相互再帰的クラスタリングによる場所と来訪者のベクトル表現

三浦　理緒† 　加藤　大受†† 　遠藤　雅樹†††,†††† 　廣田　雅春††††† 　荘司　慶行†
石川　博†††

† 首都大学東京　システムデザイン学部　〒 191–0065 東京都日野市旭が丘 6–6
††† 首都大学東京大学院　システムデザイン研究科　〒 191–0065 東京都日野市旭が丘 6–6
†† ウイングアーク１ｓｔ株式会社　〒 150–0031 東京都渋谷区桜丘町 20-1 渋谷インフォスタワー
†††† 職業能力開発総合大学校　基盤ものづくり系　〒 187–0035 東京都小平市小川西町 2–32–1
††††† 大分工業高等専門学校　情報工学科　〒 870–0152 大分県大分市大字牧 1666
E-mail: †miura-rio@ed.tmu.ac.jp, ††kato.d@wingarc.com, †††endou@uitec.ac.jp, ††††m-hirota@oita-ct.ac.jp,
†,†††{y_shoji, ishikawa-hiroshi}@tmu.ac.jp

あらまし　本稿では，類似した場所や訪問客をまとめ，任意の長さのベクトルとして表現する手法を提案する．場所や人の特徴を表現する際,「渋谷は若者の街である」,「彼は秋葉原に行くようなオタクだ」というように，場所を人で，人を場所で表現する場合がある．本手法ではこれに倣い，1) ある場所の性質はその場所を訪れた人たちによって決まる，2) ある人の性質はその人が訪問したことのある場所によって決まる，という 2 つの仮定に基づき，相互再帰的にクラスタリングを行なう手法を提案した．Twitter のデータを用いた被験者実験により，同じ性質を持った場所同士が正しくクラスタリングされるかを評価した．

キーワード　ジオタグ，ベクトル，相互再帰，クラスタリング

1. はじめに

近年，わが国の景気の低迷や，人々の消費活動の停滞を受けて，マーケティングの重要性が高まっている (注1)．人々の消費活動を促進するような出店やマーケティングなどを実現するために，場所や人の特徴を分析することが重要であると考えられる．たとえば,「高円寺はどんな街であるか」など場所に合わせた新規出店であったり,「新宿の東側と西側は違うのか」といった地域性の分析であったり，場所の性質は様々な場面で必要とされる．また,「古着好きな学生にはこのジャケットがおすすめ」といったアイテム推薦や,「主婦が多いので日用雑貨を店頭に置く」といった購買促進において，人の性質を分析することは必要不可欠である．近年では，こういった目的の分析のために，サポートベクターマシンなどの機械学習手法や，情報推薦技術が用いられている．これらの手法では，オブジェクトを任意長の特徴ベクトルとして表現し，分類や類似度計算を行なう．人や地域を適切に特徴量化することは，マーケティングや地域分析を行なううえで，大きな課題の一つになりつつある．

場所や個人を分析する上で，有用な情報源の一つとして，ソーシャルメディアが挙げられる．スマートフォンや IoT（Internet of Things）の普及により，位置情報が付与されたソーシャルメディアの情報が著しく増加している．これらの情報は，従来の，政府や企業が行なうアンケート調査による情報に比べ，より，人々の意見や感情を直接的に，かつリアルタイムに反映している．また，得られるデータ量も大規模であり，同等の量のデータをアンケート調査によって収集するのに比べてコストが低い．一方で，ソーシャルメディアの情報は，プロフィール，テキストの有無など，サービスによって得られるデータが異なる．そのため，データ量が多くても，それらを統合的に扱うことは現時点ではできていない．豊富な情報源を用いた分析を実現するためには，様々なサービスから得られる情報を統合的に扱うことができる，地域と人の特徴化の手法が必要である．また，日本におけるインバウンドの増加や，2020 年の東京オリンピック開催に向けて，外国人向けのマーケティングがの重要性が増しつつあり，外国人ユーザのデータを対象とした分析も必要になっている．しかし，外国人ユーザのデータでは，プロフィールやテキストのほとんどは外国語で記述されており，言語の種類も多いので，それらの複数の言語を同時に解析するコストは高く，実現が困難である．そのため，外国人ユーザのデータを対象とした場合においても適用可能である，言語依存性の低い分析手法が必要である．

