

## カテゴリーに特徴的な名詞の抽出と利用による Web サイトの自動分類

### Automatic Website Classification by Extracting and Using Characteristic Nouns of Category

本田 崇智† 山本 雅人† 大内 東†  
 Takatomo Honda Masahito Yamamoto Azuma Ohuchi

#### 1. はじめに

Web サイトは Yahoo! Japan のようなディレクトリ型検索エンジンで見られるように様々なカテゴリーに分類されている。カテゴリー中の Web サイトは類似の内容を示しているため、カテゴリーごとに特徴的な名詞が存在すると思われる。そのようなカテゴリーに特徴的な名詞を利用することによって、任意の Web サイトをカテゴリーへと自動分類することが可能になると考えられる。

任意の Web サイトをカテゴリーへと分類する際、分類したい Web サイトは任意のものであらかじめ決められたカテゴリーに属さない場合も存在する。そのため本研究では、カテゴリーに特徴的なキーワードを用いてあらかじめ決められたカテゴリー以外のサイトであるかどうかを判定可能な分類手法を提案し、同様にテキスト情報を用いた関連手法との比較実験を行う。

以下、2 節では関連手法について述べる。3 節では提案手法の詳細について述べる。4 節では提案手法と関連手法との比較実験を行い、有効性を検証する。5 節では結論を示す。

#### 2. 関連研究

本節では、本研究と同様にテキスト情報を用いた分類手法の関連研究について述べる。

テキスト情報を用いた分類手法について、 Bayesian Classifier を用いた手法が提案されている[1]。

Bayesian Classifier は分類したい未知のサイトが与えられたときに、カテゴリーに対する事後確率を求ることでカテゴリーへと分類する手法である。事後確率は名詞の出現によって計算され、そのうち Multi-variate Bernoulli Model は名詞が出現するかどうかによって推定する手法である。

しかしこの手法はあらかじめ与えられたカテゴリーのうちどのカテゴリーに属するか決定する手法であり、 Web サイトがどのカテゴリーにも属さないということが判定できない。そのため本提案手法によって Web サイトがどのカテゴリーにも属さないという判定ができることが検証されれば、関連研究と比較して有効性があると考えられる。

#### 3. Web サイトの自動分類手法

本研究では、現存するディレクトリ型検索エンジンに存在するカテゴリーを選択し、カテゴリー内の Web サイトを対象に本提案手法を適用し、それらを教師集合とした学習問題を扱う。またそれらの Web サイトの一部を教師集合ではなくテスト集合として用い、本手法の有効性の検証に用いる。

##### 3.1 キーワードの抽出

本研究では、カテゴリー中のほとんどの Web サイトに出現する名詞で、かつ他のカテゴリーの Web サイトにはあまり現れない名詞をカテゴリーに特徴的なキーワードと

呼び、このようなキーワードを抽出する。まず対象とするカテゴリー  $i$  ( $i=1, \dots, n$ :  $n$  はカテゴリー数) 中の Web サイトに出現する名詞をすべて抽出する。名詞の抽出にあたっては形態素解析システムである茶筅を用いた[2]。

最初にカテゴリー  $i$  に含まれる名詞  $w$  のスコアとして  $F(i, w)$  を以下のように定義する。

$$F(i, w) = \frac{df(i, w)}{|S_i|} \quad (1)$$

ただし、 $S_i$  はカテゴリー  $i$  に属する Web サイト集合であり、  $df(i, w)$  はカテゴリー  $i$  中で名詞  $w$  が出現する Web サイト数である。

次に、カテゴリーごとに Web サイトの総数や出現する名詞の総数が異なるために  $F(i, w)$  の値に偏りがみられるので、カテゴリーごとに  $F(i, w)$  の相対値をとる以下の変換を行う。

$$F_{rel}(i, w) = \frac{F(i, w)}{\sum_{w' \in W_i} F(i, w')} \quad (2)$$

ただし、 $W_i$  はカテゴリー  $i$  に属する Web サイトに現れる名詞集合である。

ここで  $F_{rel}(i, w)$  の値が大きな名詞はカテゴリー内の多くの Web サイトに出現する名詞であるため重要であると考えられるが、"お客様"などどのカテゴリーにも現れる一般的な語である場合が多い。

