基般

データマイニングと社会的公正性・中立性

神嶌 敏弘　産業技術総合研究所

データ分析における社会的問題の側面を扱う本特集の中で，本稿では，データマイニングにおける公正性・中立性について述べる．前半では公正性に関するWeb広告での事例を取り上げ，このような問題に対処するための公正配慮型データマイニングの技術を紹介する．後半では，個人化技術の中立性に関する指摘であるフィルタバブル問題を取り上げ，個人化技術コミュニティでの議論と調査研究を紹介する．

データ分析での公正性

データマイニングは意思決定支援技術の１つでもあるため，その分析結果に社会的公正性を考慮すべき場合があるだろう．個人属性情報（デモグラフィック情報），金融取引履歴，通信履歴，税務記録など膨大な個人データの集積が進むと同時に，データマイニングを利用するためのデータベースや，分析ソフトウェアなどの環境も整備されている．そのため，与信，採用，保険などの重要な決定にもデータマイニング技術が利用されるようになった．このとき，社会的・法的な公正さへの配慮，すなわち，性別，人種，ハンディキャップなど先天的要因に対する差別がない判断をすることは重要であろう．

データマイニング技術を用いた個人化によって，個人にとって不利な結果がもたらされる可能性のSweeneyによる指摘を紹介する[7]．読者は，多くの文書やWebページから，必要な情報を見つけ出す情報検索サイトを毎日利用していることと思う．これらのサイトで，キーワードを入力すると，キーワードに関連した項目に加え，そのキーワードに関連した広告も併せて表示される．Sweeneyはこの広告の選択が，人種に対する偏見に基づいている可能性について調査した．マサチューセッツ州の新生児の記録から，アフリカ系とヨーロッパ系の間で偏りが大きい2,000種以上の名前を選び，これらの名前で検索サイトとニュースサイトで検索した．

Sweeneyは，逮捕歴などを検索するInstant Checkmateなどのサイトに関する広告に注目した．アメリカでは，各州で公開されている逮捕歴の情報があり，それらを収集して検索できるサービスを提供しているサイトが存在する．**図-1**(a)は，アフリカ系に多い名前“Latanya Farrell”で検索した場合に表示された広告の例である．1行目の広告は『Latanya Farrellは逮捕されたか？』という逮捕歴を示唆するような広告文になっている．一方で，ヨーロッパ系の名前“Jill Schneider”で検索した図-1(b)では，2行目の『Jill Schneiderを見つけました』のように，特に逮捕歴を示唆しない中立的な広告文であった．より詳しく調べると，実際のリンク先のサイトで逮捕歴があるか，また，その名前のレコードが存在するかに基づいて広告文が選択されているわけではなかった．アフリカ・ヨーロッパ系の区別と，広告文が中立かどうかの独立性を統計的に検定したところ，有意にアフリカ系で逮捕歴を示唆する広告文が多かったと報告している．また，著名人・政治家を選んで検索した場合は，中立なものも，そうでないものも表示されたとのことであった．

Instant Checkmate 社に対してインタビューによる調査を行った経過も報告している．その結果，複数の広告文のテンプレートから，広告のクリック率を最大化するようなものを選択しており，恣意的な操作は認められなかった．このテンプレートは姓のみに基づいて選択しており，他の規準はないとのことであった．いわば社会の悪意が，意図せず反映されてしまった状況ともいえるであろう．

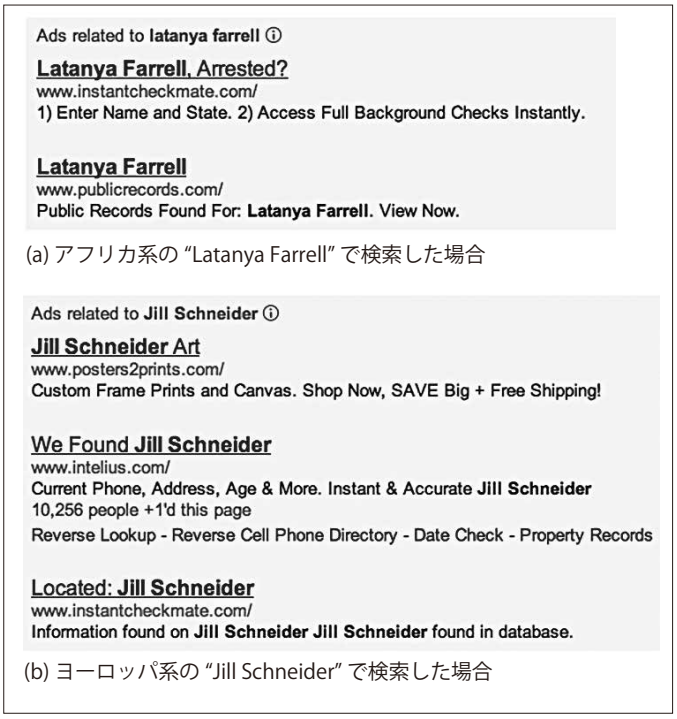

図 -1　人名で検索した場合に表示される Web 広告の例[☆1]

