論　文

# 日本語ウェブページを主観的か非主観的かに分類する分類器のジャンル領域拡大化能力の改善：実用的な分類器へ向けて

大森　晃†

Improving Genre Expansion of a Classifier that Classifies Japanese Web Pages as Subjective or Non-subjective: Toward a Practical Classifier

Akira OHMORI†

**あらまし**　本論文では，機械学習法を利用して日本語ウェブページのセンチメント分類に取り組む．分類のためのカテゴリーは「主観的」と「非主観的」である．交差検定用データセットは，限られたジャンル群に分布する日本語ウェブページからなる．まず，交差検定用データセットをほぼ確実に分類する分類器を生成できることを示す．その分類器の実用性を評価するために，本論文ではジャンル領域拡大化能力という概念とジャンル領域拡大データセットを導入する．ジャンル領域拡大データセットは，交差検定用データセットを構成するジャンル群を含む，より多様なジャンル群に分布する日本語ウェブページからなるデータセットである．ジャンル領域拡大化能力は，交差検定用データセット上で訓練・生成された分類器がジャンル領域拡大データセットを分類する能力である．本論文では，交差検定用データセット上で訓練・生成された分類器のジャンル領域拡大化能力が，低いことを示す．一方で，分類器のジャンル領域拡大化能力を改善するための方法として，遺伝的アルゴリズムを利用した POS フィルタリングに基づく素性選択法を提案し，その方法によって分類器のジャンル領域拡大化能力を改善でき，ある程度実用的とみなせる分類器を生成できることを示す．

**キーワード**　センチメント分類，libSVM，POS フィルタリング，素性選択，遺伝的アルゴリズム

## 1. まえがき

インターネットの普及に伴い，インターネット上のコンテンツであるウェブページは非常に多数かつ多様になってきている．こうした情報源の拡大は，ユーザにとっては良い傾向といえよう．その反面，現在の検索エンジンがもつ機能的限界もあって，ユーザが自分の検索要求に合致する有用なウェブページを効率良く入手することが困難になってきている．

例えば「Linux」という OS 商品について，広告・宣伝を記載するウェブページを入手したいにもかかわらず，用語説明，報道記事，ウェブ掲示板，マニュアル，使用経験談，質問・回答，などを記載するウェブページが検索結果に混在して提示されることは，よくあることである．また，例えば「東京」について，生活環境や自然環境に関する個人的な意見や印象を主として記載するような主観的なウェブページを入手したいにもかかわらず，東京に存在する施設や企業，あるいはそれらが提供するサービスに関する客観的情報を記載するようなウェブページが検索結果に混在することも，よくあることである．いずれにせよ現在の検索エンジンには，提示されたウェブページの中から上記のような検索要求に合致するウェブページを分離する機能はなく，分離作業には依然として手間がかかるのが現状である．

近年，ウェブページ検索におけるこのような状況を改善してユーザの検索要求により合致する検索結果を生み出すために，現状の検索エンジンを補完する技術として，主に機械学習法を用いたウェブページの自動的分類技術の研究が質的に異なる二つの研究分野で行われている．一つは，ジャンル分類技術の研究分野である．この研究分野では，従来から行われてきたウェブページの主題（topic あるいは subject）に従う分類ではなく，ジャンルに従う自動的分類技術が研究され

† 東京理科大学工学部第二部経営工学科，東京都
Dept. of Management Science, Faculty of Engineering, Tokyo University of Science, 1–3 Kagurazaka, Shinjuku-ku, Tokyo, 162–8601 Japan

 

ている [1]〜[5]．もう一つは，センチメント分類技術の研究分野である．この研究分野では，必ずしも機械学習法を利用するわけではないが，同じく従来から行われてきたウェブページの主題に従う分類ではなく，センチメントカテゴリーに従う自動的分類技術が研究されている [6]〜[11]．

センチメント（sentiment）という用語の学術的定義は明確ではない．定義を与えることは困難であるが，テキスト分類という枠組み内で，本論文ではセンチメントという用語の意味を次のように解釈する．つまり，センチメント（sentiment）という用語は，(a) テキスト記述者がある事態（例えば，憲法改正），ある対象（例えば，ある特定の商品やサービス）について記述する際の心的態度（例えば，主観的/客観的，楽観的/悲観的），あるいは，(b) ある事態，ある対象に関するテキスト記述者の評価（例えば，肯定的/否定的，安全/危険）を意味するものとする．そして，主観的/客観的，肯定的/否定的という心的態度あるいは評価のカテゴリーを，センチメントカテゴリーと呼ぶことにする．

現状では，センチメント分類技術の研究は，用いられるセンチメントカテゴリーが「肯定的/否定的」である場合 [6]〜[8] と，「主観的/客観的」である場合 [8]〜[11] に分けることができる．以下では，本論文と関連の深い「主観的/客観的」をセンチメントカテゴリーとして扱う先行研究について述べる．

Finn ら [8] は，フットボール，政治，金融という主題ごとに，英語で書かれたニュース記事から構成されたデータセットを，「主観的」と「客観的」という二つのセンチメントカテゴリーに分類している．BOW（Bag of Words：単語）を素性（feature）(注1) として，各データセットに対して 10-fold Cross Validation (注2) を実行し，分類器の訓練とテストを行っている．3 種類の主題について，分類性能は平均で約 87%の accuracy (注3) を達成しており，主題に依存しつつも，主観的か客観的かへの分類は可能であるとしている．

Yu ら [9] は，英語で書かれたニュース記事を「事実（fact）」と「意見（opinion）」という二つのカテゴリー (注4) に分類している．訓練・テスト用のデータセットを Wall Street Journal (WSJ) の Editorial，Letter to editor，Business，News という 4 種類の記事カテゴリーから収集したニュース記事から構成し，Editorial と Letter to editor を意見，Business と News を事実としている．BOW を素性として，分類器の訓練にデータセットの半分を，分類器のテストに残りの半分を使っている．彼らのデータセットに限れば分類器の分類性能は極めて高く，97%の F-measure (注5) を達成している．ただし，彼らも認識していることであるが，WSJ 以外から収集されたデータセットに対しては分類器の分類性能が低下する可能性がある．また，その他のカテゴリーに属する記事（例えば，論評，スポーツ，政治）が混在するようなデータセットについては，分類性能は更に低下する可能性がある．

Wiebe ら [10] は，英語で書かれたニュース記事を「主観的」と「客観的」という二つのセンチメントカテゴリーに分類している．訓練・テスト用のデータセットは，Yu ら [9] のものとは若干異なり，Wall Street Journal (WSJ) における記事カテゴリーである Editorials，Letters to the editor，Arts & Leisure review，Viewpoints から収集した記事を主観的とし，その他の記事カテゴリーから収集したものを客観的としている．素性としては単語 $n$-gram (注6) ($n = 1$〜4) などを利用して，leave-one-out CV によって分類器の訓練・テストを行っている．彼らのデータセットに限れば分類器の分類性能は極めて高く，約 94%の accuracy を達成している．Yu らに比べて主観的な記事のカテゴリーに Arts & Leisure review，Viewpoints が追加されているとはいえ，これもまた Yu らの研究と同様に，WSJ 以外から収集されたデータセットに対しては分類性能が低下するかもしれないという問題を抱える．また，Wiebe ら [11] は英語で書かれた文を，文レベルで「主観的」と「客観的」という二つのセンチメントカテゴリーへ分類することも試みている．

我々の最終目標は，機械学習法を利用して日本語ウェブページを主観的か非主観的かへ分類するための実用的な分類器を簡潔な手続きで生成することである．ここで，実用的な分類器とは，その初期状態において，ウェブページの主題やジャンルとは無関係に，諸種の主題と諸種のジャンルに分布する未知のウェブページ群を，ある程度高い分類性能で主観的か非主観的かへ分類できる分類器を指す．また手続きの簡潔性は，分

(注1)：素性値は，素性の有無（あれば 1，なければ 0）を採用している．
(注2)：Cross Validation は，以下 CV と略記する．
(注3)：accuracy については付録 **1.** を参照されたい．
(注4)：「事実」は「客観的な記事」，「意見」は「主観的な記事」に対応しており，実質的には「主観的」と「客観的」という 2 種類のセンチメントカテゴリーへの分類である．
(注5)：F-measure については付録 **1.** を参照されたい．
(注6)：文中における連続する $n$ 個の単語を意味する．

