文末表現を利用したウェブページの主観・客観度の判定

松本　章代† 　小西　達裕†† 　高木　朗††† 　小山　照夫†††† 　三宅　芳雄†††††
伊東　幸宏††

† 青山学院大学　〒 229–8558 神奈川県相模原市淵野辺 5–10–1
†† 静岡大学　〒 432–8011 静岡県浜松市中区城北 3–5–1
††† 言語情報処理研究所　〒 184–0014 東京都小金井市貫井南町 3–6–30
†††† 国立情報学研究所　〒 101–8430 東京都千代田区一ツ橋 2–1–2
††††† 中京大学　〒 470–0393 愛知県豊田市貝津町床立 101
E-mail: riir@inf.shizuoka.ac.jp

あらまし　ウェブ検索を行う際，客観的な事実，もしくは主観的な意見のどちらかのみを必要とするケースがある．このような場合において，不要な方のページを検索結果から排除できるようにすることには意義がある．我々は，主観／客観のどちらの情報を中心としてページが構成されているかを推定するために文末表現に着目した．Google が提供する API を利用し，既存のサーチエンジンの検索結果を「主観情報が主体のページ」と「客観情報が主体のページ」とに分類して表示するユーザインタフェース部として実装した．評価実験の結果，文末表現を用いることで訓練データのトピックに依存しにくい分類が可能となることが確認された．
キーワード　評判情報，意見，テキスト分類，テキストマイニング，SVM

Judgment of Subjectivity Using Sentence Final Expression in Web Pages

Akiyo MATSUMOTO†, Tatsuhiro KONISHI††, Akira TAKAGI†††, Teruo KOYAMA††††, Yoshio MIYAKE†††††, and Yukihiro ITOH††

† Aoyama Gakuin University　5–10–1 Fuchinobe, Sagamihara-shi, Kanagawa 229–8558 Japan
†† Shizuoka University　3–5–1 Johoku, Naka-ku, Hamamatsu-shi, Shizuoka 432–8011 Japan
††† NLP Research Laboratory　3–6–30 Nukuiminami-cho, Koganei-shi, Tokyo 184–0014 Japan
†††† National Institute of Informatics　2–1–2 Hitotsubashi, Chiyoda-ku, Tokyo 101–8430 Japan
††††† Chukyo University　101 Tokodachi, Kaizu-cho, Toyota-shi, Aichi 470–0393 Japan
E-mail: riir@inf.shizuoka.ac.jp

Abstract　When using a web search engine, one may require either objective facts or subjective opinions. Therefore, it would be helpful for users if unnecessary pages were to be eliminated from the search results. We analyze the final expression of a sentence to determine the “subjectivity” or “objectivity” of a web page. We have used an API provided by Google and implemented our proposed method in the user interface of a typical search engine. Our method categorizes the obtained search results as being subjective or objective. The results of our experimental evaluation have revealed that our proposed method can be used to categorize web pages without requiring domain-specific training data by making use of the final expression of a sentence.
Key words　Reputation, Opinion, Text Classification, Text Mining, SVM

1. はじめに

ウェブ検索を行う際，ユーザが持つ意図は様々であり，検索キーワードのみでは判断ができないことがある．例えば，「相模原にある美容院」についての情報が欲しいと考え，「相模原　美容院」で検索するとする．そのとき，カットの料金や場所を比較したい場合もあれば，実際に利用した人の感想など，評判が知りたいこともありうる．前者の場合は各美容院の公式サイト，後者の場合はある程度まとまった量の様々な評判情報を閲覧したいと推測できる．すなわち，同じキーワードであっても，検

索者の意図によって欲しい情報は異なるため，一方を無条件に排除する戦略は好ましくない．ただし，客観的な事実，もしくは主観的な意見のどちらかのみを必要としているユーザーに対して，不要な方のページを排除することを選択できるようにすることには検索効率の向上という面において意義がある．

