人気によるバイアスを考慮した推薦システム

村中　悠里†　　杉山　一成††

† 京都大学 情報学研究科　〒 606–8501 京都市左京区吉田本町
†† 京都大学 情報学研究科　〒 606–8501 京都市左京区吉田本町
E-mail: †muranaka.yuuri.76s@st.kyoto-u.ac.jp, ††kaz.sugiyama@i.kyoto-u.ac.jp

あらまし　現在推薦システムの分野では、商品の人気によるバイアスが問題になっている。具体的には、推薦システムのアルゴリズムによって人気のある商品が頻繁に推薦されることで、一部の商品がユーザの興味のほとんどを占めて大多数の商品は目を向けられず、商品の人気の格差が更に広がっていくという現象が発生する。この人気バイアスによって、知名度はないが品質が良い商品が埋もれてしまったり、人気に品質が伴っていない商品がユーザに行き渡ったりしてしまうという事態が起こりうる。本研究では、これらの問題を解決するために、「商品に対する評価の質」に注目した手法を提案する。

キーワード　推薦システム、人気度、バイアス

1 はじめに

近年、EC サイトや配信サービスでは、ユーザの嗜好に適合した商品やコンテンツを提供するシステムの一つとして推薦システムが導入されている。この推薦システムによって、ユーザは膨大な数の商品の中から自分の好みや目的に合致した商品を見つけ出すことができるようになった。しかし昨今では、そのような推薦システムにおいて、人気によるバイアスというものが問題視されている。

例えば、現在推薦システムに用いられている手法の一つである協調フィルタリング [1] [2] では、推薦対象であるユーザが過去に購入、もしくは評価した商品を参照することで推薦対象のユーザと好みが類似しているユーザを見つけ、そのユーザの利用履歴から推薦商品を決定している。この手法では、ユーザの嗜好に適合した商品を、比較的高い精度で推薦することが可能である。

しかし、この手法では既に他のユーザが購入、もしくは評価した商品群の中から推薦する商品を決定するという前提があるため、必然的に人気があったり、話題になったりすることにより、ユーザの目につきやすい商品が推薦されることが多い。

また、その推薦された人気商品をユーザが購入すると、そのユーザの利用履歴に人気商品が追加されるという連鎖が起こり、いわゆるフィルターバブル現象 [3] に繋がる。その結果、推薦システム内で扱われる商品が人気商品ばかりで溢れ、人気のない商品が推薦されにくくなるという事態が発生する。

このように既存の推薦システムでは人気商品が推薦されやすく、不人気商品が推薦されにくいという傾向によって、商品の出現頻度に偏りが生じることがある。この出現頻度の偏りは、市場において売り上げの大多数を人気商品が占める「ロングテール」と呼ばれる状態や、一部の商品がユーザの興味のほとんどを占めて大多数の商品は目を向けられず、商品の人気の格差が更に広がっていく「マシュー効果」[4] に繋がる。これらの諸問題は、知名度はないが品質が良い商品が人気を獲得する機会を失って市場に埋もれてしまったり、人気に品質が伴っていない商品がユーザに行き渡ったりしてしまうという事態を引き起こすことになる。このような事態は、生産者側のより良い商品を提供するというモチベーションが損なわれる一方で、消費者側は多様な商品に触れる機会や、優れた商品やコンテンツを見極める能力を失ってしまう、などといった問題に発展することも想定される。結果として市場や文化そのものの停滞にもつながり得る。

そこで本研究では、人気によるバイアスから生じる諸問題に対して、「商品に対する評価の質」に注目した手法を提案する。

「 商品に対する評価の質」とは、ユーザが商品に付けた評価の正確さや説得力の高さを指す。ある商品に対して同じ評価を付けたユーザでも、客観的な感想を述べるユーザと主観的な感想を述べるユーザでは評価の説得力が異なり、推薦に利用するという点では前者の方がより適した評価であるといえる。本研究では自身の嗜好に偏らない客観的な評価を「質の高い評価」と表現する。このように評価の質が高いユーザを推薦する際の判断材料の一つとすることで、各ユーザの主観により過ぎず、商品そのものの品質を重視した推薦が提供できるのではないかと考える。

以上の観点から、本研究では、以下の内容について扱う。

- 商品に付けられる評価に注目し、推薦システムにおける商品の人気や出現頻度の偏りを緩和するアルゴリズムを提案する。
- 提案手法によって推薦システムの推薦精度や出現頻度の公平性がどの程度変化するのかを検証する。

2 関連研究

2.1 協調フィルタリング

協調フィルタリングは、多くの推薦システムで広く導入されている。協調フィルタリングとは、推薦対象のユーザと嗜好や

興味が類似したユーザによる購入や評価といった行動履歴を参照することで、そのユーザに適合する商品を推薦する手法である [1] [2]。すなわち、この手法では人々は似たような嗜好や興味を持っており、それらの嗜好は安定しているという前提に基づいている [5]。しかし、Su ら [6] は、「灰色の羊」と呼ばれる、協調フィルタリングによる推薦の恩恵を受けないユーザの存在について言及している。例えば、映画の推薦を行う際、推薦対象のユーザが、アクション、ホラー、コメディなど、多岐にわたるジャンルを好んでいる場合、このユーザと嗜好が合致している他のユーザを見つけることは極めて困難である。このように、協調フィルタリングでは、推薦する際にユーザの嗜好のみを利用しているため、それが複雑であったり、他者と大きく異なっていたりする場合、各ユーザの嗜好に沿った適切な推薦ができなくなる可能性がある。本研究では、評価の質が高い、すなわち、評価が客観的であり、かつ自身の嗜好にあまり左右されていないユーザに注目することで、「灰色の羊」のように嗜好を捉えることが難しいユーザに対しても、適切な推薦を行うことを目指す。

