❶嗜好抽出・情報推薦の基礎理論

1. 嗜好抽出と情報推薦技術

土方嘉徳
大阪大学大学院基礎工学研究科

本稿では，嗜好抽出技術および情報推薦技術に関して，研究の歴史的な発展の経緯，基本方式，評価指標，代表的な課題について紹介する．また，今後の方向性として，人間を中心とした推薦系の構築を提案し，そのような研究の一事例を紹介する．

嗜好抽出・情報推薦の現在

近年，ユーザの嗜好に応じた推薦サービスが注目を集めている．特に，本やCD，ビデオなどを扱う世界最大のオンラインショップに成長したAmazon.com☆1の影響が大きいといえる．Amazon.comでは過去の購買履歴やアンケートなどから，ユーザが好みそうな商品を推薦してくれる．このような推薦サービスは，オンラインショップにとどまらず，現在ではニュースのポータルサイトにおけるニュース記事の推薦（MSN Newsbot☆2，Google News☆3）やハードディスクレコーダにおける番組推薦（ソニーのスゴ録☆4）などでも，実現されている．

また，上記のサービスは，ユーザの過去の閲覧（行動）履歴を用いて，ユーザの嗜好をモデル化しているが，ユーザがそのときに選択しているアイテム（商品やニュース記事の総称）に関して，他のユーザがそのアイテムと同時に見た（購入した）アイテムを推薦するサービスもある．こちらは，文房具販売（アスクル☆5），CD，ビデオ，ゲームソフトのレンタルと販売（TSUTAYA online☆6），カジュアルウェアの販売（ユニクロ☆7），総合通信販売（ニッセン☆8）などで導入されている．

上記のような推薦は一般には，「情報推薦」あるいは「レコメンデーション」と呼ばれる．情報推薦を実現するシステムは，「推薦システム」あるいは「レコメンダ」と呼ばれる．また，広義には，アイテムを何らかの情報（ユーザの嗜好情報やセキュリティ情報など）を元に取捨選択する技術のことを「情報フィルタリング」と呼ぶ．これらを実現する技術には大きな差がないため，特に使い分けすることなく，同じ意味で用いられることも多い．推薦システムでは，ユーザの嗜好を抽出する必要があるが，これに必要な技術を「嗜好抽出技術」あるいは「ユーザプロファイリング技術」[1]と呼ぶ．

また，インターネットビジネスの世界では，パーソナライゼーション[2]という言葉もよく用いられる．この定義は，「ユーザに適した情報をユーザに適した形式で提示する技術やサービス」をいう．情報推薦は，ユーザに適した情報を選択することであり，パーソナライゼーションの一手法と呼ぶことができる．情報のユーザ適応という観点では，さらに提示する情報の内容そのものも，嗜好に合わせて書き換えることも考えられるが，実用的な手法が提案されるには至っていない．

一方，情報の提示形式をユーザに適応させることに関しては，研究レベルではいくつかシステムはあるものの（主に適応型ハイパーテキスト[3), 4)]の研究分野で，教育向けに開発されてきた），実用化された例は少ない．実用化に至っていないのは，提示形式に関しては，計算機と人間とのインタラクションについての深い分析と設計が求められるからだと思われる．商用レベルでは，iGoogle☆9やMy Yahoo!☆10など，ポータルサイトのトップページなどのカスタマイゼーション（ユーザ自身による提示形式の変更）で実現されている．ビジネスの世界では，カスタマイゼーションをパーソナライゼーションに含めて考えることも多い．これらのサービスをまとめると，**図-1**のようになる．本稿では，まず嗜好抽出・情報推薦の研究の発展過程を概観することとする．次に，情報推薦の基本方式と情報推薦に必要となる嗜好情報抽出技術（ユーザプロファイリング技術）について説明する．さらに情報推薦の評価指標と情報推薦の課題について述べ，今後の情報推薦技術の方向性について述べる．

☆1 http://www.amazon.com
☆2 http://newsbot.msnbc.msn.com/
☆3 http://news.google.com
☆4 http://www.sony.jp/
☆5 http://www.askul.co.jp/
☆6 http://www.tsutaya.co.jp/
☆7 http://store.uniqlo.com
☆8 http://www.nissen.co.jp/index.htm
☆9 http://www.goole.co.jp/ig/
☆10 http://my.yahoo.co.jp

