LD-001

# 印象と興味に基づくユーザ選好のモデル化とニュースポータルサイトへの応用

## Modeling of User Preferences for News Portal Site System Based on Article Impressions and Interest

河合 由起子 † 熊本 忠彦 † 田中 克己 †,††
Yukiko Kawai Tadahiko Kumamoto Katsumi Tanaka

## 1. まえがき

近年，情報統合に関する研究はますます盛んになっており，複数の Web サイトにまたがって存在している同一テーマの Web コンテンツをまとめて提示するシステムも数多く提案されている [1][2]．ニュース記事を対象とする場合，大量の記事をどのように分類するかが重要であり，(1) 収集した全ての記事から出現頻度の高い単語を抽出して利用する [3]，(2) ユーザが閲覧した記事から出現頻度の高い単語を抽出して利用する [4][5][6]，(3) 収集した記事のリンク構造を解析する [7][8]，といったキーワードに着目した方式が提案されている．

筆者らは，複数のニュースサイトから収集した大量の記事をユーザの閲覧履歴に基づいて分類し，そのユーザが使い慣れているニュースサイトのトップページに写像して提示するという新しいタイプのニュースポータルサイトシステム「My Portal Viewer（MPV）」を提案している [4][5]．MPV は，ユーザの閲覧した記事から出現頻度を用いて興味語（ユーザが興味のあるキーワード）を抽出し，その興味語の有無に基づいて収集した記事をいくつかのカテゴリに分類する．このとき，新しく生成したカテゴリの名称を興味語そのものとすることにより，それぞれのカテゴリにどのような記事が含まれているかを判別しやすくしている．また，ユーザが普段利用しているニュースサイトのトップページの元々のレイアウトを維持しつつ，新しく生成した各カテゴリを再配置することにより，読みたい記事がどこにあるか効率的に探し出せるようになっている．しかしながら，その一方で，興味語の有無という分類基準だけでは，ユーザの好む記事と好まない記事をうまく分離できないことがあった．例えば，「阪神」というカテゴリには，「阪神，首位をキープ」のような記事だけでなく，「阪神，初の完封負け」のような記事も混在しうる．そこで，本論文では記事の印象という今までにない分類基準を導入し，ユーザの記事に対する選好を印象と興味の両面からモデル化するとともに，提案モデルを MPV に実装し，ユーザが共感（感情移入）しやすい記事を優先的に提示するシステム MPV Plus を提案する．一方，興味語の有無という分類基準では，一つの記事が複数のカテゴリに分類されうるため，記事構成にほとんど差異のないカテゴリが複数作成されることがあった．そこで，カテゴリ間における記事の重複割合に応じて，カテゴリを統合したり，分解したりする機能を MPV Plus に実装し，無駄なカテゴリの生成を抑制する．

† 独立行政法人情報通信研究機構けいはんな情報通信融合研究センターメディアインタラクショングループ，〒 619-0289 京都府「けいはんな学研都市」光台 3–5，{yukiko, kuma}@nict.go.jp
†† 京都大学大学院情報学研究科社会情報学専攻，〒 606–8501 京都市左京区吉田本町，tanaka@dl.kuis.kyoto-u.ac.jp

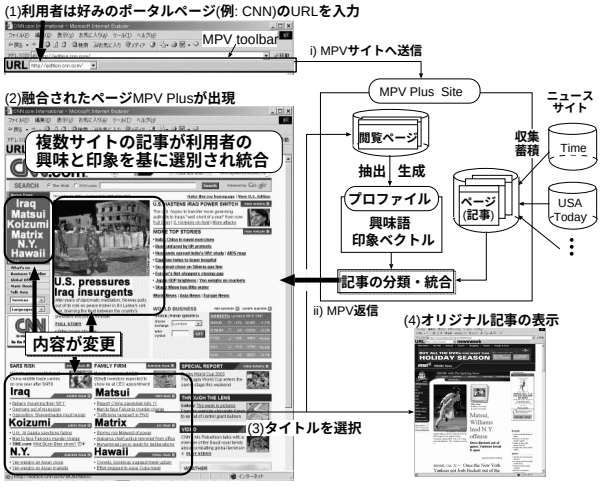



図 1: My Portal Viewer Plus の基本概念

本論文の以下の構成を示す．2. で MPV Plus の設計コンセプトを述べ，3. でユーザの閲覧履歴からユーザの記事に対する選好を動的に学習する手法を提案する．4. でユーザ選好に基づいて記事を選別し，提示する手法を，5. で記事の重複割合に応じて，カテゴリを統合したり分割したりする手法を提案する．6. でユーザの閲覧行為に対し，印象と興味に基づくユーザ選好がどのようにモデル化されるかを示し，MPV Plus の有効性を検証する．