類器の訓練・テスト用にウェブページ群を収集する手続き，素性を抽出する手続き，素性ベクトルを作成する手続きなど，分類器生成までの手続きが人手と時間の面で実際的であり，簡単な手法によって構成されていることを意味する．

本論文では，こうした最終目標に向けて，第 1 段階として，手続きの簡潔性の点から限られたジャンル群に分布するウェブページからなる交差検定用データセットを用意し，それに対して分類実験（10-fold CV）を実行し，分類可能性をチェックする．結果として，簡単な素性と素性値を利用するだけで，用意した交差検定用データセットを，非常に高い性能で分類できることを示す．

機械学習法を利用したテキスト分類技術の研究は，この段階でいったんは終結させることがほとんどである．一方 Finn ら [8] のように，生成した分類器の実用性という観点から，主題領域転化能力（domain transfer）の評価を取り上げる研究者たちもいる．彼らは，センチメント分類の枠内で，ウェブページの主題に依存しない（主題に影響を受けない）分類器を生成することが重要であるとし，分類器の主題領域転化能力（domain transfer）を評価することの必要性を強調している．主題領域転化能力（domain transfer）は，ある特定の主題について書かれたテキスト群を利用して訓練された分類器が，それ以外の主題について書かれたテキスト群を分類する能力である．彼らは 3 種類の主題（フットボール，政治，金融）のうち，一つの主題に関するデータセットを訓練用とし，それとは異なる別の主題に関するデータセットをテスト用として，全部で 6 通りの場合について主題領域転化能力を検討している．POS（Part of Speech：品詞）を素性とした場合に平均的には 78.5%の accuracy を達成しているが，高い主題領域転化能力を有する分類器を生成することは比較的困難であると彼らは結論づけている．

主題領域転化能力は，我々が最終目標とする実用的な分類器を探求していく上で非常に参考になる概念である．本論文では，主題領域転化能力に類似する分類器の能力として，ジャンル領域拡大化能力（genre expansion）を導入する．これは，ある限られたジャンルに分布するウェブページ群からなるデータセットを利用して訓練・生成された分類器が，それらのジャンルを含む，より多様なジャンルに分布する未知のウェブページ群からなるデータセット（以下，ジャンル領域拡大データセットと呼ぶ）を分類する能力である．能力の測定は，分類器の分類性能を示す指標である accuracy または F-measure による(注7)．ジャンル領域拡大化能力は，ある限られたジャンルに分布するウェブページ群からなるデータセット上で訓練された分類器の実用性を推し量る一つの指標になる．

そこで，第 2 段階として分類器のジャンル領域拡大化能力を検討する．そのために，第 1 段階で用意した交差検定用データセットとは質的に異なるジャンル領域拡大データセットを用意し，第 1 段階で生成した分類器をジャンル領域拡大データセットに適用する．結果として，分類性能がかなり低下することを示す．つまり，第 1 段階で生成される分類器は，そのままではジャンル領域拡大化能力という点で実用的とはいえないことを示す．

実用的な分類器を探求するために，第 3 段階として素性選択によるジャンル領域拡大化能力の改善を試みる．素性選択には POS レベルで素性を取捨する POS フィルタリング [12] という簡単な手法を用いる．こうした POS フィルタリングに基づく素性選択を，形式的に POS 組合せ最適化問題として扱い，その解を求めるために遺伝的アルゴリズム（Genetic Algorithm：以下 GA と記すことがある）を適用する．結果として，POS フィルタリングに基づく素性選択に GA を適用することによって，ジャンル領域拡大化能力という点で，ある程度実用的な分類器が生成できることを明らかにする．

ジャンル分類技術の研究に関連して，英語ウェブページ，日本語ウェブページに関するジャンル体系がいくつか設定されている [1]〜[5]．しかしながら，英語ウェブページのジャンル体系 [1], [2], [4], [5] については，各ジャンルについて十分な説明が与えられていない．日本語ウェブページのジャンル体系 [3] については，各ジャンルの説明は与えられてはいるものの，ジャンル名及び各ジャンルの説明が適度に洗練されておらず，ジャンル間の境界線があいまいである．現状としては，日本語ウェブページについて再利用できる適当なジャンル体系はない．本論文では，付録 **2.** に示すような独自に設定したジャンル体系を利用することとする．また，本ジャンル体系の設定法等については付録 **3.** に示す．

以下，**2.** では分類器の交差検定用データセットの構

(注7)：accuracy と F-measure については付録 **1.** を参照されたい．

成について述べる．**3.** では，機械学習で利用する素性と素性値について述べる．**4.** では交差検定用データセットについて分類実験を行い，非常に高い性能で分類できることを示す．**5.** ではジャンル領域拡大データセットの構成について述べ，**6.** では分類器のジャンル領域拡大化能力が低いことを明らかにする．**7.** では，POS フィルタリングに基づく素性選択に GA を適用することによって，ジャンル領域拡大化能力を改善でき，ある程度実用的とみなせる分類器を生成できることを示す．**8.** では，本論文のまとめと今後の課題を述べる．なお，本論文では機械学習法として，汎化能力が高いといわれている SVM [13] を用いる．また，SVM の実装は libSVM2.82 (注8) を用いる．

## 2. 交差検定用データセット

### 2.1 主観性と非主観性

ウェブページの主観性/非主観性を厳密に定義することは困難である．本論文では，その記載内容が，現実世界で客観的に認識されている事実に関するよりもむしろ，個々人の心の中に存在する考え・意見，感じ方に関する記述に偏るウェブページを，主観的ウェブページと解釈する．また，主観的でないものを非主観的ウェブページと解釈する．

これによって，主観的の解釈（ある意味で一つの分類基準）だけによって二つのカテゴリーである主観的/非主観的を設定できる．一方，先行研究に倣って主観的/客観的というカテゴリーを設定するならば，「客観的」の解釈を与えなければならない．それを与えることは可能ではあるが，分類基準が二つになる．二つの分類基準でウェブページを分類するよりは，一つの分類基準でウェブページを分類する方が人間にとっては分類しやすい．このことから，本論文では先行研究とは異なり，主観的/非主観的というカテゴリーを用いる．なお，非主観的ウェブページの中には「客観的」と明らかに判定できるものも含まれるが，「客観的」とは判定しにくいようなものも含まれる．

### 2.2 ウェブページの収集方針

ウェブページの主題やジャンルとは無関係に，諸種の主題と諸種のジャンルに分布するウェブページ群を主観的か非主観的かへ分類する分類器を訓練・生成するためには，諸種の主題と諸種のジャンルに分布する主観的あるいは非主観的なウェブページを適当な数ほど収集し，それらによって交差検定用データセットを構成することが理想的である．しかしながら，そうした理想的なデータセットを構成することは，人手と時間がかかりすぎて，実際的な方法とはいえない．本論文では次善の策として，主観的ウェブページが属する傾向の強いジャンル，非主観的ウェブページが属する傾向の強いジャンルに絞ってウェブページを収集することとした．

本論文で対象にするジャンルは，予想，標語，批評，解説，報道，独白・感想，情宣（情報宣伝），質問・相談・依頼（回答付き限定），記録（現代文限定），商品広告・宣伝，マニュアル，用語説明，案内・紹介，その他の 14 ジャンル中，その他を除く 13 ジャンルである（付録 **2.** 参照）．このうち独白・感想は主観的ウェブページが属する傾向の強いジャンルである．一方，報道，記録，商品広告・宣伝，用語説明は，非主観的ウェブページが属する傾向の強いジャンルである．このことから，主観的ウェブページとして，独白・感想のジャンルに属する日記や独り言を記載したウェブページを収集することとした．また，非主観的ウェブページとして，ニュース記事本文（報道）(注9)，政府機関の広報（記録），各種商品の広告（商品広告・宣伝），各種用語の説明（用語説明）を記載したウェブページを収集することとした．なお，ウェブページを収集する際，画像は除外した．

### 2.3 ウェブページの収集

作業者 P が WWW から主観的ウェブページの候補を 1,100 件収集した．続いて作業者 P と作業者 R が個別に 1,100 件のウェブページを読み，実際に主観的であるかどうかを判定した．その結果，両作業者の判定が一致したものは 1,055 件であった．一方で，作業者 Q が WWW から非主観的ウェブページの候補を 1,100 件収集した．続いて作業者 Q と作業者 R が個別に 1,100 件のウェブページを読み，実際に非主観的であるかどうかを判定した．その結果，両作業者の判定が一致したものは 1,058 件であった．

判定が一致した主観的ウェブページ 1,055 件，非主観的ウェブページ 1,058 件から，それぞれランダムに 1,000 件のウェブページを選択し，分類器の交差検定用データセットを構成した．主観的ウェブページはすべて独白・感想のジャンルに属する．一方，非主観的ウェブページのうち 771 件は報道のジャンル，140 件は記録のジャンル，78 件は用語説明のジャンル，11 件

(注8)：libSVM については以下の URL を参照.
http://www.csie.ntu.edu.tw/∼cjlin/libsvm/
(注9)：括弧内はジャンルを示している.