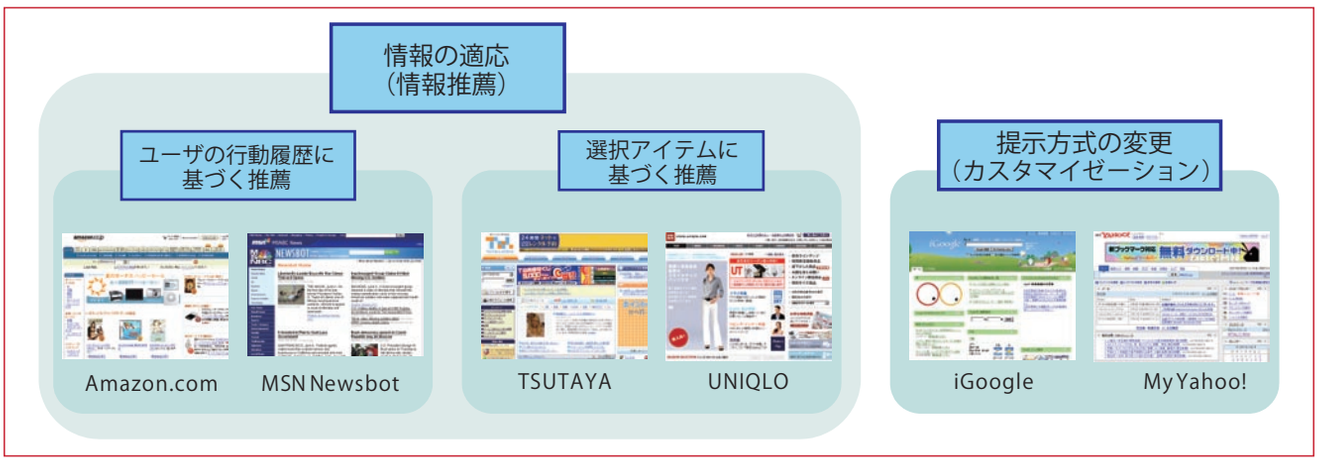

図 -1　実用化されたパーソナライゼーション技術

	1980年代後半	1990年代前半	1990年代後半	2000年代前半
研究のトレンド	インターネットプロトコルを利用した情報フィルタリングシステムの提案	半構造化文書と情報検索からのアプローチによる情報フィルタリングの研究	協調フィルタリングおよび機械学習による内容に基づくフィルタリング方式の研究	実用的な協調フィルタリング方式に関する研究
代表的研究と学術誌特集	● Information Lens (87)	● Fortzらの研究 (90) ● GroupLens (94) ● Moritaらの研究 (94) ◆ CACM特集号 (92)	● Ringo (95) ● WebWatcher (97) ● ANATAGONOMY (97) ◆ CACM特集号 (97)	● Item- based方式 (01) ● 推薦根拠の提示 (00) ● TextExtractor (02) ◆ CACM特集号 (00)
時代背景	・研究機関へのインターネットの普及	・WWW創成期	・一般ユーザへのWWWの普及 ・Amazon.comの誕生	・一般企業のWWWの本格的ビジネス利用 ・ネット企業における推薦サービスの導入

図 -2　情報推薦に関する研究の歴史

嗜好抽出・情報推薦の過去

現在，注目を集めている推薦サービスであるが，研究対象と見た場合には，意外と歴史は古い（**図 -2** 参照）．情報科学の研究者が情報のフィルタリングに注目し始めたのは，研究機関でインターネットが利用できるようになり，研究者が電子メールやネットニュース☆11 を利用するようになった 1980 年代後半である．先駆的な情報フィルタリングシステムは，1986 年に MIT の Malone らが開発した Information Lens[5] である．彼らは，電子メールやネットニュースのメッセージのヘッダに，配送や返信に関する情報だけでなく，内容に関連する場所や時間，トピックなどの情報も構造化した．フィルタリングは，読み手が作成したルールに基づいて行われる．半構造化文書の特性を知的処理に応用するという考え方は斬新なものであった．

1990 年代になると，情報検索の分野の研究者が競って，情報フィルタリングの研究を始めた．代表的な研究としては，Foltz らが 1990 年に行ったものが挙げられる．彼らは，適合性フィードバック（詳しくは，後述）が，ユーザが明示的に興味のあるキーワードを列挙するよりも，長期間利用すればユーザの興味をより正確にモデル化できることを示した[6]．また，Loeb らをゲストエディタとして，CACM（ACM の学会誌で“Communications of ACM”の略）にて特集号も組まれた[7]．

1990 年代後半になると，情報推薦・情報フィルタリングの研究が爆発的に人気を集めることとなる．このころ，Windows95 の発売で，一般のユーザにもインターネットが普及し始めたことも一因と思われる．その火種となった研究が，Resnick らが 1994 年に行った GroupLens という研究[8] である．この研究で初めて協調フィルタリングの考え方が厳密にアルゴリズムとして定式化された．それ以降，Ringo[9] をはじめとして，GroupLens のアルゴリズムの改良に関する研究が盛んに行われた．また，情報検索の延長としての情報フィルタリングではなく，情報フィルタリングを一種の文書分類問題と捉え，さまざまな機械学習アルゴリズムを適用す

☆11 Web 上の掲示版やニュース記事のポータルサイトとは異なる．複数のサーバで主にテキストデータを配布・保存するコミュニケーションツールで，現在はあまり使われなくなっている．

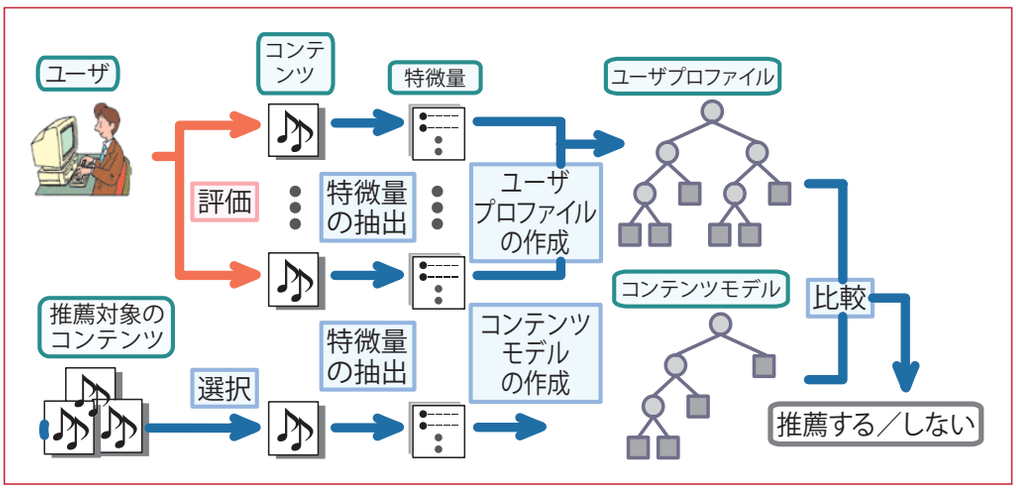

図-3　コンテンツに基づくフィルタリングの概要

る研究（代表例としてはWebWatcherが挙げられる[10]）も多く行われるようになった．情報フィルタリングというアプリケーションが，機械学習の実験場のようになっていたともいえる．また，関連する国際ワークショップや国際会議も行われ，情報検索，人工知能，データ工学，ヒューマンインタフェース，CSCWなどの伝統的な研究分野においても盛んにセッションが組まれるようになった．また，Resnickらをゲストエディタとして，CACMにて特集号も組まれた[11]．

