LD_004

Fair News Reader:バランス感覚のある記事推薦方式の提案

Fair News Reader: Recommending News Articles with Different Sentiment Aspects for a Topic

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してユーザへ推薦するかが重要であり，(1) ユーザごとに閲覧した記事から出現頻度の高い単語を抽出して利用 [3][4]，(2) 協調フィルタリングによる他ユーザの閲覧状態の利用 [5]，といった閲覧履歴に着目した分類および推薦方式が提案されている．

筆者らは，ユーザの閲覧履歴に基づいてそのユーザの興味と印象をモデル化し，複数のニュースサイトから収集した大量の記事を分類し，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 Plus （MPV Plus)」を提案している [6]．MPV Plus 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分類する．このとき，興味語ごとに新しいカテゴリを生成し，そ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，興味語（カテゴリ）に分類された各記事の印象（記事内容から人が感じる印象．例えば，「明るい」，「暗い」，「楽しい」，「悲しい」など）をベクトル形式（印象ベクトルと呼ぶ）で記述し，ユーザが閲覧した記事（群）の重心印象ベクトルと未読記事の印象ベクトルとの距離を測ることにより，興味語と記事印象の両面においてユーザの選好に合った記事の推薦を可能にしている．MPV Plus の新規性は，記事の印象という今までにない選択基準を用いて記事を分類した点にあった．

しかしながら，その一方で，ユーザの好みの印象の記事だけを推薦するという現行の方式では，似たような印象の記事ばかりを閲覧する機会が多くなり，その結果，アンバランスな閲覧しかできず，視野（閲覧範囲）を狭めてしまう可能性がある．例えば，任意のユーザの閲覧履歴から興味語として「イラク」が抽出された場合，「イラク」という単語を含む記事がカテゴリ「イラク」に分類され，各記事の印象ベクトルが算出される．その際，

† 京都産業大学理学部コンピュータ科学科，〒 603-8555 京都市北区上賀茂本山，kawai@cc.kyoto-su.ac.jp
†† 独立行政法人情報通信研究機構知識創成コミュニケーション研究センター自然言語グループ〒 619-0289 京都府「けいはんな学研都市」光台 3-5，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

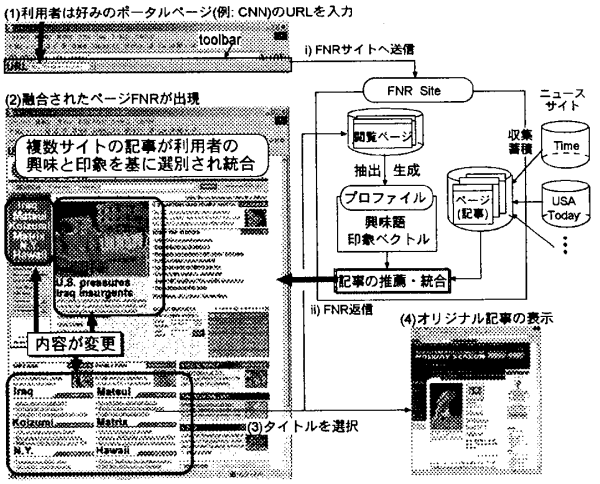


図 1: Fair News Reader の基本概念

ユーザの閲覧履歴が批判的な内容の記事に偏っていた場合，MPV Plus が推薦する記事も批判的な記事に偏り，異なった印象を持つ記事（肯定的な記事や明るい記事など）を閲覧する機会そのものが奪われてしまっていた．

そこで，本論文では興味語毎にユーザが閲覧する度に変化する閲覧特性（重心印象ベクトル）の推移を観測するとともに，その閲覧特性と未読記事の印象ベクトルとを比較することで，バランスの良い情報閲覧を可能にする記事推薦方式（Fair News Reader:FNR）を提案する．

2. システムの設計コンセプト

バランスの良い情報閲覧を可能にする記事推薦方式を MPV Plus をベースに実装する．これまで筆者らが開発してきた MPV Plus 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）を抽出する，(2) 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し，さらに興味語との関連度に応じて取捨選択する，(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）のレイアウトを統合結果表示ページとしてそのまま利用する，(4) ユーザの閲覧履歴から興味語に対するユーザの重心印象ベクトルを算出し，記事の印象ベクトルとのコサイン距離（類似度）が大きい記事を優先的に提示する，という特徴を有する．

図 1 に FNR 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 FNR のレイアウトとして指定すると，FNR 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，カテゴリごとの重心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，重心印象ベクトルも再計算される．

FNRでは，MPV Plusの特徴を継承しつつ，ユーザが閲覧した記事の印象面おける分散状況と未読記事の印象とを考慮して，バランスの良い情報推薦を図っている．すなわち，ユーザが閲覧した記事の印象を興味語（カテゴリ）ごとに調べ，揺らぎの小さい印象に対してはアンバランスな閲覧傾向にあると位置づけ，逆に揺らぎの大きい印象に対してはバランスの良い閲覧傾向にあると位置づける．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 ⇔ 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，グッドバランスとし，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，アンバランスとし，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重心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印象バランスと定義し，ユーザプロファイルに記録する．また，収集された記事の印象ベクトルも算出し，印象プロファイルとして記録する．そして，各カテゴリごとにユーザの重心印象ベクトルと未読記事の印象ベクトルを比較し，重心印象ベクトルの標準偏差がもっとも大きくなる記事を推薦する．

