感情語の色表現を用いた文書クラスタリング

中山 記男† 江口 浩二†‡ 神門 典子†‡

†総合研究大学院大学情報学専攻 〒101-8430 東京都千代田区一ツ橋 2-1-2
‡国立情報学研究所 〒101-8430 東京都千代田区一ツ橋 2-1-2
E-mail: norio@grad.nii.ac.jp, {eguchi, kando}@nii.ac.jp

あらまし 文書クラスタリングに関する従来の研究では，通常，文書の話題の類似性に着目し，文書中の内容語の出現頻度を用いる．一方，我々が提案する文書クラスタリング手法では，主に文書にあらわれた書き手の意見や見方の類似性に焦点を当てる．このような手法についてはこれまで十分に検討されてこなかったが，Web コンテンツの検索では不可決な要素である．そこで本稿では，書き手の意見や見方が利用者にとって特に重要な情報となる文書の例として Web 上の書評を取り上げ，予め作成した辞書を用いて書評のテキストから感情を示す語を抽出し，その語から連想される色を付与した感情語の色表現によって索引を作成することによって，書評に対して感情語の色表現を用いた文書クラスタリングを実現した．

キーワード 文書クラスタリング，情報抽出，感情抽出，評判情報

Document Clustering using Color representation of emotional Words

Norio Nakayama† Koji Eguchi†‡ and Noriko Kando†‡

†Department of Informatics, The Graduate University for Advanced Studies, Tokyo, 101-8430 Japan
‡National Institute of Informatics, Tokyo, 101-8430 Japan
E-mail: †norio@grad.nii.ac.jp, ‡{eguchi, kando}@nii.ac.jp

Abstract Most of the previous research on document clustering concentrated on similarity of topics discussed in documents, using the frequency of occurrence of content words. On the other hand, our document clustering method focuses on similarity of authors' perspectives or opinions that were expressed in texts. This kind of methods has not been thoroughly investigated, however, we consider that it is one of the indispensable factors for actively utilizing documents such as Web documents. In our proposed method document representation was prepared by identifying emotional expressions in texts using a predefined dictionary, and assigning a color associated by each of the expressions. The proposed method performs document clustering using the color expression of emotional words.

Keyword Document Clustering, Information Extraction, Emotion Extraction, Opinion Information

1. はじめに

本稿では，Web より収集した書評を対象として，書評に含まれる感情語を同定し，色を用いた索引付けによって文書クラスタリングを行った．

近年日常生活において Web 上の情報の重要性が増している．中でも，個人が自分の日常や購入物の感想などを記録している Web サイトの増加が著しい．そのような Web サイトの形態のひとつに Blog がある．Blog とは，専用のツールなどによって利用者がある事柄についてのコメントを記述していく Web サイトの形態である．このような形態の Web サイトが流行したことにより，様々な製品や事象に対して，非常に多くの意見が Web サイトを通して得られるようになった．この Web サイト上の意見には利用者側，開発者側ともに注目しており，実際に購入物の決定や製品開発などに利用され始めている．

ここで問題となるのが，どこに意見が書かれているか，その意見がどのような内容であるかを理解・判断する手法が必要となることである．その問題の解決策の一つとして，意見や評判，感情などの主観的情報の情報抽出と呼ばれる手法がある．情報抽出とは，そもそもテキストからあらかじめ決められたイベントや事柄に関する情報を抽出することである[1]．人名や事柄，期日などの抽出する項目をあらかじめ決めておき，そして機械学習等の方法によってその項目を抽出する技術が代表的な情報抽出である．

近年，この情報抽出の分野で評判情報検索と呼ばれる研究がある．評判情報検索とは，あるテキストから，その事柄に関する評判を検索する技術である[2]．この検索された評判から意見を抽出し，事柄自身の評価に用いる．この際対象となるものは，ある単一のモノの

