旅行ブログ記事における観光地の傾向の自動分析

高原明日美　徳久雅人　木村周平
鳥取大学大学院 工学研究科 情報エレクトロニクス専攻
{s112029, tokuhisa, kimura}@ike.tottori-u.ac.jp

1 はじめに

旅行者に観光地を推薦するシステムが必要とされている．そうしたシステムを開発する際，観光地の特徴を把握する必要がある．例えば，温泉で有名であるや，自然散策と飲食の行動が目立つという傾向の把握である[1]．

観光地の特徴を把握する方法として，アンケート分析が挙げられる．しかし，アンケート分析はコストが掛かることがデメリットとして挙げられる．そこで，低コストで観光地の特徴を把握する必要がある．

低コストで把握する方法として，旅行日記として書かれたブログ記事を参考にすることが挙げられる．しかし，ブログ記事は日々多くの人に書かれているため，特定の観光地に絞ったとしても，手動で全文を読むことは困難である．

ある観光地について書かれたブログから自動的に観光地の特徴を把握するために，文単位で主成分分析やクラスタ分析 (Ward 法) を行った．あまり良い結果は得られなかった．その結果，教師なしでの分類は難しいことがわかった．先行研究 [1] では教師ありとして SVM を使用して文単位でヒント文 (観光地の様子を説明する文) か否かを判定していたが，文単位ではあまりいい評価が得られなかった．両者の方法では，文を単位としていたことが問題であった可能性がある．

そこで，本稿では文脈単位でヒント文を判定し，観光地の特徴を抽出する．文脈単位では，文脈の幅が伸び縮みするので，島 (連続したヒント文) という概念を使用する．島を使用してカテゴリ推定を行うことで，様々な観光地において，どのカテゴリの特徴が強く現れているのかを分析することができると考える．文脈の幅を上手く設けることで，低コストで観光地の特徴を簡単に把握することができると考える．

2 使用するデータ

本稿では，コーパスとして糸魚川ブログデータ (3,222 文)，江ノ島海岸ブログデータ (660 文)，三陸海岸ブログデータ (6,356 文)，若狭湾ブログデータ (5,027 文) を使用する．カテゴリは 17 分類である．コーパスの一部を表 2 に示す．

表 2 では，id000607〜id000611 の文のうちに連続してカテゴリ「温泉」が付与されている，id000608〜id000610 までの文を 1 つの島とみなす．

表 1: カテゴリ名一覧

#	カテゴリ	#	カテゴリ
1	自然散策	10	音楽
2	動植物	11	スポーツ・アウトドア
3	飲食	12	釣り
4	買い物	13	交流
5	街並み	14	産業
6	施設	15	交通
7	温泉	16	行事
8	神社仏閣	17	その他
9	文化歴史		

表 2: コーパスの例

id	文	カテゴリ
I000607	天気も良く最高の気分です。	ヒントなし
I000608	檜風呂が隣に有ります。	温泉
I000609	このユニークな洗い場は室内に有ります。	温泉
I000610	野天風呂からの眺め、前は山、後は早川の激流です。	温泉
I000611	住所南巨摩郡早川町湯島 1780-7	ヒントなし

3 提案手法

旅行ブログ記事のカテゴリを推定する手法として，文脈単位で分類を行う手法を提案する．その手順を以下に示す．

手順 1 入力される 2 文に対して，島となるか否かを判定する SVM を構成する．島の判定器の学習では，1 文目と 2 文目を比較し，カテゴリが同じであれば +1，違う場合は −1 を付与する．ただし，どちらもカテゴリが付与されていない場合はトレーニングデータから除く．

手順 2 手順 1 の判定器を用いて，入力テキストから島を抽出する．

[1]温泉・自然散策などをカテゴリと呼ぶことにする (表 1)

