解　説

情報推薦・情報フィルタリングのためのユーザプロファイリング技術

User Profiling Technique for Information Recommendation and Information Filtering

土方 嘉徳
Yoshinori Hijikata
大阪大学大学院基礎工学研究科
Graduate School of Engineering Science, Osaka University
hijikata@sys.es.osaka-u.ac.jp, http://www.nishilab.sys.es.osaka-u.ac.jp/people/hijikata/index.html

keywords: user profiling, information filtering, recommender system, relevance feedback

1. 個人化する Web

パーソナライゼーションと言う言葉をよく目にするようになりつつある．世界中でアクセス可能な Web ページ数は，99 年の時点で 8 億ページ（Lawrence らの推定）であったのが，2003 年の時点で 33 億ページ以上（Google の公表値）に増加している．どういう情報がどこにあるのかという Web の全体像を，一個人が把握することは極めて困難になりつつあり，この中から個人に適した情報を選択し，個人に適した方法で表示することが重要となりつつある．このようなサービスをパーソナライゼーションと呼ぶ．パーソナライゼーションの解説は CACM の特集号でも採り上げられており [Riecken 00]，このことからもその注目度がうかがえる．パーソナライゼーションを行うためには，ユーザの興味や目的，おかれたコンテキストなどの情報を獲得する必要があるが，本稿ではこのうちユーザの興味に関する情報をいかにして獲得するかについて解説する．本稿では，この技術のことをユーザプロファイリング技術と呼ぶ．

ユーザプロファイリング技術の説明をする前に，パーソナライゼーションの仕組みについて説明する．パーソナライゼーションの核となる機能はユーザに適した情報の選択である．そのための技術として情報推薦システム（Recommender system)[Resnick 97] あるいは情報フィルタリングシステム（Information filtering system)[Loeb 92] がある．それぞれの違いは，情報フィルタリングは文書を対象としたフィルタリングを意味し，情報推薦は文書を含めて様々な商品やサービスのフィルタリングをも含むニュアンスが強い．Web ページを推薦の対象とするのであれば，ほぼ両者には違いはないと見て良い．

それぞれにおいて情報を選択する方式には，(1) コンテンツに基づくフィルタリング（Content-based filtering) と，(2) 協調フィルタリング（Collaborative filtering）の 2 種類がある [Riecken 00]．これらの定義は多くの解説記事で取り上げられているので，ここでは簡単な説明にとどめるが，前者は，推薦する情報の内容に基づき，情報の取捨選択を行う．後者は，ネットワーク上に存在する同じ好みを持ったコミュニティを発見し，そのコミュニティが共通して好む情報を選択する．それぞれ情報選択の基本的な考え方は異なるが，いずれにしてもユーザはどの情報が好きなのか，あるいはどの情報に興味があるのかという情報が必要となる．ただし，このような嗜好・興味に関する情報をどの程度の粒度で必要かは，それぞれの手法で異なる．コンテンツに基づくフィルタリングでは，テキスト情報の解析をキーワード単位で行うことが多いため，キーワード単位で興味の有無が分かる方が良いと言える．協調フィルタリングでは，推薦する情報単位，つまり Web であればページ単位で，複数のユーザの興味に関する情報を解析するため，ページレベルで興味の有無が分かれば十分なケースが多い．コンテンツに基づくフィルタリングの方が，より細かい単位で興味を推定する必要があるため，ユーザプロファイリング技術の開発も困難なものになると言える．

本稿では，情報推薦・情報フィルタリングに必要な，特にコンテンツに基づくフィルタリングに使えるようなユーザプロファイリング技術について解説する．ユーザプロファイリング技術の多くは，情報閲覧時におけるユーザからの何らかのフィードバックを用いていることが多いため，適合性フィードバックとの関連が深い．そこで，最初に適合性フィードバックの基礎について説明する．ついで，ユーザプロファイリング技術の分類と，各分類の具体的な手法について説明する．さらに，各分類の手法の長短所について解説した後，筆者の開発した TextExtractor というシステムについて解説する．このシステムは，各分類の手法の短所を克服した実践的なシステムである．最後に，ユーザプロファイリング技術の課題とその方向性を示す．

2. 適合性フィード バック

コンテンツに基づくフィルタリングにおいて，興味に関する情報がどのように使われるかを，適合性フィードバック（Relevance feedback）（適合フィードバック，関連 (性) フィードバックとも呼ぶ）を例に説明する．適合性フィードバックは，コンテンツに基づくフィルタリングの手法として最も代表的な手法で，もともとは情報検索において確立された手法である [Meadow 92]．適合性フィードバックの定義は，情報検索において検索結果として出力された文書の内容に基づいて，検索質問や検索戦略，検索式を修正することを指す．最も分かりやすい例を挙げると，検索エンジンの検索結果において興味のあるページをユーザが指定すると，そのページの内容に基づき，それらのページ近いページを再度検索してくれるというものである．

