2 Web上の情報推薦システム

寺野　隆雄　筑波大学ビジネス科学研究科
terano@gssm.otsuka.tsukuba.ac.jp

情報推薦システムはWeb上のよろず屋さん

情報推薦システム―リコメンダシステム（Recommender System）あるいはリコメンデーションシステム（Recommendation System）―は，Web上で情報発信する人々にとって，利用者に適切な情報を提供する手段を与える．この典型的な例はe-コマースである．

商品を消費者に売るために情報を提供する場合，やらなくてはならない作業は，以前から実はあまり変わっていない．それは次の3つにまとめられる．売り込みたい商品の性質を知ること，買いたいと思っている消費者の性向を知ること，そして，消費者―お客様―に気に入られることである．これは，昔から「よろず屋さん」や「御用聞き」が無意識にやっていたことである．消費者の側でも，近所づきあいの中から「おいしい」情報を仕入れて，それに従った行動をとるのは当たり前のことであった．絶えず変化する膨大なWeb環境の中で，このような仕組みをコンピュータやネットワーク上に構築するのが情報推薦システムのねらいである．**図-1**にe-コマースと情報推薦システムの関係を示す[6]．

本稿では，情報推薦システムの基本的な考え方と技術的な実現方法を紹介する．また，我々が開発に携わった最近のシステム例についても述べる．

典型的な情報推薦手法：協調フィルタリングとコンテンツ分析

現在利用されている情報推薦の方法は，チェックボックス方式・ルールベース方式・リコメンダシステム方式

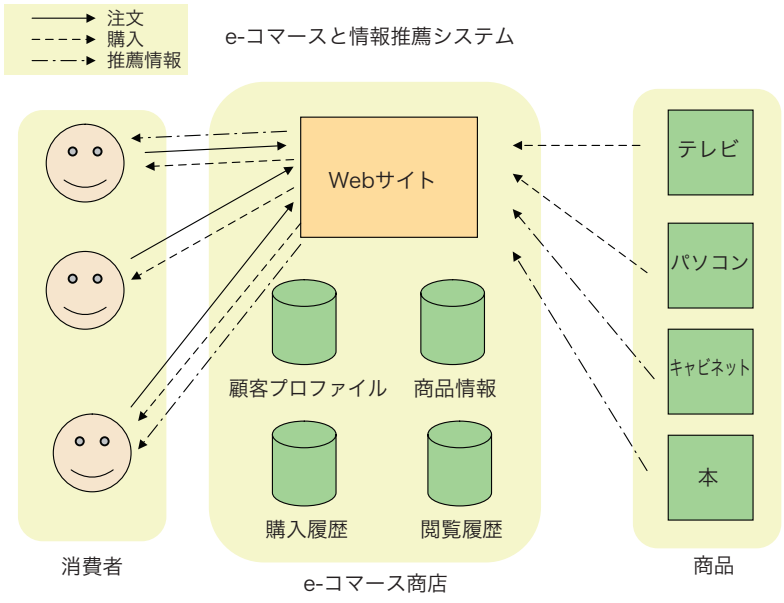

図-1　e-コマースと情報推薦システムの関係

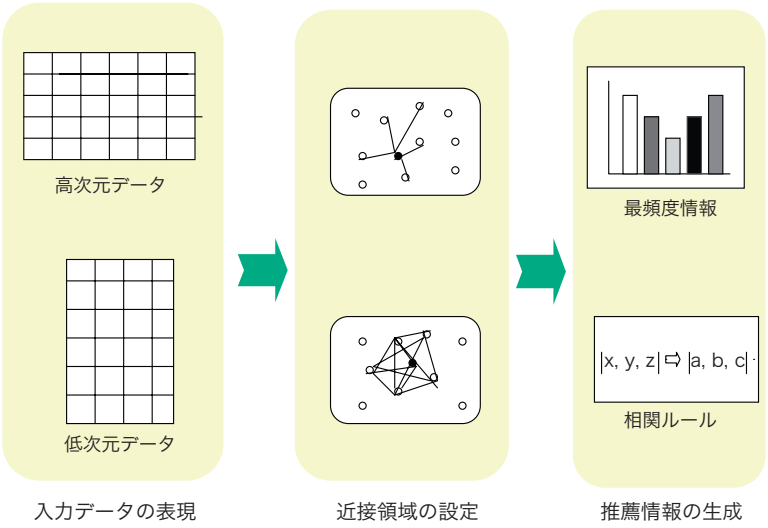

図-2　リコメンダシステムによる推薦情報生成の３フェーズ

の３種類に大別することができる．

最初のチェックボックス方式は，従来からダイレクトメールサービスなどで用いられているのと同様な方式である．すなわち，アンケート情報などを用いて消費者から直接顧客情報を取り込んでそれを情報推薦に利用するものである．