2000年代になると，情報推薦ブームもやや落ち着きを見せた．しかし，前章で紹介した推薦サービスのほとんどで使われていると思われるアイテムベースの協調フィルタリング[12]（詳細は，後述）が提案されたり，推薦の根拠を示す試みがなされたり[13]，より正確にユーザの興味に関する情報を獲得する方法が提案されたり[14]と，より実用指向の堅実な研究が行われるようになった．また，嗜好抽出・情報推薦が中心ではないが，CACMにてパーソナライゼーションの特集号（比較的実用寄りの特集）が組まれ[2]，着実に実用化が行われてきたことが，改めて認識させられた．これらの努力が今日の情報推薦の理論体系の礎になり，このときの成果が今改めて情報推薦が注目を集めている理由であると考えられる．

情報推薦の基本方式

本章以降では，ユーザの過去の行動履歴を基に推薦を行う方式に焦点を当てて解説していく．この方式では，過去の行動履歴からユーザの嗜好に関する情報を獲得し，それをモデル化する必要がある．このモデル化した嗜好情報をユーザプロファイル（user profile）と呼ぶ．また，嗜好情報を獲得しモデル化することを，嗜好抽出または

☆12 厳密には，ユーザプロファイルとそれを用いる情報推薦方式には密接な関係があり，方式によっては両者を明確に切り分けられないこともある．

ユーザプロファイリングと呼ぶ☆12．

情報推薦の方式には，一般的には，(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．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．

【コンテンツに基づくフィルタリング】

コンテンツに基づくフィルタリングは，従来の情報検索技術の影響を濃く受けている．その基本的な考え方は，情報検索の分野で提案された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）にある．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，それらのページに近いページを再度検索してくれるというものである．コンテンツに基づくフィルタリングにおいては，ユーザからの行動履歴を基に，ユーザプロファイルを変更することになる．

コンテンツに基づくフィルタリングの概要は**図-3**で表すことができる．基本的な考え方としては，推薦対象のコンテンツからコンテンツの特徴量を抽出する．コンテンツがテキストの場合は，キーワードの出現頻度などで表される．音楽データや映像データなどのマルチメディアコンテンツであるときは，テンポや周波数成分，色情報，差分画像情報などになる．抽出した特徴量は，以下で説明する方式に合わせてモデル化しておく．このモデル化したものをコンテンツモデルと呼ぶ．ユーザからも，そのコンテンツに対する評価やアンケートな

どから，コンテンツの特徴量に関する嗜好情報を抽出し，モデル化する．これがユーザプロファイルとなる．推薦は，コンテンツモデルとユーザプロファイルを比較することで行われる．コンテンツに基づくフィルタリングは，大きく分けると，(1) ルールベース方式（rule-based method），(2) メモリベース方式（memory-based method），(3) モデルベース方式（model-based method）の 3 種類に分けられる．

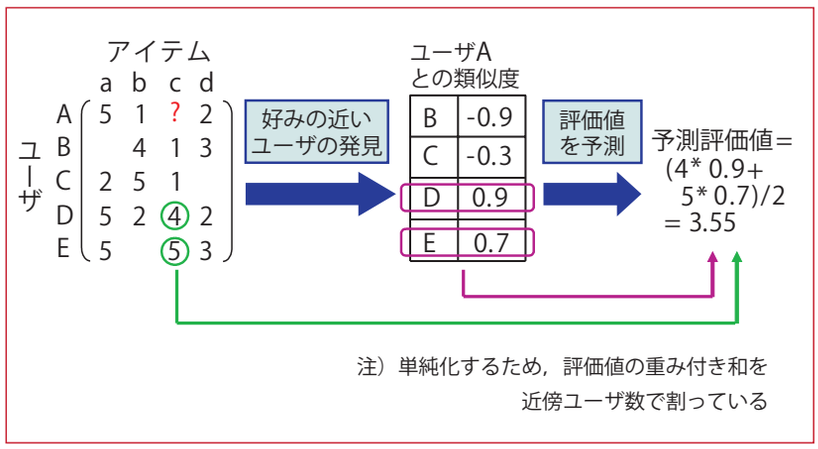

図 -4　協調フィルタリングの概要

●ルールベース方式

ドメインにおける常識やあらかじめ得られた知見，ビジネスルールなどに基づき，人手でフィルタリングのルールを設計しておく方式である．前記の知見は，データマイニング（たとえば，POS データのデータマイニング）などにより獲得することも多い．推薦対象のコンテンツがテキストである場合は，ユーザプロファイル中で，あるキーワードに高い重みが付いている場合に，特定のコンテンツを提示するようなルールが設定される．実現したい推薦サービスがあらかじめ決まっている場合，少ないコストで実現することができる．

●メモリベース方式

コンテンツモデルとユーザプロファイルの両方をベクトルで表し，ベクトル空間上での距離により，推薦するか否かを決定する方式である．情報検索におけるベクトル空間モデルと同じ考え方である．推薦対象のコンテンツがテキストである場合は，ベクトルはキーワードの出現頻度で表され，tf・idf などのキーワードに重みを付ける方法が適用される．機械学習の観点から言うと，k-nearest neighbor 法（k-NN 法）が用いられることが多い．ユーザプロファイルの更新は，

