

平成21年3月31日現在

研究種目：若手研究（B）
 研究期間：2006～2008
 課題番号：18700110
 研究課題名（和文）個人適応型バーチャルポータルサイト実現のためのユーザ利用特性抽出技術に関する研究
 研究課題名（英文）A Development of Personal Virtual Portal Web Site based on User Characteristic
 研究代表者
 河合 由起子（KAWAI YUKIKO）
 京都産業大学・コンピュータ理工学部・講師
 研究者番号：90399543

研究成果の概要：本研究では、Webニュース記事の特性ならびにユーザのもつ複数の利用特性の抽出を実現し、ユーザの特徴に基づいた柔軟な記事カテゴリの生成と分類、およびユーザの使い易いインタフェースを提供し、個人の利用特性に適応した仮想的なニュースのポータルサイトの構築を行った。

交付額

(金額単位：円)

	直接経費	間接経費	合計
2006年度	1,300,000	0	1,300,000
2007年度	900,000	0	900,000
2008年度	1,200,000	360,000	1,560,000
年度			
年度			
総計	3,400,000	360,000	3,760,000

研究分野：総合領域

科研費の分科・細目：情報学・メディア情報学・データベース

キーワード：Web, 情報統合, パーソナライゼーション, レコメンデーション, 印象ベクトル

1. 研究開始当初の背景

近年、Web情報の指数関数的増加は留まるどころを知らず、我々は日々それらの溢れる情報の中から必要な情報の検索に明け暮れている。このような中、Webにおける情報融合技術は私たちの生活に深く浸透しつつある。例えば、普段目にはしているgoogleやyahoo!などの検索サイトはもちろんのこと、ニュースや旅行などのポータルサイトにおいて、各Webサイト内の多くの情報をユーザへ分かり易く提供するための情報融合技術はなくてはならない技術となっている。このように情報融合技術が急速に我々の生活にとって必

要不可欠なものとなるのに伴い、複数のWebサイトから自律的に情報を収集し、収集した情報を統合し提供する研究も近年著しく進歩している。その結果、分散している同一のテーマに関する情報に対して、複数のWebサイトへいちいちアクセスする必要がなくなり、面倒なWebサイトの検索やアクセスといった操作が不要になってきている。しかしながら、収集された情報量は膨大なため、それらの情報をユーザの立場にたって効果的に分類して提供できる、統合手法の確立は大きな課題となっていた。

2. 研究の目的

本研究の目的は、既存の情報統合技術が抱えている問題点である、収集した膨大な情報の柔軟な分類が困難であること、分類・統合した情報の効果的な提供方法が確立されていないこと、の解決を目指す。特に、本研究ではニュースサイトをターゲットとする。既存のニュース記事の情報分類では、特定のカテゴリを設定しており、収集した個々の Web ページの記事から単語を抽出し、抽出した単語を用いてページの特徴を算出し、特徴間の類似度を測ることで記事を各カテゴリへ分類する。この場合、カテゴリ設定は統合システムの管理者によって決定されるため、ユーザはシステム側で一意に決められた数項目のカテゴリの中から、どのカテゴリに閲覧したい記事が含まれるかを予想し、さらにカテゴリ内の記事の一覧から再検索する必要がある。また、既存の統合システムでは、分類し統合した情報を提供するためのインタフェースは各システムに依存している。そのため、ユーザは収集された情報がどこにどのように分類されているかを把握するために、システムごとに提供しているインタフェースを使い慣れる必要がある。

そこで、本研究では、記事の特性のみならず、ユーザのもつ複数の利用特性を検討することで、ユーザの特徴に基づいた柔軟なカテゴリの生成と分類、およびユーザの使い易いインタフェースの提供を実現できる、個人の利用特性に適応した仮想的なニュースのポータルサイトの構築を目指す。具体的な研究開発目標としては次の項目を定める。

目標 1: ユーザの記事に対する印象というこれまでにない新たな利用特性の抽出

目標 2: ユーザの印象と興味に基づく動的なカテゴリの生成と記事の分類および推薦

目標 3: ユーザが普段より閲覧しているニュースサイトのトップページへの統合情報の写像

目標 4: システムの検証による各利用特性の評価と利活用への再検討

3. 研究の方法

(1) ユーザの閲覧履歴を基に、ユーザの記事に対する興味を抽出すると共に、記事に対する印象ベクトルの抽出(目標 1)を行う。以下に具体的な研究方法を述べる。

①興味語抽出の改善

研究代表者がこれまで行ってきた基礎研究(実験および評価)により、ユーザが事前に個人情報を記録することなく、記事に対するユーザの閲覧履歴を用いることで興味語(ユーザが興味があると推測される単語)を抽出し、ユーザプロファイルとして自動作成できることは既に確認されている。しかしながら、これまでの興味語の抽出では、閲覧回数

