東京大学大学院工学系研究科システム創成学専攻

博士論文

オンラインプラットフォームにおける バイアスを考慮したコンテンツの質の定量化

Quantifying the Unbiased Quality of the Contents in Online Platforms

指導教員 鳥海不二夫 准教授

学籍番号 37-187051

福馬 智生

概 要

本研究ではオンラインプラットフォームにおけるコンテンツの質の評価を，様々なバイアスの影響を考慮することで，より正確に推定する試みを行った．コンテンツの質を正しく推定できることは，プラットフォーム運営者にとって有益な情報をユーザーに提供しやすくなり，プラットフォームの情報過多の解決のために重要である．オンラインプラットフォームには様々なバイアスが内在しており，それら影響によって既存のコンテンツの質の自動評価技術をナイーブに適用させるだけでは，真にコンテンツの質を表現または推定が行うことができない．

我々の研究では，一元的なコンテンツの質の評価 (一位から最下位までコンテンツの質を元に推定しランキングを作成) と多元的な評価 (個人ごとにそれぞれのコンテンツへの嗜好度の予測を元にランキングを作成) を対象に，バイアスの影響を取り除いた上でのコンテンツの質の自動評価技術について提案・検証を行った．具体的に前者における既存の課題として二つを取り上げた．一つ目は認知バイアス情報によってユーザーの評価傾向が歪められてしまい，ユーザーが公平に評価を行えていない点を取り上げた．二つ目はユーザーの意見にはやらせやアンチといった観点から信頼性に偏りがあり，意見の集約の過程でコンテンツが公平に評価されない点を取り上げた．後者における課題として，ユーザーやアイテムの履歴数には偏りがあり，履歴数が少ないユーザーやアイテムに関して嗜好度の予測性能が大幅に低下する点を取り上げた．

我々の提案手法・分析によって，大きく以下のことが解決・明らかになった．
1) 事前バイアスがなければ人はどのように評価したか基づいて，コンテンツの質を認知バイアスの影響を取り除いた形で推定が可能になった．2) コンテンツの質を測るにあたり，個人の意見の集約は，評価者の過去の評価履歴などから判断した信頼性に基づいた形で行うことでより有識者に近い意見が得られる．3) 推薦システムにおいて履歴数が少ないユーザーやアイテムに関しての嗜好度のモデリングを以前より高精度に学習することが可能になった．

本研究により，オンラインプラットフォーム上におけるコンテンツの価値を全員にとって，または個人に対してより正確に推定することが可能になった．

目 次

図 目 次

表 目 次

第1章　序論

1.1　背景

1.1.1　オンラインプラットフォームの台頭

インターネット技術の爆発的な発展に伴い，様々な種類のウェブサイトやフォーラムが誕生している．それらには個人の意見，コメント，レビューが大量に集積されており，その結果，情報の収集や他者との情報交換は以前より遥かに容易になった．本論文では非常に広い範囲のウェブサービスを指して「オンラインプラットフォーム」と呼ぶ．例えばユーザーの検索クエリに基づき最適な情報を提供する検索サイト (Google[1]，Yahoo[2])，購入体験といったサービスの消費に基づく体験を書き込むレビューサイト (Amazon[3]，TripAdvisor[4]，IMDb[5]，Yelp[6])，特定のトピックについて不特定多数で議論や意見交換をする掲示板サイト (Stack Exchange[7]，Reddit[8])，個人の意見やつぶやきを特定の人の中で共有するソーシャルメディア (Twitter[9]，Facebook[10]) などが挙げられる．

これらのオンラインプラットフォームの普及は我々の情報収集行動に大きな変化を与えた．最も大きな変化は，消費者が自ら進んで情報を取得する従来のプル型の情報収集に加えて，検索の結果やサイト上の導線に従って自動的に情報が提示されるプッシュ型の情報収集が拡大した点である [1]．プル型の情報収集の代表的例としては，インターネット以前の情報源の主流であるテレビやラジオ，新聞などからの収集が挙げられる．昨今ではプッシュ型の収集における新たな試みとして，推薦アルゴリズムなどを活用した，ユーザーの興味に合わせた情報を提示する Gunosy[11]や