Rocchio の式

$$\boldsymbol{q}' = \alpha \boldsymbol{q} + \frac{\beta}{|D_R|} \sum_{\boldsymbol{d}_i \in D_R} \boldsymbol{d}_i - \frac{\gamma}{|D_N|} \sum_{\boldsymbol{d}_i \in D_N} \boldsymbol{d}_i$$

が用いられることが多い．ここで，$\boldsymbol{q}'$ は，ユーザのクエリ（情報フィルタリングではユーザプロファイル），D_R と D_N は，それぞれユーザが閲覧した文書のうち興味があるとした文書，興味がないとした文書である．

●モデルベース方式

過去に閲覧／購読したアイテムに対する評価値から，一般的な興味の傾向をモデル化し，ユーザプロファイルとする方式である．新たなコンテンツが発生すれば，それをベクトル形式などでモデル化し，上記ユーザプロファイルと比較する．一般的な興味のモデルと書いたが，実際には閲覧／購読を機械学習の教師信号と考え，コンテンツモデルとそれに対する正負の判断という組を，機械学習のアルゴリズムに入力して学習することで得られる．すなわち，機械学習における学習後のモデルがユーザプロファイルとなる．

アイテムが文書の場合は，本方式は一種の文書分類と捉えることができる．具体的には，文書から文書ベクトルを生成する．ユーザは興味あり／興味なしという評価を文書に対して付けているとする．この評価は，文書のカテゴリと考えられる．これらの組を機械学習アルゴリズム（たとえば，ベイズ分類子，ニューラルネットワーク，SVM など）にかけることで，興味あり／興味なしを判断するモデルを生成する．

学習には時間がかかるが，推薦実行時には高速な処理が可能である．

【協調フィルタリング】

協調フィルタリングの概要を**図 -4** に示す．ただし，この図では後述するメモリベース方式のうちのユーザベース方式の概要を示している．協調フィルタリングでは，アイテムの内容を見ない．持っている情報としては，ユーザがどのアイテムにどのような評価値（図では 5 段階）を付けたかという情報だけである．ここではユーザ A のアイテム c に対する予測評価値を求めている．協調フィルタリングでは，まず対象ユーザと好みの近いユーザ（ここでは，ユーザ D とユーザ E）を特定する．好みの近さは，図の行列の行をベクトルとしたベクトル間の類似度として計算される．次に，好みの近いユーザが，対象のアイテムにどのような評価値を付けていたかに基づいて，予測評価値が計算される．アイテムベース方式では，上記類似度の計算が，アイテム間で計算される．また，上記の例では，ユーザ間の類似度の計算を，すべて

のユーザ同士で比較しているが，コンテンツに基づくフィルタリングと同様，これらの関係を一般的な形式でモデル化しておく方法論もある．これをモデルベース方式と呼ぶ．以下では，メモリベース方式におけるユーザベース方式とアイテムベース方式の厳密な定義と，モデルベース方式の基本的な考え方について述べる．

●ユーザベース方式（メモリベース）

ユーザ集合を $A = \{a_1, a_2, \ldots, a_n\}$，アイテム集合を $B = \{b_1, b_2, \ldots, b_m\}$ とし，ユーザ a_i がアイテム b_k につけた評価値を $r_i(b_k)$ とする．ユーザベース方式のアルゴリズムは以下のようである．

- 近傍形成 a_i を注目しているユーザ（active user）としたとき，すべての $a_o \in A \setminus \{a_i\}$ に対する類似度 $s(a_i, a_o)$ が，r_i と r_o の類似度に基づいて計算される．一般的には，$s(a_i, a_o)$ の計算にはピアソン相関やコサイン距離が用いられる．最も似ているユーザ上位 M 人が a_i の近傍メンバになり，その集合を $neighbor(a_i) \subseteq A$ と表す．
- 評価値予測 $a_o \in neighbor(a_i)$ が評価をつけており，かつ a_i が未評価であるアイテム b_k すべてに対して，嗜好の予測値 $p_i(b_k)$ が計算される．

$$p_i(b_k) = \overline{r_i} + \frac{\sum_{a_o \in A'_i} s(a_i, a_o) \cdot (r_o(b_k) - \overline{r_o})}{\sum_{a_o \in A'_i} |s(a_i, a_o)|} \quad (1)$$

$$A'_i := \{a_o | a_o \in neighbor(a_i)\}$$

$$\overline{r_i} = \sum_{i=1}^{n} r_i(b_k)/n$$

最終的に，予測評価値 p_i に基づいて上位 N 個の推薦リスト $L_{p_i} : \{1, 2, \ldots, N\} \rightarrow B$ が計算される．関数 L_{p_i} は最も高い予測値を持つアイテムを 1 位とした降順の推薦ランキングを示す．

●アイテムベース方式（メモリベース）

アイテムベースの協調フィルタリングは，アイテム間の類似度 s が計算される．2 つのアイテム b_k，b_e に対して各ユーザが近い評価値を付けているとき，これらのアイテムの類似度 $s(b_k, b_e)$ は高くなる．類似度の計算にはコサイン距離を用いることが多い．各 b_k に対して最も似ているアイテム上位 M 個が近傍 $neighbor(b_k) \subseteq B$ と定義される．予測値 $p_i(b_k)$ は次のように計算される．

$$p_i(b_k) = \frac{\sum_{b_e \in B'_k} (s(b_k, b_e) \cdot r_i(b_e))}{\sum_{b_e \in B'_k} |s(b_k, b_e)|} \quad (2)$$

$$B'_k := \{b_e | b_e \in neighbor(b_k)\}$$

上位 N 個の推薦リスト L_{p_i} の最終的な計算は，ユーザベースの協調フィルタリングの手順に従う．

●モデルベース方式

モデルベースの方式は，ユーザやアイテム間の関係をあらかじめ一般化してモデル化しておく．代表的な手法としては，クラスタリングを用いる手法，ベイジアンネットや EM アルゴリズムなどの確率モデルを用いる手法，マルコフモデルなどの時系列モデルを用いる手法などがある．紙面の都合上，すべてを紹介することはできないので，最も基本的なクラスタリングを用いる手法について説明する．

