4ZD-01

不完全なPOI名称に対する機械学習によるカテゴリ推定手法

金平卓也† 荒川豊† 安本慶一†

†奈良先端科学技術大学院大学 情報科学研究科

1 はじめに

近年，Google Maps等の地図サービスの普及に伴い，ユーザが興味のある場所（POI : Point of Interest）を1つの地図にまとめた図1のようなマイマップの利用が拡大している．マイマップはユーザが視覚的かつ直感的に空間情報を共有することに有用であり，盛んに共有が行われている．2014年にはマイマップをフリーキーワードで検索可能なサービスが開始している．しかし，大半のマイマップに登録されているPOIにはカテゴリ情報がなく，どのようなカテゴリのPOIをまとめたマイマップかわからないのが現状である．また，マイマップはユーザが自由に作成しており，不完全なPOI名称も多数含まれている．また，同一POIに対して異なる名称がつけられる，名称のゆらぎ（例：奈良先端と先端大など）も多い．

そこで，本研究では不完全なPOIの名称からPOIのカテゴリを推定することを目的とする．この目的を達成するために，Foursquare[1]から得られる膨大なPOIの名称とカテゴリの関係をサポートベクター回帰（SVR）ベースの機械学習をする．本稿では機械学習を行う際の特徴抽出方法を4種類提案し，それぞれの推定精度と既存研究[1]の推定精度を比較する．特徴抽出方法の違いにより推定精度がどのように変化するかマイマップに登録されているPOIデータを用いて示す．

本稿の2章では，関連研究について述べる．3章では，本研究で用いる機械学習フレームワークについておよび機械学習を行う際の4種類の特徴抽出方法について述べる．4章では，各特徴抽出方法の機械学習によるPOIのカテゴリ推定結果と結果に対する考察について述べる．5章では，本研究のまとめについて述べる．

2 関連研究

POIに関する情報を用いてカテゴリを推定する研究として，Choi[1]らのPOIカテゴリ推定方法がある．彼らはYelpから得た3499件のPOIデータをトレーニングデータとしてSVMベースの機械学習をさせ，同所から得た別の941件のPOIデータをテストデータとしてPOIカテゴリの推定精度を示している．POIデータの内，POI名称とカテゴリの関係を機械学習させ，20種類のPOIカテゴリを推定した結果，45.84%の精度であった．機械学習させる際の特徴抽出方法としては，POIの名称の1文字ずつを特徴量とするunigramである．そこで，本研究では，機械学習させる際の特徴抽出方法を変更することにより，Choiらより高い推定精度を目標とする．また，高い推定精度を得るために，SVMに比べて多クラス分類に向いているSVRをベースとした機械学習を用いる．

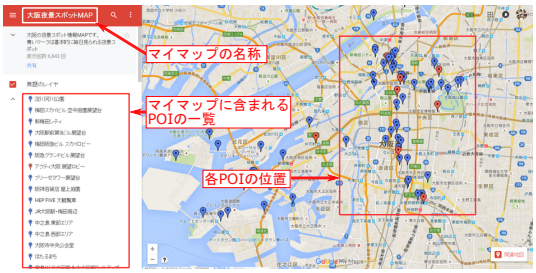

図1: 大阪市の夜景スポットをまとめたマイマップ

3 提案手法

3.1 機械学習フレームワーク

本研究では，POIの名称とカテゴリの関係を機械学習するためフレームワークとしてJubatus[2]を用いる．このフレームワークには多クラス分類器を構築するために，SVRのオンライン版であるPassive Aggressiveアルゴリズム[2]を採用している．このため，リアルタイムに機械学習が可能となっている．

3.2 トレーニングデータ

POIの名称とカテゴリの関係を機械学習するために必要なトレーニングデータはFoursquareから得る．トレーニングデータとして使用するPOIデータは2万5000件であり，25種類のカテゴリのPOIデータが各

Category Estimation Method by Machine Learning for The Incomplete Point-of-Interest Name

Takuya KANEHIRAI†, Yutaka ARAKAWA†and Keiichi Yasumoto‡

†Nara Institute of Science Technology

[1]Foursquare : https://ja.foursquare.com

[2]Jubatus : http://jubat.us/ja/

表 1: 各特徴抽出方法における平均カテゴリ推定精度

機械学習アルゴリズム	SVR				SVM*
特徴抽出方法	origin	unigram	mecab	hybrid	unigram
推定精度 [%]	11.54	60.28	67.04	70.27	45.84

