Vol.2019-MPS-123 No.45
Vol.2019-BIO-58 No.45
2019/6/19

飲食店分布を用いた地域コミュニティに対する特性ラベリング

吉田 純[1,a)]　伏見 卓恭[1,b)]

概要：本研究では，各地域に偏って分布する飲食店カテゴリを用いて地域をラベリングすることを試みる．飲食店の少ないエリアでは，多いエリアと比べると各カテゴリの飲食店の絶対数は少なく，特性が埋もれてしまう．そこで，飲食店が地域とカテゴリに独立して分布している場合の期待値と実際の分布数である実測値の差を用いて Z スコア計算し，出店数の少ない地域やカテゴリであっても，期待値に比べて偏って多くなっていれば，出店数の格差に隠れた特性を強調する手法を提案する．ぐるなび API を用いて取得した飲食店データを用いた評価実験により，絶対数では抽出できなかった各地域特有のカテゴリを提案手法により抽出できることを確認する．

Jun Yoshida[1,a)]　Takayasu Fushimi[1,b)]

1. はじめに

近年，様々な地域で数多くの飲食店が分布しており，そのカテゴリも多種多様である．例えば，駅周辺のエリアでは，帰宅時などの会社員がたくさんいるため居酒屋が多く，地方では，名産物などを使用した郷土料理屋が多く分布している．しかし，規模の小さいエリアでは飲食店の絶対値が少なく，大きいエリアと比較して特性が埋もれてしまう場合がある．本研究では，各地域に偏って分布する飲食店カテゴリを用いて地域をラベリングすることを試みる．単純に出店している店舗を地域ごとカテゴリごとにカウントし，分布数の多いカテゴリを用いて各地域をラベリングすると，居酒屋やラーメンなど，どの地域でも同じようなラベルになってしまう．本研究では，絶対数の規模の格差に隠れてしまう各地域の特性を顕在化させラベリングする．提案手法により，抽出した特性によって，小さな地域をラベリングすることで，地域活性化につながると期待できる．

地域特性を抽出する研究として，SNS 投稿写真の画像特性を抽出し，地域の特徴記述を行い地域間の類似度を求める研究がある [1]．各地域の特性を写真の特徴量を要素としたベクトルを用いて表現している．提案手法では各地域の特性を出店している飲食店のカテゴリ分布から算出した Z スコアで表現している点で異なる．また，位置情報付きコンテンツから地域限定語句を抽出する研究がある [2]．TFIDF をベースとした指標を用いて，特定の地域にのみ出現する単語を抽出している．

[1] 東京工科大学　コンピュータサイエンス学部
School of Computer Science, Tokyo University of Technology
[a)] c0116288fb@edu.teu.ac.jp
[b)] fushimity@stf.teu.ac.jp

2. 提案手法

提案手法は，飲食店 v を集めた集合 $\mathcal{V}$ が与えられた際，その出店地域 $v.area$ と カテゴリ $v.cate$ に基づき，地域ごとカテゴリごとの飲食店数を集計することで Mixing Matrix $\mathbf{F} = [f_{a,c}]$ を構築する [3]．ここで，$f_{a,c} = |\{v \in \mathcal{V}; v.area = a \wedge v.cate = c\}|/N$ は，全飲食店数 $N = |\mathcal{V}|$ に対する地域 a に出店するカテゴリ c の飲食店数の割合である．前述したように，地域 a において $f_{a,c}$ の値が大きなカテゴリ c により地域 a をラベリングするのは適切ではない．そこで，飲食店は地域とカテゴリに独立に分布していると仮定した場合の分布数の期待値 $e_{a,c}$ と実際の分布数の差を用いて Z スコア $z_{a,c}$ を計算する．まず，地域の集合を $\mathcal{A}$，カテゴリの集合を $\mathcal{C}$ とし，地域分布 $p_a = \sum_{c \in \mathcal{C}} f_{a,c}/N$ とカテゴリ分布 $q_c = \sum_{a \in \mathcal{A}} f_{a,c}/N$ を計算する．そして，飲食店の出店数は地域とカテゴリに独立であると仮定して地域 a にカテゴリ c の飲食店が存在する確率は $e_{a,c} = p_a \cdot q_c$ で計算できる．したがって，Z スコアは以下のように計算される：

$$z_{a,c} = \frac{N f_{a,c} - N e_{a,c}}{\sqrt{N e_{a,c}(1 - e_{a,c})}}. \quad (1)$$

Vol.2019-MPS-123 No.45
Vol.2019-BIO-58 No.45
2019/6/19

表 1 分布数によるラベリング（関東の県）

$Nf_{a,c}$	1 位	2 位	3 位
茨城県	居酒屋:1538	定食:761	そば:649
栃木県	居酒屋:1105	定食:720	そば:704
群馬県	居酒屋:1006	定食:679	そば:620
埼玉県	居酒屋:3908	そば:1236	カフェ:1135
千葉県	居酒屋:3331	カフェ:1096	定食:1025
神奈川	居酒屋:5782	カフェ:2100	中華:1589