クラスタリングを用いる手法では，ある特徴を有するユーザ集合（あるいはアイテム集合）を，事前にクラスタ化しておき，そのクラスタの特徴を表す代表的なベクトルを生成しておく．たとえば，100 万人のユーザがいたとしても，音楽であれば，洋楽を主に聴くグループ，演歌を主に聴くグループ，懐メロを主に聞くグループなどの数個のグループに分割される☆13．推薦の実行時にはその数個のグループとの類似度を計算するだけでよい．そのため，実行時の速度が速い点が特徴である．

クラスタリングのアルゴリズムとしては，従来から存在する K-means 法や凝集法などが用いられる．これらのアルゴリズムでは，クラスタ数を決定する必要があるが，この数が推薦の質にも影響を及ぼしてしまう．そのため，実際の推薦のパフォーマンスを測り，試行錯誤的に決定する必要がある．

嗜好抽出技術

本章では，嗜好抽出技術（ユーザプロファイリング技術）において，考慮すべき問題を述べた後に，代表的な手法を紹介する．

【抽出粒度と適合性フィードバックとの関係】

どのぐらいの粒度で興味に関する情報を獲得する必要があるかは，用いるフィルタリング方式によって異なる．協調フィルタリングでは，アイテムを処理の単位としているので，アイテム単位でよい．たとえば，アイテムが新聞記事であれば，その記事の単位でよい．しかし，コンテンツに基づくフィルタリングでは，キーワードやトピックレベルで処理を行っている．そのため，なるべく

☆13 手法としてはこれらに分類を行うのではなく，ベクトルの特徴から機械的にグループ化を行うだけである．グループ化されたものを人が後から見たときに，そのジャンルや特徴に気付くものである．

興味の対象を限定できる方がよい．

コンテンツに基づくフィルタリングは，適合性フィードバックとの関連が強いが，従来の適合性フィードバックには大きな問題が２つある．１つ目は，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ことである．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推薦の精度を低下させるということが懸念される．もう１つの問題は，ユーザに興味を持った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．これらの両方を考慮することが，嗜好抽出技術を開発する際のポイントとなる．

【明示的手法と暗黙的手法】

嗜好抽出技術には，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．以下では，それぞれの手法の詳細を述べる．

（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付けさせる方法の２種類に分類できる．短期間のうちに効果的な情報推薦を行うには (i) の方法は有用である．(ii) は間接的に，興味を表すキーワードやトピックを推定することになるが，長期間評価付けを続けていれば，(i) を上回る効果が得られることが知られている．

（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，閲覧した情報のすべてにユーザが興味を持ったと仮定する手法と，何らかの手がかりからユーザが閲覧した情報に興味があったかなかったかを判定する手法の２種類がある．前者には，(i) Web ページのアクセス履歴が用いられる．後者で用いられる手がかりとしては，(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，などが挙げられる．

(i) の方法は，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．ユーザの興味に関する情報を取得するためのシステムを，Web サーバやプロキシサーバなど，１カ所で動作させることができるため，最も現実的な方法である．具体的には，Cookie やユーザ認証，IP アドレスなどを用い，個人を識別する．

(ii) の方法では，アクセス履歴よりも精度良く，ユーザの興味に関する情報を取得することができる．なおかつ，閲覧文書に評価付けさせる手法と違い，ユーザの負荷をなくしている．文書に対する閲覧時間とユーザの文書に対する興味の度合いとの相関があることを初めて示したのは，Morita らである [16]．閲覧時間を用いることで，どれだけ明示的な手法に近づけるかが重要となる．Morita らは，ネットニュースを用いた評価実験で，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% と報告している．まったく同じ条件での実験ではないが，閲覧文書に明示的に評価付けをさせる Foltz らの手法では，再現率 25% に対して，精度はおよそ 67% となっている [6]．閲覧時間を用いた推定でも，まずまずの結果になることが分かる．

(iii) の方法では，ブラウザ上でのマウスポインタの特殊な挙動から，ユーザの興味の対象を推定する．閲覧時間よりも，よりユーザの認知負荷の高い行動を用いることで，より精度良く興味を持ったか否かを推定している．マウスポインタの挙動を初めてユーザの興味の推定に用いたのは，Sakagami らの ANATAGONOMY[17] である．彼らは，ニュース記事に対する拡大表示の操作とスクロールの操作を基に，その記事に興味を持ったかどうかを推定している．しかし，マウスポインタを用いる最大の利点は，文書に対して興味を持ったか否かだけでなく，文書の部分に対して興味を持ったか否かを推定できる可能性がある点である．これを試みたのは，土方らの TextExtractor である．土方らは，ユーザのブラウザ上での意識しないマウス操作として，なぞり読みとリンクポインティング，リンククリック，テキスト選択を挙げ，これらの操作の下に存在したテキストを文または行の単位で抽出している．文書によっては，tf・idf よりも高い精度でキーワードを抽出できると報告している．