手順 3 島に対してカテゴリ推定を行う SVM を構成する．One vs Rest 法とする．

手順 4 手順 3 の判定器を用いて，手順 2 で得た島のカテゴリを推定する．

使用する品詞は名詞 (サ変接続・ナイ形容詞語幹・一般・形容動詞語幹・固有名詞・接続詞的・接尾・副詞可能)，動詞 (自立)，形容詞 (自立・非自立) である．

使用する素性は (1) 1 文目と 2 文目の単語ペア，(2) 2 文目の単語 Unigram，(3) 2 文間の共通単語，(4) 共通単語の有無 (共通あり/共通なし)，(5) 特定の名詞の意味属性コード，(6) 単語 Unigram である．

使用する名詞の意味属性コードを以下に示す．日本語語彙大系 [3] のコードである．NI は一般名詞意味属性，NK は固有名詞意味属性であることを表す．

飲食 NI:838〜862(食料)

施設 NK:55〜59，NK:61〜65(施設名)

温泉 NI:505(泉)，NI:749(湯)，NK:60(鉱山・泉等名)

神社仏閣 NI:453〜454(神社・寺院名)

文化歴史 NI:1146(伝承)

釣り NI:542〜547(魚介名)

交通 NI:986〜990(乗り物 (本体))，NK:123(乗り物名)

行事 NI:1229(祭)，NI:1230(行事)，NI:1233(催し)

手順 1，2 では (1)〜(5) の素性を用いる．手順 3，4 では (6) の素性を用いる．

4 比較手法

文単位での自動分類を比較手法として使用する．すなわち提案手法の手順 3，4 にて，長さ 1 の島 (1 文) を入力する手法である．

5 評価方法

- カテゴリ毎の適合率

  比較手法においては 1 島＝ 1 文として扱う．

$$\text{カテゴリ毎の適合率} = \frac{\text{カテゴリ毎の一致島数}}{\text{カテゴリ毎の推定島数}} \tag{1}$$

- 適合率 P，再現率 R，F 値

$$\text{適合率 } P = \frac{\text{正解カテゴリと一致した推定島数}}{\text{カテゴリが推定された島の数}} \tag{2}$$

$$\text{再現率 } R = \frac{\text{正解カテゴリと一致した推定島数}}{\text{正解の島の数}} \tag{3}$$

$$F \text{ 値} = \frac{2PR}{P+R} \tag{4}$$

- z 得点

  z 得点は，平均が 0，標準偏差が 1 となるようにカテゴリ毎の島数を変換して得られる点である．

  以下にカテゴリ i の島数 x_i に対する z 得点 z_i を求める方法を示す．ここで，N はカテゴリ数とする ($N = 17$)．

  1. 島の平均 $\bar{x}$ を求める．

$$\bar{x} = \frac{\sum_{i=1}^{N} x_i}{N} \tag{5}$$

  2. 分散 s^2 を求める．

$$s^2 = \frac{\sum_{i=1}^{N} (x_i - \bar{x})^2}{N} \tag{6}$$

  3. カテゴリ i に対する z 得点を求める．

$$z_i = \frac{x_i - \bar{x}}{\sqrt{s^2}} \tag{7}$$

6 実験

糸魚川コーパス (3,222 文) をトレーニングデータとする．江ノ島海岸コーパス (660 文)，三陸海岸コーパス (6,356 文)，若狭湾コーパス (5,027 文) それぞれをテストデータとして実験を行う．

6.1 比較手法：文単位でのカテゴリ推定

3 つの海岸コーパスの推定結果を確認し，それぞれの推定カテゴリの得られた文数 (推定文数)，一致文数 (正解カテゴリと推定カテゴリの一致した文数)，および，カテゴリ毎の適合率を算出した．その結果を表 3 に示す．

カテゴリ毎の適合率のマクロ平均は 0.2 以下となり，低い結果となった．正解文数が少ないカテゴリに対して，過剰に推定してしまうことがあった．例えば，表 3 の江ノ島海岸の#12 では，正解分数が 72 であるのに対して，182 件の推定があった．

