神嶌 敏弘, "協調フィルタリングの課題：プライバシー，サクラ攻撃，評価値のゆらぎ", 情報処理, vol.48, no.9, pp.966-971 (2007)ドラフト版

Vol. 48 No. 9 情報処理学会論文誌 Sep. 2007

協調フィルタリングの課題：プライバシー，サクラ攻撃，評価値のゆらぎ

神 嶌 敏 弘†

あまりに多すぎる情報に埋もれて，欲しい情報を見つけ出せなくなる情報過多の問題が顕在化して久しい．この問題に対する処方箋の一つが推薦システムである．90 年代後半から，商用化がはじまり，現在では広く普及するようになった．しかし，まだ課題は多く，これらのうちプライバシー，サクラ攻撃，および評価のゆらぎについて紹介する．

1. 推薦システムの課題

Web 上の書店には，何十万冊という本が販売されている．論文や技術資料も Web 上にあふれている．自分が欲しいものや情報はきっとこの中にあるだろうが，あまりに多すぎる候補に埋もれて，目的のものを見つけ出せなくなっている．こうした問題は**情報過多 (information overflow)** と呼ばれている．これに対処すべく**推薦システム (Recommender System)** が考案された．推薦システムは，利用者が探している物や情報（アイテム）を予測し，それらを利用者に提示するシステムである．90 年代後半から，商用化が始まり，現在では広く普及するようになったが，まだまだ課題もある．本稿では，次の三つの課題について述べる．

- プライバシー：個人情報保護法の制定などプライバシーへの関心が高まっている．購買履歴などの個人情報を取り扱うための，協調フィルタリングの手法を紹介する．
- サクラ攻撃：これは，自身の製品がより推薦されるようにするため，通常の利用者に偽装して自身に有利な評価をする行為のことである．こうした行為は，他の利用者には不利益となる．
- 評価値のゆらぎ：よい推薦のためには，利用者の嗜好の情報を正確に知ることが重要である．しかし，尋ねるたびに利用者が違う評価をするなど，正確な計測は難しい．

2. プライバシー保護協調フィルタリング

協調フィルタリングでは，標本利用者の嗜好データを収集する．このデータが商品の購入履歴などであれば，これらの情報は利用者が秘匿したい個人情報となるであろう．さらに，断片的な情報を集積することで，より重大な個人情報を得ることもできるようになってきている．例えば，Sweeney は，87%の US 在住者が，性別，5 桁郵便番号，生年月日の情報だけで一意に特定できること[1] を示した．個人情報保護法の施行などに伴い，このプライバシーの問題はますます重視されるようになっている．だが，協調フィルタリングは，こうした個人情報なしには実現できない．

現状では，この個人情報の収集に伴う問題には，プライバシーポリシーを公開し，それを遵守することを誓約することで，社会的手段によって対処している．実際にポリシーを遵守しているかを監査する民間企業もあるが，監査を受けていたにも関わらず，後に違反していることが発覚した事例もある．これに対し，技術的手段によってこの問題に対処するのが**プライバシー保護協調フィルタリング (privacy-preserving collaborative filtering)** である．プライバシー保護協調フィルタリングの技術は，プライバシー保護データマイニング (privacy-preserving data mining)☆の技術を踏襲している．複数の計算機で構成されている分散環境で，各サイトに分割されて保持されているデータがあるとしよう．各サイト内の個々のデータは個人情報で，サイトの外部には公開できない．しかし，データ全体の傾向など，そこから個人情報を完全には復元できない情報は個人情報ではないと考える．この考えに基づき，各サイト内のデータを自分以外のサイトには秘密にしたまま，全てのサイトのデータを集めた集合に対する解析結果を計算する技術が，プライバシー保護データマイニングである．この計算を実現するアプローチとしては後述の安全な計算を使う方法と，個人データを

† 産業技術総合研究所, National Institute of Advanced Industrial Science and Technology (AIST),
http://www.kamishima.net/

