レビューテキストの観点情報に基づく残差補正型推薦手法

栗巣野陽太† 落合 桂一† 戸田 浩之†

† 横浜市立大学データサイエンス学部 〒 236–0027 神奈川県横浜市金沢区瀬戸 22-2
E-mail: †{d214025a,ochiai.kei.dk,toda.hir.xg}@yokohama-cu.ac.jp

あらまし EC サイト等でユーザの選択肢は拡大し推薦システムの重要性が高まる中，ユーザの購買・閲覧履歴等に基づく協調フィルタリングが広く用いられてきた．しかし，それらの従来手法は推定根拠の説明が難しく，観測の乏しい状況での精度は不安定になりやすい．そこでレビューを活用する研究が進み，履歴だけでは捉えにくい嗜好の手掛かりを取り込むことで精度改善を達成し，説明可能性への期待も高まった．その中でもレビュー中の観点に着目して推薦へ反映する観点ベース推薦が注目を集めている．しかし，各観点に対するユーザの評価には評価基準の個人差が混入しやすく，ユーザが重視する観点は対象のアイテムによって変化し，必ずしも観点に関する言及があるわけではないことから，情報の欠落も生じやすいという課題が残る．そこで本研究は，観点情報を単独モデルの中核として用いるのではなく，既存推薦モデルの予測を必要な場面で押し上げる「補助的な補正情報」として位置付け，予測残差を観点ごとのスコアに基づく特徴で補う外付けモジュールを提案する．複数データセット・複数アーキテクチャで一貫した精度改善を確認した．

キーワード 推薦システム，協調フィルタリング，レビュー活用，観点，説明可能性

1 はじめに

推薦システムにおいて，ユーザが投稿するレビューテキストは嗜好を理解する上での有力な情報である [1][2]．レビューテキストには，購入目的や評価の理由，着目した特徴など，単純な購入履歴などからは直接観測できない情報が含まれる．特に，レビュー中でユーザが言及する具体的な属性を観点として抽出し，観点ごとの評価を活用する観点ベース推薦が注目されている [3][4][5]．例えば，ヘッドホンのレビューにおける「音質」や「装着感」といった言及が観点に該当する．

しかし，観点情報の活用には複数の課題が存在する．まず，同じ観点であってもユーザごとに評価基準が異なるという主観性の問題がある [6]．また，ユーザが重視する観点は評価対象アイテムに応じて変化するため，ユーザやアイテムを固定的なベクトルで表現する従来手法では対応が困難である [4]．さらに，レビューテキスト中の観点に関する言及は偏っており，多くのユーザ・アイテムの組み合わせにおいて共通する観点情報が得られないという観点のスパース性も問題となる [7]．これらの不確実性により，観点情報をモデル内部に直接統合すると学習が不安定化しやすい．加えて，これらの不確実性を考慮すると，観点情報が十分に得られる場合にのみ活用し，不十分な場合にはベースモデルの予測を優先するといった柔軟な制御が求められる．しかし，既存手法では観点情報がモデル内部に統合されているため，このような選択的な適用は困難である．

そこで本研究では，観点情報をモデル内部に統合せず後段からの残差補正として扱い，収縮推定により観点のスパース性を，ゲート機構により観点情報の不確実性を抑制する手法を提案する．図 1 に従来手法と提案手法の違いを示す．従来手法では観点情報をモデル内部に直接統合するため，観点情報の不確実性がモデル全体の学習に影響を与える．これに対し提案手法では，ベースモデルによる予測と観点情報に基づく補正を分離し，補正モジュールが後段から残差を加える構成をとる．この分離設計により，観点情報の不確実性がベースモデルの学習に波及することを防ぐとともに，ベースモデルの種類に依存しない形で観点情報を活用できるため，既存の推薦システムに追加モジュールとして適用可能となる．

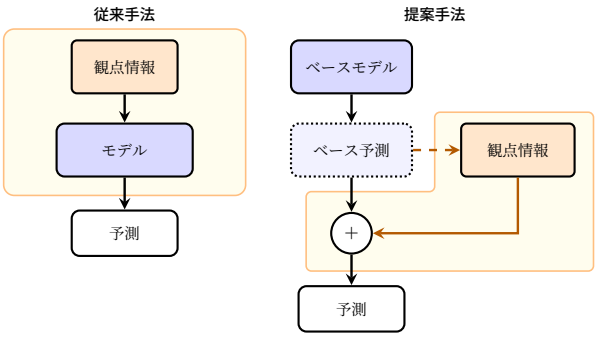

図 1 提案手法の概要

本研究では，観点情報の不確実性に対処するため，複数の設計要素を組み合わせる．具体的には，収縮推定により観点が少数しか観測されない場合の評価値推定を安定化し，ユーザとアイテムの観点プロファイル間の相互作用項により動的文脈を表現する．さらに，観点情報の可用性に基づくゲート機構により補正の適用を制御することで，観点情報が不十分な場合にはベースモデルの予測を維持する．本研究では，性質の異なる複数のベースモデルに対して一貫した改善が得られることを示すとともに，ablation study により各設計要素が予測誤差の改善に寄与することを検証する．また，履歴の少ないユーザやアイテムを対象とするコールドスタート条件下でも改善が得られる

ことを確認する．

2 関連研究

2.1 行動履歴に基づく情報推薦

推薦システムは，ユーザの過去の行動履歴から将来選好されうるアイテムを推定する基盤技術である．中でも，ユーザとアイテム間の評価・購買・閲覧などの履歴から潜在表現を学習する協調フィルタリング（Collaborative Filtering; CF）は，高い予測性能と実装容易性から広く用いられてきた．行列分解 [8] は，ユーザとアイテムをそれぞれ低次元の潜在因子ベクトルで表現し，両者の内積によってレーティングを予測する枠組みを確立した．また，クリックや閲覧など明示的な評価値が付与されていない暗黙的フィードバックに対しては信頼度重み付けが導入され [9]，NCF [10] はニューラルネットワークを用いることでユーザとアイテムの複雑な関係の学習を可能にした．