の増加に伴い興味語の数が増加してしまうという問題点があった。その主な原因として、同一記事に対して多数の興味語が抽出されることと、各興味語の関係性を考慮していなかったことなどが考えられる。そこで、初年度は、記事ごとに抽出される複数の興味語の関係性を明らかにし、階層構造化することで、閲覧回数の増加に伴う興味語の増加を抑えつつユーザの興味を包含できる興味語の抽出を目指す。

②ユーザの印象ベクトル抽出

印象ベクトルの抽出は、記事の印象を学習により自動作成した印象辞書を用いて複数の印象尺度に対する評価値として算出し、各印象尺度値を印象ベクトルとして生成する。印象辞書は、市販の全文記事データベースを用いて作成する。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)を基に分類する。また、生成された記事ごとの印象ベクトルを利用して、興味語に分類されている閲覧記事の印象ベクトルの揺らぎを標準偏差より調べることで、ユーザの興味語に対する印象ベクトルを算出する。さらに、興味語ごとの抽出された印象ベクトルをユーザの利用特性として取り扱う。以上のことより、ユーザの記事に対する印象という概念のモデル化と、ユーザが共感できる記事の選択機構の実現を目指す。

(2) 初年度において、ユーザの利用特性のうちの興味と印象の抽出が終了するため、次にユーザの印象と興味に基づく動的なカテゴリの生成と記事の分類(目標 2)、およびユーザが普段より閲覧している既存のニュースサイトのトップページへの統合情報の写像(目標 3)を行う。以下に具体的な研究方法を述べる。

①興味語と印象ベクトルに基づくカテゴリ生成と記事の分類および推薦

収集した大量の記事の分類手法は、ユーザの利用特性である興味語と印象ベクトルを用いることで実現する。まず、分類するためのカテゴリは、興味語であるキーワードを用いて動的に生成する。興味語はユーザの閲覧履歴より生成されランキングされるため、閲覧する度に変化する。そのため、同類の記事をまとめて閲覧すると、類似する意味の興味語が上位を占めるため、他の内容の記事を分類するためのカテゴリ生成が困難となる。そこで、次年度では、各興味語に分類された記事の統計を基に興味語間の類似性を検出することで興味語を統合および分離し、ユーザの利用特性を柔軟に取り入れたカテゴリの生成を目指す。

さらに、生成されたカテゴリへ分類された記事の推薦は、ユーザの興味語に対する印象

ベクトルを用いる。初年度で既に興味語ごとに分類された記事のうち、閲覧記事に対する印象ベクトルの揺らぎを基にユーザの興味語に対する印象ベクトルが抽出されている。この興味語ごとの印象ベクトルと未読である記事の印象ベクトルとのコサイン相関を算出し、コサイン相関値の高い記事を類似性が高い記事として上位へランキングして推薦する。

②統合情報の写像

統合情報とは、生成されたカテゴリとカテゴリごとに分類された記事集ならびに分類された記事のランキングの情報である。これらの情報を既存のWebサイトのトップページの限られたエリア内に、効率よく統合して写像する手法の検討ならびに開発を目指す。

(3)(2)までの研究成果より、個人適応型情報融合システムの実装が終了するため、最終年度は、システムを検証することで、各利用特性の評価と利活用への再検討(目標4)を行う。以下に具体的な研究方法を述べる。

①利用特性の評価

実装した個人適応型の情報融合システムを用いて、ユーザの各利用特性を評価する。評価実験では、被験者に統合結果を提示し、閲覧履歴とユーザプロフィールをモニタリングする。モニタリングの結果より、閲覧した記事の内容の時間的な経過と、ユーザプロフィールの興味語ならびに印象ベクトルの変化を比較し検証する。

②評価に基づく改善ならびに利活用

評価実験より得られた知見を基に、興味語と印象ベクトルの抽出制精度の更なる改善を目指す。本研究では、ニュースサイトに限定したシステム設計を行っている。しかしながら、Webを対象としているため、他のテーマでの利用が期待されることは必至である。そこで、ニュースに留まらず、新たなテーマのコンテンツを対象とした個人適応型の情報融合システムの適応を目指す。

4. 研究成果

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)を構築した(図1)。構築したシステムでは、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ことにより、記事を選択する際に記事に出現する単語そのものだけでなく、その単語の印象という新たな選択基準を用いることでユーザが共感(感情移入)しやすい記事を優先的に選択し推薦できる。



図1: My Portal Viewer

(1) システムの概要

構築したシステムは以下の特徴を持つ。

- ①ユーザの閲覧履歴からキーワード(ユーザが興味のある語)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分類
- ②ユーザが指定した好みのニュースサイト(トップページ)のレイアウトをそのまま利用
- ③印象という分類基準の導入より高精度なユーザ選好のモデル化

