D-2 Support Vector Machineを用いた地域情報ページの自動分類

Automatic Classification of Location Information Web Pages using Support Vector Machines

河合英紀† 山田寛康‡* 中川哲治‡** 佐々木寛‡ 赤峯享† 松本裕治‡ 福島俊一†
Hideki KAWAI Hiroyasu YAMADA Tetsuji NAKAGAWA Hiroshi SASAKI
Susumu AKAMINE Yuji MATSUMOTO Toshikazu FUKUSHIMA

1.はじめに

近年、WWW の発展と利用者の全国的な拡大に伴い、ホテルやグルメ情報など各地域に特化した情報（地域情報）が数多く流通するようになり、まち goo やぐるなびなど地域情報を専門に扱う地域情報ポータルが出現している。地域情報ポータルでは、利用者は住所と同時に自分の検索目的に該当するカテゴリや条件を指定して検索できる。一方、これら地域情報ポータルでは、データの収集を人手に頼っているため、コストやデータの網羅性の面で問題があり、地域情報ページの自動収集と分類技術が望まれる。

地域情報の自動収集・分類の従来技術としては、モバイルインフォサーチ[1]や空間情報抽出システム「芭蕉」[2]など、Web ページから住所を自動抽出する研究は数多くなされている。しかし、利用者の検索目的であるカテゴリを軸に分類することはあまりなされていない。大槻らの研究[3]では、自治体のページを自動的に収集しカテゴリ固有語辞書を用いて自動分類を行っているが、ページの収集範囲や分類可能なカテゴリの種類に限界がある。

本研究では、住所とカテゴリの 2 軸で地域情報ページを自動分類する。住所の抽出は、辞書を用いた単純なパターンマッチで行った。一方、辞書やルールの作成が難しいカテゴリの分類には、統計的学習理論に基づく Support Vector Machine (SVM)[4]を用いた。また、住所情報を利用することでカテゴリの分類精度が向上した。

2．地域情報検索システム

表 1 に、Web サーチエンジン（BIGLOBE サーチ[5]）における利用頻度上位の地名 3 語（大阪、北海道、京都）と同時に検索に使われた単語の頻度の割合を示す。表 1 を見ると、地名が単独で入力されるよりも、他に検索目的となる単語が同時に入力されている場合が大半であることが分かる。また、地域によって多少ばらつきはあるが、レジャーやグルメに関する語（温泉、ホテル、ラーメンなど）が多く入力されている。そこで本研究では、大分類として「レジャー」「グルメ」「非地域情報」の 3 カテゴリに、また、「レジャー」のサブカテゴリとして、「観光案内」「宿泊施設」「娯楽施設」「文化施設」の 4 カテゴリ、「グルメ」のサブカテゴリとして、「和食」「洋食」「中華」「飲み屋」「喫茶店」の 5 カテゴリに分類することを目標とした。

住所とカテゴリの 2 軸は、本来独立のはずである。しかし、実際には住所が抽出できたページは何らかの地域情報ページである可能性が高い。そこで、住所を抽出できたページに限定して上記のような地域情報固有のカテゴリに分類するならば、分類精度を向上させられる可能性がある。

† NEC インターネットシステム研究所
‡ 奈良先端科学技術大学院大学 情報科学研究科
‡* 現在は北陸先端科学技術大学院大学に所属
‡** 現在は沖電気工業株式会社に所属

表 1 検索利用頻度が高い地名とクエリ中の共起語

	大阪	北海道	京都
単独で入力	6%	24%	23%
レジャー・グルメ関連語	41%	35%	30%
公共施設・交通関連語	12%	7%	9%
他の地名（梅田、札幌など）	9%	13%	8%
イベント・出会い関連語	8%	6%	3%
その他	25%	16%	26%

3. 住所抽出

住所は、丁目番地の表記に多少揺らぎがあるものの、県名・市区町村名レベルではある程度統一された表記方法がとられている。そこで、地名辞書を強化した日本語形態素解析システム「茶筌」[6]を用いて、下記に示す比較的単純なパターンマッチで抽出を行った。

(1)HTML タグを除去し、空白文字で区切って文に分割する
(2)文を「茶筌」で形態素に分割する
(3)品詞細分類情報に地域タグを持つ形態素が 3 つ以上連続して現れている文を住所候補とする
(4)住所候補には、ニュース記事などの文中に現れる住所も含まれている。そのため「で」や「から」などの助詞が含まれていればその文を棄却し、含まれていなければその文を住所として抽出する。ただし、助詞が「の」だけの場合、「高の原」のような地名の可能性があるのでその文は住所とする。
(5)すべての文について(2)～(4)を繰り返す。