どのように近さを計算するかは，完全照合方式（exact match method）と部分照合方式（partial match method）の二通りがある [Belkin 87]．完全照合方式の代表的手法は，タウベらが開発したブール代数（Boolean algebra）に基づく検索理論 [Taube 55] であり，数個のキーワードを取り出し，それらを AND や OR，NOT でつないで，適合する文書を探し出す．数学的な表現をすれば，文書集合を $D=\{d_1,d_2,...,d_N\}$，検索質問の集合を Q，検索過程で各文書に付与される値の範囲を V，ある検索質問 $q\in Q$ に関する写像 $f_q : D\rightarrow V$ としたときに，情報検索のモデルは $<D,Q,V,f_q>$ と表現できる．もし，単純に検索質問 q が一つの単語から構成されているとして，その単語が文書 d_i に含まれているか否かだけを見るのであれば，

$$f_q(d_i)=\begin{cases}1, & q\subseteq d_i\\ 0, & \text{それ以外}\end{cases}$$

と定義して，$f_q(d_i)=1$ のときに文書を出力し，$f_q(d_i)=0$ のときには文書を出力しない検索システム（$V=\{0,1\}$）を考えることができる．さらに，複数のキーワードと AND, OR, NOT を使う場合は，$q\subseteq d_i$ 部分を，それぞれのルールで置き換えれば良い．これがブール代数に基づく検索理論の基礎となる．

部分照合方式の代表的手法は，サルトンらによるベクトル空間モデル (vector space model)[Salton 83] となる．ベクトル空間モデルは，その名の通りキーワードによるベクトルを生成して，これを文書の近さの計算に利用するものである．こちらも数学的な表現をすれば，文書 d_i $(i=1,...,N)$ における単語 t_j $(j=1,...,M)$ の重みを w_{ij} として，文書 d_i を M 次元ベクトル $\mathbf{d}_i=(w_{i1},w_{i2},...,w_{iM})^T$ として表現する．検索質問 $q\in Q$ における単語 t_j の重みを w_{qj} とすれば，検索質問も M 次元ベクトル $\mathbf{q}=(w_{q1},w_{q2},...,w_{qM})^T$ として表される．すると，その適合度はそれらベクトル $\mathbf{d}_i$ と $\mathbf{q}$ の類似度として表される．代表的な類似度としては以下のようなコサイン尺度（cosine measure）が使われる．

$$S(\mathbf{d}_i,\mathbf{q})=\frac{\mathbf{d}_i^T\mathbf{q}}{\|\,\mathbf{d}_i\,\|\,\|\,\mathbf{q}\,\|}=\frac{\sum_{j=1}^{M}w_{ij}w_{qj}}{\sqrt{\sum_{j=1}^{M}w_{ij}^2}\sqrt{\sum_{j=1}^{M}w_{qj}^2}}$$

検索質問 $\mathbf{q}$ としては，何も加工しない基となる文書 $\mathbf{d}_o$ となることもある．

ブール代数に基づく検索理論では，それに用いる単語の選択が，ベクトル空間モデルでは単語の重み付けが重要となるが，それには統計的な手法が多く用いられる．その代表的な手法として tf・idf がある．tf・idf は，単語の重み付けを次式のようにモデル化したもので，文書内で繰り返し使われ，かつ他の文書にはあまり見られないような単語は文書の内容をよく表しているという考え方に基づいている．

$$w_{t,d}=tf_{t,d}\cdot idf_t$$

$$tf_{t,d}=f_{t,d}/\sum_{t=1}^{M}f_{t,d}\qquad idf_t=\log N/df_t$$

$f_{t,d}$ は文書 d 中の単語 t の出現回数，M は文書 d 中の単語の種類の総数，df_t は単語 t の一回以上生起する文書の数，N を総文書数を示している．

以上が適合性フィードバックの基本的なメカニズムであるが，適合性フィードバックには問題が 2 つある．ブール代数に基づく検索理論とベクトル空間モデルのどちらの方法を用いた場合でも，キーワードの選択や重み付けをユーザが選択した文書全体のテキストから行っている．そのため，なかにはユーザの興味に関係しないものも含まれてしまい，それらのキーワードが検索の精度を低下させると言うことが懸念される [杉本 99]．もう一つの問題は，いずれの手法においても，ユーザに検索の基となる文書を選択させるために，閲覧操作以外の手間をユーザにかけさせることである．情報推薦・フィルタリングシステムにおけるユーザプロファイル作成においても，この 2 種類の問題を考慮する必要がある．

3. ユーザプロファイリング技術

ユーザプロファイリング技術には，大きく分けると明示的（直接的）手法（explicit method）と暗黙的（間接的）手法（implicit method）の 2 種類が存在する．これらは，ユーザの手間という観点から分けられたものである．

3･1 明示的（直接的）手法

ユーザから直接に，興味に関する情報を入力してもらう方法である．大きくは，(i) ユーザの興味に関してト

ピックやキーワードの形でアンケートに答えさせる方法，または (ii) 閲覧したページにどれだけ興味があったかを数段階で評価をつけさせる方法の 2 種類に分類できる．(i) は最も単純な方法であるため，研究対象として採り上げられることは少なく，初期の頃の推薦システムで見られるものの，最近の研究では中心の研究テーマとして取り上げられることはない．例としては，SIFT[Yan 95] というシステムが挙げられる．SIFT では，ユーザは興味のあるトピックをユーザプロファイルとして記述して，それをメールでサービス提供者に送り，情報推薦サービスを受ける．ユーザプロファイルは一般的なベクトル空間モデルで表される．