6.2 提案手法：文脈単位での島の推定

表 4 に島判定およびカテゴリ推定の例を示す．島の大きさが 2 以上の場合は，文境界を/で表した．id が S00156，S00157 の島では，正解カテゴリは「文化歴史」だが，推定結果は「交通」となった．文の内容は

表 4: 三陸海岸の島の具体例

島の id	文	正解	推定結果
S00156，S00157	凡そ、３６０年前の江戸初期（１６４３）年、水と食料を求めたオランダ商船・ブレスケンス号が、山田湾に入り大島に停泊した。/知らせを聞いた大槌代官所の奉行はとりあえず給水を許可し盛岡藩主に報告、盛岡藩は幕府に指示を仰ぐ。	文化歴史	交通
S00181，S00182	お刺身付蕎麦か海鮮ラーメン、どちらか選べます。/わたしはおそば、主人はラーメンにしました。	飲食	飲食
S00187	船の後をウミネコが追ってきます。	動植物	交通

表 3: 1 文単位・比較手法による推定の適合率

#	江ノ島海岸	三陸海岸	若狭湾
1	0.36 (8/ 71)	0.32 (23/ 781)	0.33 (24/ 591)
2	0.38 (3/ 41)	0.36 (9/ 478)	0.19 (8/ 413)
3	0.16 (5/115)	0.20 (21/1223)	0.24 (26/ 946)
4	0.11 (1/ 49)	0.00 (0/ 507)	0.04 (2/ 348)
5	0.00 (0/ 7)	0.00 (0/ 58)	0.00 (0/ 36)
6	0.00 (0/ 14)	0.31 (17/ 369)	0.19 (5/ 196)
7	0.00 (0/ 35)	0.00 (0/ 406)	0.00 (0/ 342)
8	0.00 (0/ 2)	0.00 (0/ 29)	0.00 (0/ 18)
9	0.00 (0/ 2)	0.73 (8/ 80)	0.33 (1/ 23)
10	0.00 (0/ 0)	0.00 (0/ 7)	0.00 (0/ 3)
11	0.17 (1/ 31)	0.03 (1/ 270)	0.19 (4/ 216)
12	0.00 (0/182)	0.00 (0/1379)	0.02 (2/1289)
13	0.00 (0/ 0)	0.20 (1/ 17)	0.00 (0/ 8)
14	0.00 (0/ 0)	0.00 (0/ 0)	0.00 (0/ 0)
15	0.35 (13/ 92)	0.26 (10/ 543)	0.08 (3/ 421)
16	0.00 (0/ 2)	0.00 (0/ 71)	0.00 (0/ 41)
17	0.25 (3/ 17)	1.00 (12/ 138)	0.09 (21/ 136)

歴史について述べているが，“船” という単語が含まれていることで，カテゴリが「交通」となってしまったと考えられる．次の s00181 と s00182 の島は正しく推定できている．

手順 2 で抽出された島の数を表 5 に示す．また，抽出すべき島の数および構成する文数を参考のために示す．手順 4 で江ノ島海岸・三陸海岸・若狭湾コーパスでの島のカテゴリ推定を行った結果を表 6 に示す．

提案手法と比較手法の評価値を表 7 に示す．合計はマイクロ平均である．提案手法の方が優れた結果となった．

表 7 によると，マイクロ平均で高めの適合率であったが，マクロ平均をとると 0.17〜0.32 であった．原因は適合率 0 が多いためである．これの良し悪しについては次節で述べる．