しかし，行動履歴のみに基づく手法では，推定根拠が潜在表現へ内在化され，推薦理由を解釈可能な形で説明することが難しい．さらに，ユーザがどの観点に基づいて評価したかという内訳が明示されないため，観点別の嗜好を直接扱うことも困難である [3][11]．このため，予測性能の向上に加えて，推定根拠の提示や解釈可能性の向上を目的として，レビューなどのテキスト情報を補助的に活用する研究が発展してきた．

2.2 レビューに基づく情報推薦

レビューテキストには，購入の意図や評価の根拠，着目した特徴など，行動履歴からは直接観測できない情報が含まれる．この性質を推薦に取り込む研究として，HFT [11] はトピックモデルと行列分解を組み合わせ，レビューから抽出したトピック分布を潜在因子と対応付けることで，予測結果がどのような話題と関連するかを解釈できる枠組みを示した．深層学習を用いた手法では，DeepCoNN [1] が畳み込みニューラルネットワーク（Convolutional Neural Network; CNN）によりユーザとアイテムのレビューテキストから特徴表現を抽出してレーティングを予測する手法を提案した．また，NARRE [2] はレビュー単位で重要性を学習する Attention 機構を導入し，予測に寄与したレビューを特定することで推薦理由の提示を可能にした．

これらの研究は推薦理由の提示や解釈支援に道を開いた一方で，レビュー全体を一つの特徴として扱う設計では，どの観点がどの程度予測に寄与したかを明示的に制御しにくいという課題が残る．特に，ユーザとアイテムの組み合わせに応じて着目すべき観点が変化する場合，レビュー表現を一括で扱う設計では観点単位の寄与を扱いにくい．この課題を背景として，レビューを観点単位に分解して扱う観点ベース推薦が発展してきた．

2.3 観点に基づく情報推薦

観点ベース推薦は，レビューから「どの観点について肯定的／否定的に述べているか」を抽出し，観点ごとの評価情報として予測に反映する枠組みである．レビュー情報を観点単位に分解することで，総合的な評価では見えにくい差異を観点別に比較でき，どの観点が予測に寄与したかを推薦理由として明示できる．

ANR [5] はレビューから観点ごとの表現を学習し Attention 機構で寄与した観点を特定することで，観点に基づく推薦理由の提示を可能にした．CARP [12] は Capsule Network を用いて観点レベルでユーザの肯定的・否定的な嗜好を分離して表現し，観点ごとの選好構造を明示的にモデル化する方向性を示した．MA-GNNs [13] はグラフニューラルネットワーク（Graph Neural Network; GNN）により観点間の関係性を学習し，観点を独立に扱う場合に比べて文脈を考慮した表現学習を可能にした．APH [14] は嗜好と感情の表現を分離しつつ，ハイパーグラフで観点間の高次の依存関係を捉える枠組みを提案している．

観点ベース推薦は観点単位での比較や理由提示を可能とするが，第 1 章で述べたように観点情報には主観性やスパース性といった不確実性が伴う．既存手法は観点情報をモデル内部に統合する設計が主流であり，これらの不確実性がモデル全体の学習に影響を与えやすい．本研究は，観点情報を既存推薦モデルの予測誤差を補正する信号として扱うことで，不確実性の影響を抑えながら精度改善を目指す．

3 問題設定

本研究は，ユーザ集合 $\mathcal{U}$ とアイテム集合 $\mathcal{I}$ に対して，ユーザ $u \in \mathcal{U}$ がアイテム $i \in \mathcal{I}$ に付与するレーティング y_{ui} を予測する評価値予測問題を扱う．レーティングは 1 から 5 の整数値として与えられ，予測値は区間 $[y_{\min}, y_{\max}]$（本研究では $[1, 5]$）上の実数として出力する．

推薦システムの実運用では，評価履歴のみから高精度な予測を行える場合が多い一方で，レビューテキストに含まれる「どの属性が良い／悪いか」という観点別の情報は，レビューテキスト全体を一括で扱う手法では捉えにくい詳細なユーザ嗜好を表現でき，予測精度のさらなる向上に有用である．しかし，第 2 章で述べたように，レビュー由来の観点情報には主観性，動的文脈，観点のスパース性という 3 つの課題が伴う．これらの課題により，観点情報を予測の主な入力として直接利用すると学習が不安定化し，予測精度がかえって低下する可能性がある．そこで本研究は，観点情報を「レーティングそのものを予測する主な情報」として用いるのではなく，ベースとなる推薦手法を前段に置き，その予測値に対して後段から補正量を加える構造を採用する．この構造では，観点情報の不確実性によってベース予測を誤った方向へ修正してしまう過補正が懸念されるが，本研究ではゲート機構や収縮推定によりこれを抑制する．以下，この構造を実現する提案手法の詳細を述べる．

4 提案手法

4.1 全体像

図 2 に提案手法の処理フローを示す．提案手法は，ベースとなる推薦手法による予測値に対し，レビューテキストから抽出した観点情報に基づく補正量を加算することで，最終的な予測

表 1　主要な記法

記号	説明
$\mathcal{A}$, A	観点集合，観点数
s_{ui}^{a}	ユーザ・アイテムペア (u,i) における観点 a の極性（−1〜+1）
c_{ui}^{a}	観点抽出の信頼度
p_u, q_i	ユーザ・アイテムの観点プロファイル
$\hat{y}_{ui}^{\text{base}}$	ベース推薦手法の予測値
r_{ui}	残差（$y_{ui} - \hat{y}_{ui}^{\text{base}}$）
Δ_{ui}	観点特徴から推定される補正量
g_{ui}	補正適用を制御するゲート
α	補正強度

を得る．具体的な処理の流れを以下に示す．(1) レビューテキストから観点と極性を抽出し語彙を選定，(2) 抽出された観点情報を集約し収縮推定により観点プロファイルを構築，(3) 観点プロファイルから特徴を構成しベース予測の残差を回帰，(4) 観点情報の可用性に基づくゲートで補正適用を判定，(5) 補正強度を調整し最終的な予測値を得る．