☆ C. Clifton の Web ページ[2] の解説を参考にされたい．

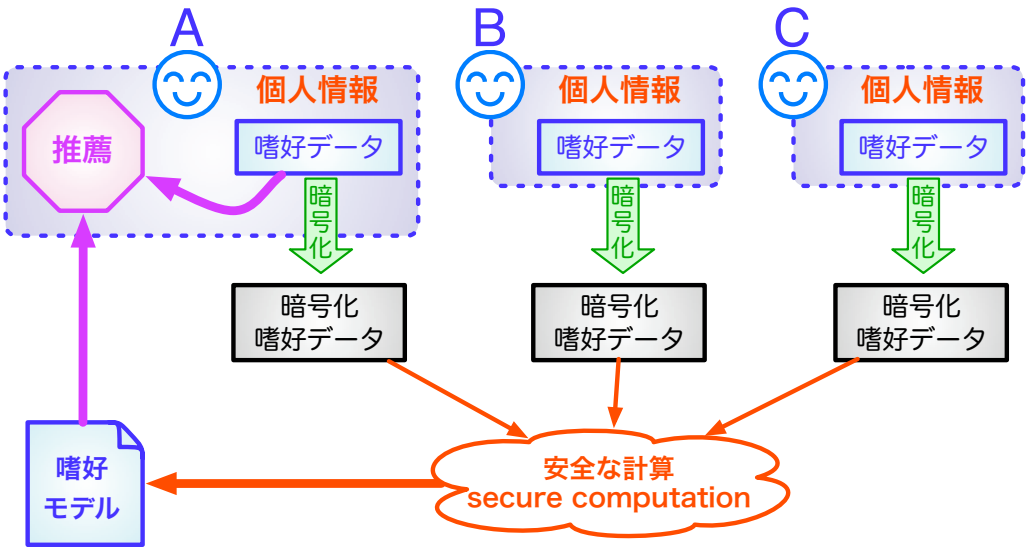


図 1 プライバシー保護協調フィルタリング

乱数で変化させてからサイトの外部に出すランダム化による方法がある．前者の方法は，通常の推薦アルゴリズムと厳密に同じ解が得られるが，利用者に後述のような制限がある．後者のランダム化では，利用者に制限はないが，近似的な計算しかできず，また，個人データの厳密な値は保護されるが，だいたいの値は予測できる場合もある．プライバシー保護協調フィルタリングにおいても，この二つのアプローチがある．ここでは，Canny[3] が提案した前者の安全な計算を使う枠組みを紹介する．

協調フィルタリングの原理を簡単に復習すると，他人の嗜好パターンと，自分の嗜好の似た部分を探しだし，それを元に自分が好むアイテムを見つける方法である．ここで，他の個人の嗜好のデータそのものではなく，それらを集めた嗜好パターンさえ分かれば推薦が可能であることが重要である．個々の商品についての好き嫌いや，何を購入したかといった，個人の嗜好データはもちろん個人情報である．だが，全体の嗜好パターンは，特定の個人の情報ではなく，これを知ってもプライバシーの問題は生じない．また，次元縮約や確率モデルを使えば，個人の情報を回復できない形式で，嗜好パターンを表現した嗜好モデルを構築できる．このモデルを，個人の嗜好データ自体は秘密にしたままで計算できれば，プライバシーを保護しつつ推薦ができる．この計算のため，個人のデータを暗号化したまま，総和などの基本計算をする**安全な計算 (secure computation)**☆を導入する．

このアイデアに基づく枠組みを図 1 に示す．この推薦システムは，図の上部にあるように A，B，C ‥‥ など多くの人が利用している．破線で囲んだ部分は，各個人が管理するローカルマシンであり，それ以外はどこかの共有サーバで実行される推薦システムを表す．このとき，A さんに推薦する場合を考えよう．それぞれ，好き嫌いや，商品を購入したかどうかといった嗜好データを，自分のローカルマシンに蓄積している．嗜好データは個人情報なので，そのままではローカルマシンの外部に取り出すことはできない．そこで，全利用者は自身の嗜好データをローカルマシンで暗号化して推薦システムに送信する．ここで，送信されたデータは暗号化されているため，個人情報は外部には漏洩しない．これらの暗号化嗜好データに，安全な計算の技術を適用して，個人の嗜好データを復号することなく，嗜好モデルを獲得する (図下)．ここでも，個人情報は暗号化されたまま計算されるため，利用者の個人情報は漏洩していない．また，この嗜好モデルを知っても，利用者全体の嗜好の傾向が分かるだけで，個人情報の漏洩にはならない．最後に，A さんは，この嗜好モデルを受け取り，自身の嗜好データと合わせれば，嗜好モデル中の他の利用者の嗜好パターンを参考にしながら，自身への推薦を計算できる．この推薦の計算も，ローカルマシン内で行われるので，個人情