(ii) は NewT[Shech 93]，GroupLens[Resnick 94]，Lira[Balabanovic 95]，NewsWeeder[Lang 95]，Syskill & Webert[Pazzani 97]，SIFTER[Mostafa 97]，ClixSmart[Smyth 00]，AntWorld[Kantor 00], Foxtrot[Middleton 03] といったシステムで用いられている．本手法では，閲覧したページに対して評価付けさせた後に，その評価値をどう処理するかが問題となる．コンテンツに基づくフィルタリングでは，ベクトル空間モデルに基づくものが多いが，その処理方式においてそれぞれ工夫がなされている．この処理の例として，Syskill & Webert と AntWorld, Foxtrot を挙げる．

Syskill & Webert は，興味の有りそうなページを推薦するソフトである．ユーザプロファイルは，一人のユーザに各トピックごとに作成される．各ページは 2 値の単語ベクトルで表され，単語ベクトルの要素に割り当てる単語は，単語 W の存在がページ集合 S の分類に与える予測情報利得 (expected information gain)$E(W,S)$ によって選択される．

$$
\begin{aligned}
E(W,S) = I(S) & \\
& -[P(W=present)I(S_{w=present}) \\
& +P(W=absent)I(S_{w=absent})] \\
I(S) = \sum_{c\in\{hot,cold\}} & -p(S_c)\log_2(p(S_c))
\end{aligned}
$$

ここで，$P(W=present)$ は単語 W のページに出現する確率，$S_{w=present}$ は単語 W が少なくとも 1 回は含まれるページの集合，$S_{w=absent}$ は単語 W が 1 回も出現しないページの集合，S_c はクラス c に含まれるページの集合を指す．クラスはユーザの興味のあるページ（hot) とユーザの興味のないページ（$cold$）の 2 値である．ページごとに，このような単語ベクトルと，興味があるか否かの教師信号を持っておけば，任意の学習アルゴリズムを用いることができる．ここでは，ベイズ分類子 (naive Bayesian classifier) と，最近傍検索 (nearest neighbor algorithm)，PEBLS，ID3，Rocchio のアルゴリズム，ニューラルネットを比較しており，ベイズ分類子が最も精度*1が良かったことを報告している．

AntWorld では，ユーザはあるキーワードで検索エンジンに検索をかけて探索を始めると，過去に同じようなキーワードで探索した他のユーザが最終的に辿り着いたページで，なおかつ良い評価がつけられたもののみを推薦する．AntWorld の推薦方法では，ブール代数に基づく検索理論とベクトル空間モデルのそれぞれの長所をうまく結合している．その結合役として，LAD 理論 (Logical Analysis of Data) を用いている．LAD の基本的な考え方は，ユーザはあるルールに基づいてページの評価を行っており，それらのルールは不完全に定義されたブール関数で表されるというものである．適合性フィードバックにこのブール関数を用いることとしている．具体的には，ページに付けられた”meet my need”や”adds information”，”not relevant”などの評価を個別のブール関数で表現し，すべてのページに対するバイナリの関数値を表として保存する．また，単語でベクトル空間モデルを構築し，頻度の閾値によりバイナリ化して表として保存する．これらの表に LAD マシンを実行し，基も経済的なブール代数のルールを生成している．

Foxtrot は，計算機科学の分野の論文（HTML，PS，PDF 形式のファイル）を推薦してくれるシステムである．ユーザの閲覧行動をプロキシで監視しており，ユーザは 5 段階で興味の程度を入力する．最大の特徴は，ユーザプロファイルにオントロジを利用している点である．ユーザが新しく発見した論文はデータベースに格納され，この時論文トピックオントロジ（計算機科学の分野を階層的に分類したもの）に分類される．分類には，IBk 分類子という k-nearest neighbor 法の類似アルゴリズムが用いられている．ユーザプロファイルはオントロジ中の概念とその重みで表される．論文中に直接に書いてある単語でユーザプロファイルを構築するよりも，より広い意味から論文を推薦できる．最も大きな利点は，オントロジの概念的階層を持ちいた方が，新しく利用し始めたユーザにも，最初から比較的高い精度の推薦が可能な点にある．情報フィルタリング・情報推薦システムにおいては，新しいユーザが入ってきたり，新しいアイテムが出現したりすると，それらに対してまだ多くの情報が集まっていないために推薦の精度が落ちるという問題，いわゆる cold-start 問題がある．オントロジを用いる方法は，この cold-start 問題の解決方法の一つとして期待される．

3･2 暗黙的（間接的）手法

ユーザの Web 閲覧時の挙動から，ユーザの興味に関する情報を取得する方法である．本手法には，閲覧したページのすべてにユーザが興味を持ったと仮定して，(i) Web ページのアクセス履歴（Web サーバまたはプロキ

*1 情報検索・フィルタリングの分野では，評価の指標として精度 (precision) と再現率 (recall）を用いることが多い．精度は検索・推薦された文書のうち，ユーザの要求に適合する文書の割合を意味し，再現率は検索・推薦対象の文書集合中におけるユーザの要求に適合する文書のうち，検索・推薦された文書の割合を意味する．