■公正配慮型・差別配慮型データマイニング

前節のような問題に対処するため，データマイニングの分析過程で，公正性や中立性の問題を扱うのが公正配慮型・差別配慮型データマイニング (fairness-aware / discrimination-aware data mining) である．この公正配慮型データマイニングには，不公正検出と不公正防止の 2 種類のタスクがある．不公正検出では，データや決定ルールの中に不公正なものがあれば，それを検出する．不公正防止では，潜在的に不公正なデータから，公正な決定をするような予測器を学習する．

最初に提案された，**図 -2** のような相関ルールに対する不公正検出の研究を紹介する[5]．相関ルールは，『⇒』の左辺の前提部の条件が成立するとき，右辺の結論部の条件が成立することを表すルールである．図 -2 のルール (a) は，対象者の居住地を表す属性 city が NYC であるとき，ローンの可否を表す属性 credit が bad となるルールの例である．ルール (a) の最後の『confidence』は確信度と呼ばれ，左辺の条件が成立するとき右辺も成立する比率で，ルールの確実性を示す．

相関ルールに対する公正性の概念として，あるデータ集合から抽出した相関ルールの集合に，公正性の尺度が α より悪いものがない状態を公正なものとする α 保護を提案した．公正性の尺度には，拡張リフト (extended lift) という尺度を用いている．図 -2 のルール (a) の左辺は社会的に配慮が不要な条件，右辺はローン利用者に不利な条件である．このルールの左辺に，センシティブ属性が社会的に保護が必要となる条件を加える．図 -2 (b) の例では，人種 race が保護すべき状態 African である条件を加えている．このとき，拡張リフトは，センシティブな条件を加えたルールの確信度の，元のルールの

(a) city=NYC ⇒credit=bad ..... confidence=0.25
(b) race=African,city=NYC ⇒credit=bad ..... confidence=0.75

図 -2　相関ルールの例

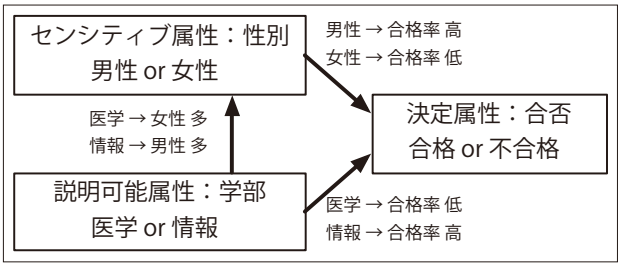

図 -3　センシティブ属性の情報が決定に影響しているが，不公正とはいえない場合の例

確信度に対する比率である．この図の例では，ルール (b) の確信度 0.75 を，ルール (a) の 0.25 で割った 3 が拡張リフトの値となり，人種がアフリカ系であることで，ローンが認められなくなる割合が 3 倍になることを示している．もし $\alpha=2$ としたならば，α 保護の条件を満たさず，決定を不公正とみなす．

しかし，社会的にセンシティブな属性に決定が影響されていても，不公正とはいえない状況も多く存在する．**図 -3** は，このような状況の，大学入試での例である [8]．この例では，男性の方が合格率が高く，上記の拡張リフトの指標では不公正と判断される．ここで，先天的な要因ではないなど，決定に影響を与えても特に不公正とはいえない説明可能属性を導入する．この例では，学部を表す変数で，女性は合格率の低い医学部を志願する割合が多いが，男性は合格率の高い情報学部を志願する割合が多い．すると，女性の合格率の低さは，難易度の高い学部を志願する割合が多いためであると説明できるため，必ずしも不公正な決定であるとはいえない．こうした場合では，合格という決定が，性別というセンシティブな要因と独立であるかどうかを単に検証するのではなく，学部のような説明可能な要因が与えられたときの条件付き独立性を検証することで，決定の公正性を判定できる．