☆ 安全な計算には秘密計算という訳語もある．だが，ここではデータが秘匿されるだけではなく，計算の結果が正しいことも考慮にいれるという意図で安全な計算としておく．

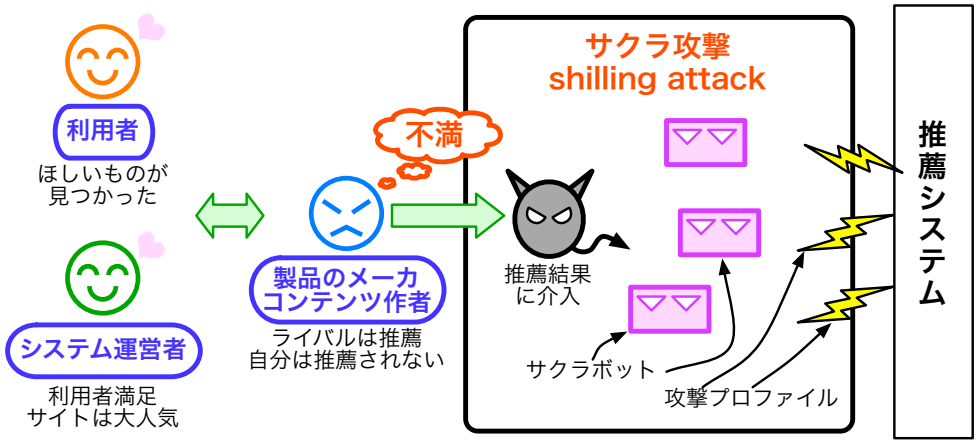


図 2 サクラ攻撃とその背景

報は外部には漏洩していない．

こうした手続きでプライバシーと有用な推薦を両立できそうである．しかし，現状のプライバシー保護協調フィルタリングでは，社会的手段が補助的に必要となる．計算結果に改竄がないかを保証するには，協調フィルタリングに参加する利用者は，互いに，個人情報を明かすほど信頼はできないが，計算の手続きは遵守する程度には信頼できるという semi-honest という前提が必要になる☆．しかし，この前提の技術的手段による保証は難しく，不正な計算手続き行った利用者は，協調フィルタリングのグループから除名されるとか，罰金を科すといった社会的手段によって保証しなければならない．そのため，匿名で参加できる peer-to-peer ネットワークなどでの実現は難しい．ソーシャルネットワークサービスなど個人認証がなされるサービスでの推薦や，複数の企業が自身の顧客の情報を秘匿しつつ，多数の利用者のデータに基づく推薦をしたい場合などに限定されるだろう．

3. サクラ攻撃

推薦システムの効用について考えてみよう．図 2 左のように，利用者は知りたい情報を入手できるようになり，システムに満足するようになるだろう．これにより，システムの運営者にも，システムの利用が促進されるという利点がある．では，利用者に推薦される製品や情報の提供者にとってはどうであろうか？自身の製品の代わりに競合他社の製品が推薦されたり，たとえ自社の製品がある程度は推薦されていても，さらに多くの自社商品が推薦されることを望んだりするだろう．しかし，これらの要求は必ずしも満たされるとは限らない．そのため，推薦結果を自身に有利にする目的で推薦システムに干渉すること☆☆が行われ始めている．