場所や人を表現する際，様々な要素に基づく，様々な表現が存在しうる．たとえば，場所をランドマークや用途などに基づいて表現したり，人をその趣味や見た目に基づいて表現したりする．その様々な要素に基づく表現のなかで，次のような 2 つの表現に注目する.「渋谷は若者の街である」や,「新橋はサラリーマンの街である」などと言うように，その場所によく訪れる人に基づいて，場所を表現することがある．逆に，人の説明をする際,「彼は秋葉原に通うようなオタクである」や,「彼女は丸の内系のオフィスレディだ」などと言うように，特徴的な場所を挙げて表現することもある．このような場所と人の相互表

(注1)：日本銀行　消費活動指数：https://www.boj.or.jp/research/research_data/cai/index.htm/

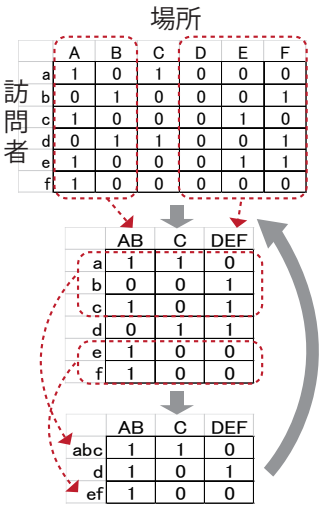

図 1　相互再起クラスタリング

現を，ソーシャルメディアから得られるデータに応用することで，地域と人を相互再帰的に特徴化することができると考えられる．つまり，

- ある場所の性質はその場所を訪れた人たちによって決まる，
- ある人の性質はその人が訪問したことのある場所によって決まる

という仮定を置くことができる．このような仮定を置くことで，相互再起処理によって人を場所，場所を人で表せる．

このような場所と人物の相互再帰的な表現手法の検証のために，ソーシャルメディアの一つである Twitter(注2) のデータを用いて，場所を分析する実験を行った．まず，東京都 23 区内で投稿された位置情報付きツイートを用いて，人がいつ，どこに行ったかを表すベクトルを生成する．次に，そのベクトルをクラスタリングによって似た場所に行っている人のベクトル同士をまとめる．まとめたベクトルから，場所クラスタをそこに来た人で表したベクトルを生成する．そして，クラスタリングにより似た人が来ている場所のベクトルをまとめる．これらを図 1 に示すように，相互再帰的に繰り返すことで，似た人が来た場所のクラスタリングを行なう．

このような相互再帰による手法には，いくつかの利点が挙げられる．第一に，位置情報しか必要としないため，多くのサービスを統合的に取り扱うことができる．地理情報サービスには，写真共有ソーシャルメディアやレストランへのチェックインなど様々なものがある．サービスの種類によって使える情報には差があるが，地理情報サービスであれば位置情報は必ず含まれる．そのためサービスの種類を問わず，横断的に情報を使うことができる．また，使う情報が位置情報と時間情報のみであるため，言語依存性が低いことが挙げられる．地理情報サービスは国際的な利用のされ方が多く，海外旅行などのデータを多く含む．言語に依らない手法にすることで，それらのデータの分析を容易にする．最後に，手法が単純で場所と人それぞれの処理に分かれているため，その都度手法に工夫を行うことが容易である．例えば投稿頻度による重み付けの方法を変えたり，人と場所でそれぞれ異なるクラスタリング手法を適用するなどの拡張手法が作成しやすい．

本論文の構成は次の通りである．まず 2. 章で関連研究について述べる．次に，3. 章で，本研究で提案する手法を述べ，そして 4. 章では，実際のデータを用いて提案手法の性能を評価する実験を行い，結果を示し 5. 章で考察を述べる．最後に，6. 章で，本研究のまとめと今後の課題を述べる．

2. 関 連 研 究

Web 上の情報を用いて，地域特徴やユーザ属性を抽出する研究は，近年盛んに行われている．ソーシャルメディアのデータを用いて地域特徴を抽出する研究 [1], [2], [3] や，観光スポットに着目しその特徴を抽出する研究 [4], [5]，ユーザ属性を抽出する研究 [6], [7], [8] などがある．

李ら [1] は，Twitter から収集したジオタグ付きツイートから，ユーザ ID，位置情報，時間情報を用いて地域ごとに群集行動をモデリングし，特定の時間帯ごとに集約した群衆行動特徴をベクトルで表現した．また，そのベクトルに NMF（非負値行列因子分解：non-Negative Matrix Factorization）を適用し分析することで，特徴的な行動パターンを抽出し，都市の特徴付けを行なう手法を提案している．この研究では，地域ごとの群衆行動をベクトルで表現して都市の特徴付けを行なっているが，本研究では位置情報と時間情報からユーザが訪れた場所と時間をベクトルで表現するという点で異なる．