評判である場合もあれば，特定の話題分野といった大きな枠組みの中での評判が対象となる場合もある．本研究ではこれら評判を処理するための基礎的な検討として，書評を処理の対象とし，「書評に含まれる書評の書き手の感情から目的に即した本を探す」といった要求を仮定して，テキストから抽出する情報を書評の書き手の感情とした．これは，書評の感情を抽出することにより，その対象となっている書籍自体がどのような内容を持つのかを知ることができるであろうと考えたためである．

本研究では感情の表現の特色として，色を用いることを提案している．感情は非常に抽象的な概念であり，自分自身の感情でさえ特定の言葉にすることはとても難しい．本稿では，感情を示す語から連想される色を用いることによって，感情を言葉に変換して扱うよりも直感的な処理が出来るのではないかと考えている．

2. 関連研究

評判情報や感情情報を扱う研究はいくつかに大別できる．1 つ目のグループは，複数の文あるいは文書において，著者が記述対象の事物に対して好意的または否定的な意見・評価をしているかどうかを判断している研究である[2~6]．ある記述対象に対してどのような語や文・パターンが好意的または否定的であるかを事前に学習し，あてはめることで文書の種類の判断を行う．この手法では対象となる事象やモノが利用者によって好意的あるいは否定的のどちらで捉えられているのか，ということを主に判断している．これに対し，本研究では好意的・否定的といった大別はせず，その文書がいったいどのような感情をもっているのかといった基準で文書分類を行う．

2 つ目のグループは，常識知や統計的手法によって，単一の語または文をあらかじめ決められた感情カテゴリーへと分類していく研究である[7]．この分類は非常に効果的であるのだが，常識知の作成コストが高いことが挙げられる．これは，常識知はある分野に特化したものである場合が多く，その場合に対象となるドメインや主題分野ごとに作成しなければならない．本研究で用いた辞書は対象領域に依存しない，汎用性を持つものを目指しているので，常識知に比べ作成コストは低い．

3 つ目のグループは，機械学習によって主観的表現や主観を表す名詞などが表れるであろうパターンを辞書として作成し，それに基づいてテキストから主観を抽出する研究である[8]．機械学習については，学習させる教師データを用意する必要があり，また対象となるドメインや主題分野ごとに学習させる作成コストが問題となる．

3. システム

本節では用いたデータや手法について説明する．特に手法の中でも索引付けに関して，最初に用いた手法と改良された手法があるので、その両方について説明する。

3.1. 適用データ

杉田ら[9]が収集したあるコミュニティに参加している Web サイトの書評に対し，本研究のシステムを適用した．文書は HTML 形式のファイルひとつで 1 文書である．文書は主に書籍等に対しての書評で構成されており，単一の書籍に対してだけではなく複数の書籍に対しての書評も含んでいる．また，書評は HTML 形式のまま収集されている．本研究ではデータに対し，単一の書籍に関する書評のみを含む文書を手作業で選別した．最終的にデータとして扱う文書数は 487 件である．

3.2. 辞書

事前に感情を示す語を登録した辞書を作成した．語は長町[10]の収集した感性ワード集と筆頭筆者が収集した語からなる 1171 件である．各語には，連想される色を肯定表現と否定表現の場合に各々 1 色ずつ対応付けている．表 1 は辞書の一部である．この際，肯定表現とは辞書に登録された形で語が現れた場合を指し，否定表現とはそれを打ち消す形で語が現れた場合を指す．例えば登録された形が「楽しい」だった場合，「楽しい」は肯定表現であり「楽しいわけではない」は否定表現である．

表 1　作成した辞書の一部

語	肯定表現のときの色	否定表現のときの色
熱い	#FF0000(明るい赤)	#FF0000(明るい青)
楽しい	#FFFFCC(明るい黄)	#333300(薄暗い緑)

3.3. 色

本研究で利用した色は，小林[11]が定義した，人間がイメージとして感じる色 130 色から類似色を排した 36 色である．また，語と色との対応付けは，36 色のカラーテーブルと語を提示し，一番妥当であると感じた 1 色を選んだ[12]．[1]