[1] https://google.com/
[2] https://yahoo.com/
[3] https://amazon.com/
[4] https://tripadvisor.com/
[5] https://imdb.com/
[6] https://yelp.com/
[7] https://stackexchange.com/
[8] https://reddit.com/
[9] https://twitter.com/
[10] https://facebook.com/
[11] https://gunosy.com/

Smartnews[12]などのサービスも成長を遂げており，プッシュ型の情報収集の流れはさらに加速している．

同様にオンラインプラットフォームの発展は，我々の購買行動にも変化を与えている．日本国内の BtoC-EC（消費者向け電子商取引）市場規模は，19.4 兆円（前年 18.0 兆円，前年比 7.65%増）に拡大しており，更なる拡大が見込まれている [2]．

このようなオンラインプラットフォームの台頭に従い，インターネット上に投稿される不特定多数の意見「レビュー」が持つ役割も大きくなっている．現在では，97 %の消費者が購入の意思決定をする際に製品のレビューを参考にしており，2014 年の 95 %からも更に増加している [3]．このようにインターネット上に投稿される情報や意見は我々の日々の行動に大きな影響を与えることが知られている．

1.1.2 コンテンツの質に関する自動評価技術

一方でユーザーの拡大に伴い，インターネット上には日々膨大な量の情報が追加されている．その結果，個人が探したい情報が見つからない，または分からないといった弊害が発生しており，それらは「情報過多 (information overload)」といった名前で広く知られている．このような弊害によって生じるユーザー体験の低下はオンラインプラットフォームの存続に致命的であり，昨今ではプラットフォームの運営側で様々な対策が行われている．

これらの背景より，膨大な情報源の中からユーザに適した情報を見つけ出し，それらを提示するためのランキング技術や推薦技術に注目が集まっている．これらを統括して本論文では「コンテンツの質に関する自動評価技術」と呼称する．検索サービスにおける代表的な事例として，Google の中心的技術である PageRank [4] アルゴリズムが挙げられる．PageRank は「重要なページにリンクされているページは重要である」という発想に基づき，ウェブページの重要度をハイパーリンク構造に基づき計算し，検索エンジンの性能を従来より大きく向上させた．

同様にコンテンツの質の自動評価技術は，嘘や悪質な情報をフィルタリングする観点からも関心が高まっている．近年のフェイクニュースは公の言論，人間社会，民主主義を脅かす問題となっている [5]．それら情報選別の目はユーザーが持つべきだとする考え方もある一方で，近年フェイク情報への対策はプラットフォーム側で行うべきだという認知も広まっている．Facebook 社ではロイター通信[13]と提携して，

[12] https://smartnews.com/
[13] https://www.reuters.com/

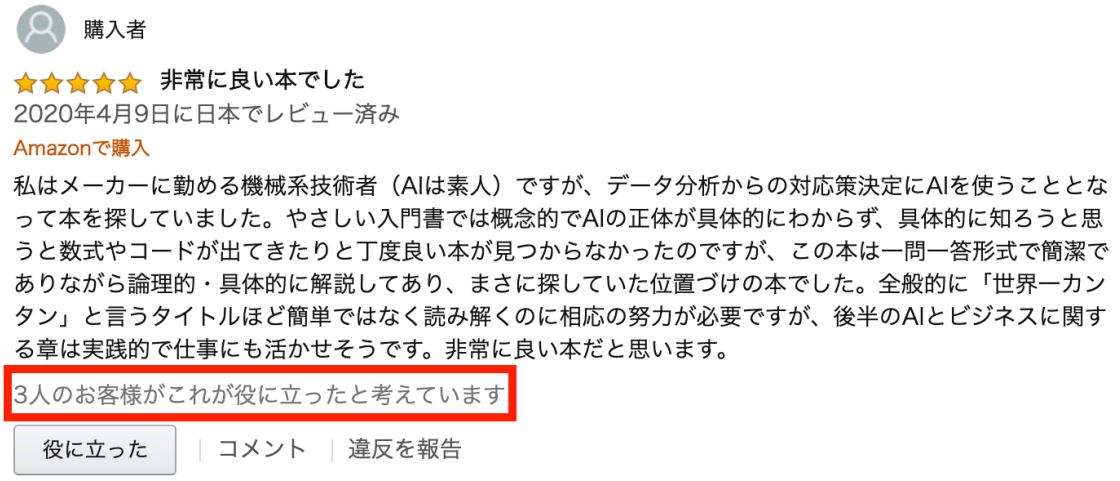


図 1.1: Amazon のレビューにおける投票形式で行われる有用度測定の例 [17]