上野ら [2] は，ジオタグと時間情報が付与されたツイートに着目し，実世界上のスポットと関連付けることで，スポットの時間帯別の特徴を抽出する手法を提案している．Tezuka ら [3] は，ウェブ上のテキストにおける場所の説明文を解析し，場所の特徴を抽出する手法を提案している．中島ら [4] は，観光ツイートからカテゴリごとに観光地を分類し，分類別のスポットの人気度を算出することで観光ルートを推薦する手法を提案している．新井ら [5] は，リアルタイムに投稿された観光体験に関わるツイートの時間帯分布，観光ルート内における観光スポットの共起頻度を用いて，観光スポットのスコア付けを行っている．これらの研究では，特定のスポットのみに着目し，その周辺でユーザが発信したジオタグ付きコンテンツを用いた地域特徴化の手法を行っている．しかし，観光地以外でのジオタグ付きコンテンツの少ない地域の特徴抽出という点では，これらの手法は実用的ではない．そこで，本研究ではユーザが訪れたすべての地域を考慮した地域特徴化の手法を提案する．

榊ら [7] は，ソーシャルメディアユーザの投稿やプロフィール文から各ユーザの特徴量を生成し，サポートベクターマシンにより分類することで職業推定を行なう手法を提案している．Burger ら [9] は，ブログの本文やメタデータからユーザの年齢を予測する手法を提案した．田中ら [10] は，マイクロブログの投稿内容と投稿時間との関係から抽出したライフスタイルに基づきユーザを類型化し，職業推定を行なう手法を提案している．これらの研究ではウェブ上のテキストを解析し，場所や人の特

(注2)：https://twitter.com/

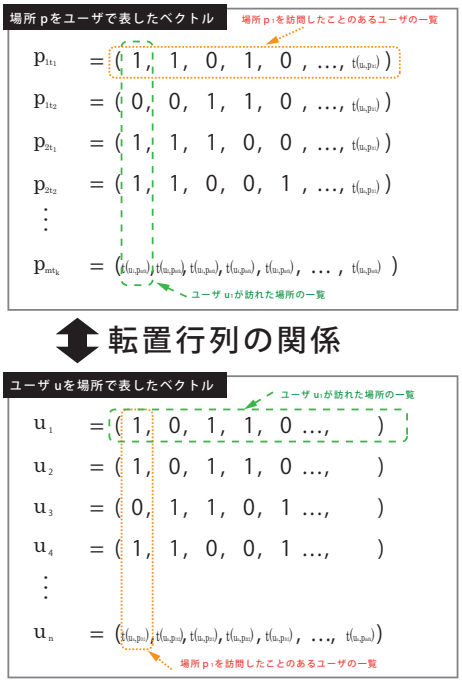

図 2　場所ベクトルと人ベクトル

徴を抽出する手法を行っているが，テキストに用いられている言語ごとに異なる言語処理を行なう必要があり，言語依存度が高い．そこで，本研究では言語依存度の低い情報として，ソーシャルメディアの投稿に付与されたジオタグと時間情報を用いて，地域を特徴化する手法を提案する．

3. 提 案 手 法

本章では，場所を人で，人を場所で表したベクトルを相互再帰的にクラスタリングを行い，同じ性質を持った場所同士を同一のクラスタにまとめる手法について述べる．3.1 節ではベクトル生成を，3.2 節で相互再起的にクラスタリングを行なう手法について述べる．

3.1 人と場所の訪問関係を表す初期ベクトルの生成

本節では，Tweet に付与された位置情報を用いて，場所を人，人を場所でベクトルで表現する方法について述べる．また，この際，同じ場所であっても時間帯によってその特徴が変化することを考慮し，「ある時間帯のある場所」というように，1 つの場所を時間帯によって複数に分けて計算を行なう．ここである時間帯にある場所に訪れた人を表す初期ベクトルを図 2 に図示する．

はじめに，クラスタリングに用いる初期ベクトルを作成する．この際，人をその訪問場所で表したベクトルと，場所をその訪問者で表したベクトルは，並びを固定して行列を作成した際，転置行列の関係にある．そのため，場所ごとに訪問者が来たかどうかを集計して，場所を表す初期ベクトルとして扱う．ツイートに付与されている位置情報に基づいて，対象となるエリアで投稿されたツイートを抽出し，ツイートの発言者のユーザ ID に基いて，重複しないようにユーザ u_n を決定する．また，場所ごとに来訪者を集計するために，対象となる領域全体を任意の大きさのグリッド p_m に区切る．今回は，1km 四方の大きさのグリッドになるように，緯度と経度の範囲を設定する．$T_{user}(u_n)$ をあるユーザ u_n によるすべてのツイート，$T_{place}(p_m)$ をある緯度経度の範囲で表せる場所 p_m 内で投稿されたすべてのツイートとおく．あるユーザ u_n を，そのユーザがどこでツイートしたかによって表した初期ベクトル $\boldsymbol{v}_{user}(u_n)$ は，

$$\boldsymbol{v}_{user}(u_n) = (t(u_n, p_1), t(u_n, p_2), \cdots, t(u_n, p_m)) \qquad (1)$$