2.2 公 平 性

推薦システムの分野における公平性は、主にユーザに着目した公平性と商品に着目した公平性の 2 種類に分類される。ユーザの公平性を扱った研究では、人種や国籍、性別といったユーザの属性が偏った推薦結果をもたらす可能性を指摘している [7]。ユーザの公平性の中でも、Li らの研究 [8] では、属性や嗜好が類似したユーザ間では推薦結果が類似することを担保する個人の公平性と、異なるグループ間では推薦結果の分布が等しくなることを重視する集団の公平性の 2 つの概念を掲げている。すなわちユーザの公平性とは、異なる集団や個人においても一貫した推薦のパフォーマンスを提供することを意味する。商品の公平性を扱った研究では、Abdollahpouri ら [9] は、商品の公平性が欠けた推薦システム内ではユーザのインタラクションが人気の偏りに影響され、ユーザの嗜好を損なった推薦結果や更なる人気の偏りの増加に繋がると述べている。本研究では推薦システムにおける公平性の中でも商品の公平性に注目し、既存研究に即して人気の偏りを考慮した手法を考案する。

2.3 人気バイアスに関する研究

推薦システム内の人気による偏りが及ぼす悪影響については数多くの文献で議論されている。Zheng らの研究 [10] では、人気の偏りが強い推薦システムでは、ユーザは多くの人が興味を持っているという理由で、推薦された商品をクリックしたり購入したりすることによってユーザのインタラクションにユーザ自身の純粋な興味や嗜好が失われている可能性を指摘している。また、推薦結果の意外性や新規性によってユーザに予期せぬ体験をもたらすことを意味する概念である「セレンディピティ」の損失についても言及されている [11] [12]。

人気による偏りを取り除く方法の一つとして Inverse Propensity Scoring (IPS) という、ユーザと商品のインタラクションに再度重み付けすることで，データ分布が均等になるように調整する手法が用いられている [13] [14]。しかし、この手法では、傾向の推定が難しくモデルの分散が大きいため、実際にはうまく機能しない。全ての人気バイアスが推薦において悪い影響を与えているとは限らないため、このような人気の偏りを闇雲に取り除く方法は、必ずしもうまくいかないことが多い。

Causal Embedding [15] という手法では、偏りの少ない均質なデータ上でユーザの暗黙的なフィードバックに関する行列分解を行うことで商品の人気の偏りを取り除く。しかし、この手法ではモデルが有用な情報を学習しづらく、偏りの少ないデータセットのみに依存した学習を行うため、安定したパフォーマンスが発揮できないという欠点がある [16]。

また、Zhang ら [17] は、人気の偏りを取り除く従来の研究とは異なり、人気の偏りを活用することで推薦精度を向上させている。具体的には、推薦を商品集合、ユーザ集合、商品とユーザのインタラクション確率、商品の人気という 4 つのノードからなる因果グラフとみなし、商品の人気からインタラクション確率へとつながる経路を断ち切ったうえでのインタラクション確率を求めることで、人気の偏りを活用したランキング付けを行っている。

しかし、これらの研究では「商品に対する評価の質」などを考慮していないため、推薦精度や商品の人気の偏りに関して、改善の余地があると考えられる。

3 提 案 手 法

本研究では、「商品に対する評価の質」に注目することで人気の偏りを考慮した手法を提案する。

3.1 語句の定義

まず、本研究で扱う語句について、その定義を明確にする。

質の高い評価

商品に対してユーザが与える評価のうち、ユーザ自身の嗜好や都合に左右されず、客観的で品質を重視した正当な評価のことを「質の高い評価」と表現する。

人　　気

商品の人気に関して、

- 単に商品の売り上げが大きいこと、
- 多くのよい評価をもらうこと、

のどちらも人気があると解釈できる。しかし本研究では、評価の高い商品が本当に品質の高さを評価されて高評価を獲得しているのか、それとも品質が伴っていないにも関わらず高評価を得ているのかを区別することに注目することにする。したがって、本研究では、評価の内容や質の高さを問わず高い評価値を集めていることを「人気」と定義する。

商品のカテゴリー／ジャンル

本研究では、ユーザが商品を探している領域を「カテゴリー」、カテゴリー内で細分化された分野を「ジャンル」と使い分ける。例えば、ユーザがおすすめの映画を求めている場合、「映画」が