シサーバのログから取得可能な「どのページを閲覧したか」という履歴）を用いる方法と，何らかの方法でユーザが閲覧した情報に興味があったかなかったを判定する方法の 2 種類に大別できる．後者において，閲覧した情報に対する興味の有無を推定するのには，(ii) ユーザが閲覧に費やした時間（閲覧時間）や，(iii) 閲覧中におけるマウス操作，(iv) 閲覧中の視線，などの方法がとられる．以下，それぞれの手法における例を示す．

(i) の方法を用いた代表的なシステムとしては，WebWatcher[Joachims 97] や WebMate[Chen 98]，Crabtree らの研究 [Crabtree 98]，橘高らの研究 [橘高 99] などがある．閲覧したページを平等に「興味のあったページ」として扱うため，明示的手法で説明した閲覧したページに対して評価付けを行う方法において，ページに対する重みが等しいケースに相当する．評価付けされたページからいかにユーザプロファイルを構築するかは，前節にて説明済みであるので，ここでは各システムの詳しい説明は割愛する．

(ii) の方法では，森田らがネットニュースの記事の閲覧時間を観測している [Morita 94]．この閲覧時間とユーザにとっての記事の有用さの度合いと不要さの度合いとの相関関係の調査結果を報告している．また，情報フィルタリングのプロファイルとして閲覧時間から有用だと推定した記事を用いた場合のフィルタリング精度についても報告している．その結果，閲覧時間と実際のユーザの評価には，明確な相関関係が有ることを示している．また，閲覧時間に影響する要素として，記事に対するユーザの興味の有無以外にも，(1) 記事の長さ，(2) 記事の読みやすさ（行中の文字数と記事中の空行の割合），(3) ネットニュースを立ち上げた時の未読の記事の数などが考えられるが，これらも調査した結果，いずれも閲覧時間との相関は低い（相関係数 0.08〜0.25）ことが分かっている．フィルタリング精度については，Sub-string indexing method と呼ばれる独自のフィルタリングの方法で評価をしているが，再現率 20%に対して，興味なしと推定した記事に対する精度で 59.5%，興味ありとして推定した記事に対する精度として 48.7%となっている．ユーザが実際に有用だと答えた記事を用いてユーザプロファイルを構築した場合との比較についても知りたいところであるが，森田らはそこまでは行っていない．しかし，LSI(latent semantic indexing)[Deerwester 90] を用いた Foltz らの情報フィルタリングの方法では，ユーザの明示的な評価をプロファイルとして用いているが，彼らの結果では再現率 25%に対して，精度はおよそ 67%となっており [Foltz 92]，まずまずの結果となっていると言える．

(iii) の方法では，ANATAGONOMY[Sakagami 97] と TextExtractor[土方 02] が挙げられる．ANATAGONOMY は，ユーザのページへの興味を推定するのに，マウスの操作を使うことを試みた初めての研究である．マウスの操作としては，ページに対してユーザが拡大表示するボタンを押したか否かや，スクロールをしたか否かをチェックしている．ANATAGONOMY は，任意の Web コンテンツを対象としたシステムではなく，Web ブラウザをインタフェースとした独自のニュースシステム上におけるマウス操作を取得している．ユーザに，閲覧したページに 5 段階の評価を明示的につけてもらい（10-90 点の間でスコアをつけている），全てのページの平均のスコアと，拡大操作をしたページのみの平均のスコア，スクロール操作をしたページのみの平均のスコア，上記の両方の操作をしたページの平均のスコアを比較している．その結果，13 人の被験者の平均を取ると，それぞれ 1.4 倍，1.7 倍，1.9 倍となっており，明らかな差があることがうかがえる．また，これらの操作があったページに対して重みをつけて，適合性フィードバックを行った時の推薦精度を評価している．その結果，ユーザの明示的な評価からページに対して重みをつけた場合ほどの精度は得られなかったが，システムの記事の推薦順位と推薦されたページに対してユーザが実際に付けた評価の間には，高い関連が見られたということを報告している．これに対して TextExtractor は，筆者の開発したシステムであるが，ANATAGONOMY と違い任意の Web ページに対してマウス操作からユーザの興味の対象を推定するシステムである．TextExtractor については，次章で詳しく説明する．

(iv) の方法では，Digital Reminder[吉田 00] と IMPACT[大野 00] が挙げられる．Digital Reminder は，ページに興味を持ったか否かを推定しているわけではないが，視線により文章に注目している状態を検出している．横書きの文章を読んでいる場合，視線は左から右への跳躍運動が繰り返され，その行を読み終わると一気に次の行の先頭に移動する．この考え方を基本に跳躍の幅を閾値として与えることで，上記状態を検出している．その結果，88.2%の精度と，90.9%の再現率で，検出可能と報告している．また，IMPACT は，ページ中の部分領域への注目度を算出している．部分領域は，HTML タグに相応する領域としている．注目度の算出方法は，次のヒューリスティックスを用いている．

（1）視線の移動距離が画面上で 15 ピクセル以内の状態が 100ms 以上続いた時を視線の停留点と定義する．
（2）停留時間が長くなるにつれて注目度は増加するが，極端に増加することはない．
（3）停留点の移動方向が，横方向の場合に，より多く注目度を増加する．
（4）各領域の注目度は，ページを表示している間，時間の経過と共に徐々に減衰する．