最終的な予測値は次式で定義する．

$$\hat{y}_{ui} = \text{clip}\big(\hat{y}_{ui}^{\text{base}} + g_{ui}\,\alpha\,\Delta_{ui},\ y_{\min},\ y_{\max}\big). \tag{1}$$

ここで $\text{clip}(x,a,b) = \max(a, \min(x,b))$ は値を区間 $[a,b]$ に制限する関数であり，予測値がレーティングの有効範囲外へはみ出すことを防ぐ．$\hat{y}_{ui}^{\text{base}}$ はベース推薦手法の予測値，Δ_{ui} は観点情報に基づく補正量，g_{ui} は補正適用を制御するゲート，α は補正強度である．本研究は，観点情報をベース推薦手法内部へ直接統合しない．代わりに後段補正として切り出し，観点情報の不確実性がベース推薦手法全体の学習へ波及することを避ける．

4.2　記法と定式化

ベース推薦手法は LightGCN [20] や NARRE [2] など任意のレーティング予測器を想定する．表 1 に本研究で用いる主要な記法を示す．

4.3　観点抽出と語彙構築

ユーザ・アイテムペア (u,i) に付随するレビューテキストから，依存関係解析に基づくルールベース手法により観点と極性を抽出する．極性とは，特定の観点に対するユーザの評価が肯定的か否定的かを数値化したものであり，-1.0 から $+1.0$ の範囲の実数値をとる．-1.0 は完全に否定的，$+1.0$ は完全に肯定的であることを示す．例えば，「画質が素晴らしい」というレビューテキストからは観点「画質」に対して正の極性（例：+0.8）が，「バッテリーが持たない」というレビューテキストからは観点「バッテリー」に対して負の極性（例：−0.7）が抽出される．

具体的には，形容詞修飾（amod），連結動詞構文（acomp），直接目的語（dobj）等の構文パターンで観点語と感情語を対応付ける．表 2 に各パターンの構造と抽出例を示す．複合名詞（compound 修飾）は結合して単一の観点語とし（例：“battery life”），否定語（not, never 等）が検出された場合は感情極性を反転させる．極性 s_{ui}^{a} は感情語辞書により -1.0〜$+1.0$ で付与し，信頼度 c_{ui}^{a} は同一レビュー内での言及回数とする．複数回言及された観点はより確実な嗜好を反映していると考え，観点プロファイル計算時に重みとして用いる．

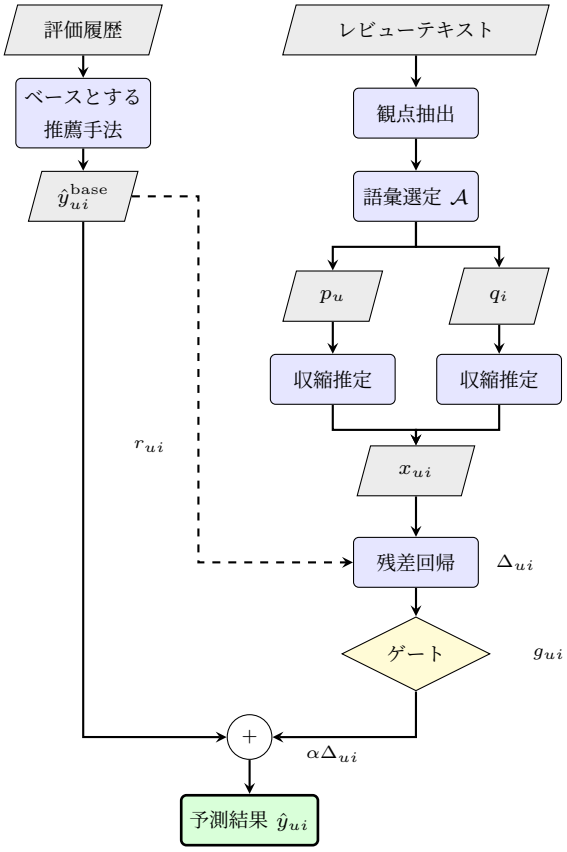

図 2　提案手法の処理フロー

表 2　観点抽出に用いる依存関係パターン

パターン	依存関係	構造	抽出例
形容詞修飾	amod	Adj ← Noun	“great quality”
連結動詞構文	acomp	Noun ← be → Adj	“screen is clear”
直接目的語	dobj	Verb → Noun	“love the design”

抽出された観点候補から，頻度と評価との相関に基づき語彙を選定する．観点 a のスコアを次式で定義する．

$$\text{Score}(a) = \log(\text{freq}(a)+1) \cdot |\text{corr}(a,y)|. \tag{2}$$

$\text{freq}(a)$ は訓練データ中での観測回数，$\text{corr}(a,y)$ は極性とレーティングの Pearson 相関係数である．この設計は，観点の有用性を「統計的な信頼性」と「レーティングとの関連性」の両面から評価する．第一項の $\log(\text{freq}(a)+1)$ は，観測頻度に基づく信頼性を表し，対数変換により高頻度観点への過度な偏りを抑制する．第二項の $|\text{corr}(a,y)|$ は，観点とレーティングの関連性を表し，絶対値を用いることで正負いずれの相関も評価対象とする．$\text{Score}(a)$ が閾値以上の観点のみを語彙 $\mathcal{A}$ として採用し，希少観点による推定不安定化を防ぐ．

4.4　観点プロファイル推定と収縮による安定化

観点スコア s_{ui}^{a} は，各ユーザ・アイテムの組 (u,i) ごとに抽出されるが，すべての観点がすべての組で言及されるわけでは

ない．例えば，ユーザ u が過去に評価した 10 個のアイテムのうち，「画質」に言及したレビューは 2 件のみ，「価格」に言及したレビューは 1 件のみといった状況が考えられる．このように，観点ごとの観測値は非常に限られているため，個別の組 (u,i) における観点スコアをそのまま特徴として用いると，大半の観点で値が得られず予測に十分な情報を提供できない．