[1] 現状の辞書は筆頭筆者一名の選択で作成されている．現在複数名にて色の選択の一致度を測る実験を行い，結果を分析中である．

3.4. 索引付け

3.4.1. 索引付け A

最初に索引付けを行なった手法（以下：手法 A）について説明する．3.1 節にて選別を行った文書 487 件から，3.2 節で作成した辞書に基づいて索引を作成した．この際，索引付けは辞書に登録されている語と文書に出現する語とのマッチングにて行われ，形態素解析などは行っていない．また，索引を付した語の前後 40 バイトに「ない」という表現が出現した場合にのみ，否定表現として語を扱っている．その後，同じく 3.2 節の辞書を参照して，語に対応している色を文書ごとに索引付けした．のべ索引語数は 10287 件である．

手法 A の索引付けは誤りが多く，特に 10287 件中の総否定表現数 3818 件から 100 件を無作為に選択しその精度 p_{index} を求めたところ，0.10 であった．同様に適用文書から 10 件を無作為に選択し，本システムが索引付けすべき否定表現の発見率 d_{index} の平均も求めたところ，こちらは 0.90 であった．精度と発見率を求める式は以下の通りである．

$$p_{index} = \frac{正しく同定できた否定表現の数}{同定された否定表現の数}$$

$$d_{index} = \frac{a - |b - a|}{a}$$

a : 否定表現と判断すべき数

b : 否定表現と判断された数

加えて単純なマッチングであるため，例えば“反面白く（はんめん・しろく）”といった文に対し，辞書に登録された“面白く”で索引付けしてしまうなどの問題が見られた．この問題をふまえ，次節では形態素解析を用いた改善手法による索引付けを行った．

3.4.2. 索引付け B

手法Aを改善した手法（以下：手法B）について説明する．3.1 節にて選別を行なった文書 487 件のテキストと 3.2 節で作成した辞書に登録されている語の両方を,形態素解析[2]を用い形態素ごとに区切った．また，その際形態素解析によって品詞が助詞と判断されたものについては，任意の 1 語に対応するワイルドカードに置き換えた．形態素解析と置き換えの結果の例を表 2 に示す．共に形態素に区切られた，文書と辞書に登録された語のマッチングにより，索引付けを行った．マッチングした語の後ろX文字以内（含む，ワイルドカード）に基本形「ない」となる語が存在した場合，否定表現として索引付けを行なった．マッチングした語の後ろの文字数であるが，前節で精度を求める際に用いた適用文書 10 件について同様に精度を求めた．表 3 に示す結果から，最も精度が高い後ろ 15 語を採用した．のべ索引語数は 6371 件である．手法Aと同様に，6371 件中の総否定表現数 761 件から 100 件を無作為に選択しその精度 p_{index} を求めた．結果、精度は 0.80 であった．また，こちらも同様に発見率 d_{index} を求めたところ，0.875 であった．

表 2　形態素解析と置き換えの結果

元の文	誰の目にも触れないよう		
出現形	基本形	品詞	置き換えの有無
誰	誰	名詞	無し
の	の	助詞	有り
目	目	名詞	無し
に	に	助詞	有り
も	も	助詞	有り
触れ	触れる	動詞	無し
ない	ない	助動詞	無し
よう	よう	名詞	無し
置き換えの結果	誰 * 目 * * 触れる ない よう[3]		

表 3　後ろ何文字までを見るか

後ろ何文字までを見るか	否定表現判断の精度
10 文字	0.70
15 文字	0.80
20 文字	0.68

3.5. 文書間類似度の計算

3.4 節の各索引付けから，色数に基づき 36 次元のベクトルで以下の式のように文書間類似度を計算した．*tf* の計算式[14]に倣って *cf*:*ColorFrequency* を定義して *cf/idf* を計算し，後に各文書内の色の重み *w* をベクトル要素として内積の値を求め，最終的にコサイン類似性尺度によって各文書間の類似度を求めた．この際，索引付け A の結果から得られたものを文書間類似度 A，索引付け B の結果から得られたものを文書間類似度 B とする．