次のルールベース方式は，e-コマースのサイト運営者が消費者の特性情報（消費者プロファイル）の各分類属性（セグメント）に対応する推薦ルールを設定する方式である．最近は機械学習あるいはデータマイニングの手法を利用して情報推薦システムを実現することが可能になっている．

最後のリコメンダシステム方式は，研究開発の中心的な話題となっている方式である．この方式は，消費者プロファイルや行動・購買履歴情報を収集分析し，別に収集した商品情報と組み合わせて，適切な推薦情報を生成するものである．これは**図-2** に示すように入力データの表現，近接領域の設定，推薦情報の生成の３つのフェーズから構成される[5]．

入力データの表現においては，商品種類数と顧客数が増加した場合の処理時間の拡大の程度，商品種類数と顧客数の組合せの疎密度，商品種類や顧客の類似度を的確に表すことが重要である．これらのデータは基本的には表形式の関係データベースに表現されていることを仮定する場合が多い．近接領域の設定においては，顧客間の距離を適切に設定し適度のクラスタを作ることが重要である．これはどの顧客とどの顧客が類似しているか，またどの商品とどの商品が類似しているかを測定する尺度を定めることになる．推薦情報の生成においては最頻度の情報を推薦する方法やデータマイニングにおける相関ルールを利用した情報を推薦する方式などが知られている．しかし，意外性のある情報を推薦するためにはこれだけでは不十分な場合も多い．情報推薦システムにおいては，顧客のニーズに完全に一致した情報はむしろ好まれず，どのような情報に興味を持たれるのかを分析することは重要である．

リコメンダシステム方式はさらに３種類に分類される．(1) 消費者情報を重視して同じ好みを持つコミュニティを生成・利用する方式を協調フィルタリング法（Collaborative Filtering Method），(2) 商品データの分析情報を重視して適切と思われる商品を推薦する方式をコンテンツ分析法（Contents Analysis Method），(3) 両者を融合した方式をハイブリッド法（Hybrid Method）．

実際にはこれらの３種類の推薦情報生成方式は補完的に用いられることが多い．消費者がe-コマースサイトに登録した直後は，チェックボックス方式で情報提供を開始し，サイトの利用が進み履歴情報が蓄積されるとリコメンダシステムによって推薦情報を自動生成してサービスを提供する．また，これらの推薦情報を分析して専門知識を抽出し，ルールベース方式でより知的な情報提供を行う．

協調フィルタリングによる推薦生成

協調フィルタリングは，消費者の行動や購買履歴から顧客の好みを推測し，膨大な種類の商品の中から顧客の嗜好に合ったものを推薦情報として提供する手法の１つである（**図-3**）．この方式の特徴は，商品情報に関する消費者の嗜好を評価し，類似した嗜好を持つ消費者のグ