## 2. システムの設計コンセプト

提案する印象に基づく記事の分類および統合方式を MPV に実装し，MPV Plus とする．既存の MPV は，(1) ユーザの閲覧履歴から興味語（ユーザが興味のあるキーワード）を抽出し，その興味語の有無に基づいて複数のニュースサイトから収集した記事をカテゴリに分類する，(2) 記事が分類された各カテゴリの名前を興味語そのものとする，(3) ユーザが指定した好みのニュースサイト（トップページ）のレイアウトを MPV のページとしてそのまま利用する，という特徴を有する．MPV Plus では，この 3 つの特徴を継承しつつ，印象という分類基準を導入することにより，より高精度なユーザ選好のモデル化を図る．すなわち，ユーザが閲覧した記事の印象をカテゴリごとに調べ，揺らぎの小さい印象に対してはユーザの選好ありと位置づけ，逆に揺らぎの大きい印象に対しては選好なしと位置づける．具体的には，記事の印象を 4 つの印象尺度*（「明るい ⇔ 暗い」，「承認

*この 4 つの印象尺度は，Plutchik の提案する 8 個の基本感情（joy, acceptance, fear, surprise, sadness , disgust, anger, anticipa-

⇔ 拒否」,「緩和 ⇔ 緊張」,「怒り ⇔ 恐れ」) に対する評価値 (1〜0 の実数値) として記述し，各記事から各尺度値を要素とする印象ベクトルを生成する．ユーザが閲覧した記事の印象ベクトルをカテゴリごとにまとめ，各カテゴリの要素ごとに平均値と標準偏差を求める．この標準偏差が閾値以上のとき，ユーザの選好なしとし，対応する平均値を *don't care* 項として扱う．逆に閾値未満のとき，ユーザの選好ありとし，対応する平均値をそのまま利用する．この操作の結果生成される各平均値を要素とする平均印象ベクトルとそのカテゴリの興味語の組み合わせを，本論文では，ニュース記事に対するユーザの選好と定義し，ユーザプロファイルに記録する．収集された記事は，このユーザプロファイルに基づいて各カテゴリに分類される．

図 1 に MPV Plus の基本概念を示す．ユーザが好みのニュースサイトのトップページを MPV Plus のレイアウトとして指定すると，MPV Plus は，そのページの HTML ドキュメントを取得後，置換対象となるカテゴリ名を同定し，ユーザの閲覧履歴に基づいて抽出された興味語と置換する．複数のニュースサイトから収集した記事は，興味語の有無に基づいて各カテゴリに分類され，興味語との関連度に応じて取捨選択される．さらに，カテゴリごとの平均印象ベクトルとのコサイン距離に応じて，記事の優先順位を決定する．ユーザがある記事を閲覧すると，閲覧履歴が更新され，興味語が再抽出されるとともに，平均印象ベクトルも再計算される．また，記事の重複割合が高いカテゴリは統合され，一つのカテゴリとして扱われる．但し，平均印象ベクトルは興味語ごとに管理され，コサイン距離も興味語ごとに計算される．

## 3. ニュース記事に対するユーザ選好の学習

本章では，ユーザの閲覧履歴を基に興味語の抽出と印象ベクトルの生成を行い，ユーザの記事に対する選好を興味と印象の両面からモデル化する手法を提案する．

### 3.1 閲覧履歴に基づく興味語の抽出

本節では，ユーザの閲覧履歴を基に興味語を抽出するための手順を示す．

1. 複数のニュースサイトから収集したページ $P_1$〜$P_n$ のメタデータとなるタイトルと概要を取得．
2. メタデータを形態素解析し，$P_i$ に出現する単語 $j$ (固有名詞，一般名詞) の重み $w_{ij}$ を $tf \cdot idf$ で定義し，以下の式より算出．

$$w_{ij} = \frac{log(P_i\text{中の単語 } j \text{ の出現頻度} + 1)}{log(P_i\text{中の異なり単語数})} \times log\frac{\text{記事の総数 } n}{\text{単語 } j \text{ が出現する記事の総数}}$$