(iv) の方法では，ユーザに特殊なハードウェアを装着し，ユーザの視線を計測することで，注視したテキスト部分を推定している．マウスポインタの操作はユーザの個人差が大きいが，それよりは個人差なく興味に関する情報を獲得できると思われる．代表的な手法には，大野の IMPACT がある [18]．大野は，視線の滞留や横方向への移動などから，注視した度合いを算出している．

上記の嗜好抽出技術の長短所をまとめると，**表 -1** のようになる．表中で，構築時間とは，ある程度の精度の推薦を提供可能なユーザプロファイルを構築するのに必要な時間である．興味を推定する単位が文書単位であれば，ページ全体に存在するノイズとなるキーワードの影響のため，ユーザプロファイルの構築に時間がかかってしまう．実現性とは，ネットビジネスへの適用の実現可能性を示している．特殊なハードやソフト，プラグインを必要とすると，利用してくれるユーザが限られてしまう．

手法	分類	負担	正確性	興味粒度	構築時間	実現性
アンケート	明示的	×	◎	—	○	○
ページ評価	明示的	×	◎	×	×	○
アクセス履歴	暗黙的	○	×	×	×	○
閲覧時間	暗黙的	○	△	×	×	○
マウス操作	暗黙的	○	○	○	○	△
視線	暗黙的	○	○	○	○	×

表 -1　嗜好抽出手法の比較

情報推薦の評価指標

評価指標は推薦システムの質と性能を判断するために重要である．これまで提案されてきた評価指標の多くは推薦の正確さの測定に関するものである．しかし，最近では他の要因，たとえば推薦の新規性や掘り出し物を見つける性能などに関する評価指標も提案されている．以下の節では，正確さの指標と正確さ以外の指標を分けて紹介する．なお，評価指標については文献 15）の論文が詳しいので，参考にされたい．

【正確さの指標】

正確さの指標は，個々の予測の正確さを判断するものと，推薦リストを評価するものの 2 種類がある．予測の正確さの指標は，予測評価値が実際のユーザの評価値にどれだけ近いかを測定する．代表的なものに平均絶対誤差（MAE）がある．MAE は，アイテム集合 B_i 中のアイテム b_k の予測 $p_i(b_k)$ の正確さを統計的に測定する有効な手段である．MAE は以下の式で計算される．

$$|\overline{E}| = \frac{\sum_{b_k \in B_i} |r_i(b_k) - p_i(b_k)|}{|B_i|} \tag{3}$$

推薦リストの正確さの指標としては，精度（precision）と再現率（recall）がある．これらは推薦リストの集合が，注目しているユーザにとって必要かどうかを表したものである．これらの説明の前に，以下の定義をしておく．ユーザ a_i によって評価付けされたすべてのアイテムの集合を R_i とする．R_i をなるべく等サイズの互いに素な K 個のスライスに分割する．この結果，ランダムに選ばれた $K-1$ 個のスライスが，推薦のための訓練事例 R_i^x として用いられる．残りのスライス $R_i \setminus R_i^x$ が生成された推薦リストの評価に使われるテストセットとなる．テストセット中，ユーザが好きであるアイテムの集合を T_i^x と表す．

再現率は，テストセット中の好きなアイテムの総数 $|T_i^x|$ に対する，推薦リスト L_i^x 中に含まれる好きなアイテム $b \in T_i^x$ の割合として定義される：

$$Recall = \frac{|T_i^x \cap \Im L_i^x|}{|T_i^x|} \tag{4}$$

記号 $\Im L_{xi}$ は写像 L_i^x の像であり，推薦リストの全アイテムを示す．

精度は推薦リストの大きさに対する L_i^x 中に含まれる好きなアイテム $b \in T_i^x$ の割合として定義される：

$$Precision = \frac{|T_i^x \cap \Im L_i^x|}{|\Im L_i^x|} \tag{5}$$

【正確さ以外の指標】

これまでに提案された正確さ以外の指標には，以下の 5 種類がある．

（1）Coverage

coverage は，システムがどれだけのアイテムを予測可能であるかを測定する指標である．特に協調フィルタリングでは，まだ誰も評価を付けていないアイテムは推薦できないため，すべてのアイテムが推薦候補になるわけではない．システムがより多くのアイテムの予測を行えることは，そのシステムがユーザの好みのアイテムをより多く見つける可能性があることを示している．coverage は，全アイテム数に対する予測がなされるアイテムの個数の割合として測定される．

（2）Novelty と Serendipity

novelty と serendipity は推薦されたアイテムの新規性や意外性を示す指標である．推薦されたアイテムがユーザの知らない好みのものであるとき，この推薦は novelty であるという．novelty の精度と再現率をテストセット中の知らない好みのアイテムの集合 C_i^x を用いて式で表すと以下のようになる：

$$Precision(novelty) = \frac{|C_i^x \cap \Im L_i^x|}{|\Im L_i^x|} \tag{6}$$

$$Recall(novelty) = \frac{|C_i^x \cap \Im L_i^x|}{|C_i^x|} \tag{7}$$

人だけでは，予測・発見することが難しかったような，意外性のあるアイテムが推薦されることを意味する．Herlocker らは serendipity の検出のためには，推薦されたアイテムがユーザをどの程度引き付け，驚きを与えたかを測定すればよいと述べている．

（3）多様性（Diversity）

多様性（diversity）は，推薦リストのトピックに関する多様性を測定する目的で考えられた指標である．実際には，リスト内のアイテム間のトピックの類似度を計算

し，それを合計したものをリスト内の類似度として考える．アイテム間のトピックの類似度は，ジャンルや作者，その他の特性に基づいて計算される．リスト内の類似度が高いことは，多様性が低いことを示す．