は商品広告・宣伝のジャンルに属する．ジャンルの多様性は確保できていないが，ウェブページの総件数が2,000 件であるので主題の多様性は確保されていると考えてよかろう．

## 3. 素性と素性値

日本語ウェブページの素性として，ウェブページに出現する単語（つまり BOW）と，その POS（Part of Speech：品詞）との組合せを利用した．素性を求めるために，交差検定用データセット内のウェブページに日本語形態素解析システム Mecab0.93 (注10) を適用した．Mecab は自然言語で書かれた文を分析して単語分割し，各単語に対応して 1 行に，表層形（出現する単語そのもの），品詞，品詞細分類 1，品詞細分類 2，品詞細分類 3，活用形，活用型，原形，読み，発音を出力する．このような出力のうち，単語としては表層形を，POS としては品詞，品詞細分類 1，品詞細分類 2，品詞細分類 3 からなる 4 項組みを，素性として採用した．「未知語」も POS の一つとしてとらえ，未知語と判断された表層形も素性として採用した．なお，Mecab で用いられる POS は IPA 品詞体系 (注11) に基づいており，未知語を含めて 70 種類である．以下，単語素性を word，POS 素性を pos と表記し，素性を (word, pos) と表記することがある．

従来から用いられている簡単な素性値としては，主に以下の 2 種類がある．

(a) 素性の出現頻度：素性が 1 ウェブページ内で出現する頻度．

(b) 素性の有無：素性が 1 ウェブページ内にあれば 1，なければ 0．

センチメント分類に関連する Pang らの研究 [7] によれば，BOW を素性として SVM を利用する場合，素性の出現頻度を素性値とするよりも，素性の有無を素性値とした方が，より高い accuracy（80%以上）を達成するという知見が得られている．このことは，センチメントカテゴリーに従うテキスト分類においては，素性の有無を素性値として用いる方が有効である可能性が高いことを示唆している．彼らが設定したセンチメントカテゴリーは肯定的/否定的であり，素性は BOW のみであるため本論文とは異なるが，本論文で扱う分類課題についても，その知見が当てはまる可能性は高いと予想できる．そのため本論文では，素性値としては素性の有無を用いることとした．

## 4. 分類可能性

分類器を訓練・テストして分類性能を求めるために，交差検定用データセットに対して 10-fold CV を実行した．その際，訓練用データを利用して分類器を訓練する方法としては，以下の二つを実行した．

（1） libSVM-Linear：線形カーネルを利用した libSVM2.8.2 による訓練．

（2） libSVM-RBF：RBF カーネルを利用した libSVM2.8.2 による訓練．

また，テスト用データを利用して，訓練された分類器の分類性能として accuracy を求めた．結果を表 1 に示す．分類性能である accuracy の単位は%であり (注12)，それらの値は 10-fold CV によって生成された 10 個の分類器の accuracy を平均したものである．

**1.** で言及したように，英語ウェブページを主観的か客観的かに分類する分類課題に対して，Finn ら [8] は主題に依存しつつも平均で約 87%の accuracy を，Wiebe ら [10] は約 94%の accuracy を達成している．表 1 は，これらに比べて高い数値を示している．Yu ら [9] は 97%の F-measure を達成している．本来は直接的な比較はできないが，この数値に対して Wiebe ら [10] は，自分たちが達成した accuracy とほぼ同程度の分類性能を達成していると主張している．そうであるとすれば，表 1 は Yu ら [9] の分類性能よりは良い数値を示しているといえる．

利用した素性は，BOW と POS という非常に簡単なものである．また素性値も素性の有無によって決めるという非常に簡単なものである．こうした簡単な素性と素性値を利用して，極めて高い accuracy を達成できることが明らかになった．結果として，ジャンルに制限はあるものの，日記や独り言（独白・感想）(注13) か

表 1 交差検定用データセットについての分類性能
Table 1 Performance of classifiers on the cross-validation dataset.

| 訓練法 | accuracy |
|---|---|
| libSVM-Linear | 100.0 |
| libSVM-RBF | 97.9 |

(注10)：Mecab0.93 については以下の URL を参照.
http://mecab.sourceforge.net/
(注11)：IPA 品詞体系については以下の URL を参照.
http://hal.yh.land.to/manual/ipadic/ipadic-ja.html#SECTop/
(注12)：以下，分類器の分類性能は%を単位とする．
(注13)：括弧内はジャンルを示している．

表 2　ジャンル領域拡大データセットの内訳
Table 2　Breakdown of a genre-expanded dataset.

| | NTCIR-3 WEBからのランダムサンプル | | | ジャンル領域拡大データセット | | | |
|---|---|---|---|---|---|---|---|
| ジャンル | 件数 | 比率(%) | 99%信頼区間 | 件数 | 比率(%) | 主観的 | 非主観的 |
| **予想** | 1 | 0.30 | 0.00 – 1.06 | 1 | 0.50 | 0 | 1 |
| **標語** | 1 | 0.30 | 0.00 – 1.06 | 1 | 0.50 | 1 | 0 |
| **批評** | 1 | 0.30 | 0.00 – 1.06 | 1 | 0.50 | 1 | 0 |
| **解説** | 2 | 0.59 | 0.00 – 1.67 | 2 | 1.00 | 0 | 2 |
| 報道 | 5 | 1.48 | 0.00 – 3.17 | 4 | 2.00 | 0 | 4 |
| 独白・感想 | 25 | 7.40 | 3.73 – 11.06 | 20 | 10.00 | 18 | 2 |
| **情宣** | 4 | 1.18 | 0.00 – 2.70 | 4 | 2.00 | 1 | 3 |
| **質問・相談・依頼** | 15 | 4.44 | 1.55 – 7.32 | 12 | 6.00 | 2 | 10 |
| 記録 | 37 | 10.95 | 6.57 – 15.32 | 29 | 14.50 | 3 | 26 |
| 商品広告・宣伝 | 31 | 9.17 | 5.13 – 13.22 | 25 | 12.50 | 3 | 22 |
| **マニュアル** | 18 | 5.33 | 2.18 – 8.47 | 14 | 7.00 | 0 | 14 |
| 用語説明 | 15 | 4.44 | 1.55 – 7.32 | 12 | 6.00 | 2 | 10 |
| **案内・紹介** | 96 | 28.40 | 22.08 – 34.72 | 75 | 37.50 | 13 | 62 |
| その他 | 87 | 25.74 | 19.61 – 31.87 | 0 | 0.00 | 0 | 0 |
| 総計 | 338 | – | – | 200 | – | 44 | 156 |

らなる主観的日本語ウェブページと，ニュース記事本文（報道），政府機関の広報（記録），各種商品の広告（商品広告・宣伝），各種用語の説明（用語説明）からなる非主観的日本語ウェブページを，簡単な素性及び素性値を利用して，ほぼ確実に分類することが可能である．

## 5. ジャンル領域拡大データセット

交差検定用データセットは，独白・感想，報道，記録，用語説明，商品広告・宣伝という 5 種類のジャンルに分布する．分類器のジャンル領域拡大化能力を検討するためには，原則として，これらのジャンルを含む 6 種類以上のジャンルに分布するジャンル領域拡大データセットが必要である．ジャンル領域拡大化能力は，分類器の実用性を推し量る一つの指標であり，ジャンル領域拡大データセットに含まれるジャンルの種類数が多ければ多いほど，そうした指標としての適切さを増すと考えられる．そこで，付録 **2.** に示すジャンル体系のうち，ジャンル「その他」を除く 13 種類のジャンルに分布するジャンル領域拡大データセットを設定することとした．