*既存研究 [1] の推定精度

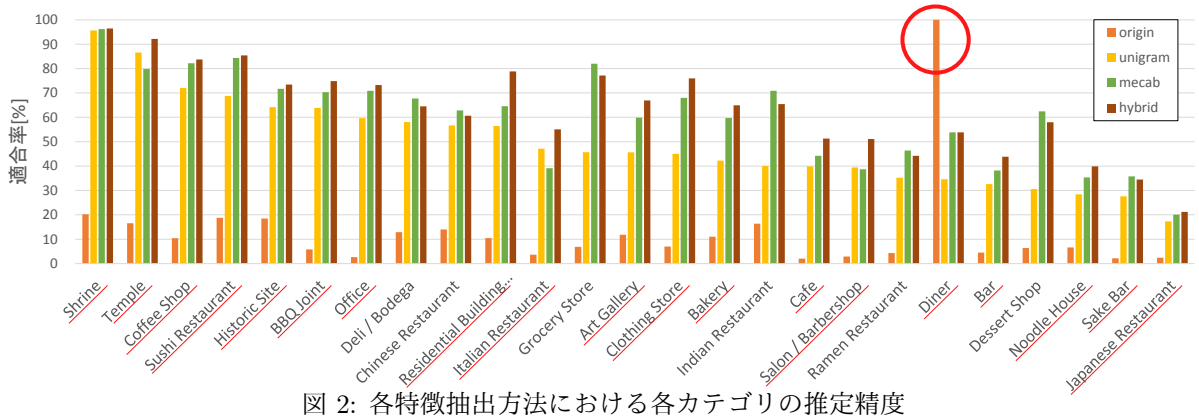

図 2: 各特徴抽出方法における各カテゴリの推定精度

1000 件ずつある．

3.3 特徴抽出方法

機械学習を行う場合，文字列などの生データは直接扱うことができないため，事前に特徴抽出を行い特徴ベクトルと呼ばれる形式に変換する必要がある．機械学習をさせるとき，この特徴ベクトルを入力として用いる．つまり，POI の名称をどのような特徴量に変換するかによって推定精度が異なるということである．本研究では以下の 4 つの特徴抽出方法を提案する．

- **origin**：POI の名称をそのまま特徴量として用いる方法．
- **unigram**：POI の名称の 1 文字ずつを特徴量として用いる方法．既存研究 [1] でも用いられている．
- **mecab**：POI の名称を形態素解析し，この結果得られる形態素を特徴量として用いる方法．形態素解析には MeCab[3] を用いる．
- **hybrid**：unigram の特徴量と mecab の特徴量の積を特徴量として用いる方法．例えば POI の名称の文字数が 3 文字で，形態素解析の結果，形態素数が 2 の場合，$3 \times 2 = 6$ つの特徴量があることになる．

4 評価

各特徴抽出方法を用いた機械学習により得られた分類器のカテゴリ推定精度を評価するために，様々なマイマップに登録されている 9200 件の POI データをテストデータとして使用する．表 1 に各特徴抽出における平均カテゴリ推定精度を示す．origin に関しては，表から分かるようにカテゴリを推定することは不可能であることが分かる．unigram に関しては Choi らの既存研究の推定精度を 14.44 ポイント上回っている．これは SVR ベースの機械学習に変更したことによる影響であると考えられる．mecab は unigram より推定精度が 6.76 ポイント上回り，hybrid に関しては約 10 ポイントも高い結果を得ることが出来ている．次に，図 2 に各特徴抽出における各カテゴリの推定精度を示す．origin に関しては，図中の赤丸のようにほとんどの POI のカテゴリが Dinner と推定され，分類できないことが分かる．図中の赤線が引かれたカテゴリは hybrid が最も良い推定結果であることを示し，unigram と mecab の両方の特徴抽出における良い部分をうまく組み合わせることが出来ていると考えられる．

[3] MeCab：http://taku910.github.io/mecab/

5 おわりに

マイマップに登録されている POI のカテゴリを平均 70.27%の精度で推定可能な分類器を構築するために，機械学習の際に用いる unigram と形態素解析を用いた新たな特徴抽出法を提案した．

参考文献

[1] Choi, Su Jeong, Seong-Bae Park, and Kweon-yang Kim. ”Estimating Category of Pois Using Contextual Information.” Indian Journal of Science and Technology 8.S7. pp.718–723. 2015.

[2] Crammer, Koby, et al. ”Online passive-aggressive algorithms.” Journal of Machine Learning Research 7. pp.551–585. 2006.