そのため他のカテゴリーには現れない特徴的なキーワードを抽出するため、カテゴリー  $i$  に現れる名詞  $w$  のスコアを以下のように定義する。

$$R(i, w) = \frac{F_{rel}(i, w)}{F_{rel}(i, w) + F_{rel}(\bar{i}, w)} \quad (3)$$

ただし、 $\bar{i}$  は  $i$  以外の全てのカテゴリーであり、それら全てのカテゴリーをまとめて一つのカテゴリーとして  $F_{rel}(\bar{i}, w)$  を計算する。

$R(i, w)$  の計算によって、カテゴリー  $i$  中の多くの Web サイトに出現しあつ  $i$  以外のカテゴリー中の Web サイトにはあまり出現しない特徴的なキーワードが抽出される。Web サイトの自動分類にあたって、あらかじめ訓練集合のサイトについてこれらの値を計算しておく。

##### 3.2 Web サイトの自動分類法

本項では、前項で述べられたカテゴリーに特徴的なキーワードを用いて、まだどのカテゴリーに分類されるべきかわからない未知のサイトをカテゴリーに分類する手法について述べる。

最初に未知のサイトに含まれる全ての名詞を抽出し、カテゴリーに特徴的なキーワードとの比較を行い、カテゴリーの類似度を計算する。ある Web サイト  $s$  とカテゴリー  $i$  との類似度  $similarity(s, i)$  を以下のように定義する。

$$\text{similarity } (s, i) = \frac{\sum_{w \in W_s} R(i, w)}{|W_s|} \quad (4)$$

ただし、 $W_s$ は Web サイト  $s$  に現れる名詞集合である。カテゴリーごとに未知の Web サイトとの類似度  $\text{similarity}$  を計算し、カテゴリーごとに設定された閾値を超えた場合にはそのカテゴリーに分類する。閾値を超えない場合は、そのカテゴリーへの分類を行わない。

### 3.3 閾値の設定

本項では、未知の Web サイトをカテゴリーに分類するにあたって必要なカテゴリーごとの閾値の設定法について述べる。

まず対象カテゴリーの抽出された特徴的なキーワードを用いて、同じ教師集合の Web サイトとの類似度を計算する。そして対象カテゴリー中の Web サイト群とその他のカテゴリー中の Web サイト群との類似度を比較することで閾値を設定する。本研究では、対象カテゴリー中の Web サイト群との類似度の最小値から最大値まで 0.01 刻みで変化させ、実際にそれらの Web サイト群を対象カテゴリーへと分類を行い、そのうち最も高い分類精度を示したときの閾値をそのカテゴリーの閾値とする。

## 4. 実験

本節では、前節で提案した自動分類手法の有効性の検証と関連研究との比較のための実験を行う。

### 4.1 実験データ

Yahoo! Japan から文献を参考に観光に関連があると考えられる以下の 10 カテゴリーを選択した[3]。各カテゴリー内の総サイト数も以下の通りである。

- 趣味とスポーツ>スポーツ>ゴルフ (549 サイト)
- 旅行と交通>宿泊施設>ペンション (1020 サイト)
- エンターテインメント>レストラン>和食 (503 サイト)
- 芸術と人文>美術館・ギャラリー (409 サイト)
- 生活と文化>祭り (485 サイト)
- エンターテインメント>グルメ>カフェ・喫茶・甘味 (358 サイト)
- 旅行と交通>宿泊施設>旅館 (1020 サイト)
- エンターテインメント>テーマパーク・遊園地 (190 サイト)
- 趣味とスポーツ>スポーツ>施設・競技場 (91 サイト)
- 趣味とスポーツ>アウトドア>公園 (181 サイト)

加えて、分類精度の検証を行うために観光とは関連の薄いと考えられる以下の 8 カテゴリーを選択した。

- 教育>高等学校 (1020 サイト)
- メディアとニュース>新聞 (173 サイト)
- エンターテインメント>音楽>アーティスト (193 サイト)
- 政治と行政>国の機関 (188 サイト)
- 生活と文化>ボランティア (252 サイト)