と表せる．ここで，$t(u_n, p_m)$ は，ユーザ u_n が場所 p_m で一度でもツイートを投稿したことがあるかを表し，その値は 0 または 1 となる:

$$t(u_n, p_m) = \begin{cases} 1 & (|T_{user}(u_n) \cap T_{place}(p_m)| > 0) \\ 0 & (otherwise). \end{cases}$$

このユーザを表す初期ベクトルは，各次元がそれぞれ場所 p_m に基づく．そのため，ある場所 p_m をベクトル $\boldsymbol{v}_{place}(p_m)$ で表した

$$\boldsymbol{v}_{place}(p_m) = (t(u_1, p_m), t(u_2, p_m), ...t(u_n, p_m)) \qquad (2)$$

に対して，並びを固定した場合に，2 次元行列の転置行列の関係になる．

また，場所について考慮する際,「この場所は昼は若者の街だが，夜はサラリーマンの街である」というように，ある場所がその時間帯によって性質が異なる場合がある．そこで，本手法では，場所を時間ごとに分割して取り扱う．ある時間帯の，あるグリッドで区切った場所を $p_m t$ と置くと，式 (2) と同様に，

$$\boldsymbol{v}_{place}(p_m t) = (t(u_1, p_m t), t(u_2, p_m t), ...t(u_n, p_m t)) \qquad (3)$$

と表わせ，$t(u, p)$ は，ある時間にある場所でツイートを一度でもしたかどうかを表す．p_{mt} で，ある時間帯のその場所を表した時，0:00 から 5:59 までを p_{m1}，6:00 から 11:59 までを p_{m2} というように 1 日を 6 時間間隔で分割することにより，一つの場所につき 4 つの時間帯にあたるベクトルを生成する．

3.2 相互再帰クラスタリング

本節では，3.1 節で生成したベクトルを，相互再帰的にクラスタリングを行なう手法について述べる．

まず，同じユーザが訪問した場所同士を一つのクラスタにまとめる．そして，それぞれのクラスタ c_{pk} に含まれる各ベクトルの平均によって，場所クラスタのベクトル $\boldsymbol{v}_{pcluster}(c_{pk})$ を求める：

$$\boldsymbol{v}_{pcluster}(c_{pk}) = \frac{1}{|c_{pk}|} \sum_{p_a \in c_{pk}}^{a} \boldsymbol{v}_{place}(p_a). \qquad (4)$$

ここで，このある場所クラスタを表すベクトル $\boldsymbol{v}_{pcluster}(c_{pk})$ は場所を人で表したベクトルと同様に並びを固定して行列を作成した場合，人を場所で表したベクトルと転置行列の関係になる．このことから，場所クラスタからなる行列を転置させて人ベクトル $\boldsymbol{v}_{user}(u_n)$ を生成する．

同様に，同じ場所に行っている人同士を一つのクラスタにま

とめる．ここで，それぞれのクラスタ c_{ul} に含まれる各ベクトルの平均によって，

$$\boldsymbol{v}_{ucluster}(c_{ul}) = \frac{1}{|c_{ul}|} \sum_{u_b \in c_{ul}}^{b} v_{user}(u_b) \qquad (5)$$

のようにクラスタから新たなベクトルを求める．このベクトルも，並びを固定して行列を作成した場合，場所クラスタを人クラスタで表した行列と転置の関係にある．転地した行列の各行を場所クラスタを表すベクトルとして用いる．このように，場所ベクトルと人ベクトルのクラスタリングを，相互再帰的に繰り返す．ここで，1 回のクラスタリングでのクラスタ数と，相互再帰計算の繰り返し数は，最終的に生成する特徴ベクトルの長さに応じて設定する．そして，最終的にできた場所クラスタをクラスタリング結果とする．

本手法は場所と人のそれぞれをクラスタリングする際に，既存の多くのクラスタリング手法と距離の計算方法を用いることができる．最終的に生成されるベクトルは用いたクラスタリング手法，距離によって異なる．本稿における実験では，Ward 法，k-means 法，k-medoids 法の 3 つのクラスタリング手法を用いた．