4. 地域情報分類

カテゴリは特定の話題を持つページの集合であり、これを精度よく分類できる規則を人手で記述するのは困難である。一方、機械学習の手法を用いて分類規則を自動学習すれば、分類規則の作成コストを削減することができる。

SVM は統計的学習理論に基づく新しい 2 値分類器であり、文書分類をはじめ自然言語処理の様々なタスクで高い認識性能を示している事から近年注目を集めている[4]。図 1 に SVM の概念図を示す。SVM は、素性空間に分布する正負例を正しく分離する数多くの超平面の中から、マージンが最大となる分離超平面を求めるアルゴリズムである。図 1 の(a)、(b)はどちらも正例（白丸）と負例（黒丸）を分離できているが、(b)の方が大きなマージンを持っており、テスト事例を精度よく分離できる。

SVM を用いた地域情報分類の手順は次のようになる。まず学習フェーズとして、各カテゴリごとに正例・負例を SVM に学習させて分離超平面を形成する。次にテストフェーズとして、分類対象データを SVM に与えて各カテゴリに該当するか否かを出力させる。分類に用いる素性は、対象とする Web ページから HTML タグを除いて形態素分割し、品

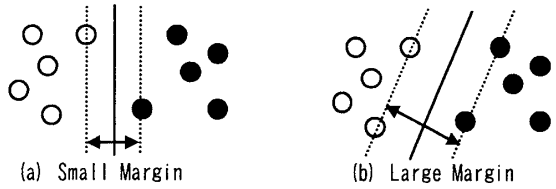

図 1 Support Vector Machine の概念図

詞が名詞、動詞、形容詞、副詞の形態素を使う。本研究では、形態素解析システムとして「茶筌」[6]を、SVM 学習ツールとして TinySVM[7]を用いた。

5. 実験結果および考察

(1) 住所抽出実験

住所抽出実験には、住所を含む Web ページ 489 件を人手で収集して用いた。自動抽出した住所の正解判定は、抽出された住所を人手でチェックし、市区町村名まで正確に取れているものを正解とした。その結果、自動抽出された住所数 1154 件に対して正しく自動抽出できた住所数は 1126 件で、適合率は 98% (1126/1154)であった。また、人手で抽出した住所数 1170 件に対して正しく自動抽出できた住所数は 1126 件で、再現率は 96% (1126/1170)であった。抽出に失敗した例は、主に下記の通りであった。

・地名を含む固有名詞や支店名が抽出された。例：「深川富岡八幡さま水掛祭」「横浜天王町店」など 28 件
・住所の途中に区切り文字が入っていたために抽出できなかった。例：「兵庫県 芦屋市 川西町」など 17 件
・茶筌の辞書に未登録の単語が住所に含まれていたために抽出できなかった。例：「さいたま市」など 9 件
・住所のすぐ後に道順などが記述されていたために抽出できなかった。例：「郵便局東へ 2 筋入る」など 6 件

(2) 地域情報分類実験

地域情報分類実験では、各カテゴリに属するページを人手で集め、合計 2763 件の Web ページを正解セットとした。「和食」でかつ、「飲み屋」など、複数のカテゴリに含まれるページには複数のカテゴリを付与した。

比較データとして、各カテゴリを代表する語を OR で結んでキーワード検索する場合の精度も求めた。OR 検索に使う単語はカテゴリに属するページに含まれる単語のうち、IDF 値が高い上位 5 語を選んだ。単語 K の IDF 値は、あるカテゴリ C 内で K が出現する文書の数を N(C,K)、全文書集合 D で K が出現する文書の数を N(D,K)として、IDF = N(C,K) / N(D,K)として求めた。IDF 値は、あるカテゴリに偏って出現する単語ほど、そのカテゴリでの値が高くなる傾向を持つ。

SVM の精度評価では、正解セットを 5 等分し、学習 4、テスト 1 の比率で交差検定を行った。表 2 に実験結果を示す。表 2 で適合率は、カテゴリ C と判定された文書数を Nr、カテゴリ C と判定され正解だった数を Np として Np / Nr で求めた。また、再現率は、正解セットのカテゴリ C に含まれる文書数を Nc として、Np / Nc で求めた。また、参考値として大分類と中分類の精度のマクロな平均値も求めた。