- 健康>保健所 (68 サイト)
- ビジネスと経済>ショッピングとサービス>おもちゃ (158 サイト)
- ビジネスと経済>ショッピングとサービス>住まい>家具・インテリア>家具 (619 サイト)

これら 18 カテゴリー、計 7284 サイトについて、各カテゴリーごとにランダムに選択した 90% のサイトを訓練集合と呼び、キーワードの抽出と閾値の設定に用いる。また残りの 10% をテスト集合と呼び、分類精度の検証に用いる。

### 4.2 関連手法による実験

本項では、提案手法による分類精度と比較するために 2 節で触れた関連手法による分類精度の実験を行う。

Bayesian Classifier では、未知のサイト  $d_i$  が与えられたとき、カテゴリー  $c_j$  に対する事後確率は以下の式で定義される。

$$p(c_j / d_i) = \frac{p(c_j) p(d_i / c_j)}{p(d_i)} \quad (5)$$

ただし、 $p(c_j)$  は全体におけるカテゴリー  $c_j$  中の Web サイト数の割合である。 $p(c_j / d_i)$  から、分類されるカテゴリーは以下の式で決定される。

$$c^*(d_i) = \arg \max_{c_j} p(c_j) p(d_i / c_j) \quad (6)$$

ここで Multivariate-Bernoulli Model では、 $p(d_i / c_j)$  が以下の式で定義される。

$$c^*(d_i) = \operatorname{argmax}_{c_j} p(c_j) \prod_{t=1}^V (B_{it} p(w_t / c_j) + (1 - B_{it})(1 - p(w_t / c_j))) \quad (7)$$

$$p(w_t / c_j) = \frac{1 + tw(c_j, w_t)}{2 + t(c_j)} \quad (8)$$

ただし、 $B_{it}$  はサイト  $d_i$  に  $w_t$  が出現したときに 1、出現しないときに 0 をとり、 $tw(c_j, w_t)$  はカテゴリー  $c_j$  に属しかつ  $w_t$  が出現する Web サイト数、 $t(c_j)$  はカテゴリー  $c_j$  に属する Web サイト数、 $V$  は全名詞数である。

また名詞の選択は以下の Feature Selection によりスコア  $I(C; f)$  の降順に選択する。今回は予備実験により最も高い精度を示した数である 1000 語を用いる。

$$I(C; f) = \sum_C \sum_{f_i \in (0, 1)} p(c, f_i) \log \frac{p(c, f_i)}{p(c)p(f_i)} \quad (9)$$

ただし、 $f_i$  は Web サイト中に  $w_t$  が出現するときに 1、出現しないときに 0 をとり、 $p(c, f_i)$  は全 Web サイト中のカテゴリー  $c$  の  $w_t$  が出現する Web サイト数の割合、 $p(f_i)$  は全 Web サイト中の  $w_t$  が出現する Web サイト数の割合、 $p(c)$  は全 Web サイト中のカテゴリー  $c$  に属する Web サイト数の割合である。

また分類精度の計算式は、以下で定義される適合率、再現率、 $F1$  という 3 つの尺度を用いた[4][5]。

$$\text{再現率} = \frac{N_p}{N_p + N_f} \quad (10)$$

$$\text{適合率} = \frac{N_p}{N_p + N_f} \quad (11)$$

$$F1 = \frac{2 \times \text{適合率} \times \text{再現率}}{\text{適合率} + \text{再現率}} \quad (12)$$

ただし、あるカテゴリーに対して分類を行う場合、実際にそのカテゴリーに属する Web サイトを正例、他のカテゴリーに属する Web サイトを負例と呼ぶ場合、そのカテゴリーに分類した Web サイト集合のうち実際に正例であった Web サイト数を  $N_p$ 、実際は不例であった Web サイト数を  $N_n$  とする。またそのカテゴリーに属さないと判断した Web サイト集合のうち、実際は正例であった Web サイト数を  $N_{pn}$  とする。