[2] 形態素解析システム茶筌[13]

[3] * はワイルドカードである

$$cf_{it} = \frac{\text{文書 } D_i \text{ における色 } C_{it} \text{ の延べ出現回数}}{\text{文書 } D_i \text{ 中の全ての色の総出現 回数}}$$

cf : 色 C_{it} の頻度　*Color Frequency*

$$idf_t = \log \frac{N}{\text{色 } C_t \text{ が出現する文書数}}$$

idf : 色 C_t が出現する文書数の文書総数 N による正規化　*InverseDocumentFrequency*

$$w_{it} = cf_{it} \times idf_t$$

w : 文書 D_i の色 C_t の重み

3.6. クラスタリング

3.5 節で求めた，文書間類似度 A・B それぞれに基づいて，文書クラスタリングを行った．文書クラスタリングのアルゴリズム[15]に関しては a)単一リンク法 b)完全リンク法 c)グループ平均法 を用いた．ここでは平均精度が最も高かった完全リンク法について結果を述べる．クラスタリング処理に関しては，各クラスタを併合する文書類似度の閾値を定め，併合するクラスタが生まれなくなった時点で処理を止めた．文書間類似度 A によるものをクラスタリング結果 A，同様に文書間類似度 B によるものをクラスタリング結果 B とする．

4. 結果と評価

4.1. 結果

図 1 は完全リンク法にて閾値を 0.5～1.0 まで 1/1000 刻みで変化させた文書クラスタリング結果 A である．図 2 は完全リンク法にて閾値を 0.5～1.0 まで 1/100 刻みで変化させた文書クラスタリング結果 B である．

4.1.1. 結果 A

a ～ e は前 15 区間での区間移動平均の近似曲線である．文書数 2 以上のクラスタをクラスタとして扱った．全クラスタ内文書数は，文書数 2 以上のクラスタ内に存在する文書数の総計であり，同様にクラスタ数は文書数 2 以上のクラスタの数，クラスタ内平均文書数は全クラスタ内文書数をクラスタ数で割ったものである．

4.1.2. 結果 B

a と c は前 3 区間での区間移動平均の近似曲線である．クラスタ，全クラスタ内文書，クラスタ数，クラスタ内平均文書の定義は結果 A と同様である．

4.2. 評価

結果 A に関して，クラスタ数に大きな変動があった 6 箇所の閾値での各クラスタ内文書に対して精度を判定した．結果 B に関しては，3 箇所で同様に精度を判定した．精度の判定は，まず筆頭筆者が各クラスタ内文書を全て読み，各クラスタの典型的な感情を理解した後に，各クラスタ内の各文書の持つ感情がそれに適しているかどうかを判断することでクラスタ精度 *p:Precision* を求めた．その閾値での全クラスタの精度を求めた後に，クラスタ全体の平均精度 $\bar{p}$ を得た．各値は下記式から得られた．表 5，表 6 は平均精度を得た閾値での各結果である．

$$p_{sk} = \frac{\text{閾値 s でのクラスタ } L_{sk} \text{ 内文書適合数}}{\text{閾値 s でのクラスタ } L_{sk} \text{ の文書数}}$$

$$\overline{p_s} = \frac{\sum_k p_{sk}}{\text{閾値 s でのクラスタ数}}$$

表 4　結果 A の精度の結果

閾値	平均精度	全クラスタ内文書数	クラスタ数
0.771	0.114	238	62
0.818	0.129	126	41
0.849	0.196	69	22
0.874	0.315	44	13
0.912	0.523	23	7
0.955	0.666	10	3

表 5　結果 B の精度の結果

閾値	平均精度 (%)	全クラスタ内文書数	クラスタ数
0.65	0.086	384	86
0.75	0.269	190	67
0.86	0.092	45	19
0.96	0.000	8	5