Ward 法は一般的な凝集型階層クラスタリング手法の一つである．対象データとなる 2 つのベクトル $\boldsymbol{x}_1$ と $\boldsymbol{x}_2$ の間の距離 $d(\boldsymbol{x}_1, \boldsymbol{x}_2)$ からクラスタ間の距離 $d(C_1, C_2)$ を

$$d(C_1, C_2) = \frac{1}{|C_1||C_2|} \sum_{\boldsymbol{x}_1 \in C_1} \sum_{\boldsymbol{x}_2 \in C_2} d(\boldsymbol{x}_1, \boldsymbol{x}_2) \qquad (6)$$

と計算し，最もこの距離の近い二つのクラスタを逐次的に併合する．

k-means 法は代表的な非階層型クラスタリング手法のひとつであり，それぞれのクラスタに属す各ノードとクラスタ重心との距離の総和を最小化することでクラスタ分割を行なう：

$$\boldsymbol{c}^* = \arg\min_{\boldsymbol{c}_i} d(\boldsymbol{x}, \boldsymbol{c}_i). \qquad (7)$$

ここで，$d(\boldsymbol{x}, \boldsymbol{c}_i)$ は，要素となるベクトル $\boldsymbol{x}$ と $\boldsymbol{c}_i$ の距離を表す．クラスタリングに用いる距離は任意であるが，ここではユークリッド距離とコサイン距離の 2 つを用いた．最も一般的に用いられる距離のひとつであり，k-means でのクラスタリングにもよく用いられるユークリッド距離 $d_{Euclidian}(x_i, x_j)$ は以下のように求められる：

$$d_{Euclidian}(x_i, x_j) = \left[\sum_{k=1}^{K} (\boldsymbol{x}_{ik} - \boldsymbol{x}_{jk})^2 \right]^{\frac{1}{2}}. \qquad (8)$$

一方で，ベクトル同士の成す角度の近さを表現すコサイン距離は，

$$d_{cosine}(x_i, x_j) = \frac{\boldsymbol{x}_i \cdot \boldsymbol{x}_j}{|\boldsymbol{x}_i||\boldsymbol{x}_j|} \qquad (9)$$

と表される．

4. 章にて，これらのクラスタリング手法と距離計算方法を用いた結果を評価する．

4. 評 価 実 験

本章では，実際に Twitter から収集したツイートを用いて，3. 章で提案した手法により相互再帰的にクラスタリングを行ない，性能を評価する．

4.1 データセット

実験のために，Twitter Streaming API(注3) を用いてジオタグ付きツイートを収集した．収集期間は 2015 年 1 月 1 日から 2015 年 12 月 31 日の全 365 日間であり，この期間に東京 23 区内で投稿されたツイートを用いた．これらの位置情報付きツイートには，海外の利用者のツイートが多く含まれた．日本でツイートをする外国人利用者には，多くの旅行客が含まれる．そのため，これらのツイートをデータに含めた場合，一部の有名観光スポットにデータが集中する可能性がある．本実験では単純化のため言語設定が日本語以外に設定されたツイートをデータセットから取り除き，ユーザ数 297,422 名，ツイート数 8,053,130 件を対象とした．

作成したデータセットに含まれる全てのツイートを用いて初期ベクトル生成を行った結果，場所の数は 3,960 個であった．これらの場所を表すベクトルは疎なベクトルが多く，ある時間帯に一人もユーザが訪れていない場所も含まれた．このような場所を表すベクトルは，零ベクトルとなるため取り除いた．また，位置情報付きで定期的に発信する BOT などの，様々な場所でツイートしているがユーザによるものではないツイートも含まれる．そこで，訪れた場所が 150 箇所以上だったユーザを取り除いた．また，計算量の削減のため訪れた場所が 50 箇所以下だったユーザを取り除いた．これらの前処理を行った結果，場所の数は 2,719 個，ユーザの数は 6,359 人となった．この 2719 個の場所を表すベクトルについて，クラスタリングを行なう．