推薦カテゴリーであり、映画の中でもホラー映画やアクション映画など、より詳細な属性をジャンルと定義する。

商品の公平性

一般的に商品の公平性とは、ロングテールな商品も人気商品と等しく推薦の機会が与えられることを指す。しかし、闇雲に推薦の機会を均等にすることは商品自体の品質を一切考慮していないことを意味する。また、商品の品質を無視した推薦は、質の高い商品を提供する企業のモチベーションを損ね、市場全体の停滞をもたらし得る。このように、推薦の機会を均等にすることは数字の上では公平とみなせるが、長期的な目で見ると必ずしも良い結果につながるとはいえない。一方で、品質が優れた商品の推薦頻度を増加させることは、推薦頻度だけを見ると公平とは言えないが、品質が正当に評価される環境は市場全体の健全化にも繋がる。したがって、本研究では、商品の公平性の中でも、(1) 人気のない商品の推薦機会が増えることと、(2) 品質の良い商品の推薦機会が増えることの 2 つを区別し、(1) を「推薦頻度の公平性」、(2) を「評価の公平性」と定義する。

3.2 評価の質が高いユーザの選定

1 章で述べたように、評価の質が高いユーザ、すなわち、評価が客観的であり自身の嗜好に左右されていないユーザに注目することで、商品そのものの品質を重視した推薦が提供でき、人気の偏りによって生じる諸問題が解決できるのではないかと考える。

具体的には、次式によって評価の質が高いユーザを選定する。

$$\frac{\sum_{i \in I_\sigma} |v_{u,i} - \overline{v_i}|}{|I_{u,i}|} \tag{1}$$

ここで、I_σ は評価値の分散が σ^2 より小さい商品の集合、i は I_σ 内の商品、u は商品に対する評価数が十分に存在するユーザ、$v_{u,i}$ はユーザ u が商品 i に付けた評価値、$\overline{v_i}$ は各ユーザが商品 i に付けた評価値の平均、$|I_{u,i}|$ は I_σ 内でユーザが付けた評価の数を表す。

この式 (1) によって、あるユーザが商品につけた評価値とその商品の平均評価値の差の総和を求め、そのユーザの評価数で割ることでユーザの 1 評価あたりの評価値の正確さを導出する。この値が小さければ小さいほどユーザが付ける評価値が客観的で正確であるとみなすことができると考える。

式 (1) に用いる商品の集合に評価値の分散の上限を設けた理由としては、評価値の分散が低い商品、すなわち、多くの人が同じような評価をしている商品は、ユーザの嗜好による評価の大きなずれが発生しづらく、純粋な品質のみを評価しやすい商品であると考えたからである。仮に評価値の分散の上限を取り払うと、ユーザが評価 “1” を付けた 100 人、評価 “5” を付けたユーザが 100 人いる商品は平均評価が “3” であるため、式 (1) によると評価 “3” を付けたユーザは客観的で正確な評価をしたとみなされる。しかし、極端な例を挙げると、この商品の品質の正当な評価が 1 であるが、100 人のユーザが、「自身の嗜好にたまたま合致した」、「自分の贔屓のブランドである」、「品薄であり、やっとの思いで手に入れた」、「有名人がお勧めしていた」など、品質とは関係のない様々な理由で評価 “5” を付けたことで、偶然、商品の平均評価が “3” になった、という状況も想定できる。この場合、評価 “3” をつけたユーザは式 (1) では正当な評価をしたとみなされるが実際には過大評価をしていることになる。

このように、評価値の分散に上限を設けることは、客観性が欠けた評価が集まって評価値がばらつき、評価値の平均が品質に対する正当な値を反映しなくなってしまった商品を除外することを目的としている。また、必然的に評価値の分散が小さい商品に対して、各ユーザは平均に近い評価を付けていることになるため、一つの商品に対する各ユーザの評価値と平均評価値とのずれは差がつかない。しかし、様々なジャンルの商品に対する評価値のずれの総和を計算することで、ジャンルを問わず満遍なく客観的な評価を付けているユーザを見つけることができるものと考える。

4 実　　験

本研究では、評価の質が高いユーザの選定に関する予備実験、提案手法の推薦精度と公平性の検証を行う本実験、ユーザに対してパーソナライズされた提案手法の検証を行う追加実験を行った。

4.1 予 備 実 験

3.2 節で述べた評価の質が高いユーザの選定について、データセットを用いて予備実験を行った。

4.1.1 実験の目的

予備実験では、推薦カテゴリー内のあらゆるジャンルに対して客観的な評価を付けるユーザの存在と、そのユーザが付けた評価値を利用することの有効性を実証することを目的としている。

4.1.2 データセット

予備実験には MovieLens 20M Dataset [18] という、ユーザが映画を評価したデータを含んだデータセットを用いた。予備実験で映画のデータセットを用いた理由としては、

- 音楽やファッションなどのカテゴリーと比較して不適切な推薦をした際のユーザの時間の損失が大きく、推薦精度がより求められること、
- ユーザの嗜好が強く表れるカテゴリーでありつつも客観的な評価を付けやすい、すなわち、主観的な評価を付けるユーザと客観的な評価を付けるユーザの両者がデータセット内に存在している可能性があること、

などが挙げられる。このデータセットでは、138,000 人のユーザが 27,000 本の映画に対して 25,000,000 件の評価を付けている。

4.1.3 実 験 手 順

以下の手順でデータセットを用いて予備実験を行った。

(i)　各映画の評価値の平均を計算

評価件数が少ない映画は少数のユーザの嗜好が強く反映され、偏った評価値になってしまう可能性があるため、本実験では評価件数が一定数を満たす映画に絞っている。

(ii)　映画を 2 つのデータセットに分割
(i) で評価値の平均を計算した映画群を 2 つのデータセットに分割する。分割は完全にランダムに行う。

(iii)　データセット内の各映画の平均評価値とユーザが与えた評価値の誤差を算出
各ユーザの評価値の誤差の総和をユーザが付けた評価件数で割ることで各ユーザの誤差の平均を算出する。