(4) 発見性（Discovery ratio）

発見性（discovery ratio）は，推薦リスト内に知らないアイテムがどれだけあるか（それらは好きである必要はない）を測定するための指標である．発見性は，リスト内のアイテム数に対するリスト内の知らないアイテム数の割合として定義される．テストセット中の知らないアイテムの集合 D_i^x を用いると以下の式で表される：

$$Discovery\_ratio = \frac{\left| D_i^x \cap \Im L_i^x \right|}{\left| \Im L_i^x \right|} \tag{8}$$

情報推薦の課題

情報推薦システムの課題を，内容に基づくフィルタリングと協調フィルタリングに分けて，説明する．内容に基づくフィルタリングは，推薦の質が利用するユーザ数に影響されない点が利点である．また，まったく新しいアイテムであっても，推薦対象に含まれる点が利点である．しかし，コンテンツのメディアの種類によっては，推薦に用いる特徴量を抽出することが困難になる．ニュース記事などのテキストで表現されたコンテンツに対しては，比較的良く機能するが，音楽や絵画，映像などのコンテンツに対しては，必ずしもメタデータが付いているとは限らないため，高い質の推薦が望めないことも多い．また，ユーザプロファイルとコンテンツモデルを直接的に比較するため，serendipity の高い推薦を行うことは難しい．

協調フィルタリングは，コンテンツを解析する必要がない点が利点である．上記のような特徴量を抽出することが困難なアイテムに対しても，高い精度で推薦を行うことができる．しかし，sparsity 問題や firstrater 問題（あるいは cold-start 問題）といった，協調フィルタリング特有の問題がある．sparsity 問題とは，推薦システム全体として，扱うアイテム数に対して，評価を付けたアイテム数が少なすぎると，推薦の質が低くとどまる問題である．first-rater 問題は，まったく新しいアイテムは，誰かが 1 人でも評価付けを行わないと，推薦候補に入らない問題である．cold-start 問題は，first-rater 問題に加えて，新たにシステムを利用し始めた利用者は，ある程度の数のアイテムに評価付けを行わないと，質の良い推薦が得られない問題も考慮したものである．

また，最近注目され始めている課題として，推薦に対するユーザの飽きの問題や，不正攻撃に対する頑健性の問題がある．協調フィルタリングは，他人の付けた評価を用いて推薦を行うため（すなわち他人のお勧めのアイテムが得られるため），当初は発見性や serendipity の高い推薦が期待された．しかし，高い推薦精度が逆に災いし，好みではあるが，似たようなアイテムばかり推薦されてしまう問題が多くの商用システムで発見されている．後者は，不正なユーザが，自社のアイテムが高く推薦されるよう，不正なユーザプロファイル（自社のアイテムばかりが高く評価されているようなユーザプロファイル）を持ったユーザのアカウントを作成する問題である．または逆に，他社のアイテムが低く評価されているようなユーザプロファイルを作成する問題である．このようなアカウントが増えると，全体の推薦の質が低下してしまう．

情報推薦の未来

これまでの推薦システムは，ユーザの過去の嗜好データに基づき，統計的に好きな確率の高いアイテムを選択するという枠組みの中で，さまざまな試みがなされてきた．しかし，ユーザの推薦に対する要求は，単に嗜好に合っているか否かという単純な問題ではない．論文や技術文書の推薦などでは，推薦を外すことのリスクは大きくなる．逆に，レストランでの昼食の推薦であれば，発見性や serendipity が重要視される．また，長期的な嗜好だけでなく，ユーザのその場のコンテキストも重要視されるであろう．筆者は，今後の推薦システムの方向性の 1 つとして，より人間を中心とした推薦システムがあるのではないかと考えている．すなわち，人間と推薦システムを切り分けて考えるのではなく，推薦のメカニズムそのものに人間の積極的なインタラクションを導入し，全体の系として推薦の質を高め，ユーザの満足度を高めていくような方向性である．

筆者は，従来から上記の観点から推薦システムを構築してきたのであるが，その一例として音楽を推薦する C-baseMR というシステム [19] を挙げる．これは，モデルベースのコンテンツに基づくフィルタリングシステムなのであるが，機械学習のアルゴリズムとして決定木を採用し，評価値データからユーザプロファイルを構築後，ユーザが自由にユーザプロファイルを編集することができるシステムである．当初は，ユーザの好みを視覚化することで，ユーザは自分の嗜好を理解するようになり，積極的にユーザプロファイルを編集するようになると考えた．しかし，対象とした音楽では，ユーザは特徴量の意味を理解するのが困難であり，ユーザプロファイルを編集したくとも，どう編集すればよいか分からないユーザが多いことが分かった．そこで，特徴量の中から 2 つを選択し，それを特徴空間として視覚化し，そこに