これらのヒューリスティックスは，観察実験の結果に基づいている．IMPACT は，情報推薦・フィルタリングというよりも，人が以前に見た情報への再アクセスというアプリケーションに注目している．ユーザ実験では，被験者は問題文をオンライン文書から探し出すというタスク

を行った後，問題に関連するキーワードでオンライン文書に対して検索を行う．注目度の高かった領域のみを再検索の対象として，視線による注目度算出の有効性を検証している．その結果，注目度の高かった領域における検索結果の数は，そうでない領域の検索結果の数の 1/6 程度で，その時の再現率は 81.0%となっている．

4. 各手法の長短所と TextExtractor

4·1 各手法の長短所

前章で分類した明示的手法と暗黙的手法には，それぞれ長短所がある（表 1）．明示的手法では，取得したユーザの興味に関する情報は，ユーザが直接答えたものであるために，信頼性が高いという利点がある．しかし，アンケートに答えさせたり，閲覧後に評価を付けさせるという手間を，ユーザ側にかけるという問題がある．また閲覧したページに評価をつけさせる方法では，ページに付けられた興味の度合いからキーワード単位での重み付けや検索式に変換する必要がある．しかし，多くの場合 tf・idf などの統計量に基づく処理をしているため，ページ全体のテキストからキーワードを選択することとなる．そのため，選択したキーワードの中にはユーザの興味と関係ないものも含まれるという問題がある．また，tf・idf は，人が記述した文書は文章から構成されており，それら文章中ではその主題に関する単語が繰り返し出現すること（ルーンの仮定 [Luhn 58]）を前提としている．これは，新聞記事や論文などでは正しいのであるが，Web ページというのは質を含めて多種多様であるため，その全てのページで有効に働くかには疑問がある．

暗黙的手法は，ユーザに明示的に興味に関する入力を強いない点が利点である．アクセス履歴・閲覧時間を使う方法の最大の利点は，その実現可能性にある．アクセス履歴は，サーバログから直接得ることができ，閲覧時間もサーバログから推定することが可能である．両手法とも，クライアント側にプログラムをインストールしたり，特殊な設定をする必要がないため，ネット上でビジネスを行うにあたっては障害が少ない．その反面，アクセス履歴を使う方法では，閲覧したすべての Web ページに興味があったと仮定しているため，そうでないページの影響が大きい．閲覧時間の方は，ユーザがあるページを表示した後に，席を外したり他の仕事を始めたりしても，それを検出することができないという問題がある．また，両手法とも Web ページのどの部分に興味を持ったのかまでは取得できない．

アクセス履歴や閲覧時間を用いる方法は，ページを表示させるという行為だけで興味があることを認識してしまうが，マウス操作を用いる ANATAGONOMY では，興味から起こるより強いアクションを必要とすることから，ユーザのより強い興味だけを検出可能と思われる．しかし，Web ブラウザ上のマウスを取得するには，クライアント側に特殊なプログラムを挿入しておく必要がある．その点で，閲覧時間よりもビジネス上の実現可能性は低くなる．また，ANATAGONOMY で取っている操作をチェックすることでは，ページ全体に興味を示したか否かはある程度判断できるが，ページのどの部分に興味を示したかということまでは判断できない．その点では，閲覧時間と変わらない．

視線を用いる方法は，ページのどの部分に興味を示したかを取得できる上，マウス操作ほど個人差がないと思われるため，最も効果の期待できる方法である．しかし，クライアント側に特殊な装置やソフトを導入する必要があり，ビジネス上の障害という観点からは，最も現実的ではない．

最後に，評価指標として重要なものとして，ユーザプロファイルの構築時間がある．多くの情報フィルタリングシステムにおいては，そのユーザプロファイルが有効となるのに，ユーザが使い始めてから時間を要することを前提としている．例えば，ANATAGONOMY のユーザ実験では，毎日同じカテゴリのネットニュースを見ているのであるが，このケースでユーザプロファイルが有効に機能する日数として 2 日必要としている．cold-start 問題に着目している Foxtrot においても，最初の 1 週間は非常に精度が悪いことを報告している．これは，ページ単位で興味の有無を推定しているためで，ノイズとなるキーワードの影響を避けられないためである．たくさんのページを閲覧していけば，それらに共通に高い頻度で表れる（tf の大きい）キーワードの重みが高くなり，結果的にこれらノイズとなるキーワードの重みが下がる．その効果を得るためには，ある程度の期間を要する．長期的に使うようなユーザプロファイルの構築であればこれでよいが，そのセッション中に有効であるような短期的な適合性フィードバックには向かない．その意味で，ページ単位で興味を推定している手法は，ページ中の部分単位で興味を推定している手法よりも，ユーザプロファイル構築に時間がかかり，短期的な適合性フィードバックに利用することはできない．

4·2 TextExtractor

TextExtractor は，前節で挙げた問題を克服したシステムで，ユーザの Web ページ閲覧中のマウス操作を利用して，ユーザが興味を持ったと思われるテキスト部分を全体のテキストから自動抽出している．Web ページ閲覧時のその場での適合性フィードバックに利用できることを目指している．具体的には，入力デバイスとしてマウスを用いて Web ページを閲覧する場合を対象として，以下の手法により上記問題を解決する（図 1）．

