6D-4

ソーシャルジオデータのクラスタリング結果に対する自動的な意味付けに関する一検討

荒川豊† 福田晃†

†九州大学・大学院システム情報科学研究院

1 はじめに

近年，Facebook や Twitter などに代表されるソーシャルネットワークサービス（SNS: Social Network Service）の普及が著しい．今や Twitter はユーザ数 5 億人を抱え，3 日で 10 億ツイートもの情報が発信されている．Facebook はユーザ数 10 億人に迫っており，1 日当たり 32 億コメント，3 億枚の写真投稿が行われている．そして，この膨大なデータから，新たなインテリジェンスを得ようとする研究が盛んに行われている．

こうした膨大なデータの中でも位置情報が含まれるデータは，**ソーシャルジオデータ**と呼ばれ，地震伝搬の分析 [2] や入力される日本語との相関関係分析 [1]，観光ルート分析 [3] など，さまざまな研究がなされており，我々も現在，各都市の人気観光スポットを自動的に抽出するシステムに関する研究に取り組んでいる．ソーシャルジオデータを観光に活用する研究は，[3] 以外にも [4][5] などがあり，これらの手法では，何らかのクラスタリング手法によって分析対象データをエリアに分割し，各エリア内におけるソーシャルジオデータの発生数をそのエリアの人気度と定義している．

ソーシャルジオデータの分析に用いられる，代表的なクラスタリング手法としては，Mean Shift Clustering[3][4] や p-DBSCAN[5] などがあるが，いずれのクラスタリング手法でも得られる結果は，クラスタの中心座標とそのメンバ情報という数値データであり，このクラスタに対応する実世界の観光情報（POI : Point of Interest）は不明である．分析結果を活用するためには，クラスタに

このようなクラスタへの名前付けに関する従来研究として，ソーシャルジオデータ内のタグ情報から推定するもの [4] や GeoNames.org などの静的な POI データベースを用いて距離の近いものを割り当てる研究があるが，ラベリングを取り扱っていない従来研究も多い．これは，実際には付与されているタグが少ない上，表記揺らぎも多く，計算時間もかかるためである．

そこで本論文では，このクラスタへのラベリングを高速かつ高精度に行う手法を提案する．提案手法では，分析対象となるソーシャルジオデータから，ラベルを推定するのではなく，別のソーシャルジオデータを用いてラベルを推定する．今回，分析対象データは，Flickr から集めたデータを用いているが，写真撮影と同様に行うチェックイン[1]という行為に着目し，ソーシャルジオデータによるソーシャルジオデータへの自動ラベリングの達成を狙っている．これは，もともとスマートフォンの GPS 精度が低いことから，チェックインサービスにおいて，スポット候補を提示する際に，距離だけではなく，人気も反映されているためである．距離に基づいた従来手法と比べて，誤差に対する耐性が強くなり，より確からしいラベルを得られるのではないかと考えている．その際，カテゴリによるフィルタリングを行うことによって，観光と無関係なスポット名を排除し，精度をさらに改善できると考えている．

従来のタグ分析による名前では，各タグの特徴量（クラスタ特有の単語か否かを判定するための値）を計算 [4] する必要があり，タグ数が多くなると計算時間が多くなるという問題があるのに対して，Foursquare を用いる提案手法は，Web API (Application Programmable Interface) に 1 度アクセスするだけでよく計算が必要ないという利点がある．

今回，Flickr を用いて収集したパリ（フランス）におけるソーシャルジオデータ 20000 件を分析対象として，クラスタリングによって得られた人気観光スポット 3 箇所に関して，Foursquare のチェックインデータを用いた名前付けにより，確からしい結果を得られるかについて検証した結果を報告する．

[1]チェックインとは，Foursquare というサービスから始まったもので，ある場所にチェックインすることでバッジやポイントを獲得できるため，ユーザが競ってチェックインをしている

Automatic name assignment mechanism for the clusters of social geo-data clustering