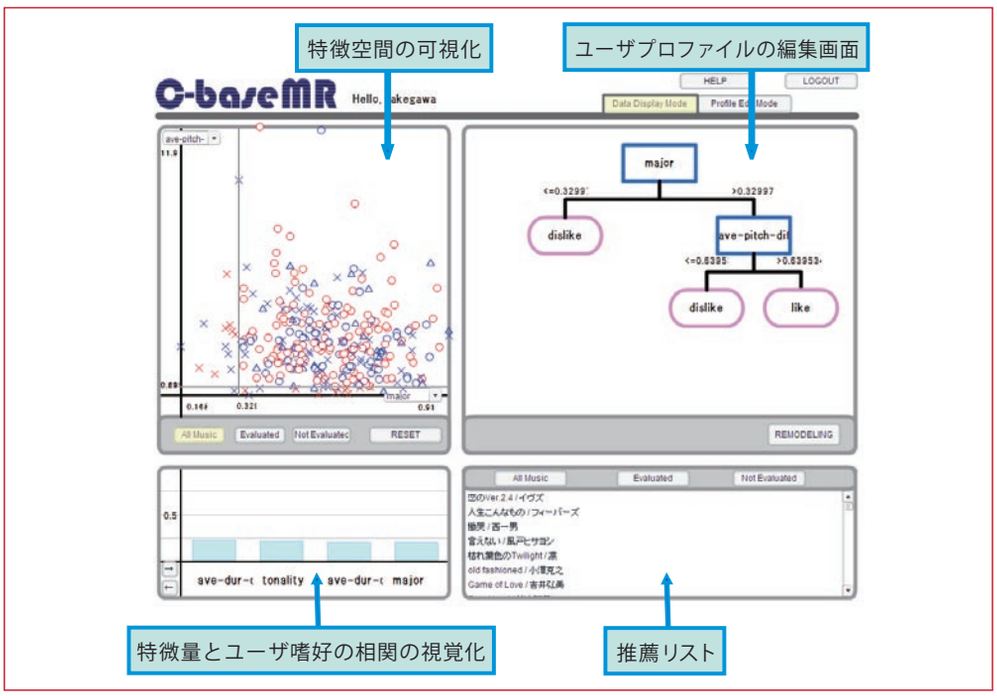

図 -5　ユーザの積極的なインタラクションを採用した音楽推薦システム（C-baseMR）

評価済みの音楽データをマッピングした．さらに，ユーザが自由に特徴空間を探索できるようにすることで，音楽特徴の理解と，自分の嗜好の理解を支援することとした（**図 -5** 参照）．この工夫により，ユーザはより積極的に音楽を発見しようとすることが確かめられた．

おわりに

研究分野としての推薦システムの歴史は長い．これまで，情報科学の観点から多くの研究がなされてきた．一方ビジネスの世界では競争の自由化が促進され，各企業からの推薦システムに対する期待は高まる一方である．しかし，積極的な推薦システムの導入には踏み切れない企業も多いのが現実である．我々，情報科学の研究者は，既存のアルゴリズムの実験場としての情報推薦からは決別すべき段階にあるのではないかと考える．人間の情報に対する要求は単純ではなく，ある程度ドメイン依存な仕組みを導入したり，ユーザをも推薦システムの系の一部として利用する必要があると考える．現実に直面し，それを乗り越える推薦システムの登場に期待したい．

参考文献

1）土方嘉徳：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，人工知能学会誌，Vol.19, No.3, pp.365-372 (2004).
2）Riecken, D. (ed.) : Personalized Views of Personalization, *Comm. of the ACM*, Vol.43, No.8, pp.26-158 (1992).
3）Brusilovsky, P. : Methods and Techniques of Adaptive Hypermedia, User Modeling and User-Adapted Interaction, Vol.6, No.2-3, pp.87-129 (1996).
4）Brusilovsky, P. : Adaptive Hypermedia, User Modeling and User Adapted Interaction, Vol.11, No.1-2, pp.87-110 (2001).
5）Malone, T. W., et al. : Semi-structured Messages are Surprisingly Useful for Computer-Supported Coodination, *Proc. of CSCW'86*, pp.102-114 (1986).
6）Foltz, P. W. : Using Latent Semantic Indexing for Information Filtering, *Proc. of ACM Conference on Office Inforamtion Systems*, pp.40-47, (1990).
7）Loeb, S. and Terry, D. : Information Filtering, *Comm. of the ACM*, Vol.35, No.12, pp.26-81 (1992).
8）Resnick, P., et al. : GroupLens : An Open Architecture for Collaborative Filtering of Netnews, *Proc. of CSCW'94*, pp.175-186 (1994).
9）Shardanand, U. and Maes, P. : Social Information Filtering : Algorithm for Automating 'Word of Mouth', *Proc. of CHI'95*, pp.210-217 (1995).
10）Joachims, T., Freitag, D. and Mitchell, T. : WebWatcher : A Tour Guide for the World Wide Web, *Proc. of IJCAI'97* (1997).
11）Resnick, P. and Varian, H. : Recommender Systems, *Comm. of the ACM*, Vol.40, No.3, pp.56-89 (1997).
12）Sarwar, B., Karypis, G., Konstan, J. and Riedl, J. : Item-based Collaborative Filtering Recommendation Algorithms, *Proc. of WWW'01*, pp.285-295 (2001).
13）Herlocker, J., et al. : Explaining Collaborative Filtering Recommendations, *Proc. of CSCW'00*, pp.241-250 (2000).
14）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 (Feb. 2002).
15）Herlocker, J., et al. : Evaluating Collaborative Filtering Recommender Systems, *ACM Transactions on Information Systems*, Vol.22, No.1, pp.5-53 (2004).
16）Morita, M. and Shinoda,Y. : Information Filtering Based on User Behavior Analysis and Best Match Text Retrieval, *Proc. of the 17th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pp.272-281 (1994).
17）Sakagami, H. and Kamba, T. : Learning Personal Preferences on Online Newspaper Articles from User Behaviors, *Proc. of WWW'97* (1997).
18）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 *WISS'2000*, pp.137-146 (2000).
19）Hijikata, Y., et al. : Content-based Music Filtering System with Editable User Profile, *Proc. of ACM SAC 2006*, pp.1050-1057 (2006).

（平成 19 年 7 月 21 日受付）

土方　嘉徳（正会員）
hijikata@sys.es.osaka-u.ac.jp
1998 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・エム東京基礎研究所に入社．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．2002 年より大阪大学大学院基礎工学研究科助手．2007 年より同講師．工学博士．