3. ユーザが $m$ 個のページを閲覧したとき，閲覧したページ全体での単語 $j$ の重み $W_j = \sum_{i=1}^{m} w_{ij}$ を算出．

tion) [9] をベースに，ニュース記事に対するユーザ選好のモデル化という観点から設計された．

表 1: MPV Plus 用に構成された印象尺度

| 印象尺度 | 印象語 |
|---|---|
| 1. 明るい⇔暗い | 明るい，うれしい，楽しい<br>暗い，悲しい，苦しい |
| 2. 承認 ⇔ 拒否 | 承認 (する)，愛好 (する)，好きだ<br>拒否 (する)，嫌悪 (する)，嫌いだ |
| 3. 緩和 ⇔ 緊張 | ゆったり (する)，のんびり (する)，ゆっくり (する)<br>緊張 (する)，緊急 (だ) |
| 4. 怒り ⇔ 恐れ | 怒る，怒号<br>恐れる，怖い，恐怖 |

4. $W_j$ 値が閾値以上となる単語 $j$ を興味語として抽出．

興味語は $W_j$ 値の大きい順に元々のカテゴリの先頭のキーワードから順に置換される．このとき，オリジナルページのレイアウトを変えずに置換することから，置換可能な興味語の数が制限される．そこで，MPV plus では「others」という名前のカテゴリを作成し，置換されなかった興味語とオリジナルページのカテゴリを格納する．ユーザは others を選択すると，格納された残りの興味語とその興味語の関連記事を別ページにて閲覧できる．

### 3.2 ニュース記事の印象ベクトルの生成

記事の印象ベクトルは，以下の手順で生成される．

1. 興味語抽出の手順 1 で取得した情報からページ $P_i$ に出現する単語 $j$ (サ変名詞，形容詞，動詞) を抽出．
2. 印象辞書 (後述) を用いて単語 $j$ の印象尺度 $e(e = 1,2,3,4)$ における尺度値 $S_{je}$ と重み $M_{je}$ を取得．
3. $P_i$ の印象尺度 $e$ における尺度値 $O_{ie}$ を以下の式より算出．

$$\sum^{j} S_{je} \times |2S_{je} - 1| \times M_{je} \Big/ \sum^{j} |2S_{je} - 1| \times M_{je}$$

但し，$|2S_{je} - 1|$ は，$S_{je}$ の値に依存する傾斜配分であり，印象尺度と関係のない一般的な単語 (印象尺度値は 0.5 に近い値をとる) が $O_{ie}$ 式の平均操作に及ぼす悪影響を軽減するために導入．

4. ページ $P_i$ の印象ベクトルを $v_i = (O_{i1}, O_{i2}, O_{i3}, O_{i4})$ と定義し，生成する．

手順 2 で用いた印象辞書は，文献 [10] の手法を用いて，日経新聞全文記事データベース [11] (1990 年版〜2001 年版，200 万強の記事) から自動構築された．文献 [10] では，印象尺度を構成する印象語は 1 語に限られていたが，これを複数語に拡張し，ある単語 $j$ が 2 つの印象語群のどちらとより共起しやすいかを定式化した．この共起のしやすさを印象の強さあるいは程度と捉え，印象尺度左側の印象語群と共起しやすい場合は，$O_{ie}$ 値は 1 に近い値をとり，右側の印象語群と共起しやすい場合は，0 に近い値をとるように設計された．表 1 に今回採用された印象尺度と各印象尺度を構成する印象語を示し，表 2 に印象辞書の一部を示す．表中，各見出し語に対し，上段が尺度値を表し，下段が重みを表す．