Yutaka Arakawa† Akira Fukuda†

†Graduate School of Information Science and Electrical Engineering, Kyushu University, 744 Motooka, Nishi, Fukuoka, JAPAN 819-039

arakawa@ait.kyushu-u.ac.jp

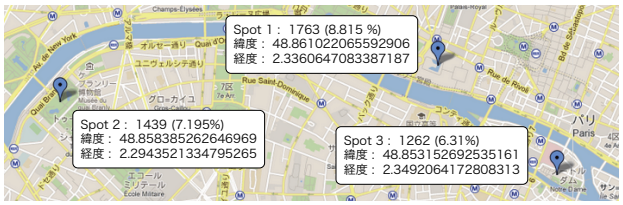

図 1: パリにおける上位 3 クラスタ

2 分析内容

分析対象データは，Flickr から収集したパリ付近のソーシャルジオデータ 830372 件のうち，無作為に選択された 20000 件のデータである．Flickr のデータには，写真だけでなく，位置情報，日時，タグ，簡易な説明，といった情報が含まれているが，本論文では，この中の位置情報とタグだけを取り扱う．

まず，緯度経度からなる位置情報に対して，Mean Shift Clustering を適用する．このとき，唯一のパラメータである *bandwidth* は，0.002 としている．*bandwidth* は，平滑化パラメータと呼ばれ，クラスタのサイズに影響を与え，今回用いた 0.002 は約 222m に相当する．

次に，クラスタリングによって得られたクラスタのうち，含まれるデータ数が多いクラスタ 3 つを人気スポットとして選択する．このとき得られる情報は，3 つのクラスタの中心座標である．

最後に，この中心座標をもとに，クラスタに対して名前付けを行い，結果を比較する．

3 分析結果

最初に，Mean Shift Clustering の結果を示す．20000 件のデータは， 351 個のクラスタに分割され，その上位 3 クラスタは図 1 に示す位置となった．この中心座標をもとに，Foursquare API にアクセスし，得られた POI 名を表 1 に示す．図 1 と照らし合わせることでわかるように，正しい POI 名が得られている．

表 1: Foursquare を用いて付けられた POI 名

Spot 1	Musee du Louvre
Spot 2	Tour Eiffel
Spot 3	Cathedrale Notre-Dame de Paris

4 まとめ

評価結果より，Foursquare のチェックイン情報により適切と思われる名付けが簡単に行えることがわかった．ただし，今回選択した 3 つの POI は，ルーブル美術館，エッフェル塔，ノートルダム大聖堂，とそれぞれ大きなものであり，周辺に他の POI が少かったために成功したとも考えられる．Foursquare におけるチェックインは，スターバックスなどのカフェなどで多くなされており，もし周辺にそのような観光に関係のない人気 POI が存在していた場合，そちらの名前が付与される可能性がある．その問題に対しては，API に指定可能なカテゴリフィルタを用いて観光に関係のあるスポットだけを選んだり，従来のタグ分析と組み合わせるといった手法により，適切な名前を抽出可能になると考えている．今後は，速度に関しての定量的な評価，名付けた名前の適切度の評価，および適切なカテゴリ選択，他の観光スポットに関する結果，などを行なっていく．また今回は基本的な Mean Shift Clustering を用いたが，将来的には，クラスタリング手法の改善にも取り組んでいく予定である．

謝辞

本研究の一部は，財団法人人工知能研究振興財団の研究助成に基づくものである．ここに記して謝意を示す．

参考文献

[1] 荒川豊, 田頭茂明, 福田晃, “[推薦論文]Twitter を用いたコンテキストと入力文字列の相関関係分析,” 情報処理学会論文誌, Vol.52, No.7, pp.2268–2276, 2011 年 7 月.

[2] T. Sakaki, M. Okazaki, and Y. Matsuo, “Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors,” In Proc. of the 19th International Conference on World Wide Web (WWW), pp. 851–860, 2010.

[3] D.J. Crandall, L. Backstrom, D. Huttenlocher, and J. Kleinberg, “Mapping the world’s photos,” In Proceedings of the 18th international conference on World wide web, pp. 761–770. ACM, 2009.

[4] T. Kurashima, T. Iwata, G. Irie, and K. Fujimura, “Travel route recommendation using geotags in photo sharing sites,” In Proc. of the 19th ACM international conference on Information and knowledge management, pp. 579–588. ACM, 2010.

[5] S. Kisilevich, F. Mansmann, and D. Keim, “P-dbscan: A density based clustering algorithm for exploration and analysis of attractive areas using collections of geo-tagged photos,” In Proc. of the 1st International Conference and Exhibition on Computing for Geospatial Research & Application, p. 38. ACM, 2010.