投稿の事実確認を実施する声明を発表している[14]．また同社は既存の画像や動画の中の人物を他者に置き換える Deepfake コンテンツの検出のためのコンペティションを開催している[15]．また Amazon 社はやらせレビューと疑わしい約 2 万件の商品レビューを削除したといった報告もある[16]．

以上より，ユーザーが迅速かつ求めている有用な情報にアクセスできるための環境の構築と，それらを達成するための技術開発の重要性は非常に高まっていると言える．

以下ではコンテンツの質の自動評価に関する既存の試みを「集合知」に基づく試みと,「機械学習」に基づく試みと二つに大別して紹介する．詳細は後述第 2 章にて述べる．

集合知に基づくアプローチ

多くのオンラインプラットフォームでは，コンテンツの有用性をユーザーが投票するシステムが実装されている．ユーザーは他ユーザーによる「いいね」や「役に立った」などの投票形式に基づく評価の合計数順に並べ替えることで，容易に有用

[14] https://www.reuters.com/article/rpb-fbfactchecking/reuters-launches-fact-checking-initiative-to-identify-misinformation-in-partnership-with-facebook-idUSKBN2061TG 最終閲覧日 2021 年 1 月 27 日

[15] https://www.kaggle.com/c/deepfake-detection-challenge

[16] Amazon deletes 20,000 reviews after evidence of profits for posts 2020 年 9 月 4 日 最終閲覧日 2021 年 1 月 24 日 https://www.ft.com/content/bb03ba1c-add3-4440-9bf2-2a65566aef4a

[17] https://www.amazon.co.jp/gp/customer-reviews/R33UUC2WKLSQDM/ref=cm_cr_dp_d_rvw_ttl?ie=UTF8&ASIN=B086YKWJVK 最終閲覧日 2021 年 1 月 27 日

性の高いコンテンツに到達できるよう誘導している．これは「群衆の叡智」という考え方に基づいており，大人数の集団の意見は，たとえ専門家であっても，個人の意見よりも正確になるといった知見から生まれている [6, 7]．例としては Amazon のレビューに対する「役に立った」(図 1.1) 機能が挙げられる．これらは実装の容易さから多くのオンラインプラットフォームで広く用いられている．

機械学習に基づくアプローチ

以下ではコンテンツの有用度を，機械学習を用いて学習する試みについて 2 種類に大別し解説する．

教師あり学習

教師あり学習は，入力と出力のペアデータから学習を行い，出力が未知の入力について出力を予測することを目的とする学習方法を指す．教師あり学習を用いる枠組みは，目標出力として人間が作成したコンテンツの質に関するアノテーションデータを用いる試みと，ユーザーの行動ログに基づく試みの二つに大別される．人間が作成したアノテーションを用いた学習は，コンテンツの質を 0 から 100 の間でアノテーターによって計測する場合 [8] や，フェイクか否かといった事実情報に基づく二値ラベル [9] を付与することが多い．

しかし人間によるアノテーションは作成コストが高く，大規模データセットの作成が困難である．それら欠点を補うために代替策として，集合知の試みで用いた “いいね” などのクリック情報を教師データとして学習し，コンテンツの有用度を推定する試みが現在は主流になっている．Airbnb 社は自社の検索エンジンを，ユーザーのクリック情報を教師データとして Deep Learning を活用することで，CTR (Click Through Rate) 性能の大幅な向上を遂げた [10]．またレビューの有用性を測る試みでは，Y 人中 X 人が役に立ったと回答した際，X/Y を教師信号として用いて学習を行うといった試みが行われている [11]．

自己教師あり学習

自己教師あり学習は，明示的なアノテーションデータを用いず，手元のデータだけから設定可能なヒューリスティックな目的関数を最適化する学習方法を指す．代表的な例として PageRank [4] は，ランダムサーファーモデルを仮定し，リンクに基