(1) Web ページ閲覧時のマウス操作から，ユーザの興味と関連して発生した可能性のある操作を抽出する．
(2) 抽出した操作の対象となるテキスト部分を文や行の単位で抽出する．

表 1 ユーザプロファイリング手法の比較

手法	分類	ユーザ負担	興味粒度	構築時間	ビジネス的実現可能性
事前アンケート [Yan 95]	明示的	×	—	○	○
ページ評価 [Shech 93]-[Middleton 03]	明示的	×	×	×	○
アクセス履歴 [Joachims 97]-[橘高 99]	暗黙的	×	×	×	○
閲覧時間 [Morita 94]	暗黙的	○	×	×	○
特殊なマウス操作 [Sakagami 97]	暗黙的	○	×	×	△
視線 [吉田 00, 大野 00]	暗黙的	○	○	○	×
TextExtractor [土方 02]	暗黙的	○	○	○	△

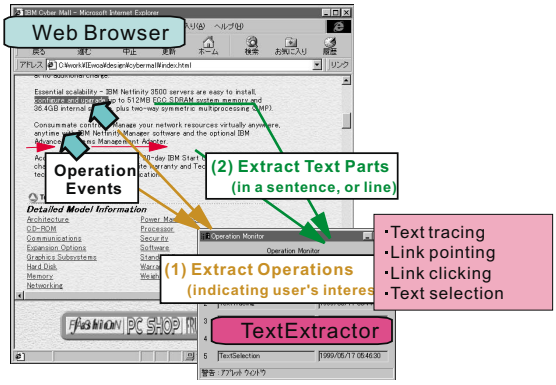

図 1 TextExtractor

ユーザの興味と関連して発生する操作として，1) なぞり読み，2) リンクポインティング，3) リンククリック，4) テキスト選択を用いており，これらは事前の観察実験で選ばれている．

マウスの操作イベントを検出するためのクライアント側のプログラムは JavaScript と Java アプレットで実装しており（図 2），W3C 標準化技術を用いて（具体的には DOM(Document Object Model)），汎用性の高いシステムを目指している*2．TextExtractor の他の手法に対する利点としては，1) ユーザに負荷を与えない，2) ページ単位でなくページ中の部分単位で興味の有無を推定できる，3) その場での適合性フィードバックに利用可能である，4) 特殊な装置がいらない，5) 標準化技術を使っており汎用性が高い，ということが挙げられる．逆に欠点としては，1) クライアント側にプログラムを埋め込む必要がある，2) 操作の個人差が大きい，ということが挙げられる．ユーザプロファイリング技術としては他の手法よりもノイズのキーワードを削減できる可能性があり，プロキシサーバやポータルサイトによる仲介サーバを置けば，十分ビジネス的な観点からも実現可能性があると考えている．その意味で，最も現実的な手法として期待できる．

*2 ブラウザにより各標準化技術の実装に差があるが，最もよく使われている InternetExplorer で，最もオープンなプログラム環境の JVM を使うという意味での汎用性

しかし，最大の問題は操作の個人差が大きい点にある．文献 [土方 02] によると，5 人の被験者実験により tf・idf と比較したところ，tf・idf よりも TextExtractor の方が，精度・再現率の高いキーワード抽出ができる可能性があることを示している（ただし，統計的な検証を得るには至っていない）．特に，ニュース記事のような一つのトピックについて文章で記述したようなページではなく，トップページ，リンク集，掲示板などの多様な形態のページにおいて，有効である可能性を示している．これは，前節で説明したように，tf・idf の tf 値が文章における重要単語の繰り返しを前提としているからである．文章で書いていないページや，様々なトピックが入り混じったページには，tf・idf はその効果を最大限に発揮できない．TextExtractor がすべてのユーザについて有効なのか，すべてのページについて有効なのかは，被験者数を増やして調べていく必要がある．現在，このような TextExtractor のユーザ・ページ種類についての，適用可能性について研究を進めている．

5. ユーザプロファイル更新と今後の課題

前章までは，何もないところからいかにユーザプロファイルを構築するかという議論をしてきたが，一度ユーザプロファイルが構築されれば，それを更新していく必要がある．特に，ユーザのある対象に対する興味の度合いは時間の経過と共に変化し，またユーザの興味の対象は，時間の経過と共に違うものに変化するからである．明示的手法やアクセス履歴を用いる手法は古くから研究が行われているため，構築したユーザプロファイルをいかにして更新していくかについても多くの研究がなされている．

代表的な研究例としては，橘高らの研究 [橘高 99] がある．この研究では，ユーザプロファイルを単語のベクトルでなく概念のベクトルとして表し，ユーザプロファイルの更新を，興味を持ったページの概念ベクトルの方向に回転させることで行っている．概念となる軸は情報提供者が決定し，さらにその概念に帰属するページ群を選択する．計算機はページごとの単語ベクトルからそのページが帰属する概念の単語ベクトルを求める．そして，新