そこで本研究は，ユーザとアイテムそれぞれについて，過去の評価履歴から観点ごとの傾向を集約した観点プロファイル $p_u \in \mathbb{R}^A$，$q_i \in \mathbb{R}^A$ を構築する．観点 a についてユーザ u が言及した評価記録の集合を $\mathcal{N}_u^a$ とし，信頼度重み付き平均を次式で算出する．

$$\bar{s}_u^a = \frac{\sum_{i:(u,i)\in\mathcal{N}_u^a} c_{ui}^a \cdot s_{ui}^a}{\sum_{i:(u,i)\in\mathcal{N}_u^a} c_{ui}^a}. \tag{3}$$

この平均は，言及回数が多い（信頼度が高い）観点評価を重視する設計となっている．

観測数が少ない場合の不安定化を防ぐため，グローバル平均 μ_a への収縮を適用する．

$$p_u^a = \frac{n_u^a \cdot \bar{s}_u^a + \tau \cdot \mu_a}{n_u^a + \tau}, \quad q_i^a = \frac{n_i^a \cdot \bar{s}_i^a + \tau \cdot \mu_a}{n_i^a + \tau}. \tag{4}$$

ここで n_u^a，n_i^a は観測数，$\tau \geqq 0$ は収縮強度である．τ=0 のときは収縮を行わず，観測がある場合は局所平均を用い，観測がない場合は μ_a を用いる．τ>0 のとき，観測数が少ないほど推定値は μ_a に近づき，観測数が十分大きいほど局所平均に近づくため，データの疎密に応じた適応的な推定が可能となる．

4.5 相互作用項と残差補正器

ユーザが重視する観点は評価対象によって変わりうる．例えば，楽器を評価する際に「音質」を重視するユーザでも，ケースを評価する際には「耐久性」を重視する可能性がある．この性質を表現するため，p_u と q_i の連結に加え，観点ごとの相互作用項 $p_u \odot q_i$（要素積）を導入する．

$$x_{ui} = [\,p_u;\ q_i;\ p_u \odot q_i\,]. \tag{5}$$

相互作用項は，ユーザとアイテムの観点プロファイルを要素ごとに掛け合わせることで，特定のペアにおいてどの観点が共に顕著であるかを表現する．p_u と q_i のみの連結では各観点についてユーザ側・アイテム側の傾向を独立に扱うことになるが，相互作用項により「このユーザがこのアイテムのどの観点で誤差を出しやすいか」を表現できる．

本研究はレーティング y を直接予測せず，ベースモデルの残差 $r_{ui} = y_{ui} - \hat{y}_{ui}^{\text{base}}$ を回帰する．残差補正とは，ベース推薦手法の予測値に対して，その予測誤差を観点情報から推定した補正量で修正することを指す．総合評価 y_{ui} だけでは「どの属性が評価に寄与したか」を観測できず，残差補正の方向性を定めにくいが，観点と極性が得られれば属性ごとに肯定／否定という形で情報を整理でき，補正器の入力として利用できる．

補正器は Ridge 回帰として次の最小化問題を解く．

$$\min_{w,b} \sum_{(u,i)\in\mathcal{D}_{\text{train}}} \big(r_{ui} - (w^\top x_{ui} + b)\big)^2 + \lambda\|w\|_2^2. \tag{6}$$

Ridge 回帰を採用した理由は，L2 正則化により高次元スパース特徴に対して推定を安定化でき，閉形式解を持つため学習・推論コストが小さく後段補正モジュールとして実用的である点にある．MLP のような非線形モデルは表現力が高い反面，残差補正という局所的役割に対して過剰適合や過補正を招きやすいため採用しない．補正器の出力を $\Delta_{ui} = w^\top x_{ui} + b$ とし，補正量をどの程度反映すべきかは自明でないため，スケーリング係数 $\alpha \geqq 0$ を導入し検証データ上で RMSE（Root Mean Squared Error）が最小となる値を探索する．

4.6 ゲート設計

観点情報がほとんど存在しない条件で補正を適用すると，観点プロファイルの推定が不安定になり過補正を引き起こす可能性がある．そこで本研究は，観点情報の可用性に基づく二値ゲート $g_{ui} \in \{0,1\}$ を導入する．

ユーザ u が訓練データで言及した観点種類数を d_u，アイテム i が言及された観点種類数を d_i とし，ゲートを次式で定義する．これらは評価数 n_u, n_i とは異なる量であり，観点情報の多様性を表す指標である．

$$g_{ui} = \mathbf{1}[d_u \geqq \text{MIN\_U}] \cdot \mathbf{1}[d_i \geqq \text{MIN\_I}] \tag{7}$$

閾値 MIN_U, MIN_I は実験設定として与える．このゲートにより，観点情報が乏しいユーザ・アイテムペアでは $g_{ui} = 0$ となり補正を適用せず，ベース予測のみを用いることで推定の安定性を確保する．

この設計は「補正を常に適用する」のではなく，「観点情報に基づく補正が成立する領域に限定して適用する」という方針を反映している．ゲートの閾値を変えることで適用率と改善幅のトレードオフを調整でき，運用上の要請に応じた柔軟な設定が可能である．

5 評価実験

5.1 データセット

本研究では，Ni ら [21] が公開する Amazon Review Data (2018) から Office Products, Musical Instruments, Toys and Games（以下 Office, Musical, Toys）の 3 ドメインを用いる．各データセットはレーティングとレビューテキストを含み，提案手法の検証に適している．表 3 にデータセット統計を示す．密度は評価数を（ユーザ数 × アイテム数）で除した値であり，ユーザ・アイテム行列のうち実際に評価が観測されている割合を表す．いずれも密度が 0.2%未満であり，現実の推薦システムが直面する疎性の課題を反映している．データ分割は 8:1:1 の分割とし，複数の乱数シードで学習・評価を繰り返して結果を集約する．

表 3 データセット統計

データセット	ユーザ数	アイテム数	相互作用数	密度
Office	7,696	4,116	51,849	0.164%
Musical	20,110	10,422	160,147	0.076%
Toys	20,090	11,044	159,814	0.072%