4.2 ベクトル生成及びクラスタリング結果

4.1 節で作成した 2,719 個のベクトルを用いて，相互再帰的にクラスタリングを行ない場所クラスタを生成した．この際，相互再帰を行う回数を 5 回とした。またそれぞれのクラスタリングにおける終了条件を，クラスタ数が元の要素数の半分になるまでと設定した。．そのため、最終的に場所，訪問者ともに 2^{-5} の大きさになる．5 回目にできた 84 個の場所のクラスタを，最終的なクラスタリング結果と見なし，人手での評価を行った．比較のため，各クラスタリング手法，各距離計算方法を用いたそれぞれの場合の場所クラスタを生成した．また，各クラスタリング手法，各計算処理方法によりクラスタごとの大きさのばらつき，同じクラスタに含まれるそれぞれの地域の類似性について考察を行った．それぞれのクラスタリング結果における，クラスタに含まれるグリッドの最大数と，含まれるグリッドが 1 個であったクラスタの数をあらわしたのものを表 1 に示す．

4.3 人手による評価

4.2 節で行った相互再帰的なクラスタリングの結果をもとに，

(注3)：https://dev.twitter.com/streaming/overview

表 1　クラスタ内の要素数

クラスタリング手法		Ward	*k*-means		*k*-medoids	
距離計算方法		Euclid	Euclid	cosine	Euclid	cosine
含まれる要素数の範囲	1〜5	77	78	27	64	20
	6〜20	5	3	23	10	26
	21〜50	1	1	18	5	24
	51〜100	0	1	8	3	12
	100 以上	1	1	8	2	2
クラスタ内の最大要素数		2,530	2,524	251	1,930	619

クラスタに含まれる地名が似た性質であるかそうでないかを人手により評価した．

まず，人手で評価を行なうために，クラスタに含まれる各グリッドに対し,「飲み屋街」,「オフィス街」,「学生街」,「住宅街」,「商業地域」,「観光地」の 6 つのラベルを付与した．ラベル付けを行なう際の基準としては，あるグリッド内の地域に「住宅」や「商業施設」などの建造物が存在するかに注目し，グリッド内の地域でその建造物が占める割合が多い場合，その建造物の属性をラベルとして付与した．たとえば，グリッド内の地域に「敷地面積の広い大学」が含まれていた場合，その地域は学校周辺であるので学生が多い地域とみなし,「学生街」のラベルを付与する．このようなラベル付けを一名の被験者が行なった．

ラベル付けを行なった後，クラスタごとに集計を行ない，各クラスタに含まれるグリッドのラベルの割合を算出した．これにより，各クラスタに似た性質を持つ場所が含まれているかどうかを人手により評価を行なった．

5. 考　　察

本節では，4.2 節と 4.3 節で得られたクラスタリング結果と，人手によるラベル付けを行った結果について考察を行なう．

5.1　クラスタリングと距離計算の手法による結果の差異

はじめに，クラスタリングに用いた手法と，クラスタリング時に用いた距離ごとの，結果のクラスタの大きさのばらつきについて考察する．表 1 で示したように，Ward 法を用いた場合，1 つのクラスタに含まれるグリッドの数の最大数が 2,530 個となった．これは入力した場所のほとんどが 1 つのクラスタにまとめられ，逆にそれ以外の場所はほとんど結合されていないことを示す．非階層クラスタリング手法である *k*-means 法および *k*-medoids 法を用いた場合でも，ユークリッド距離を用いて計算した場合，同様にほとんどの場所が 1 つのクラスタにまとまってしまう傾向が見られた．また，これらの手法ではクラスタに含まれるグリッドの数が 1 個である極端に少ないクラスタが，*k*-means 法および *k*-medoids 法でコサイン距離を用いた結果よりも多く含まれていた．クラスタリングの結果として，多くの場所が性質に関わらず 1 つのクラスタにまとまり，逆にそれ以外の場所同士がクラスタ化されないことは好ましくない．これらの手法では共通してユークリッド距離を計算に用いている．本実験で用いた訪問者と場所からなるデータはベクトル化した際に次元数が大きく，また疎な場合が多い．これは，対象となる場所が無数あるにも関わらず，一人の人が行ける場所の数には限りがあるからである．クラスタリング時にユークリッド距離を用いて計算した場合，計算結果はベクトルの零部分に大きく影響される．そのため疎なベクトル同士がまとめられ，全体として大きなクラスタに結合されてしまったことが推測される．本手法において，ユークリッド距離を用いたこれらの手法は，本目的においては不適当であると思われる．