次に，もう 1 つの公正配慮型データマイニングのタスクである不公正防止の研究を紹介する．Calders らは，潜在的に不公正な決定が含まれる訓練データから，公正な決定をする分類器を学習する 2 単純ベイズ法 (2-naive-Bayes method) を提案した [1]．通常の単純ベイズ法の分類モデルは，決定を示すクラス変数が与えられたとき，各特徴それぞれが互いに条件付き独立であると仮定する．一方，この 2 単純ベイズ法では，クラス変数 Y だけでなく，社会的な配慮が必要なセンシティブ特徴 S も与えられたときに，その他の特徴 $X^{(1)},\ldots, X^{(K)}$ が互いに独立であると仮定する．

$$\Pr[Y,\mathbf{X},S]=\Pr[Y,S]\prod_i \Pr\left[X^{(i)}\middle|Y,S\right]$$

なお，このモデルのセンシティブ特徴は，保護すべき状態と，そうでない状態の二値をとる変数に限定している．通常の単純ベイズモデルと同様に，このモデルのパラメータも訓練データ中の事例を数え上げれば容易に学習できる．しかし，この訓練データは潜在的に不公正な決定をした事例を含むため，学習したモデルも不公正な決定をしてしまう可能性がある．そこで，周辺分布 $\Pr[Y]$ の分布を訓練データのそれとあまり変えない条件の下で，センシティブ特徴 S の値が保護・非保護のいずれでも同じ決定がなされるように，$\Pr[Y, S]$ を修正する後処理を導入する．

ここで，red-lining 効果について述べておく．上記の 2 単純ベイズモデルのように，センシティブ特徴 S をモデルに含めると，センシティブな情報が決定に影響してしまうので，これは除外しておく方がよいと考えるかもしれない．しかし，たとえセンシティブな情報を直接的に用いなくても，他の特徴との間に相関があれば，センシティブな情報を排除できない．人種を直接的な理由とせず，地図上で赤枠で囲った特定の人種が住む地域の住人に貸し出しをしなかった過去の事例から，この現象を red-lining 効果と呼ぶ．このことは，センシティブな情報の漏洩を保護しても不公正な決定がなされてしまうので，逆にセンシティブな情報を収集して補正しなければならないという公正性を保証する難しさを示している．

個人化技術と中立性

以上，公正な決定に関する話題について述べた

が，ここからは個人化技術と中立性に関する話題を取り上げる．個人化 (personalization) 技術とは，情報検索や推薦システムにおいて，利用者の過去の行動に基づいて，その利用者の傾向・嗜好に，検索や推薦の結果を適合させる技術である．この個人化技術に対して提起されているフィルタバブル (filter bubble) 問題を紹介する．

フィルタバブル問題は，Pariser が，その著書[4]やプレゼンテーションイベント TED Talk などで提起した．著書では，個人情報収集の問題とフィルタバブル問題を特に区別せず論じているが，ここでは後者の問題のみを取り上げる．この問題は，個人化技術によって，利用者が接する情報の話題の範囲が狭められたり，偏ったりすることが，利用者が知らないうちに行われることに対する懸念である．Pariser が挙げた例を紹介する．ソーシャルネットサービスである Facebook では，限定された利用者間で議論などを行うために『友人』関係を明示する．この友人関係の構築に役立つように，利用者と関連がありそうな，他の利用者を推薦する機能が備わっている．Pariser がサービスを利用し始めたころは，保守派の人も，革新派の人も混在して推薦されていた．ところが，Pariser 自身が革新派の人と友人関係を実際に構築することが多いため，個人化の機能により保守派の人が推薦されなくなったと述べている．システムは利用者の断りもなく保守派を除外し，多様な選択の機会が減ったと主張している．

Pariser の指摘は大きく 2 つにまとめることができると私は考える．1 つは，利用者が多様な情報に接する機会が少なくなる問題である．このため，自身の知見を広げる芸術のようなものに接する機会が失われ，安易な娯楽情報のみを消費するようになってしまうと主張している．もう 1 つは，各人がそれぞれ異なる限られた情報にのみ接していて，互いに共有する情報が減ってしまう問題である．社会でのコンセンサスの構築には，その基盤となる情報の共有が重要であるとし，それが失われると主張している．

■ 推薦システムの研究コミュニティでの議論

このフィルタバブル問題に対する推薦システムの研究コミュニティでの議論を紹介する．推薦システムに関する国際会議 5th ACM Conference on Recommender System (RecSys 2011) では，この問題についてのパネル討論を行った[6]．このパネルでは，フィルタバブル問題を明確にする議論と，技術的な対応策についての討論があった．技術的な対応策については次節で述べることとし，フィルタバブルに関する議論をまず紹介する．個人化が利用者の経験範囲を狭めることは，1990 年代中頃にはすでに指摘されていた．しかし，特定の情報を選びとることは，他の情報を無視することを必然的に伴うため，利用者の関心に集中することと，多様な話題を提供することは本質的にトレードオフ関係になる．この問題は，個人化技術に特有のものではなく，一般のニュースにおいてもイスラム系の Al-Jazeera と，アメリカ保守系の Fox News との報道姿勢などにもみられるものである．