表 4　観点語彙の概要

順位	Office (語彙数 266)	Musical (語彙数 514)	Toys (語彙数 441)
1	price	price	star
2	product	string	toy
3	pen	product	game
4	printer	star	gift
5	quality	sound	set
6	color	quality	fun
7	ink	unit	price
8	paper	pick	piece
9	value	noise	product
10	deal	part	quality

5.2 実験設定

評価指標として RMSE を用いる．比較手法として，協調フィルタリング系の LightGCN [20] とレビュー利用型の NARRE [2] の 2 つをベースモデルとし，これらに提案手法の残差補正モジュールを後段から付与した場合の改善を評価する．提案手法の設定として，観点プロファイルは上位 K=20 観点を保持し，特徴は $x_{ui}=[p_u; q_i; p_u \odot q_i]$，収縮強度 τ=1，補正器は Ridge 回帰（λ=1.0）を用いる．スケーリング係数 α は検証集合で 0.0 〜1.5 の範囲を 0.1 刻みでグリッド探索し，観点可用性ゲート $g_{ui}=\mathbf{1}[d_u \geqq 1]\cdot\mathbf{1}[d_i \geqq 1]$（すなわち MIN_U=MIN_I=1）により補正適用を制御する．

5.3 結果と考察

5.3.1 観点語彙の定性的分析

提案手法が利用する観点語彙が各ドメインの商品特性と整合しているか確認した．観点語彙はレビューテキストから観点と極性を抽出し，頻度とレーティングとの相関に基づきフィルタリングすることで構築している．表 4 に各ドメインの観点語彙規模と上位 10 観点を示す．語彙規模は Office が 266 語，Musical が 514 語，Toys が 441 語であり，ドメインによって抽出される観点数に差がある．Office では `pen`, `printer`, `ink` など事務用品に特有な語，Musical では `string`, `sound`, `pick` など楽器固有の語，Toys では `toy`, `game`, `gift` など遊びや贈答に関連する語が上位となっており，抽出された観点語彙がドメイン特性と整合していることが確認できる．また `price`, `product`, `quality` のような共通観点も存在し，ドメイン横断的な品質要因も捉えられている．これらの共通観点と固有観点の共存は，残差補正において重要な役割を果たす．共通観点はドメイン横断的な品質要因を捉え，固有観点はドメイン特有の誤差構造を表現する手がかりとなる．

次に，ドメイン間で観点語彙がどの程度共有されているかを分析する．表 5 にドメイン間の観点語彙の共通率を示す．共通率は小さい方の語彙サイズを分母として算出しており，17.2% 〜27.8%の範囲にある．約 2 割程度の共通観点が存在する一方，語彙全体としてはドメイン固有の観点が多数を占める．Office×Toys の共通率が最も高く（27.8%），これは両ドメインで `price`, `product`, `quality` などの汎用的な評価観点が共有されているためと考えられる．一方，Musical×Toys の共通率は 17.2%と最も低く，楽器と玩具という異なる商品カテゴリの特性を反映している．この結果は，(i) 共通観点によりドメイン横断的な補正が可能であり，(ii) 固有観点によりドメイン特有の残差にも適応できるという前提を与える．

表 5　観点語彙の共通率

ドメイン対	共通観点数	共通率 (%)
Office × Musical	59	22.2
Office × Toys	74	27.8
Musical × Toys	76	17.2

以上の分析から，提案手法が利用する観点語彙が各ドメインの特性を適切に反映していることが確認できた．次に，これらの観点語彙を用いた提案手法がベースモデルの予測精度を改善するかを検証する．

5.3.2 ベースモデルに対する改善

表 6 に，ベースモデル単体と提案手法の比較結果を示す．改善率は (1 − 提案 RMSE/ベース RMSE) × 100 で算出し，正値が改善を意味する．全 6 条件で改善率 >0 かつ $p < 0.05$ の有意な改善が得られた．最大の改善は Musical/LightGCN で 0.63%，最小は Toys/NARRE で 0.23%であった．

LightGCN に比べ NARRE で改善幅が小さい傾向がある．これは，NARRE がレビューテキストを入力として活用するモデルであり，観点情報と部分的に重複する情報を既にベース予測に取り込んでいるためと考えられる．一方，LightGCN は協調フィルタリングに基づくモデルでありテキスト情報を利用しないため，観点に基づく補正がより大きな付加価値を持つ．

補正適用率は，Office で 39.3%，Musical で 52.3%，Toys で 59.6%であった．本研究は，観点情報が一定量以上存在する場合にのみ補正を適用する設計であるため，適用率は 100%ではない．この点は「補正を常に適用する」のではなく，「観点情報に基づく補正が成立する領域に限定して適用する」という設計思想を反映している．また，改善率とベース RMSE の間には正の相関（Spearman $\rho = 0.77$, $p = 0.07$）が観察され，ベース予測が困難な条件ほど補正による改善が大きくなる傾向が示唆された．

提案手法が有効であることが確認できたため，次に各設計要素の貢献を ablation study により検証する．

5.3.3 設計要素の検証（Ablation Study）

提案手法の各設計要素の有効性を検証した．

a） 残差学習の必然性

表 7 に，学習目標の異なる 3 つの学習方式を比較した結果を示す．Residual は残差を目的変数として学習する提案手法，Direct-Add は y を直接学習しベース予測へ加算, Direct-Predict は観点特徴のみで y を予測する方式である．Residual のみが全条件で一貫した改善を示した．Direct-Add では α の最適値が 0 となり補正が実質的に適用されず，Direct-Predict では大幅に劣化した．