(iv)　ユーザの平均誤差の小ささを各データセットで順位付け
各データセットにおいて平均誤差の小ささでユーザをソートし、2 つのランキングを作成する。

(v)　2 つのデータセットの順位の類似度を測る
順位の類似度の計測にはスピアマンの順位相関係数という尺度を用いる。スピアマンの順位相関係数とは、2 つの順位データの相関の強さを測る指標であり、-1 から 1 の値を取る。本相関係数は、次式で定義される。

$$\rho = 1 - \frac{6\sum_i d_i^2}{n(n^2-1)}$$

ここで、n は各データセットの要素数、d_i は i 番目の要素の順位の差を表す。この順位相関係数が 0 より大きければ 2 つのデータセット間での商品の平均評価値とのユーザの評価値の誤差の小ささの順位に正の相関が、0 より小さければ負の相関があることを示す。

4.1.4　実験結果・考察

初めに、4.1.3 節の実験手順 (i) において評価件数の下限を変更し、順位相関の計算を 5 回ずつ行った結果を表 1 に示す。また、評価件数の上限を変更した場合の順位相関について表 2 に示す。

表 1: 各評価件数下限における順位相関

評価件数下限	順位相関	映画本数
1	0.803	59047
100	0.796	10326
500	0.798	5386
1000	0.779	3794
5000	0.761	1223
10000	0.697	588

表 2: 評価件数に上限を設けた場合の順位相関

評価件数上限	順位相関	映画本数
100	0.414	48756
1000	0.477	55257

表 1 と表 2 より、評価件数の上限や下限によらず、順位相関は正の値を示している。この結果から、分割した 2 つのデータセット間でのユーザの評価と平均評価との誤差の小ささの順位にはある程度相関がある、つまり、片方のデータセットで平均に近い評価をしているユーザはもう一方のデータセットでも平均に近い評価をしている傾向にあることが分かった。これは、映画のジャンルを問わず、常に一定以上の客観性を備えた評価をしているユーザが存在していることを裏付けていると解釈できる。

表 1 と表 2 を比較すると、評価件数に上限を設ける、すなわち、評価件数が少ない映画のみでユーザの評価値の平均との誤差を求めると、順位相関が上限を設けない場合より大きく減少する。これは、片方のデータセットで誤差が少なく客観的な評価をしていたユーザが、もう一方のデータセットで必ずしも客観的な評価をしているとは限らないことを表している。このことから、データセット内で客観的な評価をしているユーザを選定する際は、評価件数の多い映画がデータセット内で占める割合に着目する必要があると考える。

表 1 に注目すると、評価件数の下限を設けない場合から徐々に下限を引き上げていくと 500 件あたりまではほとんど順位相関に変動はないが、1000 件を超えると徐々に減少する。この原因として、下限を高く設定したために条件を満たす映画の絶対数が減少したことと、評価件数の多い映画の比率が増加したことの 2 つが挙げられる。しかし、上述したように、表 1 と表 2 の比較によって、評価件数の多い映画は順位相関にプラスの影響をもたらしていると考えられるため、順位相関が減少する原因は前者であると推察できる。そこで、映画本数を 500 本に統一した上で評価件数の下限を変更して順位相関を計算した結果を表 3 に示す。これらの結果から、映画本数が同じである場合は下限が高い、すなわち、評価件数が多い映画の比率が大きいほど順位相関が高くなることが分かった。

4.1.5　予備実験のまとめ

本節では、予備実験の結果をまとめる。MovieLens のデータセットにおいて、2 分割したデータセットのそれぞれにおいて、商品に対するユーザの評価値と商品の平均評価値との誤差の小ささの順位を求めたところ、正の順位相関が確認できた。このことから、ジャンルを問わず広く平均に近い評価を付けているユーザが存在していることが推測できる。また、データセット内の映画の本数は多ければ多いほど、これに加え、データセット内で評価件数が多い映画が占める割合が高ければ高いほど、順位相関が高くなるという傾向が確認できた。しかし、データセット内の映画の本数とデータセット内で評価件数が多い映画が占める割合はトレードオフの関係にあり、評価件数の下限を高く設定すればするほどその条件を満たす映画の本数は必然的に減少してしまう。したがって、3.2 節の評価の質が高いユーザの選定を実行する際には、推薦カテゴリー内のデータセットの評価件数の分布に注目し、評価件数の少ない映画を選定に用いるデータから除外しつつ映画の本数が大きく減少することがないよう慎重に評価件数の下限を定める必要があると考えられる。

表 3: 映画本数を統一した場合の順位相関

評価件数下限	順位相関
0	0.254
100	0.295
500	0.291
1000	0.395
5000	0.534
10000	0.684

4.2 本 実 験

4.1 節の予備実験を踏まえ、実際に推薦対象のユーザにおすすめの商品を提示するという、推薦システムの一連の過程を再現する実験を、提案手法に基づいて行った。また、本研究では商品の質に人気が伴っている商品が正当に流通するような推薦を行うことを目的としているため、推薦精度に加えて商品の公平性を評価指標として用いた。