表 5: 島の判定の結果

コーパス	正解の島数	推定された島数
江ノ島海岸	32	15
三陸海岸	113	72
若狭湾	123	17

表 6: 文脈単位・提案手法による推定の適合率

#	江ノ島海岸	三陸海岸	若狭湾
1	0.0 (0/ 1)	0.3 (3/11)	1.0 (3/ 3)
2	0.0 (0/ 0)	0.0 (0/ 0)	0.0 (0/ 0)
3	1.0 (7/ 7)	0.9 (8/ 9)	0.0 (0/ 2)
4	0.0 (0/ 0)	0.0 (0/ 1)	0.0 (0/ 0)
5	0.0 (0/ 0)	0.0 (0/ 1)	0.0 (0/ 0)
6	1.0 (1/ 1)	0.9 (8/ 9)	0.8 (3/ 4)
7	1.0 (1/ 1)	0.7 (8/11)	0.0 (0/ 1)
8	0.0 (0/ 0)	0.0 (0/ 0)	0.0 (0/ 0)
9	0.0 (0/ 0)	1.0 (9/ 9)	0.7 (2/ 3)
10	0.0 (0/ 0)	0.0 (0/ 0)	0.0 (0/ 0)
11	0.0 (0/ 0)	0.0 (0/ 0)	0.0 (0/ 0)
12	0.0 (0/ 0)	0.0 (0/ 0)	0.0 (0/ 0)
13	0.0 (0/ 0)	0.0 (0/ 0)	0.0 (0/ 0)
14	0.0 (0/ 0)	0.0 (0/ 0)	0.0 (0/ 0)
15	0.6 (3/ 5)	0.7 (13/20)	0.5 (2/ 4)
16	0.0 (0/ 0)	0.0 (0/ 0)	0.0 (0/ 0)
17	0.0 (0/ 0)	1.0 (1/ 1)	0.0 (0/ 0)
平均	0.8 (12/15)	0.7 (50/72)	0.6 (10/17)

表 7: 評価値の比較

コーパス	手法	P	R	F 値
江ノ島海岸	提案手法	0.80	0.38	0.51
	比較手法	0.05	0.22	0.08
三陸海岸	提案手法	0.69	0.44	0.54
	比較手法	0.02	0.19	0.03
若狭湾	提案手法	0.59	0.08	0.14
	比較手法	0.02	0.18	0.03
合計	提案手法	0.69	0.27	0.39
	比較手法	0.02	0.19	0.04

6.3 z 得点の比較

6.1，6.2 節の結果をもとに，3 つの海岸コーパスのカテゴリ毎の正解文数の z 得点，推定文数の z 得点，および，推定島数の z 得点をそれぞれ求める．それぞれの z 得点のグラフを図 1 に示す．推定文数の z 得点では 3 つのコーパスで「釣り」の話題が多いと判定されたことがわかる．しかし，正解文数の z 得点による

と，それは誤りであることがわかる．一方，推定島数の z 得点によると，正解文数の z 得点と同様に負値であるため正しく判定されたことがわかる．

z 得点による比較では，正解文数（島数）が少ないカテゴリについて推定文数（島数）が少なくなることが重要である．ゆえに，前節での良し悪しについていうと，推定数が 0 になることはやぶさかではない．

z 得点の符号の一致を基準に全体を評価する方法が妥当といえる．そこで，符号の一致度について調べた結果を表 8 に示す．なお,「符号の一致度」＝「2 つの z 得点の符号の一致数」/「カテゴリ数」で計算する．

正解文数の z 得点と推定島数の z 得点の一致度のほうが，推定文数との一致度よりも高くなっているので，提案手法は，より正確にカテゴリ推定をすることができたと言える．機械学習としての性能評価としては甘い評価判定にみえるが，観光地の傾向を分析する上では，妥当な評価判定と，本稿では考える．

表 8: z 得点の一致度

	江ノ島海岸	三陸海岸	若狭湾
文単位の一致度	0.82	0.71	0.59
島単位の一致度	1.00	0.88	0.88

7　おわりに

文脈単位の実験と文単位の実験の結果を比較すると，文脈単位の実験の方がカテゴリ毎の適合率，適合率・再現率・F 値が上昇した．このことから，文脈単位で観光地の自動分類を行うことで，上手く観光地の特徴を抽出することができた．

参考文献

[1] 高原明日美, 徳久雅人, 村田真樹, 村上仁一: ブログから観光開発案を分析する際の分析者間の比較，言語処理学会第 21 回年次大会発表論文集, pp.155–158, 2015.