この結果は，観点情報が持つ情報量の性質を反映していると解釈できる．観点情報はユーザの嗜好の「傾向」や「偏り」を

表 6　ベースモデルと提案手法の比較

データセット	モデル	ベース RMSE	提案 RMSE	改善率 (%)	p 値	適用率
Musical	LightGCN	1.0210	1.0146	0.63	1.03e-04	52.3%
Musical	NARRE	1.0202	1.0173	0.28	3.02e-02	52.3%
Office	LightGCN	0.9819	0.9757	0.63	2.89e-05	39.3%
Office	NARRE	0.9861	0.9835	0.26	4.50e-02	39.3%
Toys	LightGCN	0.9119	0.9093	0.29	4.03e-04	59.6%
Toys	NARRE	0.9148	0.9127	0.23	8.76e-03	59.6%

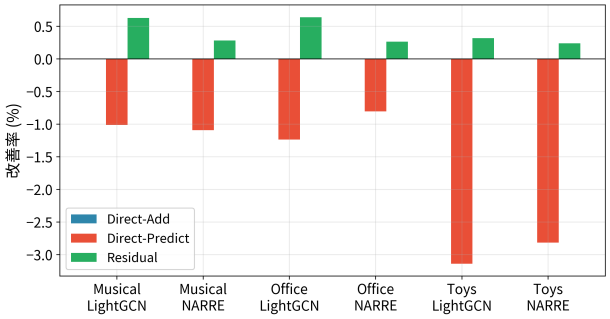

図 3　学習方式の比較

表 7　学習方式の比較

データセット	学習方式	改善率 (%)	p 値	α
Musical	Direct-Add	0.00	N/A	0.00
Musical	Direct-Predict	−1.02	1.48e-04	1.00
Musical	Residual（提案）	0.63	1.03e-04	1.10
Office	Direct-Add	0.00	N/A	0.00
Office	Direct-Predict	−1.24	7.77e-04	1.00
Office	Residual（提案）	0.63	2.90e-05	0.80
Toys	Direct-Add	0.00	N/A	0.00
Toys	Direct-Predict	−3.14	4.95e-06	1.00
Toys	Residual（提案）	0.29	4.03e-04	1.40

表 8　ゲート条件の比較

データセット	$[d_u, d_i]$	改善率 (%)	適用率	p 値
Musical	$[0, 0]$	0.41	100.0%	1.05e-04
Musical	$[0, 1]$	0.35	94.9%	1.43e-04
Musical	$[1, 1]$	0.63	52.3%	1.03e-04
Office	$[0, 0]$	0.20	100.0%	9.54e-03
Office	$[0, 1]$	0.18	87.8%	8.59e-05
Office	$[1, 1]$	0.63	39.3%	2.90e-05
Toys	$[0, 0]$	0.23	100.0%	5.21e-04
Toys	$[0, 1]$	0.20	95.4%	6.01e-04
Toys	$[1, 1]$	0.29	59.6%	4.03e-04

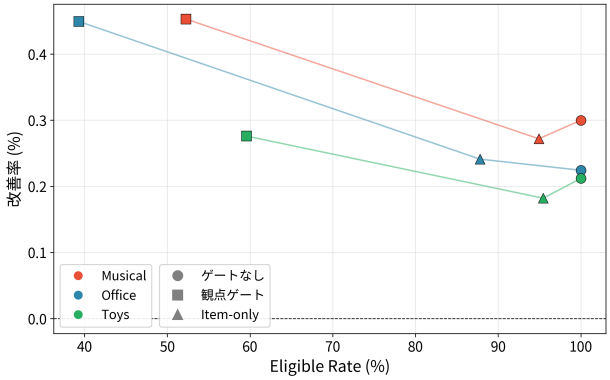

図 4　適用率と改善率の関係

捉える補助的な信号であり，絶対的な評価値を予測するには情報が不足する．Direct-Predict では観点特徴のみからレーティングを予測しようとするため，Toys で −3.14%という大幅な劣化が生じた．一方，ベース予測の残差は「なぜベースモデルが誤ったか」という相対的なずれを表すため，観点情報との対応関係が学習しやすい．この結果は，観点情報を「ベース予測の置換」ではなく「ベース誤差の補正」として学習する設計の必要性を裏付けている．図 3 は各学習方式の改善率を視覚的に比較したものであり，Residual のみが全条件で正値（改善）を示していることが明確に確認できる．

b）　ゲート設計の効果

表 8 に，ゲート条件 $[d_u, d_i]$ を変えた場合の結果を示す．$[d_u, d_i]$ はそれぞれユーザ・アイテムが訓練データで言及した観点種類数の閾値を表す．$[0, 0]$ は閾値なし（常に補正を適用），$[1, 1]$ は双方が 1 種類以上の観点に言及している場合に適用（提案手法の設定），$[0, 1]$ はアイテム側のみで判定する条件である．

結果として，$[1, 1]$ は $[0, 0]$ や $[0, 1]$ より改善率が大きい．$[0, 0]$ では観点情報が乏しいサンプルにも補正がかかり，ノイズ的な補正が混入して改善率が減少すると考えられる．また $[0, 1]$ は適用率が高い一方で改善率が小さい傾向にあり，ユーザ側の観点可用性も考慮することが重要であることを示している．特に Office では $[1, 1]$ の改善率が 0.63%であるのに対し $[0, 0]$ は 0.20%と差が大きく，選択的適用の効果が顕著である．

図 4 は，適用率と改善幅の関係を示したものである．同一条件（データセット×モデル）について，ゲート条件のみを変えた 3 点を線で結んで表示している．これにより，「同一条件内で，どの程度適用率を下げると改善幅がどう変化するか」を視覚的に比較できる．$[1, 1]$ は，適用率を適度に抑えつつ改善率を最大化する傾向を示しており，補正の選択的適用が設計上重要であることを支持する．

c）　収縮推定の効果

収縮推定について，表 9 に収縮強度 τ の比較結果を示す．データセットごとに収縮強度による違いを確認すると，Toys では τ=1 が τ=0 より改善率が大きく（LightGCN: 0.29% vs 0.13%），観測が疎な条件で収縮推定が有効であることが確認できる．この結果は，Toys が 3 データセット中で最も密度が低く（0.072%），観測数が少ないユーザ・アイテムが多いことと整合する．収縮推定により，少数観測から得られる不安定な局所平