表 2: 印象辞書に登録されているエントリー

| 見出し語 | 尺度 1 | 尺度 2 | 尺度 3 | 尺度 4 |
|---|---|---|---|---|
| 蘇生（サ変名詞） | 0.91 | 0.521 | 0.429 | 0.000 |
| | 0.464 | 0.582 | 0.732 | 0.328 |
| 出国（サ変名詞） | 0.596 | 0.209 | 0.762 | 0.201 |
| | 0.975 | 1.049 | 1.065 | 0.701 |
| 死亡（サ変名詞） | 0.28 | 0.358 | 0.260 | 0.364 |
| | 1.132 | 1.272 | 1.306 | 1.112 |
| 脱線（サ変名詞） | 0.31 | 0.546 | 0.403 | 0.291 |
| | 0.514 | 0.603 | 0.737 | 0.549 |
| 出かける（動詞） | 0.639 | 0.754 | 0.887 | 0.590 |
| | 1.430 | 1.394 | 1.304 | 1.114 |
| 挑戦する（動詞） | 0.618 | 0.687 | 0.752 | 0.500 |
| | 1.399 | 1.330 | 1.251 | 1.090 |
| 衝突する（動詞） | 0.344 | 0.353 | 0.315 | 0.529 |
| | 1.004 | 1.016 | 1.099 | 0.948 |
| 懸念する（動詞） | 0.373 | 0.319 | 0.246 | 0.293 |
| | 1.447 | 1.440 | 1.521 | 1.275 |
| 豊富だ（形容詞） | 0.597 | 0.676 | 0.761 | 0.466 |
| | 1.416 | 1.352 | 1.299 | 1.109 |
| 最適だ（形容詞） | 0.622 | 0.671 | 0.743 | 0.192 |
| | 1.185 | 1.164 | 1.145 | 0.899 |
| 困難だ（形容詞） | 0.318 | 0.305 | 0.307 | 0.317 |
| | 1.451 | 1.526 | 1.528 | 1.274 |
| 不明だ（形容詞） | 0.359 | 0.367 | 0.336 | 0.359 |
| | 1.241 | 1.337 | 1.364 | 1.18 |

### 3.3 興味語と印象ベクトルに基づくユーザ選好の学習

ユーザが閲覧した記事から抽出した各興味語に対し，ペアとなる平均印象ベクトルを求める．以下にその手順を示す．

1. 興味語 $j$ に分類された記事のうち，ユーザが閲覧した記事を $R_1, R_2, \cdots, R_m$ とし，各記事 $R_i$ の印象ベクトルを $v_i = (v_{i1}, v_{i2}, v_{i3}, v_{i4})$ とする．
2. $v_i (i = 1, 2, \cdots, m)$ に対し，各要素 $e (e = 1, 2, 3, 4)$ の平均値 $\mu_{je}$ と標準偏差 $\sigma_{je}$ を算出．

$$\mu_{je} = \sum_{i=1}^{m} v_{ie} \Big/ m \qquad (1)$$

$$\sigma_{je} = \sqrt{\sum_{i=1}^{m} (v_{ie} - \mu_{je})^2 \Big/ (m-1)} \qquad (2)$$

3. $\sigma_{je} < Threshold$ を満たすとき，揺らぎは小さいと考え，$\mu_{je}$ を興味語 $j$ に対応する平均印象ベクトルの第 $e$ 要素とし，$\sigma_{je} \geq Threshold$ の場合は揺らぎは大きいと考え，*don't care* 項を第 $e$ 要素とする．

## 4. ユーザ選好に基づく記事の分類と提示

各自のユーザプロファイル（興味語と平均印象ベクトルのペア集合）を用いて，収集した記事を分類し，また記事の取捨選択を行う．

1. 興味語 $j$ と共に出現する単語 $k$ を抽出し，単語 $j, k$ の共起度 $c_{jk}$ をすべての記事を対象に，（単語 $j$ と $k$ の共起頻度 +1）／（単語 $j$ の出現頻度 + 単語 $k$ の出現頻度）として算出．
2. ユーザが閲覧した $m$ 個のページから興味語 $j$ を含む記事を選別．
3. 興味語 $j$ に分類された記事 $P_i$ の各単語の共起度と全ページの共起度のコサイン距離を算出し，値が閾値以上の記事を選択．

表 3: 実験に利用したニュースサイト

| ドメイン名 | ポータルページの URL | 記事ページ数 |
|---|---|---|
| 朝日新聞 | http://www.asahi.com | 60 |
| 読売新聞 | http://www.yomiuri.co.jp | 21 |
| 毎日新聞 | http://www.mainichi-msn.co.jp | 42 |
| 日経新聞 | http://www.nikkei.co.jp | 30 |
| 産経新聞 | http://www.sankei.co.jp | 53 |
| Yahoo!News | http://headlines.yahoo.co.jp | 49 |