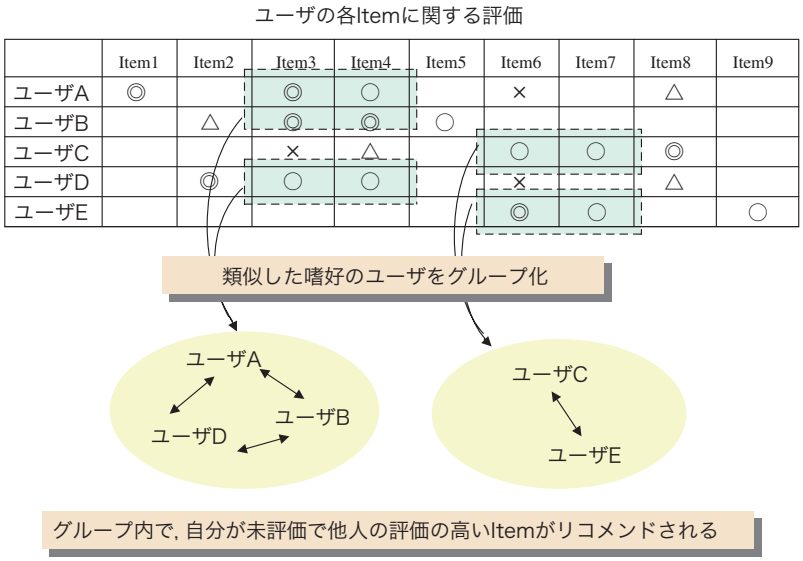

図-3　協調フィルタリングによる推薦情報の生成

ループを生成することにある．推薦情報は，類似した嗜好を持つほかの消費者が高く評価した商品のうち，自分がまだ購入していないものとして与えられる．そのためには，まず（1）多くの消費者の購買履歴を収集する．（2）特定消費者がある商品を新たに購入すると，同じ商品を購入した過去の消費者を検索してグループ化する．（3）グループ内の消費者が共通して購入した商品群をもとに，推薦情報を生成するというステップが必要である．

図-3 中では，購買履歴として５人（ユーザA,B,C,D,E）のものが集められている．◎○△×はそれぞれユーザが購入した商品（Item）であり，それぞれこの順に買った商品を評価しているものとする．表中の空欄はユーザが購入していないことを表す．この表から次のことが分かる．

まず，類似した商品を購入したユーザは類似した性向を持っていると仮定すると，購入したItemの類似性から，ユーザA,B,Dが１つのグループをユーザC,Eが２つ目のグループを形成する．ユーザAに対する推薦情報は，Aがまだ購入していないが，ユーザBが購入し○印で高い評価を与えているItem5や，Dが同様に最高の評価◎を与えたItem2となる．

この手法ではコンテンツの内容に関する分析はまったく行わない．したがってコンテンツに関して必要な情報はそのIDだけである．協調フィルタリング方式の利点は次の２点である．

- 推薦対象に関する制約が存在しない．
- 他の消費者の評価情報を用いることにより，過去に評価した情報に類似した情報のみが推薦される状況を回避できる．

すなわち，協調フィルタリングは商品のコンテンツを分析することなしに，購買情報のみから，消費者のコミュニティを生成する仕組みであるということができる．一方で次の４点が欠点として指摘されている．

- 推薦対象となるコンテンツの量に対して十分な数の消費者が必要となる．
- 消費者の誤入力による影響が他利用者に及ぶことが防げない．
- 新しい商品が登録されても誰かが評価を下すまで推薦対象にならない．
- コンテンツに対する分析を行わないため，内容が類似しているが，IDの異なる商品に関する消費者を結びつけることができない．

コンテンツ分析による推薦

コンテンツ分析による推薦では，商品の持つコンテンツ情報を分析して消費者に適したものを提供する．この基本的な方法は**図-4**のとおりである．この方式の特徴は，商品コンテンツ情報の特徴表現と消費者プロファイルの興味表現とを比較することによって推薦情報を決定する．この過程では，情報検索分野の技術が主に用いられる．そのためには，まず（1）商品のコンテンツを分析し，その特徴を表現したコンテンツデータベースを作成する，（2）消費者の購買履歴とそれに対応するコンテン

