.2 距離が移動件数に与える影響

式 (1) の距離が目的地の魅力に与える影響を決定する係数 α は表 2 の通りであった．すべて 0.85〜0.95 の範囲に収まった．これが指し示すのは，出発地から近距離にある到着地は距離が増えるに伴って大きくその魅力を減じるが，遠距離にある到着地同士を比べる場合，到着地の魅力への距離の影響は比較的緩やかになることを意味する．

6.3 地域の魅力と固有性

次に，2014 年 8 月分のデータについて，移動モデルに基づいた地域の魅力と，テキスト情報に基づいた話題の地域的固有性の値を用いて，地域のクラスタリングを

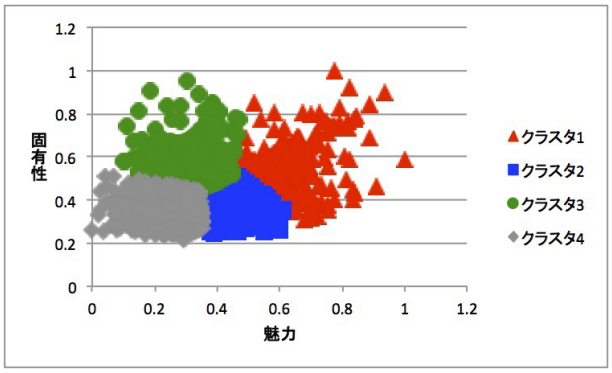

図 5　魅力と固有性に基づく地域のクラスタリング (2014 年 8 月)

表 3 各クラスタが含む地域数 (2014 年 8 月)

	地域数
クラスタ 1	353
クラスタ 2	1054
クラスタ 3	354
クラスタ 4	890

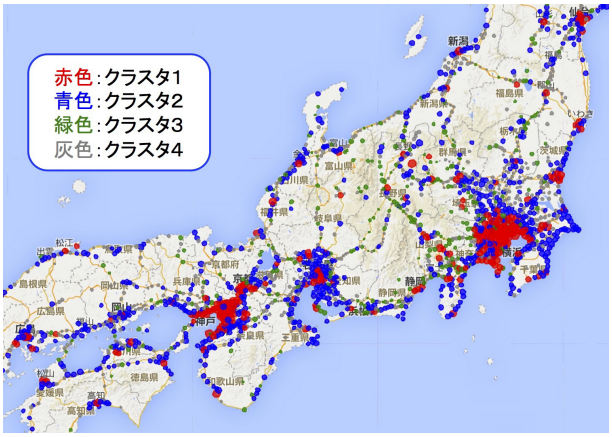

図 6 各クラスタの地図上の分布 (2014 年 8 月)

行った．クラスタリングの方法は，魅力，固有性のそれぞれについて最大値で値を割ったものを使用して，2 次元空間上に位置づけ，それを k-means 法を用いて 4 つのクラスタに分割した．その結果を図 6，表 3，図 7 に示す．

各クラスタの分類は概ね図 5 の通りのものとなった．クラスタ 1 はディズニーランド，ユニバーサル・スタジオ・ジャパン，沖縄美ら海水族館，京都伏見稲荷大社などの観光地が多く含まれたほか，この期間内に行われたロック・フェスティバルの地域が含まれていた．クラスタ 2 はイオンを含む，地方の大型商業施設，ショッピングモール，大規模店舗が多く含まれていた．クラスタ 3 はサービスエリア・パーキングエリアのような高速道路の中継地点が多く含まれており，地図上でも高速道路沿いに多かった．

7　おわりに

本研究では，Twitter 位置情報を用いて，人の高頻度滞在地と低頻度移動地を求め，移動件数と距離から地域の魅力を算出した．また，テキスト情報から地域の固有性を算出した．そして，地域の魅力と固有性の値を用いて，地域を 4 つのクラスタに分けた．その結果，魅力の高い地域の中から，その地域が持つ固有性によって魅力が高まっている場所を抽出し，利便性によって魅力が高まっている場所と区別することができた．

今後は次の課題について取り組む．まず，本研究で考案した手法が日本以外でも応用可能であるかを検証するため，他国のデータを用いた検証を行う．次に，魅力の原因が季節性のものか，通年性のものかを判断することができるようにし，季節性の点を加味したユーザ推薦ができるようにする．最後に，各住所に応じた観光地推薦をし，被験者実験によってその妥当性を検証する．

参考文献

[1] Fukunaga, K., Hostetler, L.: The estimation of the gradient of a density function, with applications in pattern recognition, IEEE Transactions on Information Theory, 1975

[2] 石川 義孝: 空間的相互作用モデル—その系譜と体系, 地人書房, 1988

[3] Ester, M., Kriegel, H., Sander, J., Xu, X.,: A density-based algorithm for discovering clusters in large spatial databases with noise, in: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, 1996

[4] Cheng, Z., Caverlee, J., Lee, K.,: You are where you tweet: a content-based approach to geo-locating twitter users, in: Proceedings of the 19th ACM International Conference on Information and Knowledge Management, 2010

[5] 若宮 翔子, 李 龍, 角谷 和俊: 位置ベース SNS を通した群衆の移動経験に基づく都市空間の近接性分析, 情報処理学会論文誌, 2013

[6] Phithakkitnukoon, S., Teerayut Horanont, T., Witayangkurn, A., Siri, R., Sekimoto, Y., Shibasaki, R.: Understanding tourist behavior using large-scale mobile sensing approach: A case study of mobile phone users in Japan, Pervasive and Mobile Computing, 2014

[7] 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

[8] 国土交通省観光庁: 平成 27 年版観光白書, http://www.mlit.go.jp/common/001095743.pdf,2015