(2) ニュース記事に対するユーザ選好の学習

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し実装した。興味語は、ユーザの閲覧した記事集合から出現頻度の高い単語(主に名詞)を抽出した。抽出された興味語は値の大きい順に、ユーザが指定したニュースサイトの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(図2)。

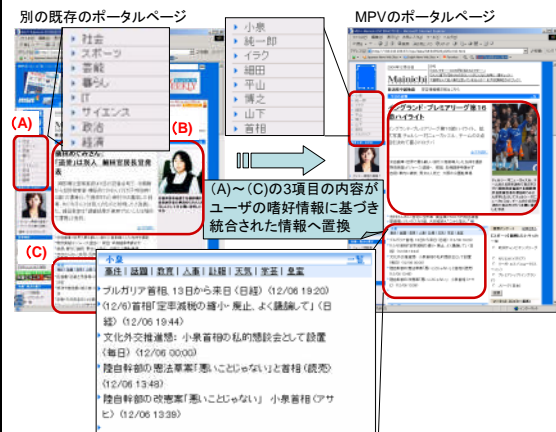


図2: 興味語の置換と記事分類

(3) ニュース記事の印象ベクトルの生成

ユーザが閲覧した記事の印象を各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⇔拒否, 緩和⇔緊張, 恐れ⇔怒り)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, その要素を 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, ユーザの記事に対する選好と定義し, ユーザプロフィールに記録する。収集された記事は, このユーザプロフィールに基づいて各カテゴリに分類される。

(4) 興味の推移分析

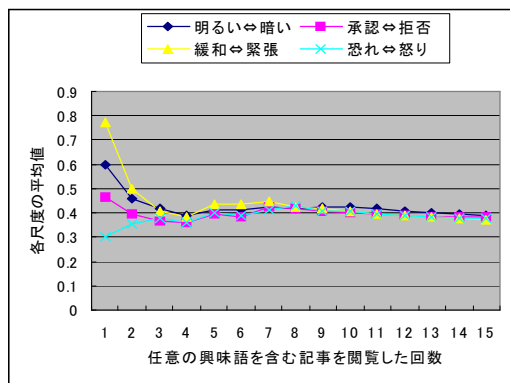
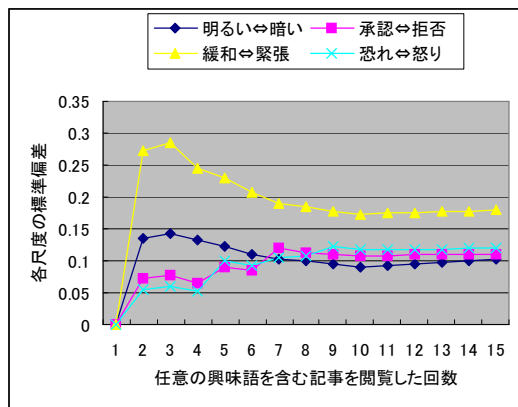


図3: ユーザの閲覧回数ごとに算出される4尺度の標準偏差(上)と平均値(下)

図3に, 記事を閲覧するたびに算出される任意の興味語に対する4尺度の標準偏差と平均値の推移を示す。最初の閲覧では, 「緩和⇔緊張」と「明るい⇔暗い」の標準偏差が高いが, 閲覧回数が増加するに従って徐々に減少している。つまり, この話題に関する記事を

閲覧するに従って, これら2つの尺度に偏りが生じていることを示している。その際の平均値は0.4付近を示しており, 「緊張」と「暗い」に偏っていることが分かる。このことよりユーザは「暗く, 緊張感のある記事を好んで選び閲覧している」という分析ができる。

(5) 得られた知見と今後の展望

新たな印象情報により, 興味語(カテゴリ)に分類された各記事に対する印象(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ベクトルとの距離を測ることにより, 興味語においてユーザの選好に合った記事推薦が可能になる。

さらに, ユーザの印象と記事の書き手の印象をベクトル形式でモデル化することによって, ユーザの好みの印象に近い記事を推薦するという方式だけでなく, 異なる印象で書かれた記事を推薦するということが可能となる。前者では, ユーザの選好に合わせて記事を提示できるという利点がある一方で, 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, ユーザはアンバランスな閲覧しかできず, 視野(閲覧範囲)が狭まってしまふという問題点がある。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, 各記事のベクトルが算出される。その際, 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, 推薦される記事も批判的な記事に偏り, 異なった印象情報を持つ記事(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う可能性がある。そこで, 後者の異なる印象情報を上手く利用することで, バランスのとれた記事推薦も可能にする5)。ただし, 興味語に対するユーザの印象情報と異なる記事を推薦するということは, そのユーザにとってはマイナーな記事であり, 本来の選好ではない可能性が高いため, 提示可能な範囲内の最下位で表示することで, 押付けとまらない情報推薦とする。