4. 記事 $P_i$ の印象ベクトル $v_i = (v_{i1}, v_{i2}, v_{i3}, v_{i4})$ と 3.3 節の手順で生成された興味語 $j$ に対する平均印象ベクトル $\mu_j = (\mu_{j1}, \mu_{j2}, \mu_{j3}, \mu_{j4})$ とのコサイン距離 $D_i$ を算出．

$$D_i = \sum_{e=1}^{4} (v_{ie} \cdot \mu_{je}) \Big/ \sqrt{\sum_{e=1}^{4} v_{ie}^2 \times \sum_{e=1}^{4} \mu_{je}^2} \qquad (3)$$

但し，$\sigma_{je} \geq Threshold$ のとき，$\mu_{je}$ は *don't care* 項なので，計算から除外．

5. $D_i$ の大きい順に表示可能な記事 $P_i$ を表示．

## 5. 興味語の統合

本システムでは，カテゴリ間の記事の重複を軽減することを目的として，カテゴリを統合する．すなわち，カテゴリ間の記事を比較し，同一記事が多い場合，その 2 つのカテゴリを統合し，新たなカテゴリを作成する．以下，その手順を示す．

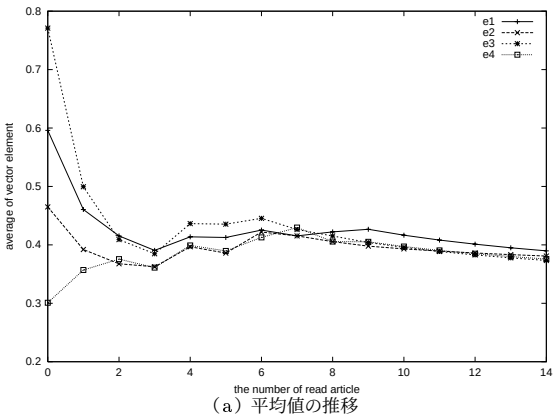
1. カテゴリ $i$ の記事集合 $I$ とカテゴリ $k$ の記事集合 $K$ の積集合 $I \cap K$ と和集合 $I \cup K$ を求め，その要素数 $|I \cap K|$，$|I \cup K|$ を算出．
2. $L = |I \cap K| \ / \ |I \cup K| > Threshold_1$ のとき，カテゴリを統合し，新たなカテゴリを生成．$L < Threshold_2$ のとき，カテゴリを分割し，元の興味語を名前とするカテゴリを作成．
3. 以上の操作を再帰的に行う．

統合された新たなカテゴリの名前は，元々のカテゴリ名を連続したものとする．例えば，カテゴリ名が $j$ と $k$ の場合は，「$j$ ／ $k$」が新たなカテゴリ名として提示される．統合後も元の興味語をそのまま提示することで，ユーザは分類された記事の具体的なトピックを容易に推測できる．

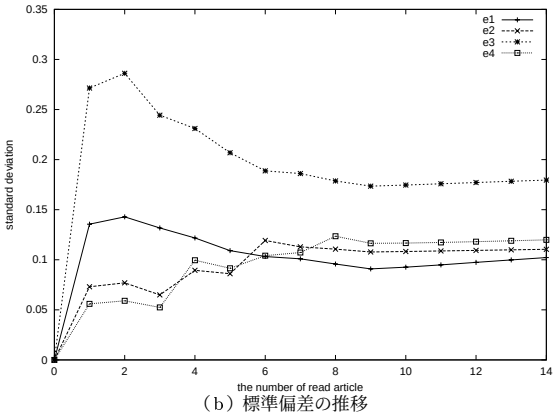
## 6. 評価実験

MPV Plus を Windows OS 上に実装し，ユーザの閲覧行為による，各興味語に対する印象ベクトルの揺らぎ（標準偏差）を評価した．なお，主メモリ 2GB のラップトップ PC を MPV サーバとし，システムの開発には Perl，Microsoft Visual Studio .Net C#，Mecab[12] を利用した．

表 3 に実験に用いた 6 つのニュースサイトを示す．記事は，2005 年 4 月 28 日の 8:50 から 9:20 までに各サイ



（a）平均値の推移



（b）標準偏差の推移

図 2: ユーザの閲覧による印象ベクトルの平均値と標準偏差の推移（興味語「中国」の場合）

トに蓄積されている 255 個を収集したものである．各記事から平均 11 個の単語が抽出され，そのうち少なくとも 1 単語以上が興味語となるよう，記事の単語抽出の閾値を 0.1，興味語の閾値を 0.06 と設定した．