そのようなジャンル領域拡大データセットの設定にあたって，NTCIR プロジェクトによって収集されたテストコレクションである NTCIR-3 WEB (注14) を利用した．具体的には，付録 **3.** で言及した，NTCIR-3 WEB からランダムに選択した 338 件のウェブページを利用した．これらのウェブページは，ジャンル体系の設定に伴って既にジャンルに分類されている．表 2 の左側にその分類結果を示すとともに，各ジャンルに属するウェブページの件数が 338 件中に占める比率 (%) (注15)，及び 99%信頼区間(注16) を示す．表中，ページ件数の少ないジャンルがあるが，これはランダムに選択した 338 件のウェブページの中に，当該ジャンルに属するものが少なかったことによる．

次に，ジャンル「その他」を除く 13 種類のジャンルすべてに分布するように 200 件のウェブページを選択し，各ウェブページを作業者 X，Y，Z が個別に主観的か非主観的かに分類した．各ウェブページが主観的か非主観的かの最終的な判定は，作業者 3 人の判定をもとに多数決で行った．このようにして構成したジャ

(注14)：主として .jp ドメインから収集された 11,034,409 件のウェブページを含む．ただし，画像は除外されている．NTCIR-3 WEB については以下の URL を参照．http://research.nii.ac.jp/ntcir/permission/perm-ja.html#ntcir-3-web
(注15)：概数であるので総計は 100%にはならない．
(注16)：NTCIR-3 WEB を母集団とした場合の母比率の 99%信頼区間である．

ンル領域拡大データセットの内訳を表 2 の右側に示す．各ジャンルに属するウェブページの件数がジャンル領域拡大データセット中に占める比率 (%) は，案内・紹介を除いて 99%信頼区間に入っている．案内・紹介についても，99%信頼区間の上限を大きく超えてはいない．このことから，ジャンル領域拡大データセットにおける各ジャンルに属するウェブページの件数は妥当と考えられる．なお，太字で示したジャンルは，交差検定用データセットには含まれていないジャンルである．

## 6. ジャンル領域拡大化能力の評価

**4.** で述べたように，訓練法 libSVM-Linear, libSVM-RBF を用いて，交差検定用データセットに対して 10-fold CV を実行し，訓練法ごとに 10 個の分類器を得た．これらの分類器がジャンル領域拡大データセットに対してどの程度のジャンル領域拡大化能力（accuracy）をもつかを分類実験によって調べた．その結果を表 3 の accuracy(10FCV) における上段の数値により示す．accuracy(10FCV) における上段の数値は，当該訓練法のもとで訓練・生成された 10 個の分類器によってジャンル領域拡大データセットを分類し，その結果得られた accuracy を平均したものである．比較のために，下段括弧内に表 1 の数値を示してある．一方，交差検定用データセットの 2,000 件すべてを訓練用データとし，訓練法 libSVM-Linear, libSVM-RBF を用いて分類器を生成した．そして，それらの分類器をジャンル領域拡大データセットに適用してジャンル領域拡大化能力を求めた．結果を表 3 の accuracy(all) に示す．

表 3 の accuracy(10FCV) における上段数値と下段括弧内数値との比較から分かるように，ジャンル領域拡大データセットについての accuracy は，かなり低下しており最高でも 75%台である．accuracy(all) については最高でも 77%台であり，訓練用データの件数を増やしてもジャンル領域拡大化能力は低いことが分かる．こうした結果から，交差検定用データセット上で訓練・生成された分類器は，ジャンル領域拡大化能力が十分ではなく，実用的であるとは判断できない．

表 3　ジャンル領域拡大化能力
Table 3　Genre-expansion.

| 訓練法 | accuracy(10FCV) | accuracy(all) |
|---|---|---|
| libSVM-Linear | 75.7<br>(100.0) | 77.5 |
| libSVM-RBF | 65.5<br>( 97.9) | 65.5 |

交差検定用データセットについては非常に高い分類性能を発揮する分類器でも，未知のデータセットについては分類性能がかなり低下するという現象はしばしば起こる．表 3 もそうした現象の現れを示すものである．ただし，ここで注意すべき点は，単に未知であるデータセットについてではなく，確実にジャンル領域を拡大した未知のデータセットについて分類性能が低下していることを示している点である．

## 7. ジャンル領域拡大化能力の改善

### 7.1 POS フィルタリングに基づく素性選択法

交差検定用データセット上で訓練・生成された分類器のジャンル領域拡大化能力の低さは，分類器の過学習に起因すると考えられる．そうであれば，素性選択を行って素性の数を減らすことにより過学習は抑えられる可能性があり [14]，結果として分類器のジャンル領域拡大化能力が改善される可能性がある．そこで，分類器のジャンル領域拡大化能力を改善するための素性選択法について検討する．

素性 (word, pos) すべてからなる集合を Feature-Set と表記し，その空でない任意の真部分集合を Feature-Subset と表記すると，素性選択は，ある Feature-Subset を選択することを意味する．全素性の数を $N$ とした場合，選択対象として意味のある Feature-Subset は，すべての素性を使う場合と，どの素性も使わない場合とを除いて，$2^N-2$ 通りある．分類器のジャンル領域拡大化能力を改善するような Feature-Subset は自明ではない．そのような Feature-Subset を探索するにあたり，交差検定用データセットから収集した全素性の数 $N$ は 77,301 であることから，我々は極めて大きな探索空間に直面することになる．

探索空間を縮小できるような素性選択の方法が必要である．本論文では，POS レベルで素性を取捨する POS フィルタリング [12] を利用する．平ら [12] は，SVM を用いたテキスト分類における素性選択法として，相互情報量フィルタリングと POS フィルタリングを比較している．彼らは，相互情報量フィルタリングよりも POS フィルタリングの方が効果的であることを示している．また，SVM には分類に無関係な品詞を有する単語を除外する能力に限界があることを明らかにし，この限界を補う意味で，単純で実際的な POS フィルタリングによる素性選択が有効であると

している．更に，我々の見解では，POS フィルタリングによる素性選択は，探索空間の縮小に役立つ．

70 種類の全 POS からなる集合を POS-Set，その空でない任意の真部分集合を POS-Subset と表記すると，POS フィルタリングに基づく素性選択の手順は以下のようになる．

（1） POS フィルタリング：POS-Set から分類器の訓練・テストに使用する POS を選択し，それらを要素とする POS-Subset を構成する．

（2） 素性選択：素性の組合せとして，

$$\text{Feature-Subset} = \{(\text{word}, \text{pos}) \mid (\text{word}, \text{pos}) \in \text{Feature-Set} \ \&\ \text{pos} \in \text{POS-Subset}\}$$

を選択する．

POS フィルタリングに基づく素性選択を行う場合，70 種類の POS のあらゆる組合せのうち素性選択の対象として意味のある組合せは，すべての POS を使う場合と，どの POS も使わない場合とを除いて，$2^{70} - 2$ 通りである．集合 Feature-Subset を素性選択の直接的な対象にすることに比べれば，探索空間の縮小にはつながる．しかしながら，分類器のジャンル領域拡大化能力を改善するような POS-Subset を探索するにあたり，探索法に何ら工夫をしないとすれば，各 POS-Subset に対応して，訓練用データによって分類器を訓練・生成し，そのジャンル領域拡大化能力を調べる必要がある．これは計算量の面で非現実的であり，探索法に何らかの工夫が必要となる．

### 7.2 POS 組合せ最適化問題

POS フィルタリングに基づく素性選択によって分類器のジャンル領域拡大化能力を改善する問題は，以下のように POS 組合せ最適化問題として形式的に表現できる．ただし，POS 組合せ最適化問題の解を得ることが目的ではない．それは，あくまでも手段であり，ジャンル領域拡大化能力が高い分類器を求めることが目的である．

［POS 組合せ最適化問題］

$$\text{Max}_{\text{POS-Subset}}\ \text{accuracy}(\text{POS-Subset})$$
$$\text{subject to} \quad \text{POS-Subset} \subset \text{POS-Set}.$$

ただし，POS-Subset は空でない真部分集合である．また，accuracy(POS-Subset) は，以下の手続きによって求めるジャンル領域拡大化能力（accuracy）である．

（1） POS フィルタリング：POS-Set から分類器の訓練・テストに使用する POS を選択し，それらを要素とする POS-Subset を構成する．

（2） 素性選択：素性の組合せとして，

$$\text{Feature-Subset} = \{(\text{word}, \text{pos}) \mid (\text{word}, \text{pos}) \in \text{Feature-Set} \ \&\ \text{pos} \in \text{POS-Subset}\}$$