そこで，ウェブ検索時に検索結果として与えられたページの内容が「主観的な感想や意見を述べたページ」なのか「客観的な事実を述べたページ」なのかを判別し，一目でわかりやすいように表示するシステムを作成する．各ページに対し「主観的な意見」が主体となっているページか否かをそれぞれ推定し，検索結果の画面に「主観情報主体」と「客観情報主体」とをグルーピングして表示する．これにより，検索結果画面からリンクをたどらないとわからなかった「主観的」か「客観的」かのおおまかな判断が検索結果画面上のみでできるようになり，求めているデータにたどりつくまでの時間や手間を短縮することができる．

ウェブページをカゴテリごとに分類する手法はこれまでにも多数提案されている [1] [2]．中でも我々と最も近い研究に,「事実」と「意見」を分類する Finn らによる手法 [3] がある．Finn らは，bag of words や品詞，記号の数や文の平均長といった統計情報，さらに人手で作成した主観性を表す語のリストなどを用い，機械学習によって分類を実現している．この手法では「フットボール」や「政治」といったトピックを限定した中では高い精度で判別できるものの，学習させたアルゴリズムを異なるトピックに適用すると，精度が著しく落ちるという問題が示されている．

この問題に対応すべく，我々は，判別に助動詞などの文末表現を用いることを検討する．体験記など（主観的な表現が多いページ）は「てみた」「てきた」という表現が多いであろうし，一方で公式サイトなど（客観的な表現が多いページ）には「ございます」「おります」などの敬語が多いなど，文末表現には何らかの傾向があり，かつトピックに依存しないと思われる．

一方，ウェブから評判情報を抽出し利用しようという試みは，近年盛んに行われている [4]．我々も，本研究の成果を基盤とし，より需要のあるシステムへとさらに発展させていきたい考えである．

本稿では，2 章でまず，主観／客観度の判定手法を検討する．3 章では，その判定手法に基づき，SVM を用いて主観／客観文書の判別を行う検索システムについて述べる．4 章では，このシステムの妥当性を検証するために行った評価実験について説明し，今回提案する手法が主観／客観度の判定に利用できること，また，形容詞・形容動詞を利用した場合と比較し，話題に依存しない判定が可能であることを示す．

2. 主観／客観度の判定手法

提案手法では，文末の助動詞／終助詞／表明思考動詞を抽出し，その頻度に基づいて主観／客観度の判定を行う．

2.1 文末に着目する理由

日本語は思考，状態，意志，態度など，文末に重要な表現が多くあらわれる [5]．日本語の文の場合，最も文末にある自立語が文の中心要素であるから，その語と結びついている助動詞や助詞も含めて，この部分は文全体の意味に深くかかわることになる．逆に，従属文中の動詞やその語と結びついた助動詞・助詞の重要度は低い．たとえば『先日発売された iPod は好評なようだ』という文をみると,「た」は過去を意味する助動詞であるが，文全体としては過去の話ではない．

これまでにも，土井 [6] がウェブページ中の文末表現に注目し，その実態調査を行っている．土井は，文末表現として句点の直前の 1 文字に着目し，ウェブページのジャンルとの関連を整理している．この結果は興味深いものであるが，文末の 1 文字だけでページのジャンルを特定できるわけではない．

そこで我々は，文の終端からさかのぼって自立語が出現するまで（自立語を含む）を「文末」と定義し，主観／客観度の判定にはこの「文末」の表現を利用する．たとえば『先日発売された iPod は好評なようだ』という文では,「好評だ」という形容動詞で表される自立語が出現するまでの「好評なようだ」の部分が文末となる．