3. ニュース記事に対する閲覧特性の学習

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閲覧特性を興味語と記事印象の両面からモデル化する手法について述べる．

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

*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, anticipation）[7] をベースに，ニュース記事に対する閲覧特性のモデル化という観点から設計された．

表1: FNR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
	暗い，悲しい，苦しい
2. 承認 ⇔ 拒否	承認（する），愛好（する），好きだ
	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）
	緊張（する），緊急（だ）
4. 怒り ⇔ 恐れ	怒る，怒号
	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．

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[8]の手法を用いて，日経新聞全文記事データベース[9]（1990年版〜2001年版，200万強の記事）から自動構築された．文献[8]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．

3.3 興味語と印象ベクトルに基づくユーザ閲覧特性の学習

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

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
死亡（サ変名詞）	0.28	0.358	0.260	0.364
	1.132	1.272	1.306	1.112
挑戦する（動詞）	0.618	0.687	0.752	0.500
	1.399	1.330	1.251	1.090
懸念する（動詞）	0.373	0.319	0.246	0.293
	1.447	1.440	1.521	1.275
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

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 $\mu_{je} = \sum_{i=1}^{m} v_{ie}/m$ と標準偏差 $\sigma_{je} = \sqrt{\sum_{i=1}^{m}(v_{ie} - \mu_{je})^2/(m-1)}$ を算出．

3. $\sigma_{je} < Threshold$ を満たすとき，揺らぎは小さいと考え，μ_{je} を興味語 j に対応する重心印象ベクトルの第 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. 閲覧特性に基づく記事の分類と推薦

各自のユーザプロファイル（興味語と重心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，さらに推薦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. 3.3 節のステップ 2. で得られた標準偏差 σ_{je} の揺らぎが閾値 T_2 より小さい場合は，下記を処理した後ステップ 4 を実行．閾値 T_2 より大きい場合は下記を処理せず，ステップ 4 を実行．

   (a) 興味語 j を含む記事集合の重心印象ベクトル $vc_j = (vc_{j1}, vc_{j2}, vc_{j3}, vc_{j4})$ を算出．

   (b) vc_j と 3.3 節の手順で生成された興味語 j に対する重心印象ベクトル $\mu_j = (\mu_{j1}, \mu_{j2}, \mu_{j3}, \mu_{j4})$ との差 $diff\_vc_j = vc_j - \mu_j$ を算出．

   (c) 差分ベクトル $diff\_vc_j$ を k 倍したベクトルと各記事の印象ベクトル $v_i = (v_{i1}, v_{i2}, v_{i3}, v_{i4})$ との最小距離を算出し，k が大きい順に順位付けする．

4. 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．

   (a) 記事 P_i の印象ベクトル v_i と 3.3 節の手順で生成された興味語 j に対する重心印象ベクトル μ_j とのコサイン距離 D_i を算出．

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

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

   (b) D_i の大きい順に順位付けする．この時，μ_{je} が全て *don't care* 項の場合には，算出した共起度のコサイン距離の大きい順に順位付けする．

5. ステップ（3）-（c）で既に記事に順位付けされている場合は，表示（推薦）可能な個数の r 割となる記事数を（3）-（c）の記事とし，残りを（4）-（b）の記事とする．（3）-（c）で順位付けされていない場合は，（4）-（b）ので順位付けされた記事を表示可能な分だけ提示する．

5. 性能評価

FNR 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）と推薦記事について検討した．

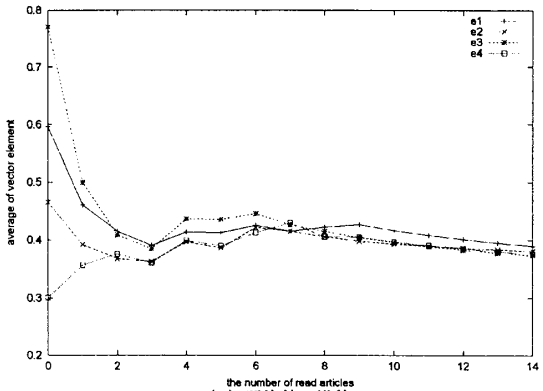
本実験では 6 つの日本語のニュースサイトから記事を収集した．記事は，2005 年 4 月 28 日の 8:50 から 9:20 までに各サイトに蓄積されている 255 個を収集したものである．各記事から平均 11 個の単語が抽出され，そのうち少なくとも 1 単語以上が興味語となるよう，記事の単語抽出の閾値を 0.1，興味語の閾値を 0.06 と設定した．

まず，ユーザの閲覧行為による重心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どのように変化したかについて考察する．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．