づき無限試行後の訪問確率をページの重要度と仮定し，最適化する点で自己教師あり学習に分類できる．

また推薦システムで広く用いられている協調フィルタリング手法も自己教師あり学習とみなされている．代表的な手法である Matrix Factorization [12] は，ユーザーとアイテムを同一の潜在空間に写像し，同空間内での距離が嗜好度とする仮定を置き，既知データについて最適化を行うことで，未知のユーザーアイテム間のインタラクションについて嗜好度を予測することを可能にしている．同手法はコンテンツの質，または嗜好度を個人単位で推定している点で本研究との関連性が高い．

1.1.3 コンテンツの質の自動評価技術に潜むバイアス

前項では現在用いられているコンテンツの質に関する自動評価技術を紹介した．しかしながら，これらの指標が真の意味でコンテンツの質を表しているとは限らない．本項では既存の試みについて「バイアス」という観点から検討する．ここでのバイアスとは，運営者や開発者が想定しているデータの分布や仮定と実際の挙動にズレや偏りが存在することを指す．本項では例としてデータバイアスとプレゼンテーションバイアスを取り上げる．

データバイアス

データバイアスとはデータ作成者の偏見や認知バイアス，不適切なデータ取得手続きなどにより訓練データ中の特徴量や目的変数に偏りが生じている場合を指す [13]．ウェブデータに潜む代表的なバイアスとして 1) デモグラフィック（性別や年齢，国籍，所得，職業など）に起因するバイアス，2) ユーザーあたりの活動頻度に起因したバイアスが挙げられる．例えば Wikipedia における女性の編集者の割合は 16%程度しかいないことが知られている [14]．また英語で書かれたウェブ上のコンテンツは全体の約 50%に及ぶ一方，英語のネイティブスピーカーは 5%程度しかいない [15] といったことが知られている．また Ricardo et al. [15] の調査によると，2009 年において約 4 万人のアクティブユーザーがいる Facebook のデータセットうち，7%のユーザーの投稿が全体の 50%を占めていることが分かった．同様に 2013 年の Amazon レビューデータセットにおいては，わずか 4%のユーザーの投稿が全体の 50%の占めていることが分かった．

このように，ウェブ上のコンテンツデータと実世界の事象や意見をそのまま反映しているとは言い難く，ズレが生じている．これらデータバイアスを含んだデータを用いて機械学習モデルを構築する場合，同様のバイアスを学習してしまう．

例として Amazon で開発されていた採用支援 AI の事例が挙げられる[18]．同サービスは候補者の履歴書から採用すべきかを予測するものである．同社はシステムが男性の評価を高く推定するケースがあることから運用を取り止めた．原因として，候補者の男性の割合よりも採用された男性の割合が大きいといった偏りがあるデータを元に学習を行っていたことが挙げられる．これらは明示的な特徴量として性別を用いなくても，男女で使いやすい単語の傾向などから性別情報がリークする点において防ぐことが非常に難しい問題として知られる．このようにデータバイアスはそれが社会的に不適切なものであっても，予測結果に反映されるため，機械学習モデルを社会応用する際には常に留意する必要がある．

プレゼンテーションバイアス

プッシュ型の情報収集の特徴として，ユーザーはウェブ上の全ての情報に目を通すことはなく，システムから提示される上位の一部しか目を通さないことが挙げられる．その結果上位に表示されたコンテンツほどクリックされやすいというプレゼンテーションバイアスが発生する．これらは有用であったり興味がある内容だとしても上位に表示されなければクリックされないという現象を招き，クリックと有用度が必ずしもイコールの関係ではなくなる．

1.1.2 項で紹介した，集合知に基づくシステムや検索エンジンについては顕著にこの問題が発生する．その結果，安易にクリック情報を元にコンテンツの質を測ると，既に上位に表示させているコンテンツを必要以上に有用と判断してしまう傾向にある．

検索エンジンでは，ユーザーは検索クエリについて上位のコンテンツしか見ず，上位に表示された結果は，関連性が高いといった理由以上に，下位のコンテンツよりも多くのクリックを集める．図 1.2 に 2005 年（左）と 2014 年（右）のウェブ検索結果ページのアイトラッキング解析のヒートマップを示す．本図よりユーザーの視線が一部のコンテンツに集中していることが分かる．

同様にレビューサイトでは，上位にランクされたレビューは，より多くの注目を集め，その結果より多くの人々の票を獲得することになる．図 1.3 は，ランキング上位 50 件のレビューの「役に立った」の投票数を示しており，上位が多くの投票を独占し，下位の商品にはレビューが集まりにくいことが知られる [16]．