2.2 助動詞，終助詞，表明思考動詞に着目する理由

助動詞は，時制（過去，未来），相（進行，完了），態（受動，使役），法（推量，命令，願望など）などの意味を文に付加する働きを持つ．推量や願望の助動詞（「らしい」「だろう」「たい」など）など，書き手の気持ちを表す語は，主観的な記述に使われる．また,「〜していた」「〜してきた」「〜してしまった」などの表現は，体験談などによく用いられると考えられる．つまり主観的な表現があるページに出現すると推測できる．一方,「くださる」「いただく」といった敬語は，企業の公式サイトなど，客観度の高いページで頻出することが予想される．

終助詞の「〜ね」「〜よ」などは，口語に頻出する表現であるため，ウェブページの場合は，日記や体験記など，つまり主観的な表現があるページで使用される傾向にあると考えられる．

また,「思う」「感じる」「コメントする」など，思考や表明を意味する動詞（以降「表明思考動詞(注1)」と呼ぶ）は，感想や意見などを述べるときに用いられるため，主観度の高いページで用いられる可能性が高い．

すなわち，これらの語は主観・客観の判定に有用であり，かつトピックに依存しないと考え，判断材料（素性）として採用する．

一方で Bruce ら [7] は，主観的な文には形容詞が含まれる可能性が非常に高いことを報告している．確かに形容詞は，主観的にとらえられた性質を表すために用いられることも多い．したがって，形容詞も素性に含めることにより，主観的な情報の判断に役立つ可能性はある．しかしながら，形容詞を含めることにより，トピックに依存する訓練データが作成されてしまう可能性が高いと考え，あえて素性に含めないことにする．

2.3 素性の工夫

素性について以下の改善案を検討する．

(1) 意味の近い語をグループ化（テ形複合動詞，当為，概言，

(注1)：引用接続助詞「と」が係り得る動詞．この品詞カテゴリは，今回利用したパーザの辞書の分類に基づいている．全部で約 250 語が登録されている．

敬意，など）する．

(2) 文末表現を形態素ごとに分割せず一塊にして素性とする．

2.3.1 形態素のグループ化

出現頻度が低い語は，判別に活かされにくい傾向にある．しかしながら，出現頻度が低い語しか存在しないページもあり，訓練データから選ばれた語だけでは，十分な判定が行えないページが実際には多いことがわかった．

そこで，実験に用いたページ中に含まれる形態素のうち，意味や働きが近い語をグループ化する．なお，助動詞と終助詞のグループ化については「基礎日本語文法 [8]」の分類に，表明思考動詞のグループ化についてはパーザの辞書に登録されたラベルに，それぞれ基づいている．

これらの形態素のグループを素性に加える．

表 1 グループ化した形態素（助動詞）

分類	形態素
受動・使役・可能	れる できる られる せる させる させられる される
過去	た
テ形複合動詞 (相に関係するもの)	ている てある てしまう ていく てくる
テ形複合動詞 (授受に関係するもの)	てもらう てやる てくれる てあげる てくださる ていただく
テ形複合動詞 (その他)	ておく てみる
許可・禁止	てよい てもよい てはいけない
依頼	てほしい
当為	べきだ ほうがよい べし なければならない なくてはいけない
意志	つもりだ ようとする
勧誘	う よう
願望	たい たいものだ
概言	らしい ようだ みたいだ はずだ だろう まい そうだ
説明	わけだ わけではない わけでもない
否定	ない
敬意	てくださる ていただく ござる いたす ておる くださる いただく なさる てまいる ていらっしゃる

表 2 グループ化した形態素（終助詞）

分類	形態素
断定	さ のさ
疑問	か のか かな のかな かしら かどうか のかい
確認・同意	ね よね な ねぇ ネ ねえ わね
知らせ	よ ぞ ぜ よっ わよ ゾ てよ
感嘆	なぁ なあ わ なあ〜
禁止	な

2.3.2 形態素の組み合わせ

例えば「てみる」という助動詞の場合,「てみよう」「てみたい」と「てみた」では，使われ方が大分異なる.「〜をやってみた」などという表現は，日記などによく用いられるが，願望の「たい」は広告のキャッチコピーなどによく利用されており，また勧誘の「よう」は広告，紹介，リンク集などに含まれ易いため,「てみよう」や「てみたい」は主観性が低いページにも頻出する．