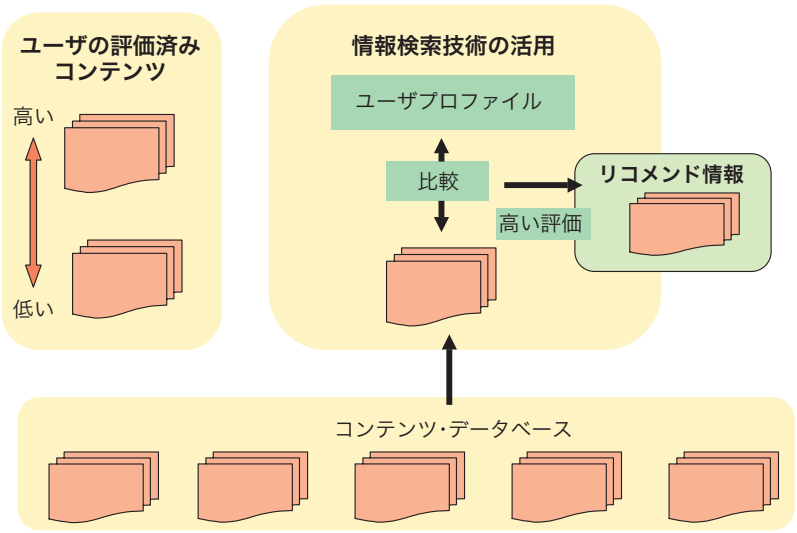

図-4　コンテンツ分析による推薦情報の生成

ツ情報を収集する，(3) 消費者が高い評価を与えた商品と類似したコンテンツを持つ商品をもとに，推薦情報を生成するというステップが必要である．

コンテンツ分析による推薦情報の生成では，協調フィルタリングの場合に見られたような商品種類数と消費者数の規模の大きさは要請されない．ところが欠点として以下の 3 点が挙げられている．

- 推薦の対象とするコンテンツの内容がコンテンツ分析の技術によって制限される．現状では情報検索とテキストマイニングの技術がその中心であるために，コンテンツとしてはテキスト情報以外のものはうまく扱えない．
- 推薦される情報は利用者プロファイルに対して高い類似度を持った情報である．したがって利用者はすでに評価した情報と類似した情報ばかりを提供されるという傾向がある．
- コンテンツ分析において利用者の明示的な評価のみが入力情報である場合が多い．したがって利用者にコンテンツ評価作業をさせるための動機付けが必要である．

TwinFinder：意外性の高い推薦情報を提供するリコメンダシステム

本章では，我々が開発した情報推薦システム TwinFinder の研究例を紹介する [2]．TwinFinder は日本における洋書販売ビジネスを対象としており，購買履歴と書籍情報を用いて，顧客ごとに適切な推薦情報を自動生成する，コンテンツ分析方式のシステムである．TwinFinder は，推薦情報の幅を広げる工夫をすることにより，コンテンツ分析方式の欠点を解消することを狙う．我々が目標とするシステムの規模は扱う商品の種類は数百万点，顧客数は数万人のオーダーである．

TwinFinderのリコメンド情報生成手順

図-5 は，TwinFinder がリコメンド情報を生成する基本手順である．

テキスト・ドキュメントを対象としたコンテンツ分析型リコメンデーションでは，消費者の興味の表現である顧客プロファイルやコンテンツの特徴表現として，キーワード・ベクトルが用いられる．TwinFinder は，消費者の購買履歴と書籍情報を用いて，顧客プロファイルのキーワード・ベクトルを生成する．その後，顧客プロファイルと書籍特徴の類似度を計算し，類似度が高い書籍を推薦情報とする．

複数の書籍を購入した消費者は，複数のカテゴリーから洋書を購入している例が多く，一般書籍に対する興味の幅は，非常に広い．インターネットの書籍販売においては，消費者が自ら目的の書籍を探す方法として検索が用いられる．消費者が複数のカテゴリーに興味を持っている場合，検索する書籍カテゴリーによって，まったく異なるキーワードを使い分ける．したがって，顧客興味を表現する顧客プロファイルを作成する場合にも，消費者の興味の幅が広い場合には，複数のキーワード・ベクトルを作成する必要がある．TwinFinder は，書籍カテゴリーを顧客興味の単位と仮定し，購買書籍カテゴリー別の複数のキーワード・ベクトルにより顧客プロファイ