4.2.1 実 験 概 要

予備実験と同様に MovieLens のデータセットを用いた。推薦対象のユーザをデータセット内からランダムに選択し、そのユーザに対して提案手法により各映画の予測評価値を算出した。また、予測評価値順に映画を並べたランキングを推薦対象のユーザへの推薦リストとした。

4.2.2 評 価 指 標

本実験では、現行の推薦システムにおける、人気による推薦頻度の格差を解消することを目的としているため、提案手法の評価指標として、純粋な推薦精度の高さに加えて商品の公平性を考慮している。

推 薦 精 度

本研究における推薦精度の測定に用いた評価指標は以下の通りである。

正 誤 評 価

正誤評価では推薦対象のユーザの評価の良し悪しを適切に予測しているかを測る。この評価方法によって、ユーザが高評価を付けている商品を推薦し、低評価を付けている商品を推薦しないようにできるか、すなわち提案手法がユーザの嗜好を正しく捕捉できているかを検証することが可能である。具体的な評価指標として、適合率 (Precision)、再現率 (Recall)、F 値を用いた。

1 章でユーザが付ける評価には客観的なものと主観的なものがあることを説明した。この性質について補足すると、高評価が付けられる理由として「品質がいいから」というシンプルな理由のほかに「人気だから」や「好きな芸能人が使っていたから」といった主観的な理由も多くある。一方で低い評価を付ける理由としては品質の悪さに由来するものが大多数であり、「人気がないから」や「気に食わないから」など、あえて主観的な理由で低評価を付けることは少ないと考える。したがって、推薦精度の評価をする際に、推薦対象のユーザが商品に対して高評価を付けていることは、必ずしも高品質であることを保証するものではない。つまり、提案手法によって推薦された商品集合に、推薦対象のユーザが高評価をつけた商品が含まれていないことは、必ずしも間違った推薦とは言えない。逆に、推薦対象のユーザが低評価を付けていた商品が提案手法によって推薦された商品集合に含まれていることは、先述の低評価を付ける際の心理が品質の悪さに由来することを考えると、可能な限り避けたい。一般的に再現率は推薦漏れの少なさ、適合率は推薦ミスの少なさを表す指標であり、この二つの関係はトレードオフであるといわれる。しかし、以上のことから、本研究では品質の割に売り上げが伸びない商品の救済を目的としているため、推薦対象のユーザが高評価をつけた商品を推薦した割合である再現率よりも、推薦した商品を推薦対象のユーザが高評価を付けている割合である適合率をより重視するべきだと考える。

順 位 評 価

順位評価では予測評価値が高い順に並べた商品の順位の正当性を測ることが可能である。スマートフォンや PC に一度に表示できる情報量には限度があるため、実際に EC サイトや配信サービスで導入されている推薦システムでは、推薦上位の商品が何件か抜粋されて画面に表示されるのみであり、ユーザには具体的な予測評価値などは公表されない。つまり、推薦システムの実用性という観点で考えると、絶対的な予測評価値の正確性よりもユーザの嗜好に即した商品が正しく上位に順位付けされることの方が重要である。具体的な評価指標として、normalized Discounted Cumulative Gain (nDCG) [19] を用いた。この評価指標ではランキングが上位のときの推薦ミスほど大きく減点されるため、一度に推薦できる商品が限られている現行の推薦システムに適した指標である。

商品の公平性

3.1 節で述べたように、本研究では商品の公平性を、(1)「推薦頻度の公平性」と (2)「評価の公平性」で区別している。(1) の指標では、ロングテールな商品の推薦頻度が増え、人気な商品の推薦頻度が抑えられることで商品ごとの推薦頻度に格差が無くなることが良い推薦とみなされる。一方で、(2) の指標では質の高い商品の推薦頻度が増え、質の低い商品の推薦頻度が抑えられることを良い推薦とみなしている。

推薦頻度の公平性

推薦頻度の公平性は推薦リスト内のロングテールな商品の割合 (Long-tail Rate) で測定する [20]。

評価の公平性

評価の公平性を正確に測るには商品の質の高さを具体的な値で定義する必要があるが、質の高さを直接定量的に評価するのは困難である。そこで、3.2 節の式 (1) によって定めた評価の質が高いユーザが付けた評価に注目し、推薦リスト内の商品に付けられた全評価のうち、評価の質が高いユーザに付けられた高評価の割合を評価の公平性の指標と定義する。評価の質が高いユーザが高評価を付けた商品は、他のユーザが高評価を付けた商品に比べて品質が伴っている可能性が高いため、この指標によって評価の公平性を評価することが可能であると考える。なお、評価の公平性を、推薦リストを導出するのに用いたユー

ザと同じユーザを用いて測ると必然的に高い値が出るため、推薦商品の決定と公平性の測定には異なるユーザを用いた。

4.2.3 比較手法

比較手法としては、商品の平均評価を高い順に推薦するという最もシンプルな方法 (以下、比較手法 1) と協調フィルタリング (CF) を用いた。人気のある商品をそのまま推薦する比較手法 1 や、ユーザの嗜好に沿っているが公平性を考慮していない推薦を提供する協調フィルタリングと比較することによって、どの程度、既存手法での精度を保ちつつ公平性を向上させることを、提案手法によってできるのかを検証する。