表 3 グループ化した形態素（表明思考動詞）

分類	形態素
表明	コメントする 表示する 言う 発表する 教える 報告する 入力する 説明する 注意する 挨拶する
思考	思う 分かる 考える 知る 感ずる 願う 期待する 決定する

そこで，文末表現を形態素ごとに分割せず一塊にして素性として追加する．たとえば,「てきました」が出現した場合は,「てくる」「ます」「た」「てくる＋ます＋た」の頻度を +1 する．

2.4 主観／客観文書の分類法

主観／客観文書の判別には，テキスト分類には，高精度の判定が期待できる SVM を用いる．

そのために，ページの内容が主観情報主体なのか，客観情報主体なのかについて，あらかじめ人手で判定を行い，正解データを作成する．

3. システム概要

本システムは，検索エンジンに「Google AJAX Search API(注2)（以下，Google-API)」を用いたウェブアプリケーションである．VineLinux 上で動作しており，開発言語は Ruby，日本語の係り受け解析には（株）CSK が開発したパーザを用いている．

処理の流れは以下のとおりである．

(1) ユーザが入力した検索ワードに対し，Google-API を用いて検索結果を取得する．

(2) ダウンロードしたウェブページの文章から，構文解析器を用いて文末の助動詞，終助詞，表明思考動詞のみを抜き出す．

(3) 抜き出した各文末表現の頻度を算出し，あらかじめ SVM を用いて求めた識別関数を利用して「主観／客観」を判定する．

(4) 検索結果画面は左右に 2 分割し，左に客観／右に主観ページが並ぶように HTML ファイルを出力する（図 1）．

4. 評 価 実 験

4.1 正解データの作成

まず，第三者の被験者 3 名が合議し「よく行う検索の代表例」として以下の 20 種類の検索ワードを設定した．

- ”ヒープソート” アルゴリズム
- 相模原　美容院
- 相模原　アルバイト
- 相模原　ラーメン

(注2)：http://code.google.com/intl/ja/apis/ajaxsearch/

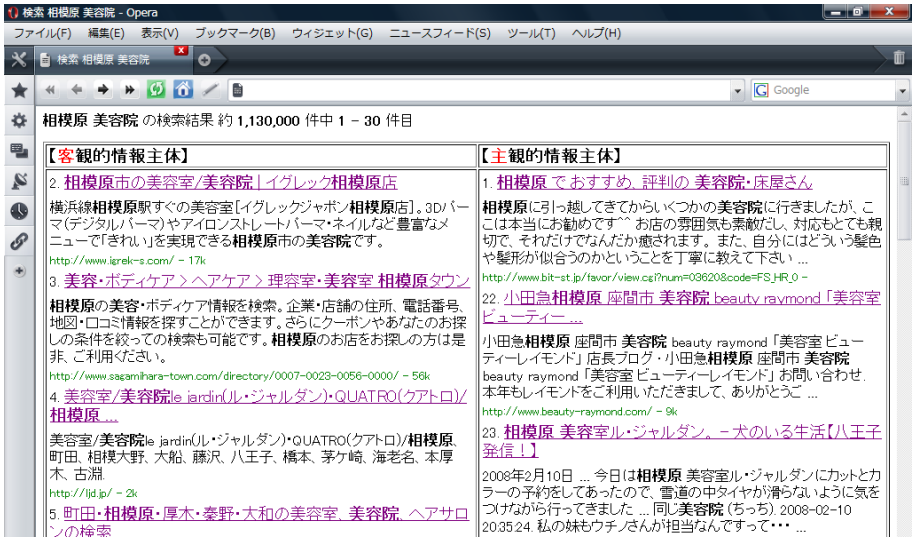

図 1　実 行 画 面

- ”自動車保険”　比較
- Android　携帯電話
- 現金　懸賞
- ”ネットブック”　価格
- 格安　”海外旅行”
- 学生　年金
- 論文　書き方
- 履歴書　ポイント
- HDD　増設
- ”ブルーレイ”　”ノートパソコン”
- TDL　アトラクション
- Ruby　イテレータ
- ”引越し”　”見積もり”
- 人気　企業
- 無料　ゲーム
- 自動車　税金

