# 密度に基づく空間クラスタリングを用いたジオタグ付きツイートからのトピック抽出

Extracting Local Topics in Geo-tagged Tweets using Density-based Spatial Clustering

酒井 達弘

Tatsuhiro Sakai

広島市立大学大学院情報科学研究科 データ工学研究室

Data Engineering Laboratory, Graduate School of Information Sciences, Hiroshima City University

**概要** ソーシャルメディア上に投稿される位置情報付きデータから有益な知識を抽出する研究が盛んに行われている．本研究では，位置情報，テキストと画像から構成されるデータをジオソーシャル画像データと呼び，ジオソーシャル画像データからのトピック抽出手法を提案する．提案手法は，最初に，密度に基づくマルチモーダル空間クラスタリングを用いて，トピックが含まれる領域をマルチモーダル空間クラスタとして抽出する．そして，マルチモーダル空間クラスタ中のトピックを定めるために，ネットワークベースの重要度算出手法を用いて代表画像データを抽出する．提案手法では，画像データ間の類似度を正確に算出するために，学習済み畳み込みニューラルネットワーク（CNN）を用いて特徴ベクトルを抽出する．実際に Twitter 上から収集したジオソーシャル画像データを用いて行った評価実験の結果，提案手法は「京都」のトピックを抽出できることを示した．

## 1 はじめに

ビッグデータへの関心の高まりとともに，ソーシャルメディア上に投稿される画像データやテキストから有益な知識を抽出する研究が盛んに行われている．また，GPS 付きスマートフォンと GPS に連動したアプリケーションの普及とともに，位置情報付きのデータがソーシャルメディア上に盛んに投稿されるようになってきている．本研究では，ソーシャルメディア上に投稿される，位置情報，画像とテキストから構成されるデータのことをジオソーシャル画像データと呼ぶ．Twitter 上に投稿されるジオソーシャル画像データには，個人的な趣味や話題だけでなく，ユーザが各地域で日々目にした事象や話題が含まれており，各地域のトピックを抽出することができれば，観光情報，マーケティングや動向分析に活用することができる．

本研究では，ジオソーシャル画像データからのトピック抽出手法を提案する．提案手法は，最初に，密度に基づくマルチモーダル空間クラスタリングを用いて，トピックが含まれる領域をマルチモーダル空間クラスタとして抽出する．そして，マルチモーダル空間クラスタ中のトピックを定めるために，ネットワークベースの重要度算出手法を用いて代表画像データを抽出する．提案手法では，画像データ間の類似度を正確に算出するために，学習済み畳み込みニューラルネットワーク（CNN）を用いて特徴ベクトルを抽出する．評価実験では，実際に Twitter 上から収集したジオソーシャル画像データを用いて行った実験結果を報告する．

## 2 提案手法

### 2.1 密度に基づくマルチモーダル空間クラスタリング

密度に基づくマルチモーダル空間クラスタリング [1] を用いてトピックが含まれる領域をマルチモーダル空間クラスタを抽出する．密度に基づくマルチモーダル空間クラスタリングでは，あるジオソーシャル画像データについて，そのデータの近傍に類似したジオソーシャル画像データが閾値以上存在する場合，その近傍は高密度と判断する．そして，高密度な近傍を接続していくことで，クラスタを形成する．内容が類似しているジオソーシャル画像データの投稿が密集している領域には，何かしらの注目されているトピックが存在していることとなる．

提案手法では，二つのジオソーシャル画像データ $gsi_i$ と $gsi_j$ 間の類似度について，テキストの類似度と画像データの類似度のトレードオフであるマルチモーダル類似度 $msim$ を用いる．

$$msim(gsi_i, gsi_j) = w \times tsim_{i,j} + (1-w) \times isim_{i,j}, \quad (1)$$

$w$ は重み ($0 \leq w \leq 1$) であり，$tsim_{i,j}$ はテキスト間の類似度，$isim_{i,j}$ は画像データ間の類似度である．テキスト間の類似度については語句ベースのコサイン類似度を用いる．画像データ間の類似度については，学習済み CNN である VGG16 へ画像データを入力し，出力層手前の値を特徴ベクトルとして用い，コサイン類似度によって求める．

### 2.2 ネットワークベースの重要度算出手法

ネットワークベースの重要度算出手法を用いてマルチモーダル空間クラスタから代表画像データを抽出する．ネットワークベースの重要度算出手法は類似度グラフ $SG$ の作成と，ノードの重要度の算出の二段階によって行う．そして，重要度の高い画像データをマルチモーダル空間クラスタの代表画像データとして抽出する．

密度に基づくマルチモーダル空間クラスタ $msc$ に所属するジオソーシャル画像データ集合を $msc = \{gsi_1, gsi_2, \cdots, gsi_{|msc|})\}$ とする．類似度グラフ $SG = (V, E)$ は，ノード集合 $V$ と辺集合 $E$ から構成される．ノード $v_i \in V$ はジオソーシャル画像データ $gsi_i$ に対応し，辺 $e = (j, k) \in E$ はノード $v_j$ とノード $v_k$ 間に辺が存在すること示す．ここで，

$$E = \{(j,k) \mid v_j \in V, v_k \in V, \\ isim(gsi_j, gsi_k) \geq \alpha\} \qquad (2)$$