これまで述べてきたように, ニュースコンテンツに対するユーザ嗜好の抽出と記事推薦を行ってきた。また, この推薦技術を実空間に適用可能な研究開発も行っている。今後, 実空間とWeb空間ではユーザとコンテンツ間のさらに複雑なインタラクションが発生することは容易に想像できる。印象情報といった多様な嗜好情報抽出技術のさらなる向上だけでなく, インタフェースを含めたユーザにとって強制的でない柔らかな情報推薦技術についての研究開発も必要であると

考える。

5. 主な発表論文等

(研究代表者, 研究分担者及び連携研究者には下線)

[雑誌論文] (計 10 件)

- ① Adam Jatowt, Yukiko Kawai, Hiroaki Ohshima and Katsumi Tanaka: What Can History Tell Us? Towards Different Models of Interaction with Document Histories, Proceedings of 19th ACM Conference on Hypertext and Hypermedia (HT 2008), ACM Press, Pittsburgh, USA, pp. 5-14 (2008) 査読有
- ② Yukiko Kawai, Yusuke Fujita, Tadahiko Kumamoto, Jianwei Zhang and Katsumi Tanaka, Using a Sentiment Map for Visualizing Credibility of News Sites on the Web, Proceedings of Second Workshop on Information Credibility, pp. 53-58(2008) 査読有
- ③ 藤田 裕介, 河合 由起子, 熊本 忠彦, 張建偉, 田中 克己, 地域性に基づく発信者の観点可視化システムの提案, 情報処理学会 Web とデータベースに関するフォーラム論文集 (2008) 査読有
- ④ Yukiko Kawai, Tadahiko Kumamoto, and Katsumi Tanaka, Fair News Reader: Recommending News Articles with Different Sentiments Based on User Preference, Springer Lecture Notes in Artificial Intelligence (LNAI) 4692, pp. 612-622(2007) 査読有
- ⑤ Adam Jatowt, Yukiko Kawai and Katsumi Tanaka, Detecting Age of Page Content, Proceedings of the 9th ACM International Workshop on Web Information and Data Management (WIDM 2007), ACM Press, pp. 137-144(2007) 査読有
- 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論文誌, Vol.18・No.2, pp. 59-69(2006) 査読有
- ⑦ 河合由起子, 嗜好情報に基づくニュースコンテンツの推薦とその応用-画一的な便利さと多様な嗜好への適応, 情報処理学会誌, Vol. 48, No. 9, pp. 979-983(2006) 査読有
- ⑧ Yukiko Kawai, Tadahiko Kumamoto, Katsumi Tanaka. User Preference Modeling Based on Interest and Impressions for News Portal Site Systems, LNCS 4080, Krakow, Poland, pp. 549-559 (2006) 査読有
- ⑨ 河合由起子, 熊本忠彦, 田中克己. Fair News Reader: バランス感覚のある記事推薦方式の提案, 情報処理学会 情報科学技術レ

ターズ

Vol. 6, No. LD-002, spp. 55-58(2006) 査読有
⑩ Adam Jatowt, Yukiko Kawai and Katsumi Tanaka, User-oriented Freshness Detection for Improved Browsing and Navigation in WWW, LNCS 4080, Krakow, Poland, pp. 832-84 (2006) 査読有

[学会発表] (計 3 件)

- ① 濱砂佳貴, 河合由起子, 熊本忠彦, 田中克己, センチメントマップによる複数ニュースサイトの差異情報可視化手法の提案, 電子情報通信学会 第 19 回データ工学ワークショップ(DEWS2008)
- ② 中谷文彦, 河合由起子, 熊本忠彦, 柔軟な Web コンテンツ検索のためのレーダーチャート検索システムの提案, 電子情報通信学会 第 19 回データ工学ワークショップ(DEWS2008)
- ③ 野澤大介, 河合由起子, 熊本忠彦, 興味に基づく音楽情報推薦システムの提案, 情報処理学会第 70 回全国大会(2008)

[図書] (計 2 件)

- ① Adam Jatowt, Yukiko Kawai and Katsumi Tanaka: Browsing Assistant for Changing Pages, In: Nguyen N. T., Jain L. C. (Eds.): Intelligent Agents in the Evolution of Web and Applications, Springer-Verlag, pp. 137-160 (2009)
- ② Adam Jatowt, Yukiko Kawai and Katsumi Tanaka: Utilizing Past Web for Knowledge Discovery, In: Krol D., Nguyen N. T. (Eds.): Intelligence Integration in Distributed Knowledge Management, IGI Global, pp. 283-301 (2008)

6. 研究組織

(1) 研究代表者

河合由起子 (KAWAI YUKIKO)

京都産業大学・コンピュータ理工学部・講師

研究者番号: 90399543