これらの各検索ワードに対して上位 50 ページ分の検索結果を取得する．次に，取得したそれぞれのページを被験者に閲覧させ，その内容に対して以下の基準で 5 段階評価を行なってもらう．

−2.　客観的情報（事実）中心
−1.　やや客観的である
0.　どちらとも言えない
+1.　やや主観的である
+2.　主観的情報（意見）中心

1 ページにつき 3 名が評価を行った．A，B，C の 3 名の各評価について，評定者間の一致度を調べるため，ケンドールの一致係数を求めた．その結果，AB 間は $W = 0.751$ ($p < 0.01$)，BC 間は $W = 0.678$ ($p < 0.01$)，CA 間は $W = 0.755$ ($p < 0.01$) であった．そこで，一致度が高かった C と A の 2 名の評定値の平均をとったものを正解データとする．なお，1,000 ページ中，バイナリファイルやページの移動などにより取得できなかった 62 ページを除いた，938 ページを実験データ（表 4）とする．

表 4　目視データ内訳

主観／客観度	ページ数	割合	
-2	300	32.0%	客観
-1.5	152	16.2%	↑
-1	147	15.7%	
-0.5	87	9.3%	
0	85	9.1%	
0.5	40	4.3%	
1	55	5.9%	
1.5	43	4.6%	↓
2	29	3.1%	主観
計	938	100.0%	

4.2　予 備 調 査

SVM では，各素性が判別にどのように寄与しているかの分析が困難である．そこで，予備実験としてあらかじめ重回帰分析を行い，主観／客観度の判定に有効な文末表現を確認しておく．

作成したプログラムを用いてウェブページ中の終助詞，助動詞，表明思考動詞の頻度(注3)をとり，目視での評価と特定の語の出現頻度に相関関係があるか否かを調査した．今回は，実験データ 938 ページを重回帰分析の訓練データとした．素性選択の条件として，終助詞，助動詞，表明思考動詞は，10 文書以上に出現している（DF 値が 10 以上の）形態素を選んだ．次に目視で評価した正解データを従属変数，ページ内に含まれる終助詞，助動詞，表明思考動詞の形態素ごとの頻度を独立変数とし，ステップワイズ法（変数投入 $p < 0.05$，変数除去 $p > 0.05$）を用いて変数を絞り込み，重回帰分析を行った．その結果選ばれたすべての変数，およびその偏回帰係数を，表 5 に示す．従属変数の値は，大きい方が主観的であることを意味する．分散分析を用いて重回帰式の検定を行った結果，自由度調整済み決定

(注3)：文献 [9] の報告を参考に TF-IDF を利用することも検討したが，今回の予備調査において出現頻度（TF）を用いた方が結果が良かったため，単純に頻度を採用することにした．

係数（重相関係数の 2 乗）は 0.510，F 値は 18.1 であり，有意水準 1%において帰無仮説が棄却されるので，ウェブページの客観，主観表現の判別に文末表現を利用することが有効であるという結果を得ることができた．

ここで，選ばれた独立変数およびその偏回帰係数を確認すると，以下に示す点については予想通りであったことから，回帰式には仮説に沿った妥当な変数が概ね選ばれていることがわかる．