と定義する．$\alpha$ はパラメータであり，類似度が高いノード間の辺のみに制限することができる．

媒介中心性を用いて各ノードの重要度を算出する．媒介中心性とは，ノードがどれくらいネットワーク上で重要な媒介を行っているかを示し，通常，ノード間の最短路が何本通っているかで算出され，$v_i$ の重要度 $BC(v_i)$ は以下の式で表される．

$$BC(v_i) = \sum_{s \neq t \neq v_i} \frac{path_{s,t}(v_i)}{path_{s,t}} \qquad (3)$$

ここで，$path_{s,t}(v_i)$ はノード $s$ とノード $t$ 間の経路でノード $v_i$ を通る経路の数，$path_{s,t}$ はノード $s$ とノード $t$ 間の経路の総数である．各ノードの $BC(v_i)$ の値を計算し，$BC(v_i)$ の値を，そのノードが示すジオソーシャル画像データの重要度とする．

## 3　評価実験

評価実験では，画像データの特徴ベクトル抽出方法として，BoF（Bag-of-Features）と学習済み CNN である VGG16 を用いた場合の比較を行う．Twitter 上に投稿されたジオタグ付きで画像 URL を持つツイートをジオソーシャル画像データとして扱い実験を行う．データセットは京都府庁から半径 30km 以内で投稿された 11,189 件（2011 年 11 月から 2012 年 2 月まで）のジオタグ付きツイートを用いる．

図 1 に VGG16 を用いた手法で抽出されたツイート数で上位 20 位のクラスタを地図上に示す．図 1 より，京都で投稿されたジオソーシャル画像データから構成されたクラスタが抽出できているのが分かる．表 1 と 2 に二手法で抽出されたツイート数で上位 5 位のクラスタのツイート数と頻出語句を示す．表 1 と 2 より，二手法で抽出されたクラスタ 1 と 2 は，「京都駅」と「清水寺」で抽出されたクラスタであることが分かる．表 2 のクラスタ 3 は「渡月橋」で抽出されたクラスタであるが，表 1 のクラスタ 3 は「南禅寺」や「平安神宮」などの様々なトピックから構成されたクラスタであった．VGG16 を用いた手法では，これらのトピックは別々のクラスタとして抽出されており，より正確に画像データ間の類似度を算出できたといえる．また，表 1 のクラスタ 4 と 5 は「天龍寺」と「渡月橋」，表 2 のクラスタ 4 と 5 は「金閣寺」と「八坂神社」に関するクラスタであった．

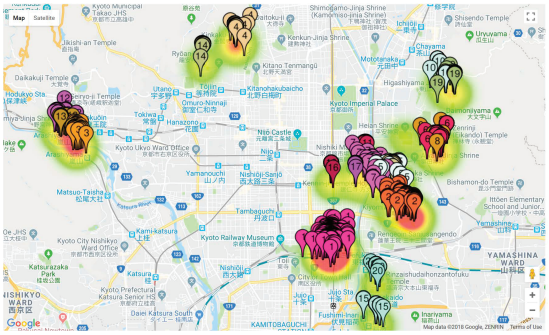

図 1: 抽出されたクラスタ

表 1: 抽出されたクラスタ（BoF）

| ID | ツイート数 | 頻出語句 |
|---|---|---|
| 1 | 220 | 京都，駅，なう，タワー，新幹線 |
| 2 | 108 | 清水寺，清水，東山，京都，舞台 |
| 3 | 84 | 南禅寺，永観堂，京都，左京，なう |
| 4 | 55 | 寺，天龍，京都，右京，嵯峨 |
| 5 | 51 | 渡月，橋，嵐山，なう，紅葉 |

表 2: 抽出されたクラスタ（VGG-16）

| ID | ツイート数 | 頻出語句 |
|---|---|---|
| 1 | 252 | 京都，駅，なう，下京，新幹線 |
| 2 | 100 | 清水寺，清水，東山，舞台，なう |
| 3 | 48 | 月橋，渡，京都，嵯峨，右京 |
| 4 | 42 | 金閣寺，鹿苑寺，京都，なう，綺麗 |
| 5 | 28 | 京都，東山，八坂神社，祇園北側，祇園南側 |

表 2 の各クラスタについて，代表画像データを抽出した．クラスタ 1 については，「京都駅」周辺で撮影された画像データが，クラスタ 2，3，4 と 5 については，清水寺，渡月橋，金閣寺と八坂神社の画像データがそれぞれ代表画像データとなった．

## 4　まとめ

本研究では，密度に基づくマルチモーダル空間クラスタリングとネットワークベースの重要度算出手法を用いたジオソーシャル画像データからのトピック抽出手法を提案した．実際に Twitter 上から収集したジオソーシャル画像データを用いて行った評価実験の結果，提案手法は「京都」のトピックを抽出できることを示した．今後の課題としては，定量的な評価を行うことや抽出したトピックを可視化するためのアプリケーションの開発が挙げられる．

## 参考文献

[1] T. Sakai, K. Tamura, H. Kitakami, and T. Takezawa, “Density-based multimodal spatial clustering using pre-trained deep network for extracting local topics,” pp. 7-12, 2018.