表 9　収縮強度 τ の比較

データセット	τ	改善率 (%)	p 値
Musical	0	0.61	1.28e-04
Musical	1	0.63	1.03e-04
Office	0	0.63	3.73e-05
Office	1	0.63	2.90e-05
Toys	0	0.13	1.68e-03
Toys	1	0.29	4.03e-04

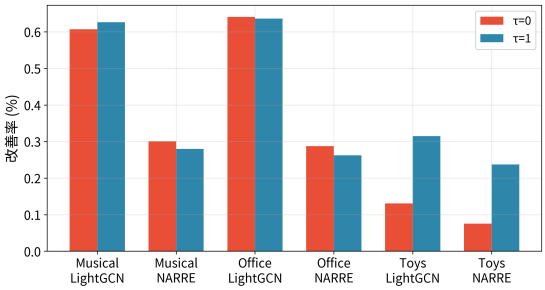

図 5　収縮強度 τ の比較

表 10　相互作用項の効果（paired t-test）

データセット	モデル	改善率差 (ポイント)	p 値
Office	LightGCN	+0.01	0.0014
Office	NARRE	+0.02	0.0072
Musical	LightGCN	+0.03	0.0003
Musical	NARRE	+0.02	0.0234
Toys	LightGCN	+0.04	<1e-04
Toys	NARRE	+0.04	0.0051

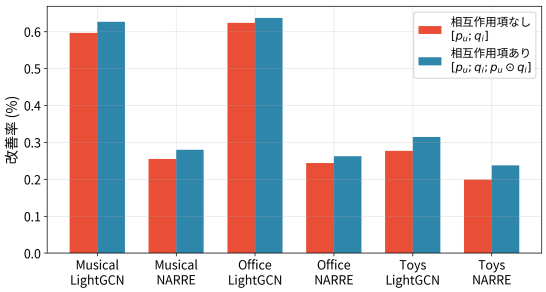

図 6　相互作用項の効果

均をグローバル平均へ引き寄せることで，推定の分散を抑制できたと考えられる．一方，Office や Musical では両者の差は小さいが，単一の収縮強度を全条件に適用する設計においては疎なデータへのロバスト性を優先することが実用上重要である．

図 5 に収縮強度 τ の比較を可視化した結果を示す．Toys では τ=1 の効果が顕著であり，密度が最も低いデータセットで収縮推定の安定化効果が大きいことがわかる．τ=0 では観測がある場合は局所平均を用い，観測がない場合は全体平均へフォールバックするが，少数観測の局所平均は偶然のばらつきの影響を受けやすい．収縮推定はこの不安定性を抑え，特に疎なデータでの補正精度を向上させる．

d）　相互作用項の効果

相互作用項については，$p_u \odot q_i$ を含む特徴が含まない場合より全条件で有意に改善した（paired t-test, $p < 0.05$）．表 10 に相互作用項の有無に関する対応検定結果を示す．6 条件すべてで相互作用項ありの方が有意に改善しており（平均差は約 0.03 ポイント），観点ごとにユーザ側傾向とアイテム側傾向を結合する設計の妥当性が確認された．

相互作用項が有効である理由として，ユーザが重視する観点は評価対象のアイテムに応じて変化することが挙げられる．例えば，音質を重視するユーザであっても，ケースを評価する際には耐久性を重視する可能性がある．p_u と q_i の連結のみでは各観点を独立に扱うことになるが，相互作用項 $p_u \odot q_i$ により「このユーザがこのアイテムのどの観点で誤差を出しやすいか」という組み合わせ依存の情報を表現できる．残差補正器は Ridge 回帰を用いた線形モデルであるため，入力特徴側でこの依存性を明示的に与えることが重要となる．

図 6 に相互作用項の有無による改善率の比較を示す．全条件で相互作用項あり（青）の方が改善が大きく，その差はおおむね 0.01–0.04 ポイント程度である．効果量は小さいが，6 条件すべてで $p < 0.05$ の有意差が得られており，相互作用項が一貫して改善に寄与していることが確認できる．この結果は，線形モデルである Ridge 回帰において，入力特徴側で組み合わせの依存性を明示的に与えることの有効性を示している．

以上の ablation study により，各設計要素の有効性が確認できた．

5.3.4　コールドスタート条件の分析

履歴数が少ないコールドスタート条件（$n_u \leqq 5$ または $n_i \leqq 5$）における提案手法の挙動を分析した．表 11 に結果を示す．アイテムコールドスタート条件では，Musical/LightGCN で 1.08% と全体（0.63%）より大きな改善が得られた．これは，履歴が少ないアイテムほどベース予測が不安定であり，観点情報による補正の余地が大きいためと考えられる．

注目すべき点として，コールドスタート条件でも補正が適用されている（適用率 30〜50%程度）ことが挙げられる．これは，履歴数が少ないことと観点情報が乏しいことは必ずしも一致しないことを示している．ユーザやアイテムの履歴が少なくても，その限られた履歴の中でレビューが書かれていれば観点情報は得られる．提案手法のゲートは観点種類数に基づくため，「レビューを書く傾向のあるユーザ」や「レビューが付きやすいアイテム」では，履歴が少なくても補正が適用される．この性質により，コールドスタート条件においても観点情報を活用した精度改善が可能となっている．

6　おわりに

本研究では，レビューテキストから抽出した観点情報を用いて，既存の推薦モデルの予測誤差を後段から補正する残差補正型推薦手法を提案した．従来の観点ベース推薦はモデル内部への統合を採るため，レビュー由来の不確実性が学習全体へ波及しやすく，モデル構造への依存が強くなる課題があった．これに対し本研究は，ベースモデルの予測を保持したまま観点情報を誤差を補う補助信号として扱う枠組みを与えた．