- 「コメントする」「評価する」はブログやレビューで使用されているので主観．
- 「発表する＋た（発表した）」は客観性の高いニュースに頻出すると考えられるので客観．
- 「てしまう＋ます＋た（てしまいました）」「てみる＋た（てみた）」といった過去の表現は体験談によく用いられるので主観．
- 「願う」「思う」「思う＋た（思った）」は気持ちを表す際に用いられる語なので主観．
- 「です＋ね」「だろう＋か＋ね」「ている＋の」「よね」などの口語表現はブログや掲示板などコミュニティサイトでよく用いられるので主観．
- 「する＋よう（しよう）」は勧誘表現であり，広告（商品販売サイト）などでよく用いられるので客観．
- 「敬意」つまり敬語表現は公式サイトに頻出するので客観．

4.3 SVM による分類

SVM を用いて，各ページに対し主観情報主体／客観情報主体の 2 値分類を行う．4.1 節で述べた正解の評価値（$-2 \leqq target \leqq +2$）については，正の値を +1，負の値を −1 に変換し，0 は判別困難なページとして除いておく．

ツールは LIBSVM を利用する．SVM のタイプの指定は C-SVC，カーネル関数は RBF を選択し，その他の設定はデフォルトのままとした．

提案手法が未知のトピックに対応できる手法であることを示すため，訓練データとテストデータが異なる検索ワードの文書となるよう交差検定（20-fold cross validation）を行う．その結果，正解率は 83.4%であった．

なお，各検索ワードの文書を半数（約 25 ページ）ずつに分け訓練データとテストデータを作成した場合，すなわち既知のトピックの場合の正解率は 84.4%であった．

4.4 形容詞・形容動詞を用いた場合との比較実験

提案手法では，主観／客観表現の判定に文末の助動詞，終助詞，表明思考動詞の出現頻度を用い，文書の内容を表す語をあえて使用しない．このことは，あらかじめ学習されていないトピックのウェブページについても適用が可能である，というメリットにつながると考えられる．そこで，文末表現に着目してウェブページの主観／客観表現を判定することが，トピックに依存しない方法として有効であることを確認するための実験を行う．主観と深くかかわる語として形容詞および形容動詞を文末の終助詞，助動詞，表明思考動詞の代わりに素性として抽出する．

4.3 節と同様に，20-fold cross validation により正解率を求

表 5　偏回帰係数

	独立変数	偏回帰係数	
1	コメントする	0.158	主観
2	う	0.106	主観
3	れる＋ます	-0.165	客観
4	ない＋だろう＋か	0.115	主観
5	ござる＋ます	0.098	主観
6	評価する	0.130	主観
7	発表する＋た	-0.149	客観
8	てしまう＋ます＋た	0.183	主観
9	願う	0.102	主観
10	おもう	-0.146	客観
11	する＋てくれる＋ます	-0.349	客観
12	てくる	0.273	主観
13	ありませぬ	0.111	主観
14	ことができる	-0.095	客観
15	ておく＋ます	0.087	主観
16	た＋かな	-0.308	客観
17	です＋ね	0.163	主観
18	だろう＋か＋ね	0.105	主観
19	てはいけない	0.174	主観
20	てみる＋た	0.056	主観
21	する＋よう	-0.094	客観
22	語る	-0.064	客観
23	ようになる	-0.085	客観
24	ことはできる＋ます＋ない	0.075	主観
25	しかない	0.299	主観
26	ている＋の	0.145	主観
27	する＋ない	-0.108	客観
28	よね	0.341	主観
29	ている＋な	-0.997	客観
30	あう＋た	0.180	主観
31	ない＋た	-0.098	客観
32	思う＋た	0.086	主観
33	取る	-0.166	客観
34	方がよい	0.078	主観
35	ようだ＋です	0.064	主観
36	ないといけない	-0.066	客観
37	わけではない	0.085	主観
38	てくれる＋ます	0.136	主観
39	できる＋ます＋ない	0.095	主観
40	敬意	-0.072	客観
41	する＋ている＋ます	0.057	主観
42	こともできる＋ます	0.116	主観
43	れる＋ている＋た	0.162	主観
44	ている＋か	-0.142	客観
45	受動・使役・可能	-0.174	客観
46	思う	0.567	主観
47	予想する	0.057	主観
48	がちだ	0.057	主観
49	てしまう＋ます	0.096	主観
50	する＋ます	-0.074	客観
51	てしまう	-0.103	客観
	定数項		