一方，コサイン距離を用いた *k*-means 法と *k*-medoids 法では，クラスタに含まれるグリッドの数の最大数がそれぞれ 251 個，619 個であり，クラスタに含まれるグリッドの数が 1 個であったクラスタがそれぞれ 13 個，6 個と，ユークリッド距離を用いた他の 3 手法と比べ少なかった．コサイン距離はベクトル間の角度を類似度の指標として用いているため，ユークリッド距離に比べて高次元で疎なベクトル間の距離の計算に適している．そのため，この 2 手法ではクラスタに含まれるグリッドの数が極端に多い，あるいは少ないクラスタが少なくなったと考えられる．また，*k*-medoids 法でのクラスタリング結果は，クラスタに含まれるグリッドの数が極端に多い，あるいは少ないクラスタを除いたクラスタの数が一番多かった．*k*-medoids 法ではクラスタの代表点として重心そのものではなく，重心に最も近いノードを用いるため，外れ値に強いという特徴がある．これらのことから、本目的においては，距離としてコサイン距離を用いた *k*-medoids 法が最も適していると考えられる．

5.2　ラベル付け結果の考察

最も適切にクラスタ分けが行われたコサイン距離に基づく *k*-medoids 法のクラスタリング結果について考察を行なう．クラスタに含まれるグリッドの数が極端に多い，あるいは少ないクラスタを除いたクラスタのなかから 5 個を無作為に選び，4.3 節の方法により，各クラスタに含まれるグリッドに付与したラベルの割合を算出した．この結果をもとに，同じ性質を持った場所がクラスタリングされているか考察を行なう．各ラベルの割合の算出結果を表 2 に示す．

表 2 より，全体的に「飲み屋街」,「オフィス街」,「商業地域」,「観光地」の含まれる割合が低く，また全体的に「学生街」,「住宅街」が含まれる割合が高いことがわかった．特に「住宅街」は，クラスタ c 以外では高い割合で含まれており，クラスタ a, d では 60%近く含まれた．ラベル付けを行なう際，学校や商業施設などのランドマークがさほど多く存在せず，マンションやアパートが多く含まれたグリッドを「住宅街」とした．このようなグリッドが全体的に多く，駅や商業施設周辺以外の地域はほとんどが居住区域であることから，全てのクラスタに高い割

表 2　各クラスタのラベルの割合（%）

ラベル	クラスタ				
	a	b	c	d	e
飲み屋街	0	0	13.3	0	0
オフィス街	8.5	8.1	6.6	2.6	7.6
学生街	28.5	18.9	36.6	26.3	38.4
住宅街	60	32.4	20	60.5	33.3
商業街	2.8	8.1	23.3	0	15.3
観光街	0	8.1	0	10.5	2.56

合で含まれたと考えられる．

「学生街」は全てのクラスタで高い割合で含まれているが，他のラベルよりも多く含まれたのはクラスタ c，e であった．また，この二つのクラスタには「商業街」が次に多く含まれた．クラスタ c には，都立大学駅から渋谷駅周辺まで，東急電鉄東横線に沿うようにグリッドが存在した．東急電鉄東横線の沿線やその周辺に学校が多く存在し，またグリッドが「若者の街」と呼ばれる渋谷駅周辺まで隣り合って存在しており，渋谷は学生向けの商業施設が多く存在していることで有名である．また，クラスタ c には他のクラスタと比べ「飲み屋街」の割合も高い．クラスタ e は，玉川通り，世田谷通り沿いにグリッドが存在し，その周辺には学校が多く存在した．玉川には商業施設が多く，学生が好むお洒落な洋服店が並ぶ街として有名である．このことから，クラスタ c，e は「学生がよくいく場所」がまとめられたクラスタであると思われる．しかし，クラスタ c は隣り合った場所，および異なる時間の同じ場所のみがまとめられ，クラスタ e にも隣り合った場所，異なる時間の同じ場所が多く，それ以外の場所は住宅街がほとんどであった．クラスタ a，b，d も「学生街」の割合が高いが，「学生街」のラベルを付与したグリッドは，ほとんどが「住宅街」のラベルを付与したグリッドと隣り合っており，学校が多い居住区域で，似た人が来ている場所がまとめられたと考えられる．これらの結果より，いくつかのクラスタには似た性質を持つ場所がまとめられたものの，性質としては似ていない場所がほとんどであったことがわかった．たとえば，オタクである学生が学校に行った後飲み屋に行っている場合と，サラリーマンである教師が学校に行ったあと飲み屋に行っている場合，彼らが訪れた場所は学校と飲み屋のラベルを与えられることになるが，彼らの性質は異なる．このように，同じラベルを付与したグリッドであっても，来た人の性質は異なる場合があり，性質が似ていない場所がまとめられた可能性があると思われる．また，5 個のクラスタ全てが隣り合った場所，および異なる時間の同じ場所が多くまとめられていたことから，本手法により似た人が来ている場所が正しくまとめられたと考えられる．