ユーザの閲覧行為を場合分けすると，主に（1）特定のトピック（例えば，尼崎脱線事件）を選択的に閲覧する,（2）複数の特定トピックをランダムに閲覧する,（3）不特定のトピックを閲覧する，の 3 通りが考えられる．しかしながら，MPV Plus は，ユーザの閲覧した記事から興味語を抽出し，カテゴリを作成するため，印象という分類基準はカテゴリごとに適用される．そこで，本実験では,（1）に対応するユーザを想定し，平均印象ベクトルにおける各要素の平均値ならびに標準偏差がユーザの閲覧行為によりどのように変化するかを調べた．具体的には,「反日運動に対する賛否両論」の記事をランダムに閲覧し，徐々に否定的な書き振りの記事に移行していくという閲覧行為について考察した．図 2 にその結果を示す．

まず，図 2(b) の標準偏差の推移を見てみると閲覧開始直後は，様々な印象の記事を閲覧したので，印象尺度 1 の「明るい⇔暗い」と印象尺度 3 の「緩和⇔緊張」に関する標準偏差が高めとなっているが，閲覧する記事が印象面において偏りを見せるにつれて，その値は徐々に減少している．このときの，印象尺度 1 の平均値は 0.4 弱となっており（図 2(a)），暗めの印象の記事が選択されるようにユーザプロファイルが更新されていることが確認された．以上より，MPV Plus が印象面でも記事の取捨選択を正しく行っていることがわかった．

## 7. まとめ

本論文では，複数のニュースサイトから収集した記事をユーザプロファイル（興味語と平均印象ベクトルのペア集合）に基づいて分類し，ユーザの興味と好みに合わせて記事を優先的に提示可能な MPV Plus を提案した．MPV Plus は，ユーザの閲覧履歴から興味語（ユーザの興味のある語）と平均印象ベクトル（ユーザの好む印象）を決定し，収集した記事のうち，この分類基準に合致する記事を優先的に提示する．

今後の課題として，印象辞書の高精度化が挙げられる．現時点では各単語の印象を尺度値と重みによって示しているが,「警報解除」など単語の組み合わせ方によっても，その印象は変わる．また，同じ単語であっても，ユーザが抱く印象には個人差があるので，これを解消するためにテキストの印象を決定する要因を調べることも必要と考えている．

## 参考文献

[1] Kathleen R. McKeown, Regina Barzilay, David Evans, Vasileios Hatzivassiloglou, Judith L. Klavans, Ani Nenkova, Carl Sable, Barry Schiffman, and Sergey Sigelman. Tracking and Summarizing News on a Daily Basis with Columbia's Newsblaster. In *Proceedings of the Human Language Technology Conference, 2002*, San Diego, USA, 2002.

[2] 渡邊拓也, 大野成義, 太田学, 片山薫, 石川博. 差異に注目した複数文書融合手法. データ工学ワークショップ (DEWS)2005, 2005.

[3] NewzCrawler. `http://www.newzcrawler.com/`.

[4] Yukiko Kawai, Daisuke Kanjo, and Katsumi Tanaka. My Portal Viewer for Content Fusion based on User's Preferences. In *Proceedings of IEEE International Conference on Multimedia & Expo(ICME)*, 2004.

[5] 河合由起子, 官上大輔, 田中克己. 個人の選好に基づく複数ニュースサイトの記事収集・閲覧システム. 情報処理学会論文誌：データベース, Vol. 46, No. SIG8(TOD26), pp. 14–25, 2005.

[6] Newsbot. `http://uk.newsbot.msn.com`.

[7] GoogleNews. `http://news.google.co.jp`.

[8] Deng Cai, Shipeng Yu, Ji-Rong Wen, and Wei-Ying Ma. Extracting Content Structure for Web Pages Based on Visual Representation. *Springer, Lecture Notes in Computer Science (APWeb2003)*, Vol. 2642, pp. 406–417, 2003.

[9] Robert Plutchik and Henry Kellerman (eds). *Emotion: Theory, Research, and Experience*, Vol. 1, pp. 3–33. Academic Press Inc., 1980.

[10] 熊本忠彦, 田中克己. Web ニュース記事を対象とする喜怒哀楽抽出システム. インタラクション 2005, Vol. 2005, No. 4(A-103), pp. 25–26, 2005.

[11] 日本経済新聞社. 日経全文記事データベース DVD-ROM 版. 1990–1995 年版，1996–2000 年版，2001 年版.

[12] MeCab. `http://chasen.org/~taku/software/mecab/`.