表 2 をみると、キーワード検索（表 2 の「キーワード」の欄）では、適合率か再現率の一方が著しく低いなど、精度のマクロ平均値は適合率、再現率ともに 50%台にとどまった。例えば、「文化施設」カテゴリでは、「開館」「閉館」「蔵書」などがカテゴリに特徴的な単語であるが、これだけでは検索漏れが多い。また、「喫茶店」カテゴリに特徴的な単語「コーヒー」「ケーキ」は他のカテゴリにも多く出現するので、適合率が低下してしまう。

表 2 地域情報分類結果（表中の各値は適合率／再現率）

	カテゴリ[事例数]	キーワード	SVM	住所+SVM
大分類	レジャー [1254]	69% / 53%	90% / 87%	98% / 93%
	グルメ [1195]	38% / 52%	89% / 90%	94% / 98%
	非地域情報 [441]	92% / 35%	87% / 81%	—
中分類	観光案内 [107]	32% / 85%	83% / 71%	93% / 72%
	宿泊施設 [246]	74% / 52%	84% / 62%	88% / 80%
	娯楽施設 [400]	85% / 30%	73% / 71%	77% / 77%
	文化施設 [501]	90% / 65%	88% / 88%	93% / 87%
	和食 [367]	64% / 49%	75% / 67%	90% / 77%
	洋食 [301]	57% / 49%	60% / 70%	76% / 54%
	中華 [198]	41% / 46%	76% / 46%	65% / 57%
	飲み屋 [208]	28% / 60%	63% / 50%	69% / 43%
	喫茶店 [102]	17% / 70%	67% / 52%	75% / 38%
	マクロ平均値	57% / 54%	78% / 70%	83% / 71%

一方、SVM を用いた分類（表 2 の「SVM」の欄）では、適合率、再現率がマクロ平均値でそれぞれ 21 ポイント、16 ポイント上昇し、大分類では 90%近い精度で分類できた。これは、SVM が決定した高次元素性空間上の分離超平面が、テスト事例をうまく分離できているからである。中分類では、カテゴリ当たりの訓練事例数が減るために精度が低下しているが、訓練事例の追加で、ある程度精度向上が期待できる。

さらに、住所抽出が可能なページのみに絞った分類（表 2 の「住所+SVM」の欄）では、全文書を対象に分類した場合よりも適合率のマクロ平均値が 5 ポイント上昇した。したがって、住所情報を利用することによって、大量のページを効率的に分類することができるといえる。

6. おわりに

本研究では、地域情報検索サービス向けデータの自動作成方法について、形態素の品詞細分類情報のパターンマッチによる住所抽出と、SVM による地域情報の自動分類を提案し、評価実験を行った。その結果、比較的単純なルールでも 98%の適合率で市区町村レベルまでの住所を抽出することができた。また、キーワード検索では検索漏れや検索間違いを排除しにくいのに対して、SVM によるカテゴリ分類では「レジャー」「グルメ」などの大分類で 90%近くの精度で分類することができた。さらに、住所情報を利用して大量のページを効率的に分類できることを示した。今後は、カテゴリを増やすとともに、営業時間や定休日などの属性情報の抽出などに取り組んでいきたい。

参考文献

[1] 横路誠司, 高橋克巳, 三浦信幸, 島健一, 位置指向の情報の収集、構造化および検索手法, 情報処理学会論文 Vol. 41, No. 7, 1987(2000).
[2] 相良毅, 有川正俊, 坂内正夫, ジオリファレンス情報を用いた空間情報抽出システム, 情処論データベース 41/SIG6(TOD7), 69(2000).
[3] 大槻洋輔, 佐藤理史, 地域情報ウェブディレクトリの自動編集, 情報処理学会論文誌, Vol. 42, No. 9, 2310(2001).
[4] Vladimir N. Vapnik, Statistical Learning Theory, A Wiley-Interscience Publication, (1998).
[5] BIGLOBE サーチ, http://attayo.jp/
[6] 日本語形態素解析システム ChaSen 「茶筌」, http://chasen.aist-nara.ac.jp/
[7] SVM 学習ツール TinySVM, http://cl.aist-nara.ac.jp/~taku-ku/software/TinySVM/