結果 A では，後にその結果から累乗の近似曲線を得ることで，クラスタ精度の曲線を得た．図 1 右上の式の上段は近似曲線を求める式であり，下段はその相関係数である．閾値が高くなり全クラスタ内文書が減少するほどクラスタ精度が上昇していることが図 1 に見られ，文書間類似度が高ければ高いほど正しく同じような感情を持つ文書が集まっていることが読み取れる．また，クラスタ数が最大になる閾値では分類が非常に粗く，精度は低かった．

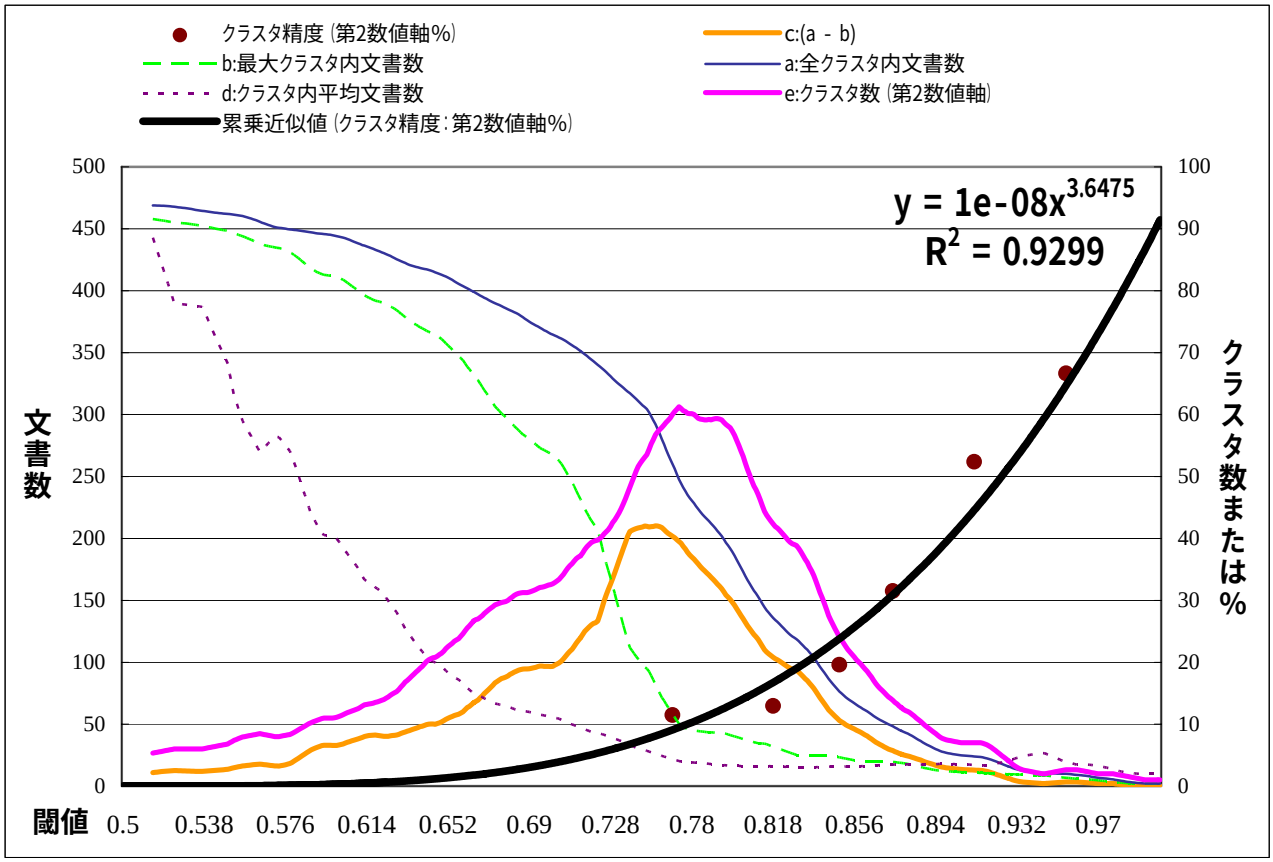

図 1　完全リンク法でのクラスタリング結果A