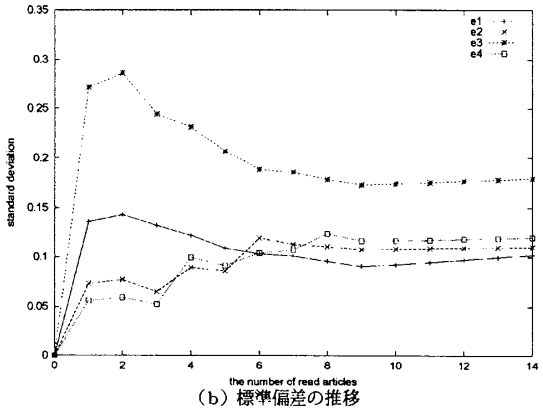
まず，図 2（b）の標準偏差の推移を見てみると閲覧開始直後は，ある程度，印象の異なった記事を閲覧したので，印象尺度 1「明るい⇔暗い」と印象尺度 3「緩和⇔緊張」に関する標準偏差が高めとなっているが，閲覧する記事が印象面において偏りを見せるにつれて，その値は徐々に減少している．このときの，印象尺度 1 の平均値は 0.4 弱となっており（図 2（a）），ユーザの閲覧対象が暗めの印象の記事に偏っているとユーザプロファイルに記録されているのが確認された．

次に，図 3 に，興味語「中国」を含む記事の印象尺度値の分散状況を，印象尺度 1「明るい⇔暗い」，印象尺度 3「緩和⇔緊張」（それぞれ $e1$，$e3$ と表記）とした 2 次元平面上にプロットしたものを示す．

図 3 より，まず，興味語「中国」に分類された記事そのものの印象は，緊張感があり暗めの記事の内容であったことが容易に確認できた．しかしながら，その中でも，明るくゆったりとした内容の記事も存在することが分かった．次に，ユーザがそれらの記事から 14 個の記事を閲覧した場合に，算出されたユーザの重心印象ベク



（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重心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）

トルと標準偏差について検討する．図 3 より，ユーザは暗く緊張感の高い記事を好んで閲覧する傾向にあり，14 回目以降も偏った印象の記事を中心に閲覧してしまう可能性が高い．この結果，FNR では，アンバランスな閲覧状況に陥ったものと判断し，重心印象ベクトルの存在する第 3 象限と異なり最も遠い第 1 象限にある記事を優先的に提示する．これにより，暗めで緊張感のある内容の記事ばかりを閲覧していたユーザに対し，明るく穏やかな内容の記事も推薦できるようになり，バランスの良い情報閲覧を可能にすることができたと言える．

6. まとめ

本論文では，ユーザの閲覧履歴から興味語（ユーザの興味のある語）を抽出し，その興味語に対するユーザの重心印象ベクトルの推移と未読記事の印象ベクトルの分散状況を分析・比較することで，ユーザがグッドバランスで情報を閲覧できる記事推薦方式を提案し新たなニュースリーダとして FNR を検討した．FNR は，ユーザがアンバランスな閲覧状態にあると判断すると，ユーザの過去の閲覧記事とは異なった印象をもつ記事を優先的に推薦し，バランスのとれた情報推薦を可能にしていることが確認できた．

但し，FNR は「政治」「経済」「社会」など多様な視点からの観察・分析を必要とする分野（ジャンル）には有効であるが，「芸能」や「スポーツ」といった嗜好性の

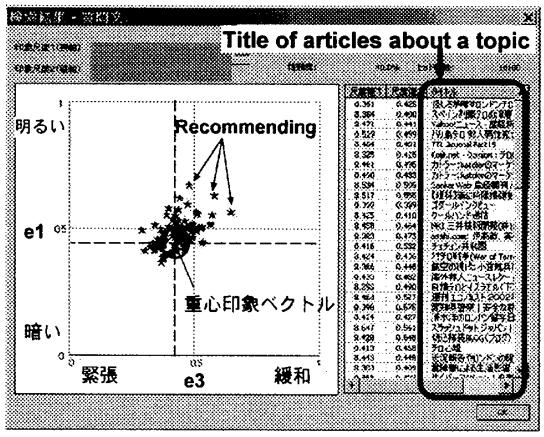


図 3: ユーザの重心印象ベクトルと相対する記事の推薦

高い分野にはあまり適さないと考えられる．今後は，長期的な実験を通した FNR の有効性の検討が必要であると考えている．

謝辞

本研究の一部は，平成 18 年度科学研究費補助金若手研究 (B)(課題番号：18700110) ならびに特定領域研究 (課題番号：18049075) の助成によるものです．ここに記して謝意を表すものとします．

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*, 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] 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.

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

[5] 森幹彦, 山田誠二. ブックマークエージェント: ブックマークの共有による情報検索の支援. 電子情報通信学会論文誌, Vol. J83-D-I, No. 5, pp. 487–494, 2000.

[6] 河合由起子, 熊本忠彦, 田中克己. 印象と興味に基づくユーザ選好のモデル化手法の提案とニュースサイトへの応用. 日本知能情報ファジィ学会誌, Vol. 18, No. 2, pp. 59–69, 2006.

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

[8] 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.

[9] 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.