名詞の学習には全 18 カテゴリーの訓練集合を用い、分類のテストに観光に関係のあると思われる 10 カテゴリーを用いたときの分類結果が以下の表 1 である。

表 1：関連手法による分類結果

カテゴリー	再現率	適合率	F1
ゴルフ	0.86	0.97	0.91
ペンション	0.77	0.94	0.85
和食	0.83	0.44	0.58
美術館	0.61	0.91	0.73
祭り	0.78	0.82	0.8
カフェ	0.59	0.62	0.6
旅館	0.74	0.83	0.78
テーマパーク	0.29	0.71	0.42
施設	0.71	0.63	0.67
公園	0.5	0.24	0.33

#### 4.3 提案手法による実験

本項では 3 節で提案した手法の有効性を示すための実験を行う。キーワードの抽出と閾値の設定には全 18 カテゴリーの訓練集合を用い、分類のテストに観光に関係のあると思われる 10 カテゴリーを用いたときの分類結果が以下の表 2 である。

表 2：提案手法による分類結果

カテゴリー	閾値	再現率	適合率	F1
ゴルフ	0.49	0.95	0.89	0.92
ペンション	0.55	0.79	0.8	0.8
和食	0.57	0.74	0.97	0.84
美術館	0.5	0.82	0.93	0.87
祭り	0.47	0.95	0.85	0.9
カフェ	0.51	0.56	0.75	0.64
旅館	0.58	0.67	0.73	0.69
テーマパーク	0.49	0.59	0.42	0.49
施設	0.49	0.57	0.8	0.67
公園	0.45	0.82	0.54	0.65

#### 4.4 考察

提案手法と関連手法との分類精度を比較したとき、カテゴリー“ペンション”と“旅館”以外では提案手法が関連手法以上の精度を示した。これら 2 つのカテゴリーでうまくい

かなかつた原因としては、提案手法では名詞を抽出して分類を行うので、カテゴリー“ペンション”と“旅館”は同じ宿泊施設のカテゴリーに属し似たような名詞が多く存在するため互いに間違って分類されているからだと考えられる。実際に提案手法では、間違って“ペンション”に分類された全 18 サイト中、“旅館”的サイトが“ペンション”的サイトと間違って分類されたサイト数が 14 サイトとほとんど全てを占めている。また間違って“旅館”に分類された全 22 サイト中、“旅館”的サイトが“ペンション”的サイトと間違って分類されたサイト数が 12 サイトとこちらも半分以上を占めている。

このことから、収集したいカテゴリー中に似たような名詞が多く出現するカテゴリーが存在するときには Bayesian Classifier を組み合わせることによってより高い分類精度を示すことも可能であると考えられる。

#### 5 おわりに

本研究では、任意のカテゴリーの Web サイトを自動収集するための Web サイトのカテゴリーへの自動分類手法について提案した。カテゴリーごとに Web サイトから特徴的なキーワードを抽出し、Web サイトとの類似度を計算し適切な閾値を設定することでそのカテゴリーに属するか属さないかの判定を可能にした。またディレクトリ型検索エンジン Yahoo! Japan を用いて、単純に分類精度の点から関連手法との比較を行い、有効性があることを示した。

今後は、ディレクトリ型検索エンジンのデータに加えて WWW 上に存在する膨大な Web サイトを対象に本手法の有効性を示していく予定である。

#### 参考文献

- [1] Daphne Koller, Mehran Sahami, “Hierarchically Classifying Documents Using Very Few Words”, Proceedings of ICML-97, 14<sup>th</sup> International Conference on Machine Learning, 1997.
- [2] 茶筅(<http://chasen.naist.jp/hiki/Chasen/>)
- [3] 岡本伸之: 観光学入門, 有斐閣アルマ, 2001.
- [4] C.J. van Rijsbergen: Information Retrieval, London: Butterworths, 1979.
- [5] D.Lewis, “Evaluating Text Categorization”, Proceedings of the Speech and Natural Language Workshop, pp.312-318, 1991.