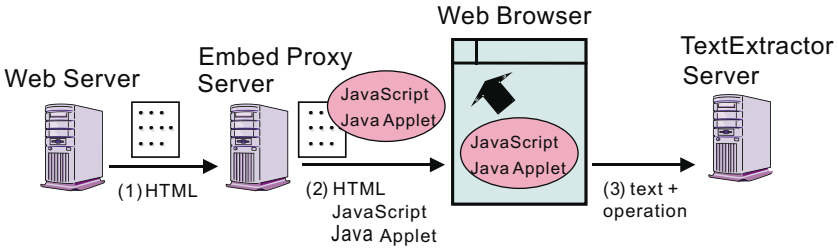

図 2 TextExtractor システム構成

しいページの単語ベクトルから，各概念の単語ベクトルとの類似度を求め，その類似度でそのページの概念ベクトルを算出する．このように，低次元のベクトル空間におけるベクトルの回転（回転の度合いはパラメータで調整）によりユーザプロファイルを更新しているため，ユーザの興味の変化に適応しやすくなっている．

この他にも様々な手法で，この問題に取り組んでいる．それらを簡単に紹介すると，宮原らはユーザープロファイルに活性度という概念を設け，活性度は興味を示したページによって刺激を受け，時間の経過と共に減衰することとしている [宮原 98]．興味の減衰はガンマ分布に従うという仮説を立てている．NewT では，遺伝的アルゴリズムを用いることで，最近に活性化された強い遺伝子を残すようにしている [Shech 93]．SIFTER は，適合性フィードバックの履歴（各カテゴリに興味を持つか否か）をベルヌーイ試行とみなし，ベイズ確率を用いて各カテゴリに対する興味の変化を検出している [Mostafa 97]．Crabtree は，閲覧した文書を一定期間ごとに決まったカテゴリに分類し，その期間の差における変化を見ている [Crabtree 98]．

しかし，これらの研究はユーザの興味をページ単位で取得しており，そのためユーザの興味の変化を長期間でしか捉えることができない．ページ単位での興味の有無から，適合性フィードバックを行っているため，ユーザがページの一部にしか興味を持たない場合には，更新したユーザプロファイルには大きな誤差を含むことになる．しかし，TextExtractor に代表されるページの部分への興味の有無を検出できるシステムでは，抽出したキーワードの有効期間については一切考慮していない．現在閲覧中のページに対する適合性フィードバックであれば，これで十分であるが，現在のセッション中のコンテキストに基づく適合性フィードバックであれば，今見ているページから抽出されたキーワードと，数ページ前のページから抽出されたキーワードに重みの差をつける必要がある．セッション中の興味の変化まで検出できるようなシステムができれば，情報推薦・情報フィルタリングの分野にとって，大きな技術革新になると言えるだろう．

6. ま と め

本稿では，情報推薦・情報フィルタリングシステムのためのユーザプロファイリング技術について解説した．特に，ユーザの興味の有無を検出する手法を，適合性フィードバックやブール代数に基づく検索理論，ベクトル空間モデルなどの基礎技術と合わせて解説した．既存のユーザプロファイリング技術には，興味について明示的に入力しなければならないユーザの負担の問題や，興味対象を絞り込む際の粒度の問題，ユーザプロファイル構築に要する時間の問題，そしてビジネス的実現可能性の問題など，多くの問題を抱えていることを示した．この中でも特にユーザプロファイル構築時間については，興味の変化に適応する必要性の観点からも，現状では満足のいくレベルにはない．今後は、ユーザの操作履歴やオントロジの利用など様々な観点からの試みが期待される．

◇ 参 考 文 献 ◇

[Balabanovic 95] Balabanovic, M. and Shaham, Y.: Learning Information Retrieval Agent: Experiments with Automated Web Browsing, in *Proc. of AAAI Spring Symposium on Information Gathering from Heterogeneous, Distributed Environments*, pp. 13–18 (1995)

[Belkin 87] Belkin, N. J. and Croft, W. B.: Retirieval techniques, *Annual Review of Inromation Science and Technology*, Vol. 22, pp. 109–145 (1987)

[Chen 98] Chen, L. and Sycara, K.: WebMate: A Personal Agent for Browsing and Searching, in *Proc. of the 2nd International Conference on Autonomous Agent (Agent'98)*, pp. 132–139 (1998)

[Crabtree 98] Crabtree, I. and Soltysiak, S.: Identifying and Tracking Changing Interests, *Iternational Journal of Digital Library*, Vol. 4, pp. 38–53 (1998)

[Deerwester 90] Deerwester, S., e. a.: Indexing by Latent Sematic Analysis, *Journal of the Society for Information Science*, Vol. 41, pp. 391–407 (1990)

[Foltz 92] Foltz, P. W. and Dumais, S. T.: Personalized Information Delivery: An Analysis on Information Filtering Methods, *Comm. of the ACM*, Vol. 35, No. 12, pp. 51–60 (1992)

[土方 02] 土方 嘉徳, 青木 義則, 古井 陽之助, 中島 周：マウス挙動に基づくテキスト部分抽出方式と抽出キーワードの有効性に関する検証, 情報処理学会論文誌, Vol. 43, No. 2, pp. 566–576 (2002)

[Joachims 97] Joachims, T., Freitag, D., and Mitchell, T.: WebWatcher: A Tour Guide for the World Wide Web, in *Proc. of IJCAI'97* (1997)

[Kantor 00] Kantor, P. B., e. a.: Capturing Human Intelligence in the Net, *Comm. of the ACM*, Vol. 43, No. 8, pp.