4.2.4 結果・考察

商品に対して一定数の評価を付けているユーザ 500 人に対して提案手法と比較手法を用いて実験を行った。評価の質が高いユーザの人数 $|u_r|$ を変更して比較した。以下にその結果を示す。

表 4: 各条件における推薦精度

(i) $|u_r| = 100, N = 100$ の場合均

評価指標	提案手法	比較手法 1	CF
Precision	0.292	0.245	0.315
Recall	0.331	0.279	0.524
F 値	0.288	0.242	0.382
nDCG	0.914	0.923	0.865

(ii) $|u_r| = 50, N = 100$ の場合

評価指標	提案手法	比較手法 1	CF
Precision	0.326	0.244	0.345
Recall	0.435	0.285	0.573
F 値	0.372	0.244	0.419
nDCG	0.910	0.925	0.947

(iii) $|u_r| = 200, N = 100$ の場合

評価指標	提案手法	比較手法 1	CF
Precision	0.287	0.245	0.345
Recall	0.321	0.278	0.573
F 値	0.281	0.242	0.419
nDCG	0.917	0.923	0.947

表 5: 商品の公平性

(i) $|u_r| = 100, N = 100$ の場合

評価指標	提案手法	比較手法 1	CF
Long-tail Rate	0.580	0.530	0.451
評価の公平性	0.274	0.373	0.0985

(ii) $|u_r| = 50, N = 100$ の場合

評価指標	提案手法	比較手法 1	CF
Long-tail Rate	0.580	0.530	0.494
評価の公平性	0.250	0.332	0.0802

(iii) $|u_r| = 200, N = 100$ の場合

評価指標	提案手法	比較手法 1	CF
Long-tail Rate	0.550	0.530	0.494
評価の公平性	0.295	0.377	0.143

初めに推薦精度の評価を述べる。

正誤評価については適合率と再現率の両指標とも、全条件において比較手法 1 の精度を上回っており、人気のある商品だけを推薦する手法に比べると推薦対象のユーザの嗜好に沿った推薦ができていることが分かる。しかし、提案手法と協調フィルタリング (CF) の正誤評価の精度を比べると、後者の方が優れている。 特に、再現率はユーザ単位でみても協調フィルタリングの精度の方が大きく上回っているため、ユーザの嗜好を捉え、ユーザが高評価を付けるであろう商品を推薦することに関しては協調フィルタリングが適しているといえる。

ただし、本実験で重視している適合率については、協調フィルタリングと比べても大きく劣っていない。ユーザが低評価を付けるであろう商品を推薦しないことに関しては協調フィルタリングと同等の性能を有していると考えられる。

nDCG に関しては大きな特徴の違いがみられなかった。数値評価と順位評価に関しては既存手法と比べても大きく精度を落としておらず、安定した推薦精度を発揮しているといえる。

評価の質が高いユーザの人数を変更した場合、nDCG は人数による変動はほぼなかったが、再現率と適合率は人数が少なくなるにつれて向上している。この結果から、評価の質が高いユーザは人数の多さよりも一人一人の評価の質の高さが精度に依存すると推測できる。

以上より、提案手法は既存手法と比べて大きく精度を落とすことはなく、条件を変更しても安定している傾向にあった。一方で、ユーザ単位でみると、精度が協調フィルタリングと比べてやや劣る箇所が見受けれられ、改善の余地があると考える。

続いて、商品の公平性の全体的な評価としては、協調フィルタリングにおける公平性と比較すると、どちらの指標も提案手

法が上回っており、一定の精度を保ったまま公平性を向上させられることが分かった。また、評価の公平性については比較手法 1 が最も高性能であったが、高評価の件数が多い商品を順に推薦するという手法であることにより、必然的に評価件数が増え、評価の質が高いユーザの評価が集まりやすかったためだと考えられる。

条件ごとの比較としては、評価の質が高いユーザが多い方が評価の公平性は向上するが、Long-tail Rate は減少した。推薦精度の結果と合わせると、評価の質が高いユーザの人数は推薦精度と公平性のどちらにも干渉し、闇雲に人数を増やせば精度が向上するとも限らず、高い水準の公平性を備えたユーザのみで構成することが重要だと推測できる。

提案手法によって推薦精度を損なわずに公平性を向上させることが可能となったが、課題として現状の手法では、ユーザに一切パーソナライズされておらず同じ推薦結果を提供していることが挙げられる。この手法のまま推薦を継続すると同じ商品のみが推薦され続け更なる推薦頻度の偏りが、生じてしまうため、短いスパンでの推薦には有効であるが、長期的な実用性は伴っていない。そこで、推薦結果に変化を付けるために提案手法に修正を加えた追加実験を行った。

4.3 追加実験

評価の質が高いユーザが付けた評価値を利用することによって向上した公平性に加えて、ユーザごとにパーソナライズされた推薦結果を提供し、かつ推薦精度の水準を高めるために、協調フィルタリングのアルゴリズムの一部を導入し、推薦リストの決定方法に修正を加えた。修正点としては、式 (1) により決定した評価の質が高いユーザ群の中で推薦対象のユーザとの類似度を求めて順位付けし、順位に応じて評価値に重みを付け、類似度の高いユーザの評価値をより重視するようにした。なお、類似度の計算にはコサイン類似度を用いた。