こうした行為の一つに**サクラ攻撃 (shilling attack)** がある．これは，サクラボット (shilling bot) と呼ばれるエージェントプログラムなどを用いて，自身に有利な推薦が他の利用者になされるような嗜好データを推薦システムへ入力するものである．サクラ攻撃は，利用者にとっては不適切な推薦がなされるため，また，運営者にとってもシステムへの信頼を失わせる行為であるため，望ましくない．以下，文献[4] をはじめとするサクラ攻撃についての研究を紹介する．

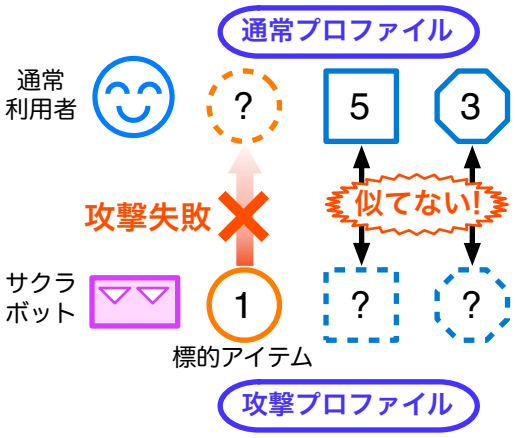
サクラ攻撃は，その攻撃意図，すなわち推薦をどのように変化させたいかという目的によって次の二つに分けられる．

- **販促攻撃 (push attack)**：本来なら推薦されないはずのアイテムを推薦されるようにする．
- **排除攻撃 (nuke attack)**：本来なら推薦されるはずのアイテムを推薦されないようにする．

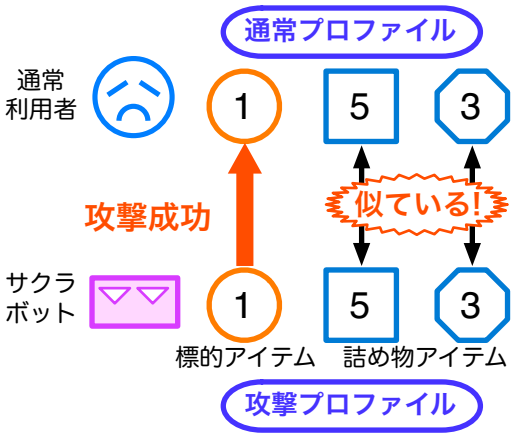
販促攻撃では自社の製品が，排除攻撃では競合他社の製品が対象となる．一般に，推薦システム中の嗜好データなどの情報を，より元のまま，また，より大量に利用した攻撃ほど効果的である．しかし，利用者の嗜好情報などは，通常の方法では外部からアクセスできないので，攻撃者は合法的には入手できない．だが，アイテムへの平均評価値といった統計情報などには合法的に入手できるものがある．推薦システムが利用者

☆ semi-honest を前提としない安全な計算もあるが，それを用いて推薦を行うのは計算量的に困難である．

☆☆ 具体的な事例は[4] を参照されたい



(a) 標的アイテムのみ



(b) 標的アイテム＋詰め物アイテム

図 3 攻撃プロファイル

の便宜のためにこれらの統計情報を公開している場合もあるし，映画における IMDB[5] のような評価サイトから，代替情報を入手できる場合もある．また，推薦アルゴリズムごとに効果的な攻撃の作戦があるので，どういったアルゴリズムを用いているかという情報も重要になる．

次に，サクラ攻撃がどのように行われる[4] かについて述べよう．図 3 は，排除攻撃の様子を示した．まず，図 3(a) に注目されたい．上側は，通常利用者の嗜好データ (通常プロファイル) を示した．この通常利用者に推薦されるアイテムを，サクラ攻撃によって変えたいとする．一方，下側は，仮想利用者であるサクラボットが，推薦システムに入力する嗜好データで，攻撃プロファイル (attack profile) と呼ばれている．図中には，丸，四角，八角で示した三つのアイテムがあり，各プロファイル中では，これらのアイテムは 5 段階で評価されている．「5」が最高評価，「1」が最低評価，そして「?」が未評価を示している．この中で，丸で示したアイテムは，排除したい競合製品を表し，これを標的アイテム (target item) と呼ぶ．通常利用者はこの標的アイテムを未評価で，今からこの利用者の標的アイテムへの評価値を予測するとしよう．