表 2 Z スコアによるラベリング（関東の県）

$z_{a,c}$	1 位	2 位	3 位
茨城県	そば:17.72	ラーメン:11.39	定食:10.07
栃木県	そば:25.47	焼きそば:17.39	ラーメン:13.68
群馬県	そば:22.91	定食:13.24	焼きそば:10.53
埼玉県	そば:19.77	中華:14.43	うどん:12.32
千葉県	ファミレス:12.12	中華:11.26	そば:9.34
神奈川	中華:15.27	イタリアン:9.04	居酒屋:8.99

表 3 Z スコアによるラベリング（大分類）

$z_{a,c}$	1 位	2 位	3 位
那覇	沖縄料理:102.16	バー:16.51	しゃぶしゃぶ:7.736
祇園・岡崎・清水寺	京料理:59.44	懐石料理:20.38	割烹:15.18
広島市	お好み焼き:53.22	広島風お好み焼き:41.39	鉄板焼き:17.99
日光・鬼怒川	湯葉料理:53.05	そば:16.67	定食:13.23
廿日市市・大竹市・宮島	あなご料理:44.78	お好み焼き:14.70	牡蠣料理:11.79
仙台市	牛タン:41.20	ショットバー:9.19	牡蠣料理:4.83

式 (1) の分母は標準偏差であり，Z スコアの値は独立性を仮定した際の期待値と比較して，どの程度有意に多くまたは少なく存在するかを表している．Z スコアが正で大きいほど，カテゴリ c の飲食店が地域 a に統計的有意に多く存在するといえる．Z スコアを用いることにより，出店数の少ない地域やカテゴリであっても，独立性を仮定した際に比べて偏って多く出店していれば大きな値となり，規模の格差により隠れてしまう地域・カテゴリ間の関係の強さを強調できる．

3. 評価実験

本研究では，ぐるなび API [*1] を用いて取得した全 565,810 件（全 178 カテゴリ）の飲食店データを対象とし，各地域の特性をラベリングする．そして，有名な地域を取り上げ，直感と合致した結果が得られているかを評価する．

表 1 に，関東地方の都県に対して単純な分布数が多い上位 3 つのカテゴリでラベリングした結果を示す．上述したように，日本全国で分布数の多い居酒屋，カフェ，定食がラベルとして用いられる傾向にあり，各地域の特性が現れていない．表 2 は，関東地方の都県に対する Z スコアの上位 3 カテゴリによるラベリング結果である．表 2 より，神奈川県が中華，群馬県がそばでラベリングされており，単純分布数による結果よりは妥当な結果が得られている．他の都県に関しては，必ずしも妥当とは言い難い．これは，地域の分類粒度が大きすぎるためと考えられる．

表 3 は，分類粒度を都道府県より細かくした地域に対するラベリング結果である．ここでは，1 位のカテゴリに対する Z スコアが高い順に表示している．表 3 を見ると，那覇が沖縄料理，京都の祇園が京料理，広島市がお好み焼き，仙台市が牛タンなど，直感に合致した結果が得られているか．さらに，日光が湯葉（単純分布数 : 16），廿日市市がなご料理（単純分布数 : 9）のように，絶対数が少なくあまり知られていないが，その地域に偏って分布するカテゴリによりラベリングできている．

[*1] https://api.gnavi.co.jp/api/

4. おわりに

本研究では，ある特定の地域に偏って分布する特徴的なカテゴリによるラベリングを目的とする．Z スコアにより地域に偏在するカテゴリを抽出し，ラベルとして付与する手法を提案した．結果の良し悪しを直感的に判断しやすい飲食店カテゴリの実データを用いて提案手法によるラベリングを評価した．よく知られる妥当なラベリング結果だけでなく，分布数自体は少なくあまり知られていないが偏って分布するカテゴリも得られることを確認した．今後の課題として，飲食店のカテゴリ情報だけでなく，ご当地ソングに使用される単語や生産される農産物，観光スポットの種別などの情報を用いて，Z スコアの有効性を確認する予定である．さらに，共クラスタリングの技術を応用するなどして，Z スコアの値が高くなるように地域・カテゴリの分類粒度を設定する手法の開発が必要である．

謝辞 本研究は，JSPS 科研費（No.19K20417）の助成を受けたものである．

参考文献

[1] 滝本広樹，川西康友，井手一郎，平山高嗣，道満恵介，出口大輔，村瀬 洋：SNS 投稿写真の画像内容に基づく地域間の類似度算出に関する検討 (マルチメディア・仮想環境基礎)，電子情報通信学会技術研究報告 = IEICE technical report : 信学技報， Vol. 116, No. 73, pp. 83–88 (2016).

[2] 奥 健太，西崎剛司，服部文夫：地域限定性スコアに基づく位置情報付きコンテンツからの地域限定語句の抽出，情報処理学会論文誌データベース（TOD），Vol. 5, No. 3, pp. 97–116 (2012).

[3] Newman, M. E. J.: Mixing patterns in networks, *Physical Review E*, Vol. 67, No. 2, pp. 026126+ (online), DOI: 10.1103/PhysRevE.67.026126 (2003).