を選択する．

（3） 分類器の訓練・生成：Feature-Subset に対応して，訓練用データセットによって分類器を訓練・生成する．

（4） ジャンル領域拡大データセットの分類：Feature-Subset に対応して，その分類器をジャンル領域拡大データセットに適用し，POS-Subset に対応する accuracy を求める．

accuracy(POS-Subset) を求める手続きから分かるように，上記問題の解を求めるためには，まず分類器を訓練するための訓練用データセットが必要である．上記問題においては，これまで交差検定用データセットと称していたものを，訓練用データセットとして利用することとした．また分類器のジャンル領域拡大化能力を求めるためにジャンル領域拡大データセットが必要である．

上記問題の解を求めることによって，上記問題の枠内でジャンル領域拡大化能力が高い分類器が得られる．しかしながら，その分類器が上記問題の枠外でも，言い換えると，上記問題の枠内で利用されるジャンル領域拡大データセットとは異なる別のジャンル領域拡大データセットに対しても，同程度のジャンル領域拡大化能力を有するかどうかは不明である．最終的に得たいものは，分類器の生成に関与しないジャンル領域拡大データセットに対して，実用的とみなせる程度のジャンル領域拡大化能力を有する分類器である．したがって最終的には，上記問題を解くことによって得られる分類器を別のジャンル領域拡大データセットに適用し，そのジャンル領域拡大化能力を求め，評価しなければならない．

このことは，2 種類のジャンル領域拡大データセットを用意する必要があることを意味している．一つは上記問題の枠内で利用するためのものである．もう一つは，上記問題を解くことから得られる分類器のジャンル領域拡大化能力を，上記問題の枠外で評価するためのものである．**5.** で述べたように，13 ジャンルに分布する 200 件のウェブページからなるジャンル領域拡大

表 4　2 種類のジャンル領域拡大データセット（A と B）
Table 4　Two kinds of genre-expanded datasets (A and B).

| | ジャンル領域拡大データセットA | | | ジャンル領域拡大データセットB | | |
|---|---|---|---|---|---|---|
| ジャンル | ページ件数 | 主観的ページ件数 | 非主観的ページ件数 | ページ件数 | 主観的ページ件数 | 非主観的ページ件数 |
| 予想 | 0 | 0 | 0 | 1 | 0 | 1 |
| 標語 | 0 | 0 | 0 | 1 | 1 | 0 |
| 批評 | 1 | 1 | 0 | 0 | 0 | 0 |
| 解説 | 1 | 0 | 1 | 1 | 0 | 1 |
| 報道 | 2 | 0 | 2 | 2 | 0 | 2 |
| 独白・感想 | 10 | 9 | 1 | 10 | 9 | 1 |
| 情宣 | 2 | 0 | 2 | 2 | 1 | 1 |
| 質問・相談・依頼 | 6 | 1 | 5 | 6 | 1 | 5 |
| 記録 | 14 | 1 | 13 | 15 | 2 | 13 |
| 商品広告・宣伝 | 13 | 2 | 11 | 12 | 1 | 11 |
| マニュアル | 7 | 0 | 7 | 7 | 0 | 7 |
| 用語説明 | 6 | 1 | 5 | 6 | 1 | 5 |
| 案内・紹介 | 38 | 7 | 31 | 37 | 6 | 31 |
| 総件数 | 100 | 22 | 78 | 100 | 22 | 78 |

データセットを既に構成した．手続きの簡潔性の面から，本論文では新たにウェブページを収集せず，**5.** で構成したデータセットから表 4 に示すような 2 種類のジャンル領域拡大データセット A，B を構成した．そして，一方を POS 組合せ最適化問題の枠内で用い，他方を POS 組合せ最適化問題の枠外で用いることとした．

### 7.3　遺伝的アルゴリズム（GA）の適用

POS 組合せ最適化問題を解くために，特に POS フィルタリングのために，メタヒューリスティクスの一つである GA を利用する．メタヒューリスティクスにはシミュレーテッドアニーリング法，タブーサーチ法など他の方法もあるが，GA を利用する最大の理由は，そのアルゴリズムが単純であるからである．更に，POS 組合せ最適化問題における目的関数 accuracy(POS-Subset) についての微分可能性や単峰性などの具体的知識が不要であり，多点探索を行うことで大域的探索が可能であることも，GA を利用する理由である．

Morariu ら [15] は既に，SVM によるテキスト分類において，素性選択のために GA を適用している．ただし，分類対象とするデータセットは新聞記事であり，分類カテゴリーは記事の主題を用いており，素性は BOW を用いている．また，素性選択では直接的に素性を選択する方法を採用している．GA の適用は，記事を表現する素性ベクトルの次元縮小を主な目的としている．

これに対して本論文では，分類カテゴリーはセンチメントカテゴリーを用いており分類課題が異なる．素性は BOW と POS の組合せであり，素性が異なる．また，直接的に素性を選択するのではなく，POS フィルタリングに基づく素性選択という方法を採用しており，素性選択の方法が異なる．これは探索空間の縮小を可能にしてくれる．最大の相違点は，本論文では分類器のジャンル領域拡大化能力の最大化を目的として GA を適用する点である．ジャンル領域拡大化能力は分類器の新しい評価指標であり，その最大化を目的として GA を適用した研究はない．

#### 7.3.1　POS-Subset のコード化

POS 組合せ最適化問題に GA を適用するためには，POS-Subset を GA のアルゴリズムにおいて操作可能なようにコード化する必要がある．本論文では，次のようにコード化する．

70 種類の全 POS からなる集合 POS-Set に属する各 POS に対して，0 から 69 の ID 番号（以下，これを POS-ID と称す）を付与する．POS-Subset をコード化するために，図 1 のように 70 ビット長のビット列を用意する．ここで，各ビットの位置を示すビット番号は，ビット列の先頭のビットから最後のビットに向けて 0, 1, . . . , 69 とする．POS-Subset に属する各 POS について，その POS-ID（0〜69）をビット番号とするビットにコード 1 を設定する．それ以外のビットにはコー

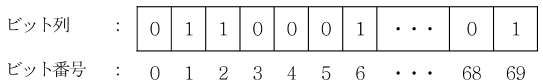


図 1　POS-Subset のコード化（個体）
Fig. 1　Coding of POS-Subset (individual).

ド 0 を設定する．このようにしてコード化された POS-Subset を個体と呼び，個体の集まりを集団と呼ぶ．

**7. 3. 2**　処理手続き

GA を利用して POS 組合せ最適化問題を解く手続きの概略を以下に述べる．

（1） 個体（コード化された POS-Subset）をランダムに 20 個生成し，第 1 世代の集団とする．

（2） 現世代の集団の各個体に対応して，以下 (2-1)〜(2-3) を実行する．

(2-1) 訓練用データセットによって分類器を訓練・生成する．ここで訓練用データセットは，**2.** で言及した交差検定用データセットの全件である．

(2-2) 分類器をジャンル領域拡大データセット（A または B）に適用し，GA における適合度としてジャンル領域拡大化能力（accuracy）を求める．

(2-3) ジャンル領域拡大化能力が許容水準以上である分類器と，それをもたらす個体を格納する．本論文では 85%を，ある程度実用的な水準であると考え，許容水準を 85%とした．ここで格納する個体と分類器は，分類器の生成に関与しない，つまり上記(2-2)で用いられないジャンル領域拡大データセット（B または A）に適用して，POS 組合せ最適化問題の枠外でのジャンル領域拡大化能力を求める対象になる．

（3） 世代数が 50 に至ったかどうかを判定し，至っていれば終了し，そうでなければ次の手続き(4)を実行する．

（4） 次世代の集団を設定し，(2)以下を繰り返す．次世代の個体は以下のように設定する．

(4-1) 現世代の 20 個体を，適合度をもとにして降順に並べる．

(4-2) 適合度が高い上位の 4 個体をエリート個体として選択する．

(4-3) その他 16 個体からランダムに選んだ二つの個体からなる組を 8 組作り，各組に交叉オペレータを適用し，子と呼ばれる 16 個体を作る．

(4-4) 16 個体の子に突然変異オペレータを適用し，新しい 16 個体の子を作る．また，4 個のエリート個体にも突然変異オペレータを適用し，新しい 4 個体の子を作る．エリート個体は，交叉オペレータ，突然変異オペレータの適用対象とせず，そのまま次世代の個体として残すのが普通である．しかしながら，より優れた次世代個体を得る上で最も有効な GA オペレータは突然変異オペレータである．エリート個体をそのまま次世代個体として残した場合，適合度の高い個体を得にくいという予想から，ここでは交叉オペレータの適用対象からは除外するが，突然変異オペレータの適用対象としている．