このとき，標的アイテムを推薦されないようにするには図 3(a) のように，サクラボットは標的アイテムに最低の評価「1」を与えさえすれば良さそうである．しかし，この攻撃は失敗する．多くのサクラボットが悪い評価を与えることで，このアイテムへの平均的な評価は確かに低下する．だが，協調フィルタリングでは，嗜好パターンが似ている利用者を参考に推薦することを思いだされたい．この利用者の通常プロファイルでは，標的アイテム以外の四角や八角のアイテムも評価されている．一方，攻撃プロファイルでは，丸の標的アイテムのみが評価され，それ以外の四角や八角のアイテムは全て未評価である．そのため，通常プロファイルと攻撃プロファイルは似ていないと判定され，サクラボットとこの利用者の嗜好パターンは違うとみなされる．そのため，標的アイテムの評価は通常利用者の推薦には反映されず，攻撃は失敗する．

そこで，他の利用者の嗜好パターンに攻撃プロファイルを似せるため，図 3(b) のように，標的アイテム以外の四角や八角のアイテム群にも評価値を与える．これらのアイテムを詰め物アイテム (filler item) と呼ぶ．これらの，詰め物アイテムへの評価が通常利用者のそれと類似していれば，通常プロファイルと攻撃プロファイルとは類似していると判断される．すると，攻撃プロファイルの標的アイテムへの評価は，通常利用者の推薦に影響し，攻撃は成功する．しかし，通常利用者の嗜好データは入手できない．よって，詰め物アイテムの評価値は，通常利用者の，一般的な評価の傾向に基づいて与える．この与え方の違いによって，次のような攻撃方法がある．

- **ランダム攻撃 (random attack)**：ランダムな評価値を，詰め物アイテムに与える．必要な追加情報はないが，あまり効果はなく，発覚し易い．
- **平均攻撃 (average attack)**：詰め物アイテムに，それらのアイテムへの評価値の平均値を与える．平均評価値は，アイテムの評価サイトなどから入手できる．嗜好が似ている他の利用者が好むアイテムを推薦する方法には効果的だが，自分が好きなアイテムと類似したアイテムを推薦する方法には限定的な効果しかない．

- **セグメント攻撃 (segment attack)**：映画のジャンルなど，アイテムの分類情報が利用できる場合の販促攻撃に用いる．標的アイテムと同じセグメント☆のアイテムには，高い評価を与えるようにする．これは，同じセグメントのアイテムには，同じような評価がなされやすいという傾向を利用したものである．平均攻撃とは逆に，類似したアイテムを推薦する手法に対して有効といわれている．

作為的な利用者の評価に影響されて，協調フィルタリングが不適切な推薦をしたとしよう．そして，他の利用者がその推薦に従ったとしても，その後，その利用者は作為のない評価をするので，自律的にこうした攻撃は無力化されるとも考えられていた．しかし，Cosley ら[6] はこれに対して否定的な調査結果を報告している．利用者が，以前に評価したことのあるアイテムについて，以前と同じ，1 段階良い，1 段階悪いの 3 種類のものを「予測評価」として利用者に提示した．すると，作為的にずらした方向に，利用者のアイテムへの評価は変化した．さらに，未評価のアイテムについて，アルゴリズムができるだけ正確に予測した評価，それより 1 段階良い/ 悪い評価を利用者に提示すると，やはり，同様の傾向がみられた．このように，利用者の評価が，提示された評価に「引きずられる」現象が報告されている．そのため，サクラ攻撃に対して推薦システムは自律的に回復することはできない．そこで，攻撃を検出して排除する必要がある．

これらの攻撃は，真のデータベースの評価値分布と，攻撃プロファイルとの統計的な分布の差をはずれ値検出の技術によって見つけることで検出する．だが，攻撃プロファイルの大きさが比較的小さければ，検出は難しい．また，攻撃が特定の時期にまとまって行われることが多いことを利用し，プロファイルが入力される時刻のパターンを監視することで，検出を試みる研究もある．