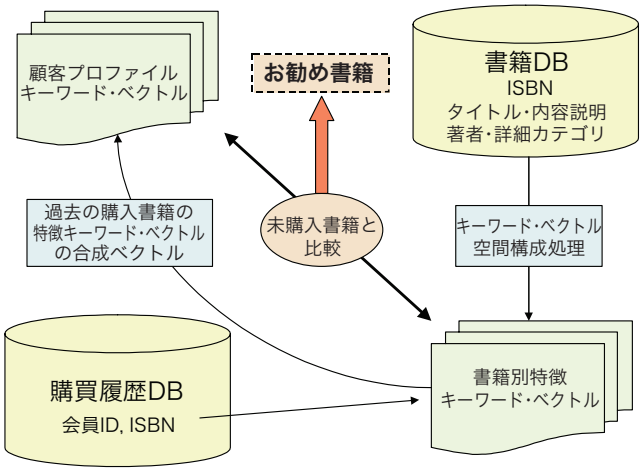

図-5　TwinFinderの推薦情報の生成手順

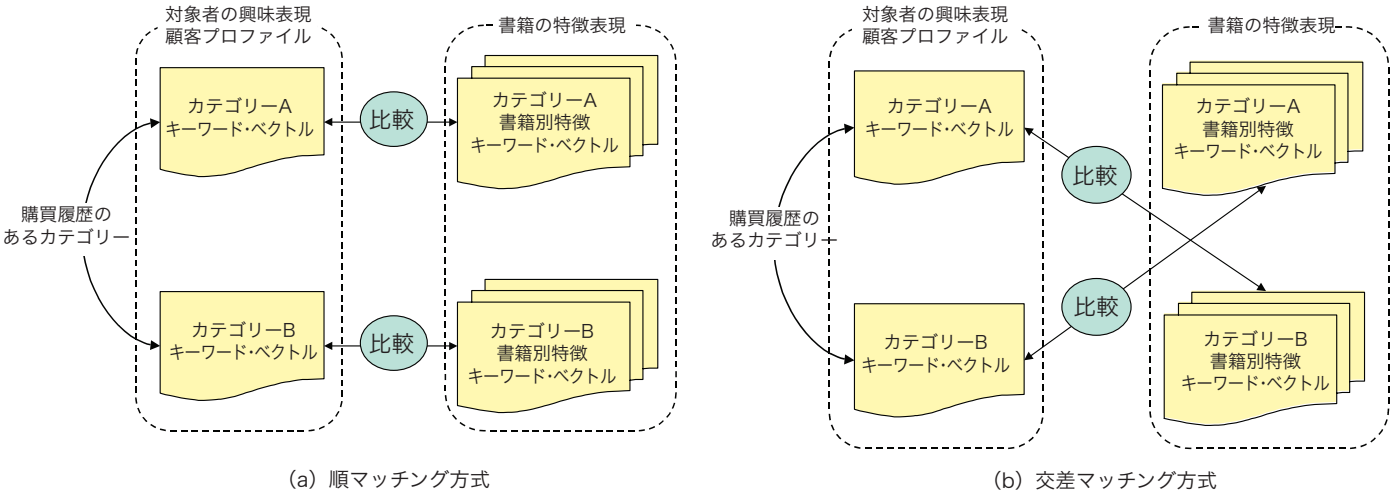

図-6　順マッチング型と交差マッチング型リコメンデーション

ルを構成する．

順マッチング型と交差マッチング型の情報推薦

順マッチング型リコメンデーションは，**図-6** (a) のように，顧客の購買書籍カテゴリーごとに作成したキーワード・ベクトルを用いて，同一の書籍カテゴリーに含まれる書籍の書籍特徴と比較を行う方法である．順マッチング型リコメンデーションでは，まったく興味のないカテゴリーの書籍が，いくつかのキーワードが一致したというだけで，推薦情報が出力されるのを防ぐことができる．

一方，交差マッチング型リコメンデーションは，**図-6** (b) のように，購買書籍カテゴリーごとに作成したキーワード・ベクトルを用いて，他の購買書籍カテゴリーに含まれる書籍の書籍特徴との比較を行う方法である．ただし，比較の対象となる書籍カテゴリーは，過去に購入実績があり，顧客が興味を持っている書籍カテゴリーに限定する．キーワード・ベクトルと書籍特徴の書籍カテゴリーの関係を交差させることにより，意外性の高いリコメンド情報が生成されることを期待する方法である．

我々は，電子メールを典型的な書店利用者に送付して実施したプッシュ型の評価実験を通して，TwinFinderの性能について，次の点を確認した．

- 意外性が高いリコメンド情報は，有益性に関しても高い評価を得た．情報推薦サービスでは，顧客の選択の幅を広げるような情報を提供することが重要である．
- TwinFinderは，書籍数に対して顧客規模が小さい場合

においても有効に機能する．

- リコメンドの正確性を高める順マッチング型リコメンデーションは，リコメンド情報の有益性において優れている．
- 利用者の持つ複数の分野の興味を結びつける交差マッチング型リコメンデーションは，意外性があっておもしろい情報を提供することができる．