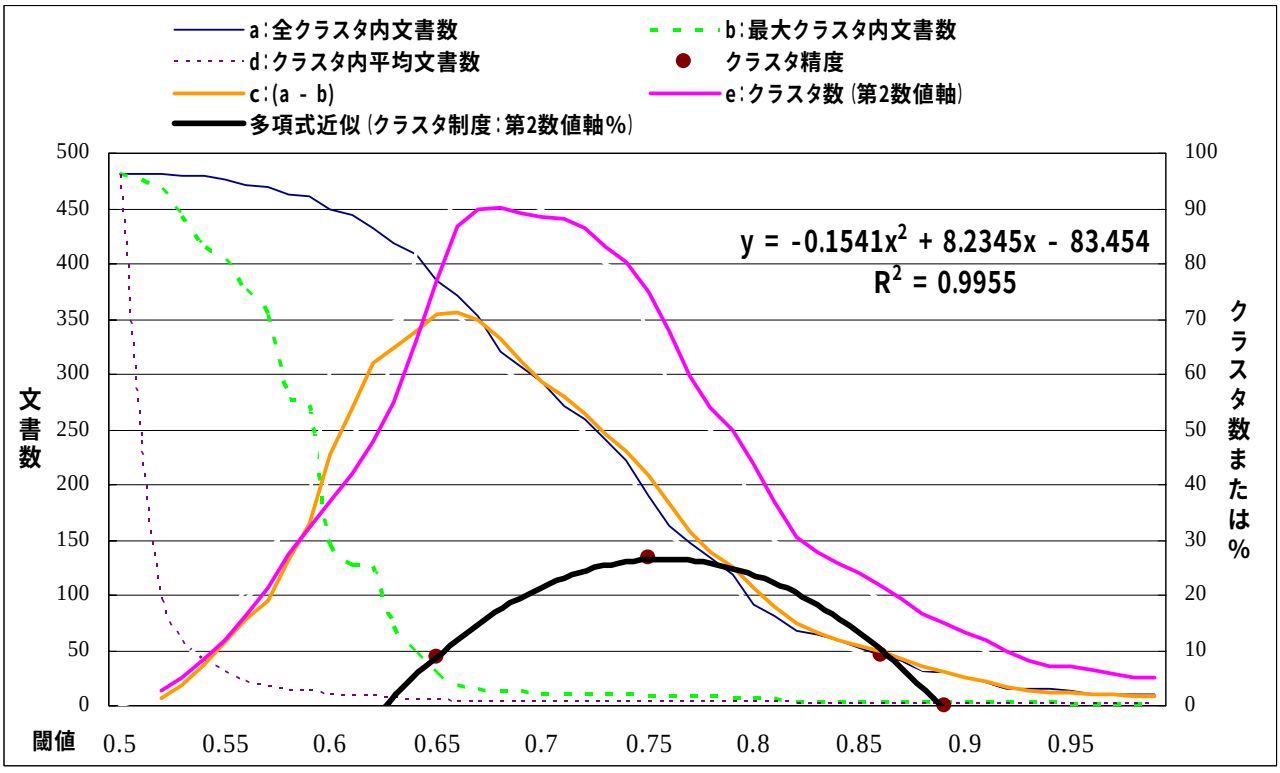

図 2　完全リンク法でのクラスタリング結果B

一方結果Bでは，閾値が低い区域に関しては結果Aよりも高い精度が得られた反面，閾値が高くなると結果Aよりも悪い結果が得られた．これは，高い閾値で同一クラスタ化した文書が含む感情が異なっていたためである．中程度の閾値での精度が向上していることから，ある程度同じような感情をもつ文書を分類する能力は向上しているといえる．両結果において，閾値が高くなるにつれてクラスタ内全文書数が減少しているが，これは1文書しか含まないクラスタが増加したことにより，全クラスタ内文書にカウントされなくなったためである．