4. 評価値のゆらぎ

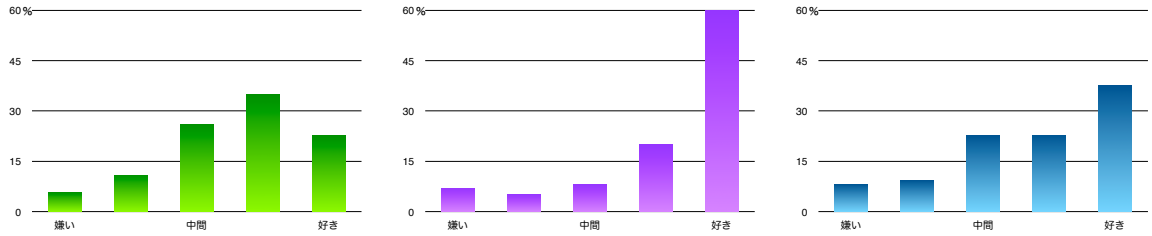
利用者が好みの度合いを答えるには，それを測る尺度が必要になる．好みの度合いを表す尺度として，0 ～ 5 や $-3 \sim +3$ のような数値尺度を使う採点法 (scoring method) や，上・中・下 や 適合・不適合 などの順序付きカテゴリ尺度を使う格付け法 (rating method) が良く利用されている．採点法や格付け法は，単純な入力フォームを用いて，比較的多数のアイテムに対する嗜好データを得られることが利点である．採点法や格付け法は多用されてきたが，当然ながら欠点もある．その例として，評価値の揺らぎや偏りがある．採点法や格付け法によって計測した絶対的な評価値が，真の評価値と乖離している間接的な証拠と，その乖離の原因を示す．

まず，評価値のゆらぎの証拠を示す．聴覚，味覚，または触覚など人間の感覚の度合いを定量的に測定する官能検査などの研究では，たとえ同じ評価値を与えていても，人によって嗜好の強さが違っていたりとか，時間がたつと一貫性が保たれなくなる問題があることが知られていた．テイスターなど訓練された被験者が，時間的に続けて評価をした場合でなければ，尺度を一定に保つことは難しいとされている．嗜好データについても，評価付けした後，日にちがたった後に，もう一度同じ被験者に同じ評価付けさせると，二つの評価値の間の相関は 0.70 であった[6] との報告がある．筆者の実験でも，寿司の嗜好について採点法で尋ね，続けて，無関係な質問を幾つかしたのち，順位法という別の方法で再び同じアイテムについて嗜好を質問すると，68.3%の被験者の回答に不整合が観測された．他に，代表的な GroupLens データを対象とした実験で，いろいろな工夫にもかかわらず，エラーの数値☆☆を 0.73 より小さくできない現象がある．このことから，評価値そのものにゆらぎがあるのではないかと示唆されている[7]．これらの報告は，嗜好データにはゆらぎがあることを示している．以上のように，絶対的な評価値を使う採点法や格付け法では，被験者は，質問時期の違いによりゆらぎが生じるといえるだろう．

次に，評価値の偏りについて述べる．図 4 に，5 段階の採点法を用いた 3 種類の嗜好データの，評価値の分布を示す．それぞれ，(a) MovieLens の 100 万要素のデータ集合[8]，(b) 電子商取引サイト Amazon.com[9]，(c) 寿司の嗜好調査[10] での分布である．どのデータでも，「好き」の方へ明らかに偏っている．評価値が同じアイテムは，同等に好まれるとみなされるため，多数のアイテムが集中している部分では，細かな嗜好の差が分からない．この偏りの原因の一つに，サンプリングの問題が挙げられる．サンプリングの偏りの原因として，図 4 の (a) や (b) では，関心があって，好きなアイテムを利用者が選択的に評価していることや，図 4 の (b) や (c) では市場の淘汰を受けて，多くの人に好かれやすいアイテムのみが候補となっていることがある．このようなサンプリングの偏りは予測誤差の過小