情報推薦システムのよろず屋さんになるために

本稿では，情報推薦システムの基本的な考え方について述べた．以下では，結論に代えて実践的なアドバイスをまとめる．

情報推薦システムをつくるために

Web上で稼働する適切な情報推薦システムを実現するためには，まず，扱う情報の種類とそれにアクセスする利用者の数とを適切に見積もることが重要である．e-コマースが扱う商品の種類は数十種類から数百万種類の範囲にある．利用者の数も数百から数千万の範囲にある．これらに同じ算法を適用することはできない．

次に考慮しなければならないのは情報提供の形態である．利用者のWeb利用を前提とするならば利用者をひきつけるプル型のシステムが適当であり，逆に利用者の興味をひきつけるならプッシュ型のシステムが適当である．また，利用者情報の処理をオンラインに行うのかオフラインで行うのかによってもシステム化の手法は異なる．

膨大なWebログを使った分析には当然ながらデータマイニング手法が不可欠である．利用者の特性を調べるにはWeb探索行動を分析することも必要である．我々の経験によると，利用者がWebページのコンテンツを調べているのか，それとも単に探索の途中で特定ページに入ったのかを知るには，ページ滞在時間１分を境に判定するのがよい．

前章で紹介したTwinFinderは，コンテンツ分析と利用者プロファイル処理をオフラインで実施することを前提としたプッシュ型のシステムであり，利用者のWeb探索行動は利用していない．

情報推薦システムを研究するために

第１の課題は情報推薦システムの評価問題である．e-コマースの推進者は自分にとって都合のよい情報を推薦したくなる．情報の享受者はどんな情報が自分にとって適切であるかは自覚していない．つまり推薦情報の良さを定義する客観的な尺度は存在しない．文献5）などの研究では情報検索で利用される尺度に基づいて適合率・再現率に基づいた尺度で推薦情報の評価を行っている．しかしこれらの尺度は利用者の興味の大きさや意外な情報に対する反応などを考慮していないので不十分である．

第２の課題は，情報推薦のコンテクスト依存性である．利用者の推薦情報に対する判断は，Web利用のコンテクストに依存して決定される．マイページなど利用者とその利用状況に依存した情報提供の方法については最近大きな注目が集まっている[3]．これには利用者の認知モデルを適切に同定することが重要である．

第３の課題はプライバシー問題である．問題は，プライバシーの保護と個人用にカスタマイズされた情報提供の両方を望んでいることにある．適切な情報提供には，利用者に関する詳細かつ大量の情報が不可欠である．Webにアクセスするたびに，どのような情報が外部に漏れているかは一般利用者では知ることができない．匿名データから得られる情報の範囲はどのようなものか，また，どのようにすべきかについてはいっそうの研究が必要である．

情報推薦システムの研究開発は非常な広がりをみせているが現在のところそれらを一覧できるようなWebサイトは存在しない．研究分野全般については少し古いが文献1）にさまざまな解説がある．コンテクストに依存したシステムの個別化については文献3）に詳しい．利用者のプライバシーに関する議論は文献4）が参考になる．今後の研究に期待する．

参考文献

1）Resnick, P. and Varian, H.R. (eds.): Recommender Systems, Communications of the ACM, Vol.40, No.3, pp.66-72 (1997).
2）Hirooka, Y., Terano, T. and Otsuka, Y.: Extending Content-Based Recommendation by Order-Matching and Cross-Matching Methods, In Bauknecht, K., Madria, S.K. and Pernul, G. (eds.): Electronic Commerce and Web Technologies, 1st Int. Conf., EC-Web 2000, Springer Lecture Notes in Computer Science LNCS-1875, pp.177-190 (2000).
3）Mostafa, J. (ed.): Information Customization, IEEE Intelligent Systems, Vol.17, No.6, pp.8-52 (2002).
4）The Platform for Privacy Preferences 1.0 (P3P1.0) Specification W3C Recommendation (16 Apr. 2002), http://www.w3.org/TR/2002/REC-P3P-20020416/
5）Sarwar, B., Karypis, G., Konstan, J. and Riedl, J.: Analysis of Recommendation Algorithms for E-Commerce, Proceedings of ACM on E-Commerce (EC-00) (2000).
6）寺野隆雄：情報推薦システム，北村泰彦，山田誠二（編）：eビジネスの理論と応用，東京電機大学出版局，pp.59-87 (2003).

（平成15年4月1日受付）