上から変数の追加順．

めたところ，75.5%となった．

一方，各検索ワードの文書を半数ずつ訓練データとテストデータとした場合の正解率は 81.4%であった．

以上の実験により，訓練データと異なるトピックの文書に適用した場合，形容詞・形容動詞を用いるよりも文末表現を用いた方が正解率の低下は少なく抑えられることが確認できた．したがって，提案手法は訓練データのトピックに依存しにくい方法であると認められた．

5. むすび

ウェブ検索を行う際，同じキーワードを用いた場合にも，状況によってニュースサイトや公式ページなどによる客観的な情報が欲しい場合と，評判や感想など，客観的な情報ではわからない主観的な情報が欲しい場合とが存在する．そこで，ウェブ検索時に検索結果として与えられたページの内容が「個人の感想や意見を述べたページ」なのか,「客観的な概要を述べたページ」なのかを判別し，一目でわかりやすいように表示するシステムを作成した．

本研究では，主観／客観度の判定に文末表現を利用するアプローチをとった．文末の終助詞，助動詞，表明思考動詞の出現頻度に基づき，SVM によって主観／客観的文書の分類を行う評価実験を行った．その結果，文末の終助詞や助動詞，表明思考動詞の出現頻度に着目することが，そのページの主観／客観度の判断に有効であることを確認できた．また，形容詞・形容動詞を用いた場合よりも文末表現を利用した方が，訓練に利用したデータのトピックの影響を受けにくいことが認められた．

今後は，実際に分類ができるとどのくらい有用かという観点からの評価実験を行う予定である．また，主観／客観以外の分類に文末表現が適用可能かどうかについて明らかにしていきたい．例えば，情報の価値が時間の経過によって変動する（古くなると役に立たなくなる）内容のページ（ニュース，オークション，求人情報など）を検索結果の画面から特定できることは，検索効率の向上に役立つと考えられる．そこで，文末表現を利用することによる「時間の経過で価値が減退する情報」の判定に，まずは取り組みたい．

文　献

[1] 江口浩二, “テキスト処理に基づく Web 情報アクセス支援”, 第 3 回情報科学技術フォーラム（FIT2004）特別企画「WEB 知的処理の基礎」, 2004.

[2] 包直也，松本章代，鈴木雅人，“文末の表現に着目した閲覧者が受ける印象による Web 文書のクラスタリング”，情報処理学会第 69 回全国大会 講演論文集，Vol.2，pp.559–560, 2007.

[3] Finn, A., Kushmerick, N. and Smyth, B., “Genre Classification and Domain Transfer for Information Filtering”, Proceedings of ECIR-02, 24th European Colloquium on Information Retrieval Research, 2002.

[4] 乾孝司, 奥村学, “テキストを対象とした評価情報の分析に関する研究動向”, 自然言語処理, Vol.13, No.3, pp.201–241, 2006.

[5] 東照二, “歴代首相の言語力を診断する”, 研究社, 2006.

[6] 土井晃一，“文末態度表現に注目した Web Page の調査”, 情報処理学会研究報告, 1998-NL-130, Vol.1999, No.22, pp.49–56, 1999.

[7] Bruce, R., and Wiebe, J., “Recognizing subjectivity: A case study of manual tagging”, Natural Language Engineering, Vol.5, pp.187–205, 1999.

[8] 益岡隆志, 田窪行則, “基礎日本語文法—改訂版—”, くろしお出版, 1992.

[9] 岡野原大輔, 辻井潤一, “レビューに対する評価指標の自動付与”, 自然言語処理, Vol.14, No.3, pp.273–295, 2007.