112–115 (2000)
[橘高 99] 橘高 博行, 佐藤 直之, 鈴木 英明, 曽根岡 昭直：パーソナライズ情報提供方式の提案と評価, 情報処理学会論文誌, Vol. 40, No. 1, pp. 175–187 (1999)
[Lang 95] Lang, K.: NewsWeeder: Learning to Filter NetNews, in *Proc. of ICML'95*, pp. 331–339 (1995)
[Loeb 92] Loeb, S. and Terry, D.: Information Filtering, *Comm. of the ACM*, Vol. 35, No. 12, pp. 26–81 (1992)
[Luhn 58] Luhn, H. P.: The Automatic Creation of Literature Abstracts, *IBM Journal of Research and Development*, Vol. 2, No. 2, pp. 159–165 (1958)
[Meadow 92] Meadow, C.: *Text Information Retrieval Systems*, Academic Press (1992)
[Middleton 03] Middleton, S., Shadbolt, N., and De Roure, D.: Capturing Interest Through Inference and Visualization: Ontological User Profiling in Recommender Systems, in *Proc. of the Second International Conference on Knowledge Capture (K-CAP'03)*, pp. 62–69 (2003)
[宮原 98] 宮原 一弘, 岡本 敏雄：Web ブラウジングに基づいた興味の定量的同定法とその協調フィルタリングへの適用, 信学技報, ET97-115 (1998-3), pp. 17–24 (1998)
[Morita 94] Morita, M. and Shinoda, Y.: Information Filtering Based on User Behavior Analysis and Best Match Text Retrieval, in *Proc. of the 17th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pp. 272–281 (1994)
[Mostafa 97] Mostafa, J., e. a.: A Multilevel Approach to Intelligent Information Filtering: Model, System, and Evaluation, *ACM Transactions of Information Systems*, Vol. 15, No. 4, pp. 368–399 (1997)
[大野 00] 大野 健彦：IMPACT：視線情報の再利用に基づくブラウジング支援法, in *Proc. of the 8th Workshop on Interactive Systems and Software (WISS'2000)*, pp. 137–146 (2000)
[Pazzani 97] Pazzani, M. and Billsus, D.: Learning and Revising User Profiles: the Identification of Interesting Web Sites, *Machine Learning*, Vol. 27, No. 3, pp. 313–331 (1997)
[Resnick 94] Resnick, P., e. a.: GroupLens : An Open Architecture for Collaborative Filtering of Netnews, in *Proc. of CSCW'94*, pp. 175–186 (1994)
[Resnick 97] Resnick, P. and Varian, H.: Recommender Systems, *Comm. of the ACM*, Vol. 40, No. 3, pp. 56–89 (1997)
[Riecken 00] Riecken, D., e. a., : Personalized Views of Personalization, *Comm. of the ACM*, Vol. 43, No. 8, pp. 26–158 (2000)
[Sakagami 97] Sakagami, H. and Kamba, T.: Learning Personal Preferences on Online Newspaper Articles from User Behaviors, *Proc. of the Sixth International World Wide Web Conference, In Computer Networks and ISDN Systems*, Vol. 29, pp.1447–1456 (1997)
[Salton 83] Salton, G. and McGill, M. J.: *Introduction to Modern Information Retrieval*, McGraw-Hill (1983)
[Shech 93] Shech, B. and Maes, P.: Evolving Agents for Personalized Information Filtering, in *Proc. of IEEE Conference on Artificial Intelligence for Applications*, pp. 345–352 (1993)
[Smyth 00] Smyth, B. and Cotter, P.: A Personalized Television Listings Service, *Comm. of the ACM*, Vol. 43, No. 8, pp. 107–111 (2000)
[杉本 99] 杉本 雅則：情報収集システムにおけるユーザモデリングと適応型インタラクション, 人工知能学会誌, Vol. 14, No. 1, pp. 25–32 (1999)
[Taube 55] Taube, M., e. a.: Storage and Retrieval of Inforamation by Means of the Association Ideas, *American Documentation*, Vol. 6, No. 1, pp. 1–18 (1955)
[Yan 95] Yan, T. W. and Garcia-Molina, H.: SIFT - A Tool for Wide-Area Information Dissemination, in *Proc. of 1995 USENIX Technical Conference*, pp. 177–186 (1995)
[吉田 00] 吉田 将志, 吉高 淳夫：Digital Reminder：ユーザの視点からの実世界指向データベースの構築とそのインタフェース, in *Proc. of the 8th Workshop on Interactive Systems and Software (WISS'2000)*, pp. 101–110 (2000)

著 者 紹 介

土方 嘉徳（正会員）

平成 10 年大阪大学大学院基礎工学研究科物理系専攻修士課程修了．同年日本アイ・ビー・エム (株) 東京基礎研究所入所．現在，大阪大学大学院基礎工学研究科システム創成専攻助手．博士 (工学)．知的 Web 技術，パーソナライゼーション，テキストマイニングの研究に従事．平成 15 年 2 月計測自動制御学会学術奨励賞，平成 15 年 8 月電気学会「論文誌 C 発刊 30 周年記念」特集最優秀論文賞，各受賞．IEEE ほか会員．