(4-5) 以上，エリート個体の選択，交叉オペレータ，突然変異オペレータの適用によって作られた 20 個体の子を次世代の集団とする．

**7. 3. 3**　適用結果とジャンル領域拡大化能力

ジャンル領域拡大データセットは A と B の 2 種類がある．これに応じて，GA を適用する POS 組合せ最適化問題としては以下の 2 種類を設定できる．

（1） POS 組合せ最適化問題 (A)：ジャンル領域拡大データセット A を用いて適合度を求める．

（2） POS 組合せ最適化問題 (B)：ジャンル領域拡大データセット B を用いて適合度を求める．

本論文では，これら 2 種類の問題に GA を適用した．その際，分類器の訓練法として libSVM-Linear と libSVM-RBF を用いた．それぞれの問題に GA を適用し，それぞれについて得られた最良の適合度 (A) と適合度 (B) を表 5 に示す．問題 (A) の libSVM-RBF と問題 (B) の libSVM-Linear については，複数の同じ適合度を示しているが，これらは最良の適合度を示す異なる個体（POS-Subset）が複数存在したことを意味している．また表 5 には，当該適合度 (A)/(B) を達成する分類器をジャンル領域拡大データセット B/A に適用して得られたジャンル領域拡大化能力を，ジャンル領域拡大化能力 (B)，ジャンル領域拡大化能力 (A) として示す．

**6.** で述べたように，交差検定用データセットの 2000 件すべてを訓練用データとして訓練法 libSVM-Linear，libSVM-RBF を用いて分類器を生成し，その分類器をジャンル領域拡大データセットの 200 件すべてに適用してジャンル領域拡大化能力を求め，結果を表 3 の accuracy(all) に示した．表 3 の accuracy(all) と表 5 のジャンル領域拡大化能力 (A)/(B) とを比較することは公平とはいえない．なぜならば，accuracy(all) は 200 件からなるジャンル領域拡大データセットを対象としているのに対し，ジャンル領域拡大化能力 (A)/(B) は，100 件からなるジャンル領域拡大データセットを対象としているからである．比較の公平性を

表 5　GA の適用結果とジャンル領域拡大化能力
Table 5　Results of GA and genre expansion.

| 問題 | 訓練法 | 適合度 | ジャンル領域拡大化能力 | 比較基準 | 有意確率 |
|---|---|---|---|---|---|
| (A) | libSVM-Linear | 適合度(A) | ジャンル領域拡大化能力(B) | 比較基準(B) | |
| | | 89.0 | 77.0 | 75.0 | 0.644167 |
| | libSVM-RBF | 適合度(A) | ジャンル領域拡大化能力(B) | 比較基準(B) | |
| | | 93.0 | 82.0 | 59.0 | 0.000003 |
| | | 93.0 | 84.0 | | 0.000000 |
| | | 93.0 | 85.0 | | 0.000000 |
| (B) | libSVM-Linear | 適合度(B) | ジャンル領域拡大化能力(A) | 比較基準(A) | |
| | | 86.0 | 79.0 | 80.0 | 0.802587 |
| | | 86.0 | 82.0 | | 0.617075 |
| | libSVM-RBF | 適合度(B) | ジャンル領域拡大化能力(A) | 比較基準(A) | |
| | | 93.0 | 79.0 | 70.0 | 0.049535 |

確保するために，交差検定用データセットの 2000 件すべてを訓練用データとして訓練法 libSVM-Linear, libSVM-RBF を用いて生成された分類器を，改めてジャンル領域拡大データセット A/B に適用して，分類器のジャンル領域拡大化能力を求めた．表 5 においては，その結果をそれぞれ比較基準 (A)，比較基準 (B) として示してある．これらの数値は素性選択を行わない場合のものである．更に，ジャンル領域拡大化能力と比較基準との統計的有意差を検討するため，$\chi^2$ 両側検定によって有意確率を求めた．それを表 5 における「有意確率」の欄に示す．

表 4 に示したとおり，ジャンル領域拡大データセット A と B については，一方だけに含まれるジャンルがある．批評の 1 ジャンルはジャンル領域拡大データセット A だけに含まれており，予想と標語の 2 ジャンルはジャンル領域拡大データセット B だけに含まれている．表 5 は，ジャンルに若干の差異があるこのようなジャンル領域拡大データセットを用いても，GA を適用した POS フィルタリングに基づく素性選択によって，比較基準に比べて改善されたジャンル領域拡大化能力を達成できること，場合によっては比較基準に比べて統計的有意差が認められる程度のジャンル領域拡大化能力を達成できること，更には，ある程度実用的な分類器を生成できることを示している．

表 5 において，POS 組合せ最適化問題 (B) については，訓練法 libSVM-Linear を利用したときは，ジャンル領域拡大化能力が比較基準を 1%下回る場合があるが，2%改善される場合もある．他の場合では程度の差はあるが，ジャンル領域拡大化能力は比較基準に比べて改善されている．このことから，GA を適用した POS フィルタリングに基づく素性選択によって，ジャンル領域拡大化能力を改善できることが明らかになった．ただし，比較基準に比べて，改善されたジャンル領域拡大化能力に統計的有意差がすべての場合に認められるわけではない．POS 組合せ最適化問題 (A) と (B) ともに，訓練法 libSVM-Linear を用いた場合には統計的有意差は認められない．一方，訓練法 libSVM-RBF を用いた場合には，統計的有意差が認められる．このうち，85%のジャンル領域拡大化能力を発揮する分類器が得られており，ある程度実用的な分類器を生成できることが明らかになった．

なお，ジャンル領域拡大データセット A と B のジャンルには若干の相違はあるものの，ジャンル分布の点では，ほぼ類似のデータセットともみなせる．この場合，POS 組合せ最適化問題を解き，最良の個体（POS-Subset）と，そのもとで生成される分類器を求めれば，ジャンル領域拡大化能力が比較基準を上回ることは当然であると考えられるかもしれない．しかしながら，表 5 から分かるように，確かに適合度はすべて比較基準を上回るものの，ジャンル領域拡大化能力が比較基準を下回る場合がある．

**7. 3. 4**　獲得した POS-Subset

上述したように，POS 組合せ最適化問題に GA を適用して，最良の適合度を与える解（POS-Subset）の探求を通じて，ジャンル領域拡大化能力がある程度実用的水準にある分類器を生成することができた．こ

表 6　獲得した POS-Subset
Table 6　Acquired POS-Subset.

| | POS-Subset |
|---|---|
| (1) | **その他-間投，助動詞，**助詞-並立助詞，助詞-係助詞，**助詞-接続助詞，**助詞-特殊，**助詞-終助詞，**助詞-連体化，**動詞-自立，名詞-ナイ形容詞語幹，名詞-一般，**名詞-代名詞-縮約，**名詞-副詞可能，名詞-動詞非自立的，名詞-固有名詞-一般，**名詞-固有名詞-人名-姓，**名詞-形容動詞語幹，名詞-接尾-サ変接続，**名詞-接尾-人名，**名詞-接尾-副詞可能，**名詞-接尾-助動詞語幹，**名詞-接尾-形容動詞語幹，**名詞-非自立-一般，形容詞-自立，**接頭詞-動詞接続，接頭詞-名詞接続，**記号-括弧開，連体詞** |
| (2) | **その他-間投，**フィラー，副詞-助詞類接続，**助動詞，**助詞-副助詞，助詞-副助詞／並立助詞／終助詞，助詞-副詞化，**助詞-接続助詞，**助詞-格助詞-一般，**助詞-終助詞，**動詞-接尾，**動詞-自立，名詞-ナイ形容詞語幹，名詞-一般，名詞-代名詞-縮約，名詞-固有名詞-人名-姓，**名詞-固有名詞-地域-一般，名詞-接尾-一般，**名詞-接尾-人名，名詞-接尾-助動詞語幹，**名詞-接尾-特殊，**名詞-非自立-一般，**名詞-非自立-形容動詞語幹，**形容詞-自立，**形容詞-非自立，感動詞，接続詞，接頭詞-数接続，**記号-括弧開，連体詞** |