前実験では評価の質が高いユーザ内では評価の重みはすべて等しく、ユーザごとに差をつけていなかったが、追加実験では推薦対象のユーザとの類似度によって差を生み出している。具体的な推薦リストの決定方法としては、まず評価値 “3” を基準値の “0” として、評価値 “3” からの距離を商品のスコアとして定義する。ここで、類似度が高いユーザの付けた高評価にはプラス方向の重みを、低評価にはマイナス方向の重みを与えたいため、類似度が高いほど評価値 “3” からの距離に傾斜をかけている。具体的には推薦対象のユーザと類似度が高いユーザが付けた評価値 5 にはスコア “4”、評価値 “1” にはスコア “-4” を付与する一方で、類似度が低いユーザが付けた評価値 “5” にはスコア “2”、評価値 “1” にはスコア “-2” を付与する、といったように、類似度によって与えるスコアを調整する。そして式 (1) によって求めた評価の質が高いユーザのスコアの累計を求め、累計値をもとに商品の推薦リストを決定する。

4.3.1 結果・考察

提案手法を修正した場合 (以下、提案手法 2) の実験結果を、以下に示す。

表 6: 提案手法 2 の推薦精度と公平性

(i) $|u_r| = 100, N = 100$ の場合

評価指標	提案手法 2	提案手法	CF
Precision	0.474	0.292	0.315
Recall	0.582	0.331	0.524
F 値	0.493	0.288	0.382
nDCG	0.908	0.914	0.865
Long-tail Rate	0.413	0.414	0.451
評価の公平性	0.145	0.290	0.0985

(ii) $|u_r| = 50, N = 100$ の場合

評価指標	提案手法 2	提案手法	CF
Precision	0.554	0.326	0.345
Recall	0.810	0.435	0.573
F 値	0.635	0.372	0.419
nDCG	0.891	0.910	0.947
Long-tail Rate	0.429	0.580	0.494
評価の公平性	0.118	0.250	0.0802

(iii) $|u_r| = 200, N = 100$ の場合

評価指標	提案手法 2	提案手法	CF
Precision	0.544	0.287	0.345
Recall	0.727	0.321	0.573
F 値	0.595	0.281	0.419
nDCG	0.912	0.917	0.947
Long-tail Rate	0.409	0.550	0.494
評価の公平性	0.167	0.295	0.143

以上の結果に関する考察を述べる。

各条件を通して、再現率と適合率が飛躍的に向上した。これは、4.2 節の本実験から質が高いユーザの評価の利用することで一定の精度が担保される上に、類似度を導入したことにより、ユーザごとにパーソナライズされたことに起因すると推測できる。また、協調フィルタリングの精度を上回った理由としては、協調フィルタリングでは嗜好が近いユーザが高評価を付けた商品が推薦されるが、嗜好から外れた商品を推薦することもあり、その商品には品質が一切保証されていない。一方で、提案手法 2 では一定の品質が保証されているため、仮に嗜好から外れた商品を推薦したとしてもユーザがその商品に低評価を付ける可能性は少ない。このように、提案手法 2 ではユーザが低評価を付ける商品を推薦するケースが少ないため結果的に推薦精度が向上したと考える。公平性については、Long-tail Rate、評価の公平性、ともに元の提案手法よりも損なわれた。推薦に用いる評価の質が高いユーザの集合自体は 4.2 節の実験と変わらないため、類似度によって評価値に重みを付けたこと、すなわ

ち、ユーザごとにパーソナライズしたことが公平性の低下の直接の原因であり、公平性と推薦精度はトレードオフの関係であると考えられる。しかし、適合率、再現率、評価の公平性については協調フィルタリングよりも向上していることから、既存手法に比べて推薦精度を向上させながら、一定の品質が備わった商品を推薦することが可能である。また、全実験を通して推薦精度を保ちながら公平性を向上させることが可能であったため、状況ごとに推薦精度と公平性のどちらを重視するべきかに応じた使い分けができると考える。例えば、一度商品を購入した後に、再度利用するリピーターが少ないような EC サイトでは、推薦精度を重視することでユーザの嗜好に適合した商品が推薦され、そのユーザの継続した利用に結び付くことが想定できる。このように、提案手法の使い分けによって様々な目的での導入が可能である。

5 結論・今後の展望

本研究では、ユーザが与える評価の質に注目し、評価の質が高いユーザの存在を利用して推薦システムにおける商品の人気や出現頻度の偏りを緩和するアルゴリズムを提案した。提案手法を用いた本実験では、条件ごとに多少の結果の差異はあるものの、全体的な傾向として、推薦精度を保ちつつ公平性を向上させられることが確認できた。ユーザに対するパーソナライズ性を強化した追加実験では提案手法よりも推薦精度を向上させることができたが、公平性はやや損なわれた。この結果からユーザの嗜好に適合した推薦と公平性を追求した推薦の両立の難しさが見受けられつつ、提案手法によって、状況に応じてユーザの嗜好の考慮と公平性の追求を切り替えながら推薦することが可能であると考えた。

今後の展望としては、評価の質が高いユーザの選定方法などに工夫を加えて、推薦精度や公平性などの各指標の更なる向上を目指しつつ、MovieLens 以外の様々なカテゴリーのデータセットを用いた実験を行ない、カテゴリーごとの人気や評価の特徴を把握することで、提案手法を改良していきたい。

文　献

[1] Paul Resnick, Neophytos Iacovou, Mitesh Suchak, Peter Bergstrom, and John Riedl. GroupLens: An Open Architecture for Collaborative Filtering of Netnews. In *Proc. of the 1994 ACM Conference on Computer Supported Cooperative Work (CSCW '94)*, pp. 175–186, 1994.