さらに，何かしらのフィルタリングに人間は常に触れていて，その影響を人間はうまく扱えていると指摘した．たとえば，バイキング形式の食事で，たとえ料理をどのように配置したとしても，その配置は客が選ぶ料理に必ず影響を与えるので，中立的な配置というものはあり得ない．このように，どんな情報も中立ではなく，人は常に何らかのバブルにとらわれているが，人はこれらの影響をうまく扱いつつ生活できている．個人化技術には確かにある種の情報の偏りはあるが，これらの影響を今までのようにどうにか扱いつつ，この技術を使いこなしてゆく以外にはないとの意見が述べられた．

■ 推薦システムにおける技術的対応策

このフィルタバブル問題に対して，個人化技術の長期的利用がどれくらい嗜好パターンに影響するかについての調査を紹介する[3]．映画の推薦システムである Movielens の利用者の行動が，31 カ月の間にどのように変化したかを調査した．まず，推薦結果を採用するグループと無視するグループどちらで

も，利用開始時より，最後の期間では提示される映画の多様性は低下するが，採用するグループの方が低下の幅が大きかった．一方，実際に鑑賞し評価した映画の多様性について調べると，推薦を採用・無視する2つのグループ間に，最初は差はないが，長期にわたって利用すると多様性は低下し，その度合いは無視するグループの方が大きかった．すなわち，推薦を採用する利用者の方が，この事例では多様な映画を鑑賞していた．

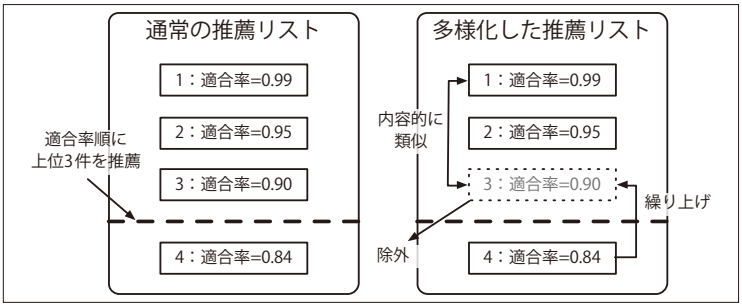

図-4　推薦リストの多様化の手順

こうした多様性の低下に対抗するための代表的手段が，推薦の多様性の強化である[2]．一般の推薦システムでは，推薦候補を実際に推薦するかどうかは個々に決定するが，推薦リスト全体の多様性を考慮して推薦するかどうかを決めることで，多様なアイテムを推薦する．通常は，利用者が最も好むであろう順に，すなわち適合率の高い順に推薦候補から選択する．一方，**図-4**のように多様な推薦をする場合には，推薦リスト中に類似したものがあれば，それを除外して，利用者の嗜好への適合性と推薦の多様性のバランスをとる．

以上，データマイニングや個人化に公正性や中立化の問題と，それを扱うための技術を紹介した．データ分析技術の利用が広まるにつれて，これらの技術の重要性は増すものと考えている．しかし，これらの研究は始まったばかりであり，これからの発展が期待される．

参考文献
1) Calders, T. and Verwer, S. : Three Naive Bayes Approaches for Discrimination-free Classification, Data Mining and Knowledge Discovery, Vol.21, pp.277-292 (2010).
2) Hurley, N. : Keynote : Towards Diverse Recommendation, In RecSys Workshop : Novelty and Diversity in Recommender Systems, p.1 (2011).
3) Nguyen, T. T., Hui, P.-M., Harper, F. M., Terveen, L. and Konstan, J. A. : Exploring the Filter Bubble : the Effect of Using Recommender Systems on Content Diversity, In Proc. of the 23rd Int'l Conf. on World Wide Web, pp.677-686 (2014).
4) パリサーイーライ，井口耕二：閉じこもるインターネット―グーグル・パーソナライズ・民主主義，早川書房 (2012).
5) Pedreschi, D., Ruggieri, S. and Turini, F. : Discrimination-aware Data Mining. In Proc. of the 14th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining, pp.560-568 (2008).
6) Resnick, P., Konstan, J. and Jameson, A. : Panel on the Filter Bubble, The 5th ACM Conf. on Recommender Systems (2011). http://acmrecsys.wordpress.com/2011/10/25/panel-on-the-filter-bubble/
7) Sweeney, L. : Discrimination in Online Ad Delivery, Communications of the ACM, Vol.56, No.5, pp.44-54 (2013). http://dx.doi.org/10.1145/2447976.2447990
8) Žliobaitė, I., Kamiran, F. and Calders, T. : Handling Conditional Discrimination. In Proc. of the 11th IEEE Int'l Conf. on Data Mining (2011).

(2014年9月1日受付)

■**神嶌 敏弘**　mail@kamishima.net
1994年京都大学大学院工学研究科情報工学専攻修士課程修了．2001年博士(情報学)．1994年電子技術総合研究所(現産業技術総合研究所)入所．推薦システム，データマイニング，機械学習に関する研究に従事．