両結果が分類した感情の例としては「戦争や争い，事件などの混沌とした重苦しさ」や,「夢や希望があふれ楽しそうだ」といったものがあった．同じような感情を持つ文書が本手法の文書クラスタリングによってひとつのクラスタに分類されたことにより，語から連想された色によるクラスタリングが有効であったといえる．精度が高かったクラスタは含まれている文書の感情が非常にはっきりしており，かつ文書から多くの語が索引付けされていた．逆に精度が低かったクラスタには，感情が弱い文書や語があまり索引付けされなかった文書が目立った．また，あるクラスタの典型的な感情を持つような文書が，別のクラスタに含まれた例も見られた．

5. おわりに

本稿は，テキスト中に表された感情という曖昧な情報に基づいて文書をクラスタリングする手法として，感情語から連想される色を用いて文書の索引付けを行う手法を提案した．話題についての従来のクラスタリングや検索と組み合わせることによる有効性も考えられる．今後は話題に基づくクラスタリングとの比較や組み合わせ，手法の精度の向上などを目指す予定である．また，連想された色のようなものを利用した場合には,どうしても個人差が議論される.これに関して，本研究で用いた辞書の妥当性や色の連想の個人差を調査する実験を行い，現在解析中である．

文　献

[1] Appelt D.E., Israel D.J., Introduction to Information Extraction Technology, A Tutorial Prepared for IJCAI-99, 1999.

[2] 立石健二, 石黒義英, 福島俊一, “インターネットからの評判情報検索, ”情報処理学会研究報告 自然言語処理, vol.2001, No.69, pp.75-82, 2001.

[3] Dave K., Lawrence S., Pennock D.M., Mining the Peanut Gallery : Opinion Extraction and Semantic Classification of Product Reviews, pp.519-528, International World Wide Web Conference, Budapest, Hungary, May 2003.

[4] Pang B., Lee L., Vaithyanathan S., Thumbs up? Sentiment Classification using Machine Learning Techniques, pp.79--86, Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP2002), Philadelphia, July 2002.

[5] 舘野昌一, “「お客様の声」に含まれるテキスト感性表現の抽出方法, ”情報処理学会研究報告 自然言語処理, vol.2003, No.4, pp.105-112, 2003.

[6] Turney P, Littman M, Measuring praise and criticism: Inference of semantic orientation from Association, vol.21, pp.315-346, ACM Transactions on Information Systems (TOIS), 2003.

[7] Liu H., Lieberman H., Selker T., A Model of Textual Affect Sensing using Real-World Knowledge, To Appear in Proceedings of IUI 2003, Miami, Florida, January 2003.

[8] Riloff E., Wiebe J., Wilson T., Learning Subjective Nouns using Extraction Pattern Bootstrapping, Proceedings of the Seventh Conference on Natural Language Learning (CoNLL-2003), Edmonton, Canada, May 2003.

[9] 杉田茂樹, 江口浩二, “目録データベースとWebコンテンツの統合的利用方式, ”情報処理学会研究報告 情報学基礎, Vol.2001, Num. 20, pp.153-158, 2001.

[10] 長町三生, 感性工学のおはなし, pp191-208, 日本規格協会, 1995.

[11] 小林重順, カラーリスト, 日本カラーデザイン研究所, 講談社, 1994.

[12] 中山記男, 大倉典子, “感性情報を用いたテキスト分類手法の検討, ”電子情報通信学会2003年総合大会, 2003.

[13] 松本裕治, 北本啓, 山下達雄, 平野善隆, 日本語形態素解析システム『茶筌』version2.0 使用説明書, 1999.

[14] Sparck Jones, K., Index term weighting, Information Storage and Retrieval, Vol9, No.11, p.619-633, 1973.

[15] Ricardo Baeza-Yates, William B.Frakes, Information Retrieval, pp.419-442, Prence Hall PTR, 1992.