[2] Joseph A Konstan, Bradley N Miller, David Maltz, Jonathan L Herlocker, Lee R Gordon, and John Riedl. GroupLens: Applying Collaborative Filtering to Usenet News. *Communications of the ACM (CACM)*, Vol. 40, No. 3, pp. 77–87, 1997.

[3] Eli Pariser. *The Filter Bubble: What the Internet is Hiding from You*. Penguin Press, 2011.

[4] Robert K Merton. The Matthew Effect in Science: The Reward and Communication Systems of Science are Considered. *Science*, Vol. 159, No. 3810, pp. 56–63, 1968.

[5] Zhi-Dan Zhao and Ming-Sheng Shang. User-based collaborative-filtering recommendation algorithms on hadoop. In *Proc. of the IEEE 2010 3rd International Conference on Knowledge Discovery and Data Mining (WKDD '10)*, pp. 478–481, 2010.

[6] Xiaoyuan Su and Taghi M Khoshgoftaar. A Survey of Collaborative Filtering Techniques. *Advances in Artificial Intelligence*, Vol. 2009, pp. 421425:1–421425:19, 2009.

[7] Ziwei Zhu, Xia Hu, and James Caverlee. Fairness-aware tensor-based recommendation. In *Proc. of the 27th ACM international conference on information and knowledge management*, pp. 1153–1162, 2018.

[8] Yunqi Li, Yingqiang Ge, and Yongfeng Zhang. Tutorial on fairness of machine learning in recommender systems. In *Proc. of the 44th international ACM SIGIR conference on research and development in information retrieval (SIGIR '21)*, pp. 2654–2657, 2021.

[9] Himan Abdollahpouri, Masoud Mansoury, Robin Burke, and Bamshad Mobasher. The connection between popularity bias, calibration, and fairness in recommendation. In *Proc. of the 14th ACM Conference on Recommender Systems (RecSys '20)*, pp. 726–731, 2020.

[10] Yu Zheng, Chen Gao, Xiang Li, Xiangnan He, Yong Li, and Depeng Jin. Disentangling user interest and conformity for recommendation with causal embedding. In *Proc. of the Web Conference 2021 (WWW '21)*, pp. 2980–2991, 2021.

[11] Qiuxia Lu, Tianqi Chen, Weinan Zhang, Diyi Yang, and Yong Yu. Serendipitous personalized ranking for top-n recommendation. In *Proc. of the 2012 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology (WI-IAT '12)*, pp. 258–265, 2012.

[12] Zihao Zhao, Jiawei Chen, Sheng Zhou, Xiangnan He, Xuezhi Cao, Fuzheng Zhang, and Wei Wu. Popularity bias is not always evil: Disentangling benign and harmful bias for recommendation. *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, pp. 1–13, 2022.

[13] Tobias Schnabel, Adith Swaminathan, Ashudeep Singh, Navin Chandak, and Thorsten Joachims. Recommendations as Treatments: Debiasing Learning and Evaluation. In *Proc. of the 33rd International Conference on Machine Learning (ICML '16)*, pp. 1670–1679, 2016.

[14] Thorsten Joachims, Adith Swaminathan, and Tobias Schnabel. Unbiased Learning-to-Rank with Biased Feedback. In *Proc. of the 10th ACM International Conference on Web Search and Data Mining (WSDM '17)*, pp. 781–789, 2017.

[15] Stephen Bonner and Flavian Vasile. Causal embeddings for recommendation. In *Proc. of the 12th ACM conference on recommender systems (RecSys '18)*, pp. 104–112, 2018.

[16] Tianxin Wei, Fuli Feng, Jiawei Chen, Ziwei Wu, Jinfeng Yi, and Xiangnan He. Model-agnostic counterfactual reasoning for eliminating popularity bias in recommender system. In *Proc. of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining (KDD '21)*, pp. 1791–1800, 2021.

[17] Yang Zhang, Fuli Feng, Xiangnan He, Tianxin Wei, Chonggang Song, Guohui Ling, and Yongdong Zhang. Causal Intervention for Leveraging Popularity Bias in Recommendation. In *Proc. of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '21)*, pp. 11–20, 2021.

[18] F Maxwell Harper and Joseph A Konstan. The movielens datasets: History and context. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, Vol. 5, No. 4, pp. 1–19, 2015.

[19] Kalervo Järvelin and Jaana Kekäläinen. Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems (TOIS)*, Vol. 20, No. 4, pp. 422–446, 2002.

[20] Yingqiang Ge, Juntao Tan, Yan Zhu, Yinglong Xia, Jiebo Luo, Shuchang Liu, Zuohui Fu, Shijie Geng, Zelong Li, and Yongfeng Zhang. Explainable Fairness in Recommendation. *arXiv preprint arXiv:2204.11159*, 2022.