本論文では，「似た場所に行った人は似ている」，「似た人が訪れた場所は似ている」という 2 つの仮説に基づいて手法を提案した．しかし実験を行った結果，訪れた人が似ていても，場所の性質としては似ていない地域同士がまとめられた．この結果には，いくつかの原因が考えられる．まず，人が外出する際，必ずしも同じ性質を持つ場所を訪れるわけではなく，日によって訪れる場所が変わる．また，一人の人が 1 つの属性しか持つとは限らない．たとえば，「オタクな学生」と「オタクなサラリーマン」などがいた場合，学校と職場のある地域同士がまとめられた場合が考えられる．ユーザの趣味や属性を強く反映するであろう訪問先に比べて，自宅などの落ち着いた場所にいる場合や，駅や電車内など移動中に多くツイートをする可能性がある．

6. おわりに

本論文では，ツイートに付与された位置情報や，投稿時間を用いて，「場所は人で表せる」，「人は場所で表せる」という 2 つの仮定に基づき，相互再帰的にクラスタリングを行なう手法を提案した．また，実際のデータを用いて提案手法を適用した実験を行った後，人手によるラベル付けを行ない，同じ性質を持った場所が正しくクラスタリングされているかを評価した．その結果，ほとんどのクラスタは，性質が似ている場所かどうかに関わらず，異なる時間の同じ場所や隣り合った場所がまとめられた．このことより，本手法により訪れた人が似ている場所がまとめられたことがわかった．

本手法では，ユーザがある場所に来たかどうかに基いたベクトルを生成し，それに対しクラスタリングを行なうため，性質が似ていない場所であるにも関わらず，来たユーザが似ていれば同じクラスタに含まれたと考えられる．今後の課題としては，居住地や職場などのよく現れる場所や，移動中やあまり訪れない場所などにおけるツイートの投稿数による重み付けや，平日と休日の差異などを考慮した改善をすることが挙げられる．また，東京都区部を 1km 四方のグリッドに区切りベクトルを生成したが，いくつかのグリッド内に鉄道の駅が 2 箇所以上含まれたり，グリッドの大部分が住宅街で占められているなど，地域の特徴が複数存在しており，1km 四方のグリッドでは範囲が広いため，グリッドに区切る際の範囲を適切に選ぶ必要がある．本論文における実験では，クラスタリングの相互再帰の回数を 5 回，1 回のクラスタリングでクラスタ数が半分になるように設定したが，クラスタにまとめられる際のベクトル間の距離の閾値や，相互再帰の回数を適切に選ぶこともまた必要であると思われる．

謝　辞

本論文 (の一部) は傾斜的研究費 (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

文　献

[1] 李龍, 若宮翔子, 角谷和俊. Tweet 分析による群衆行動を用いた地域特徴抽出. 情報処理学会論文誌: データベース (TOD54), Vol. 5, No. 2, pp. 36–52, 2012.

[2] 上野弘毅, 奥健太, 服部文夫. 1p-6 位置情報付きユーザ生成コンテンツに基づくスポットの時間的特徴化の提案. 情報処理学会第 75 回全国大会, Vol. 1, p. 6, 2013.

[3] Taro Tezuka, Ryong Lee, Hiroki Takakura, and Yahiko Kambayashi. Cognitive characterization of geographic objects based on spatial descriptions in web resources. Cite-

seer, 2003.

[4] 中嶋勇人, 新妻弘崇, 太田学. 位置情報付きツイートを利用した観光ルート推薦. 研究報告データベースシステム (DBS), Vol. 2013, No. 28, pp. 1–6, 2013.

[5] 新井晃平, 新妻弘崇, 太田学. Twitter を利用した観光ルート推薦の一手法. 第 7 回データ工学と情報マネジメントに関するフォーラム (DEIM2015), G7-6, pp. 1–8, 2015.

[6] 今井規善, 奥健太, 服部文夫. 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. 情報処理学会第 75 回全国大会, Vol. 1, p. 5, 2013.

[7] 榊剛史, 松尾豊. ソーシャルメディアユーザの職業推定手法の提案. 知能と情報, Vol. 26, No. 4, pp. 773–780, 2014.

[8] Zhiyuan Cheng, James Caverlee, and Kyumin Lee. 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*, pp. 759–768, 2010.

[9] John D Burger and John C Henderson. An exploration of observable features related to blogger age. pp. 15–20, 2006.

[10] 田中成典, 中村健二, 加藤諒, 寺口敏生. マイクロブログの投稿時間に着目したユーザの職業推定に関する研究. 情報処理学会論文誌データベース (TOD), Vol. 6, No. 5, pp. 71–84, 2013.