のことは，ジャンル領域拡大化能力の改善にとって有効な POS-Subset を獲得したことを意味するが，必ずしも言語学的に意味のある，あるいは説明可能な POS-Subset を得たわけではない．したがって，本論文でそれを明示的にする価値はないと思われる．

しかし一方で，どのような POS-Subset が獲得されたかを示すことは，今後の類似研究に貢献するものと考えられる．ここでは，あくまでも参考として，表 5 において最高値であり，比較基準との統計的有意差が認められ，ある程度実用的水準にあると考えられるジャンル領域拡大化能力 85%を達成する場合の POS-Subset を示す．そのような POS-Subset は，適合度 (A) = 93.0%，ジャンル領域拡大化能力 (B) = 85.0% を達成するもので，2 組が得られている．それらを表 6 に示す．

太字で示した POS は，双方に共通するものであり，ジャンル領域拡大化能力の改善に重要な役割を果たしたと推察できる．その他の POS については，重要であるかどうかは不明であり，ジャンル領域拡大化能力の改善にとって効果のないものが含まれている可能性を否定できない点に注意を要する．

## 8. むすび

本論文では，日本語ウェブページを対象に，機械学習法を利用して主観的か非主観的かに分類する課題に取り組んだ．簡単な素性と素性値を用いて，限られたジャンルに分布する交差検定用データセットについては，非常に高い性能で分類することができた．しかしながら，交差検定用データセット上で訓練・生成される分類器は，そのままではジャンル領域拡大化能力という点で実用的とはいえなかった．そこで，素性選択によるジャンル領域拡大化能力の改善を試みた．その際，GA を適用した POS フィルタリングに基づく素性選択を提案した．このような素性選択によってジャンル領域拡大化能力の改善を試みた結果，ジャンル領域拡大化能力を改善でき，ある程度実用的とみなせる分類器を生成することができた．GA を適用した POS フィルタリングに基づく素性選択は，限られたウェブページ資源を利用して実用的な分類器を得ようとするものであり，その工学的な有用性が明らかになったといえる．

現状では，十分に高いジャンル領域拡大化能力を有する分類器を得たわけではない．更に能力の高い分類器を得るためには，GA の適用方法に工夫が必要と考えられる．これについては今後の課題としたい．

**謝辞**　NTCIR-3 WEB は国立情報学研究所の許諾を得て使用させて頂きました．この場を借りて深謝致します．土井晃一工学博士には，本研究の初期段階での議論に参加して頂いたことに感謝致します．

## 文　献

[1] N. Dewdney, C. VanEss-Dykema, and R. MacMillan, “The form is the substance: Classification of genres in text,” Proc. Workshop on Human Language Technology and Knowledge Management — Volume 2001, pp.1–8, Toulouse, France, July 2001.

[2] Y.B. Lee and S.H. Myaeng, “Text genre classification with genre-revealing and subject-revealing features,” Proc. 25th Annual International ACM SIGIR Conf. on Research and Development in Information Retrieval, pp.145–150, Tampere, Finland, Aug. 2002.

[3] 木村託巳，山田寛康，島津　明，“WWW 探索支援のための記述意図によるテキスト分類,” 言語処理学会第 9 回年

次大会発表論文集，pp.505–508, 2003.
[4] S. Meyer zu Eissen and B. Stein, “Genre classification of web pages,” Proc. 27th German Conf. on Artificial Intelligence (KI-2004), pp.256–269, Ulm, Germany, Sept. 2004.
[5] A. Kennedy and M. Shepherd, “Automatic identification of home pages on the web,” Proc. 38th Hawaii Int. Conf. on System Sciences (HICSS’05)-Track 4-Volume 04, p.99.3, Big Island, Hawaii, United States, Jan. 2005.
[6] P.D. Turney, “Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews,” Proc. 40th Annual Meeting of the Association for Computational Linguistics (ACL-2002), pp.417–424, 2002.
[7] B. Pang, L. Lee, and S. Vaithyanathan, “Thumbs up? Sentiment classification using machine learning techniques,” Proc. Conf. on Empirical Methods in Natural Language Processing (EMNLP-2002), pp.79–86, 2002.
[8] A. Finn and N. Kushmerick, “Learning to classify documents according to genre,” IJCAI-03 Workshop on Computational Approaches to Style Analysis and Synthesis, Acapulco, Mexico, Aug. 2003.
[9] H. Yu and V. Hatzivassiloglou, “Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences,” Proc. Conf. on Empirical Methods in Natural Language Processing (EMNLP-2003), pp.129–136, 2003.
[10] J. Wiebe, T. Wilson, R. Bruce, M. Bell, and M. Martin, “Learning subjective language,” Computational Linguistics, vol.30, no.3, pp.277–308, 2004.
[11] J. Wiebe and E. Riloff, “Creating subjective and objective sentence classifiers from unannotated texts,” Proc. 6th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2005), pp.475–486, 2005.
[12] 平　博順，春野雅彦，“Support Vector Machine によるテキスト分類における属性選択，” 情処学論，vol.41, no.4, pp.1113–1123, 2000.
[13] V.N. Vapnik, The Nature of Statistical Learning Theory, Springer, 1995.
[14] F. Sebastiani, “Machine learning in automated text categorization,” ACM Comput. Surv., vol.34, no.1, pp.1–47, 2002.
[15] D.I. Morariu, L.N. Vintan, and V. Tresp, “Evolutionary feature selection for text documents using the SVM,” Trans. Engineering, Computing and Technology, vol.15, pp.215–221, Oct. 2006.

## 付　　録

### 1. accuracy と F-measure

ウェブページを分類した結果として表 A·1 に示すような分類行列(注17)（confusion matrix）が与えられた場合，accuracy と F-measure の定義を以下に示す．ここで，カテゴリー A/カテゴリー B の対には，主観的/客観的，事実/意見，主観的/非主観的などがある．

表 A·1　分類行列
Table A·1　Confusion matrix.

| | | 分類器による分類 | |
|---|---|---|---|
| | | カテゴリーA | カテゴリーB |
| 人間による分類 | カテゴリーA | a (件) | b (件) |
| | カテゴリーB | c (件) | d (件) |

$$\text{accuracy} = (a+d)/(a+b+c+d).$$

$$\text{recall: } R = a/(a+b).$$

$$\text{precision: } P = a/(a+c).$$

$$\text{F-measure} = 2PR/(P+R).$$

### 2. 日本語ウェブページのジャンル体系

以下に，本論文で利用するジャンル体系を示す．各ジャンルの説明における「物」と「事」は広辞苑に従い，「物」とは形のある物体をはじめとして，存在を客観的に感知できる対象であり，「事」とは意識・思考の対象のうち，具象的・空間的でなく，抽象的に考えられるものであるとする．

予想　　ある物・事の今後の動きや結果についてあらかじめ想像したもの．例としては，株価予測や天気予報を記載したページがある．

標語　　個人・団体のモットーやスローガン，商品のキャッチフレーズなど，人の注意をひくように工夫した簡潔な文言のみを書いたもの．例としては，「滅菌効果抜群！」というような商品のキャッチフレーズを記載したページがある．なお，簡潔な文言とともに，その説明が加えられている場合は，このジャンルには該当しない．

批評　　ある物・事の善悪・優劣・美醜・良悪・是非などについて第三者として評価し論じたもの．例としては，業況批評を記載したページがある．

解説　　ある物・事について，その内容（法文の条項，文学作品，ニュース記事，ある人の発言，株価動向など）を引用した上で，それを分析して，分かりやすいように客観的に説明したもの．例としては，ある法文を引用し解説したページがある．通常，引用は他

(注17)：confusion matrix の日本語訳として，正誤判別行列，混同マトリックスなどが用いられているが，ここでは分類行列という用語を用いる．

の情報源からとられた文章の形式をとるが，場合によっては，グラフや表の形式をとることもある．引用内容が明示的でない場合は，このジャンルに該当しない．更に，用語（単語，句）の意味を説明するページは，このジャンルに該当しない．

報道　　社会の出来事などを報道機関が告げ知らせようとしたもの．例としては，ニュースの見出し（ヘッドライン），ニュース記事本文を記載したページがある．ニュースのような情報を記載はしているが，報道機関が明示的でない場合，このジャンルに該当しない．ここで報道機関は通常のニュース配信組織（テレビ局，新聞社，雑誌発行機関など）だけを意味していない．報道機関には，他のページに記載されたニュースを改めて報道する組織・機関（例えば，http://news.google.co.jp）も含まれる．ただし，個人は含まれない．