[2] 石野亜耶, 難波英嗣, 竹澤寿幸, 知能と情報 (日本知能情報ファジィ学会誌), Vol.22, No.6, pp.667–679, 2010.

[3] 池原悟, 宮崎正弘, 白井論, 横尾昭男, 中岩浩巳, 小倉健太郎, 大山芳史, 林良彦: 日本語語彙大系, 岩波書店, 1997.

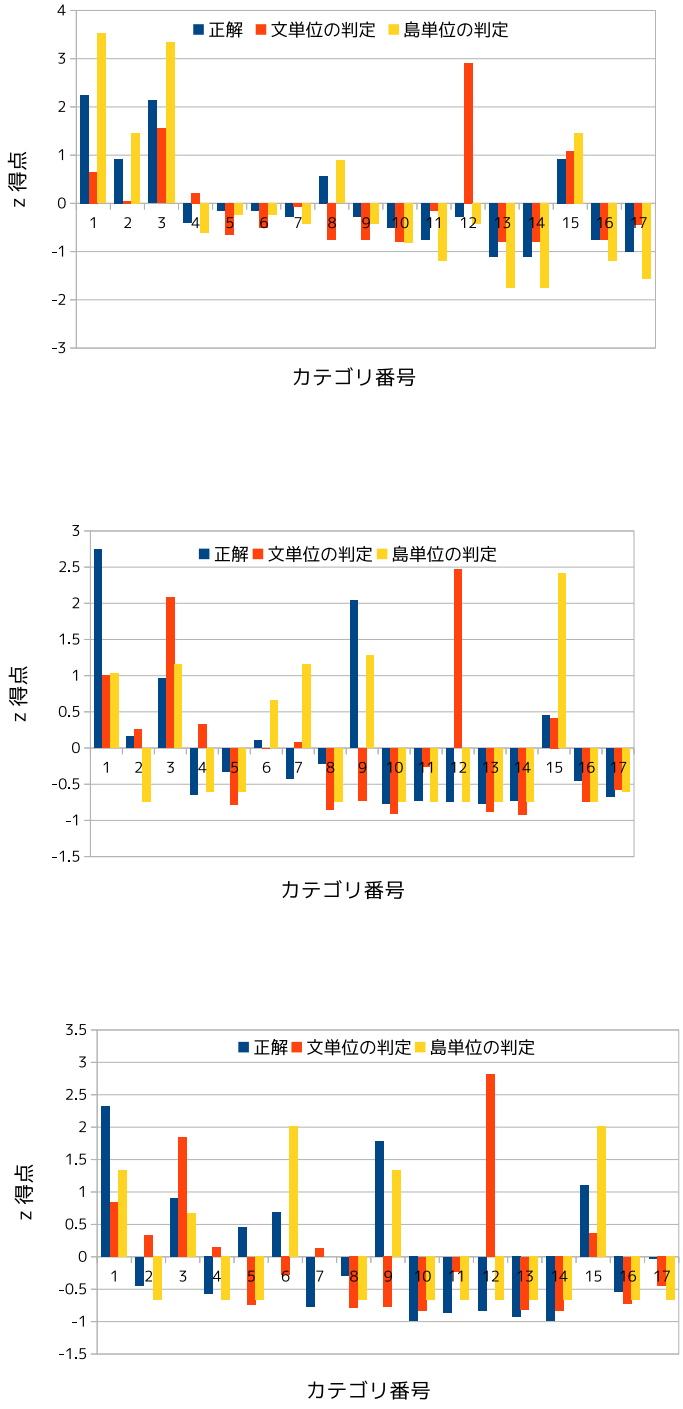

図 1: カテゴリ毎の z 得点の比較（上より江ノ島海岸，三陸海岸，若狭湾）