表 11　コールドスタート条件の結果

データセット	条件	n	適用率	改善率 (%)
Musical	Full	16015	52.3%	0.63
Musical	アイテムコールドスタート	3581	46.8%	1.08
Musical	ユーザコールドスタート	12355	40.4%	0.50
Office	Full	5185	39.3%	0.63
Office	アイテムコールドスタート	1602	31.0%	0.62
Office	ユーザコールドスタート	4293	31.2%	0.60
Toys	Full	15982	59.6%	0.29
Toys	アイテムコールドスタート	3548	51.0%	0.61
Toys	ユーザコールドスタート	11062	46.3%	0.30

実験では，3 ドメインにおいて性質の異なる 2 種類のベースモデル（LightGCN, NARRE）に対して一貫した改善を確認し，残差学習・ゲート・収縮推定・相互作用項の各設計要素がそれぞれ改善に寄与することを検証した．特に，ゲート機構による選択的適用により観点情報が乏しいサンプルへのノイズ的な補正を抑制できること，アイテムコールドスタート条件で全体より大きな改善が得られることを示した．

本研究の意義は，観点情報を説明のための付加情報に留めず，ベースモデルの予測誤差を減らすための実用的な補正信号として組み込んだ点にある．提案手法は特定のアーキテクチャに依存せず後段補正として適用可能であるため，推薦モデルの種類が変化しても観点情報の活用方針を再利用できる．

今後の課題として，(1) 観点抽出の品質と補正性能の関係の定量化，(2) ランキング指標や時系列分割での評価，(3) 補正適用領域推定の高度化が挙げられる．

文　献

[1] Lei Zheng, Vahid Noroozi, and Philip S. Yu. Joint deep modeling of users and items using reviews for recommendation. In *Proceedings of the 10th ACM International Conference on Web Search and Data Mining*, pages 425–434. ACM, 2017.

[2] Chong Chen, Min Zhang, Yiqun Liu, and Shaoping Ma. Neural attentional rating regression with review-level explanations. In *Proceedings of the 2018 World Wide Web Conference*, pages 1583–1592. International World Wide Web Conferences Steering Committee, 2018.

[3] Yongfeng Zhang, Guokun Lai, Min Zhang, Yi Zhang, Yiqun Liu, and Shaoping Ma. Explicit factor models for explainable recommendation based on phrase-level sentiment analysis. In *Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '14)*, 2014.

[4] Zhiyong Cheng, Ying Ding, Xiangnan He, Lei Zhu, Xuemeng Song, and Mohan S. Kankanhalli. Aspect-aware latent factor model: Rating prediction with ratings and reviews. In *Proceedings of the 2018 World Wide Web Conference (WWW)*, pages 639–648, 2018.

[5] Jin Yao Chin, Kaiqi Zhao, Shafiq Joty, and Gao Cong. Anr: Aspect-based neural recommender. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management (CIKM)*, 2018.

[6] Heena Lim et al. Reducing contextual noise in review-based recommendation via aspect term extraction and attention modeling. *Information Sciences*, page 123078, 2026.

[7] Xinyu Guan, Zhiyong Cheng, Xiangnan He, Yongfeng Zhang, Zhibo Zhu, Qinke Peng, and Tat-Seng Chua. Attentive aspect modeling for review-aware recommendation. *ACM Transactions on Information Systems (TOIS)*, 37(3):1–27, 2019.

[8] Yehuda Koren, Robert Bell, and Chris Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37, 2009.

[9] Yifan Hu, Yehuda Koren, and Chris Volinsky. Collaborative filtering for implicit feedback datasets. In *Proceedings of the 2008 IEEE International Conference on Data Mining*, pages 263–272. IEEE, 2008.

[10] Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. Neural collaborative filtering. In *Proceedings of the 26th International Conference on World Wide Web*, pages 173–182. International World Wide Web Conferences Steering Committee, 2017.

[11] Julian McAuley and Jure Leskovec. Hidden factors and hidden topics: Understanding rating dimensions with review text. In *Proceedings of the 7th ACM Conference on Recommender Systems (RecSys '13)*, 2013.

[12] Hongwei Wang, Naiyan Wang, and Dit-Yan Yeung. Capsule network for recommendation and explaining what you like and dislike. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, 2019.

[13] Jianing Liu, Xu Xu, Yifan Sun, Ziang Li, Linfeng Zhang, Hongbin Ye, and Xin Man. Multi-aspect enhanced graph neural networks for recommendation. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, 2023.

[14] Aihua Liu, Xiaoyan Zhao, Xiaofang Duan, Gongcheng Xu, Yuchong Yang, Weijun Zhang, and Hongxia Yang. Aspect performance hypergraph attention for explainable review-based recommendation. In *Proceedings of the ACM Web Conference (TheWebConf)*, 2025.

[15] Wang-Cheng Kang and Julian McAuley. Self-attentive sequential recommendation. In *Proceedings of the 2018 IEEE International Conference on Data Mining (ICDM)*, 2018.

[16] Fei Sun, Jun Liu, Jian Wu, Changhua Pei, Xiao Lin, Wenwu Ou, and Peng Jiang. Bert4rec: Sequential recommendation with bidirectional encoder representations from transformer. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management (CIKM)*, 2019.

[17] Hongtao Liu, Wenjun Wang, Hongyan Xu, Qiyao Peng, and Pengfei Jiao. Neural unified review recommendation with cross attention. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, 2020.

[18] Hao Wang, Naiyan Wang, and Dit-Yan Yeung. Collaborative deep learning for recommender systems. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2015.

[19] Donghyun Kim, Chanyoung Park, Jinoh Oh, Sungyoung Lee, and Hwanjo Yu. Convolutional matrix factorization for document context-aware recommendation. In *Proceedings of the 10th ACM Conference on Recommender Systems (RecSys)*, 2016.

[20] Xiangnan He, Kuan Deng, Xiang Wang, Yan Li, Yongdong Zhang, and Meng Wang. Lightgcn: Simplifying and powering graph convolution network for recommendation. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pages 639–648, 2020.

[21] Jianmo Ni, Jiacheng Li, and Julian McAuley. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In *Proceedings of the 2019 Confer-*

ence on Empirical Methods in Natural Language Processing (EMNLP), 2019.