独白・感想　　ある物・事について，個人あるいは複数の人がその人の立場で気ままに自分の考え・思いを語ったもの，あるいは個人的な体験記．例としては，経験談，私的な記録（日記など），私的なメッセージ，個人の布教文，自作の書き物（物語，雑談など），メール，を記載したページがある．

情宣（情報宣伝）　　団体（圧力団体，宗教団体，町内会，労働組合など）が推し進めようとしている考え方・思想・経典についての情報を提供し，その有効性や有用性，危険性や有害性などを説明して理解・共鳴させようとしたもの．例としては，ウェブ版ニュースレター，ウェブかわら版，機関紙を記載したページがある．

質問・相談・依頼（回答付き限定）　　ある物・事について，他人に意見を求め，返答を得たもの．例としては，FAQ，Q&A，不特定多数による一連の質問・回答を記載したページがある．質問だけ，あるいは，ある質問に対する回答だけを記載するページは，このジャンルに該当しない．

記録（現代文限定）　　過去の事実や将来の計画について，発信者あるいは受け手が後々の証拠として使い得るように，書き残されたもの．ただし，現代文に限定する．例としては，研究報告，終了イベント，競技結果，史実，議事録，都市再開発計画を記載したページがある．たとえ事実や計画を記載していても，記載内容が単純に追加されるという更新を除いて，記載内容の一部あるいは全部が更新される可能性が高いページは，このジャンルに該当しない．

商品広告・宣伝　　商品化されたものについて，人々に関心をもたせ，購買させることを目的として，最低限，その商品を特定する情報（商品名，品番など）と価格情報を知らせたもの．例としては，株式投信，講座，宿泊施設，駐車場の広告を記載したページがある．「オープンプライス」という語句は，価格情報とみなす．以下のようなページはこのジャンルに該当しない．

(a)　だれかによって既に購買された商品の情報を記載するページ．

(b)　製造中止とか販売終了とかの理由で，既に購入できなくなっている商品の情報を記載するページ．

マニュアル　　ある物・事について，その使用法，調理法，摂取法，作成法，設定法，見方，進め方などの手順に関する技術的知識・情報のすべて，ないし一部を主として現在形で説明したもの，あるいはノウハウ．例としては，PC 関連，書類届出関連のマニュアル，ノウハウを記載したページがある．

用語説明　　ある物・事について，それが何であるかの客観的説明に重点を置いたもの．例としては，PC 用語，歴史的人物，史跡，文化遺産を記述したページがある．

案内・紹介　　情報の受け手にとって未知，既知を問わず，ある物・事へ導く情報，あるいは，物事の特徴的情報を羅列して知らせたもの．例としては，イベント案内，方法の紹介，新製品の紹介を行ったページがある．

その他　　上記のどのジャンルにも当てはまらないもの．例としては，英語ページ，画像ページ，Not-Found ページがある．

### 3. ジャンル体系の設定法

ジャンル体系の設定は，2 人の作業者によって行った．この 2 人の作業者をジャンル設定者と呼ぶことにする．ジャンル設定者はまず，日本語ウェブページに関する経験に基づいて，ジャンル体系の初期暫定版を設定した．初期暫定版に含まれたジャンルは，(1) 広告，(2) 宣伝，(3) 案内，(4) 紹介，(5) 風評，(6) 批評，(7) 相談，(8) 感想，(9) 会合記録，(10) 独白，(11) 標語，(12) 解説，(13) その他，の 13 ジャンルであった．

続いて，付録 **2.** に示したジャンル体系の確定版を求めるために，ジャンル設定者が暫定版の洗練を繰り返した．暫定版の洗練化を通じて確定版を求める作業は，以下のように行った．以下の作業はすべて，ジャンル設定者（2 人）が同席して綿密な議論を通じて協同で行った．

(1) 暫定版における問題点の明確化　暫定版のジャンル体系に従って日本語ウェブページを分類する際の問題点を明らかにするために，NTCIR-3 WEB (注18) のウェブページからランダムに抽出した 200 件(注19) を閲覧し，分類した．ウェブページがどのジャンルに属するかの決定（つまり分類）は，ジャンル設定者（2 人）の合意によって行った．その結果，初期暫定版の代表的な問題点として以下が明らかになった．

- ニュース記事など，重要性の高いジャンルとして分類されるべきウェブページであるにもかかわらず，暫定版では該当するジャンルがなく，ジャンル「その他」に分類せざるを得ない．
- 広告と宣伝，案内と紹介，感想と独白のようなジャンルは，両者を明確に区別することが困難である．
- 会合記録のようなジャンル名とその定義はあまりにも狭義であり，何らかの記録を掲載するウェブページであっても，会合記録以外はジャンル「その他」に分類せざるを得ない．
- 各ジャンルの定義が十分に洗練されていないため，分類に悩む場合が多々ある．

(2) 暫定版における問題点の解消　暫定版が抱える問題点を解消するために，ジャンル名と定義を改訂した．改訂では，新しいジャンル名の追加（例えば，報道を追加する），既設ジャンル名の併合（例えば，案内と紹介を併合して案内・紹介とする），既設ジャンル名の変更（例えば，会合記録を記録に変更する）を行った．ジャンル名の改訂に伴いジャンルの定義を新設・改訂し，結果として新しい暫定版を設定した．ジャンルの定義を新設・改訂する際には，200 件のウェブページの閲覧経験をもとに，各ジャンルを特徴づけるような定義を与えた．

(3) 暫定的確定版の設定　新しい暫定版について，上記 (1) と (2) のように問題点の明確化と解消を繰り返した．この繰返しは，200 件のウェブページを分類するにあたり適当と考えられるジャンル体系を得るまで行った．そして，結果として得たそのようなジャンル体系を，暫定的確定版として設定した．

(4) 暫定的確定版における問題点の明確化　先に用いた 200 件のウェブページとは異なる未閲覧のウェブページを分類する際に，暫定的確定版のジャンル体系にどのような問題が生じるかは不明であった．問題点を明らかにするため，再び NTCIR-3 WEB のウェブページから未閲覧のウェブページ 500 件をランダムに抽出し，暫定的確定版のジャンル体系に従って，それらの閲覧・分類を試みた．その場合，ジャンル体系の問題点がほぼ明確になったと考えられ，これ以上の閲覧には意味がないと判断した時点で閲覧・分類を止めた．結果としては，138 件のウェブページを新たに閲覧・分類した．ジャンル名は適当であったが，定義を多少洗練する必要のあるジャンルがあった．

(5) 確定版の設定　暫定的確定版におけるジャンルの定義を改訂した．その後，改訂した暫定的確定版に従って，既に閲覧済みの 338 件のウェブページの分類を改めて行った．その結果，問題がないと判断し，それを確定版として設定した．

付録 **2.** に示すジャンル体系は，以上のようにして設定したものであり，ほぼ納得できるジャンルと定義を与えており，ジャンル間の境界が既存のジャンル体系より明確である．そのため，我々にとって使いやすいというだけでなく，他の研究者による再利用が可能であると考えられる．

（平成 19 年 2 月 28 日受付，8 月 18 日再受付）

(注18)：主として .jp ドメインから，画像を除外して収集された 11,034,409 件のウェブページを含んでおり，ジャンルの網羅性は高いと考えられる．NTCIR-3 WEB については以下の URL を参照されたい．http://research.nii.ac.jp/ntcir/permission/perm-ja.html#ntcir-3-web

(注19)：200 件のウェブページは，ジャンル体系の暫定的確定版を得るまでに何度か閲覧する対象である．件数の設定には客観的根拠はない．強いて理由を挙げれば，適度な多様性が確保できる一方で閲覧負荷が大きくならないこと，更に分類記憶が残りにくいことを考慮して，200 件が適当であろうと考えたことによる．

**大森　晃**（正員）

1985 広島大学大学院工学研究科博士課程後期了（システム工学専攻）．工博．1982 年 9 月より 1 年間ケースウェスタンリザーブ大学客員研究員．1985 年 4 月より富士通国際情報社会科学研究所に勤務．1993 年 10 月より東京理科大学工学部第二部経営工学科助教授（現在，准教授）．ソフトウェア工学，品質管理，言語情報処理，HCI，教育工学などの研究に従事．IEEE Computer Society，ACM，日本品質管理学会，情報処理学会，言語処理学会各会員．