☆ 映画の場合なら，ラブストーリー，ホラー，ファンタジーといったアイテムの分類カテゴリ．

☆☆ 5 段階尺度で，平均絶対誤差 (MAE) で評価



(a) MovieLens (b) Amazon.com (c) 寿司

図 4 アイテムへの評価値の分布

評価を引き起こす．その他，文献[6] では，評価尺度から，5 段階評価で 3 といった，中立の評価を取り去って，評価の目盛りの数を偶数にすると，以前は中立に評価されていたアイテムが，肯定的な方へずれて評価されるようになることを報告している．

そこで，採点法や格付け法以外の調査方法の利用が考えられる．採点法や格付け法で得られる量は，本質的には大小関係にのみ意味がある順序尺度であると指摘されている．そこで，好きなものから嫌いなものへ，順に複数の対象を並べるという順位法を利用する「なんとなく協調フィルタリング[11]」を筆者は提案した．この順位法の採用で，少なくとも調査したデータにおいて，予測精度は向上した．ただし，順位法にも，評価は常に相対的で，絶対的な評価は得られないといった問題はある．そのため，相対的に良いものを選ぶような意志決定には役立つが，絶対的な評価値を参考のために示すといった目的には向かない．

5. 推薦システムの資料

本稿では，プライバシー，サクラ攻撃，そして評価値のゆらぎの協調フィルタリングの三つの課題を紹介した．そのほかにも，利用者の現在の情報要求への適合，より信用される推薦，分散環境下での大規模化，アイテムや利用者の入れ替わりへの迅速な対処など，まだまだ課題は多い．Herlocker らは，文献[7] で，「良い推薦とは」ということについて深く考察しており，本格的に推薦システムに取り組む前に一読することを薦める．推薦システム全般についての，他の問題や話題については，拙著の解説[12] を参考にされたい．

参 考 文 献

1) Sweeney, L.: Uniqueness of Simple Demographics in the U.S. Population, *LIDAP-WP4* (2000). http://privacy.cs.cmu.edu/dataprivacy/papers/LIDAP-WP4abstract.html.
2) Clifton, C.: . http://www.cs.purdue.edu/homes/clifton/.
3) Canny, J.: Collaborative Filtering with Privacy, *Proc. of the 2002 IEEE Symposium on Security and Privacy*, pp.45–57 (2002).
4) Lam, S. T.K. and Riedl, J.: Shilling Recommender Systems for Fun and Profit, *Proc. of The 13th Int'l Conf. on World Wide Web*, pp.393–402 (2004).
5) Database, T. I.M.: http://imdb.com/.
6) Cosley, D., Lam, S. K., Albert, I., Konstan, J. A. and Riedl, J.: Is Seeing Believing? How Recommender Interfaces Affect Users' Opnions, *Proc. of the SIGCHI Conf. on Human Factors in Computing Systems*, pp.585–592 (2003).
7) Herlocker, J.L., Konstan, J.A., Terveen, L.G. and Riedl, J.T.: Evaluating Collaborative Filtering Recommender Systems, *ACM Trans. on Information Systems*, Vol.22, No.1, pp.5–53 (2004).
8) MovieLens データ：http://www.grouplens.org/node/12#attachments.
9) Weigend, A. S.: Analyzing Customer Behavior at Amazon.com, *Invited Talk at KDD2003* (2003).
10) 寿司の嗜好調査データ：http://www.kamishima.net/sushi/.
11) Kamishima, T.: Nantonac Collaborative Filtering: Recommendation Based on Order Responses, *Proc. of The 9th Int'l Conf. on Knowledge Discovery and Data Mining*, pp.583–588 (2003).
12) 神嶌敏弘：推薦システムのアルゴリズム (1)〜(3)，人工知能学会誌, Vol.22, No.6 〜 Vol.23, No.2 (2007–2008).

神嶌 敏弘（正会員）

1968 年生．1992 年京都大学情報工学科卒業．1994 年同大学院修士課程終了．同年電子技術総合研究所入所．2001 年博士 (情報学)．同年電子技術総合研究所は産業技術総合研究所へ再編．機械学習とその応用の研究に従事．AAAI, ACM，人工知能学会会員．