[18] https://jp.reuters.com/article/amazon-jobs-ai-analysisidJPKCN1MLODN 最終閲覧日 2021 年 1 月 27 日

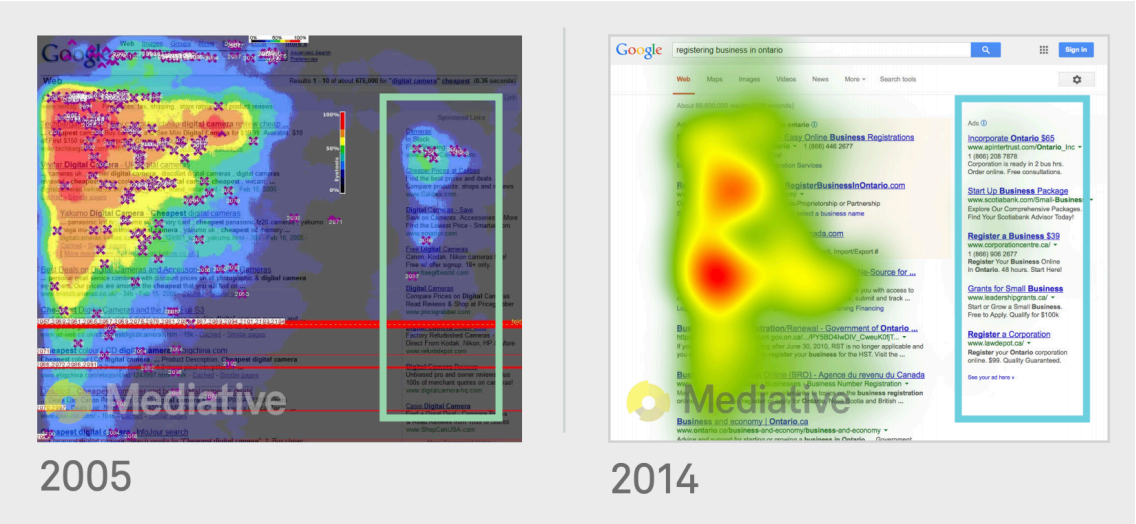


図 1.2: 2005 年（左）と 2014 年（右）のウェブ検索結果ページのアイトラッキング解析のヒートマップ．Maynes et al. の 9,12 ページより [17]

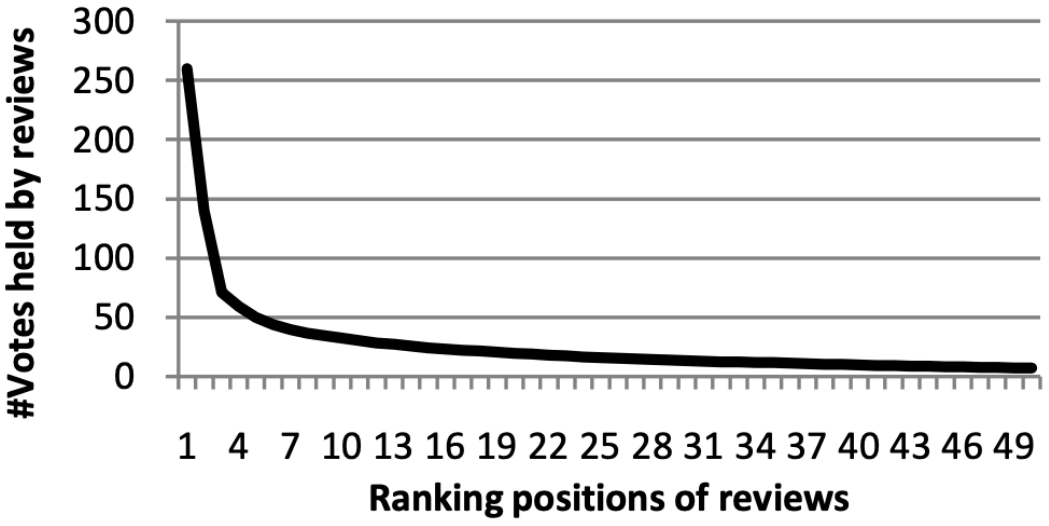


図 1.3: ランキング上位 50 件のレビューの「役に立った」の投票数．Liu et al. の論文の図 3 より [15].

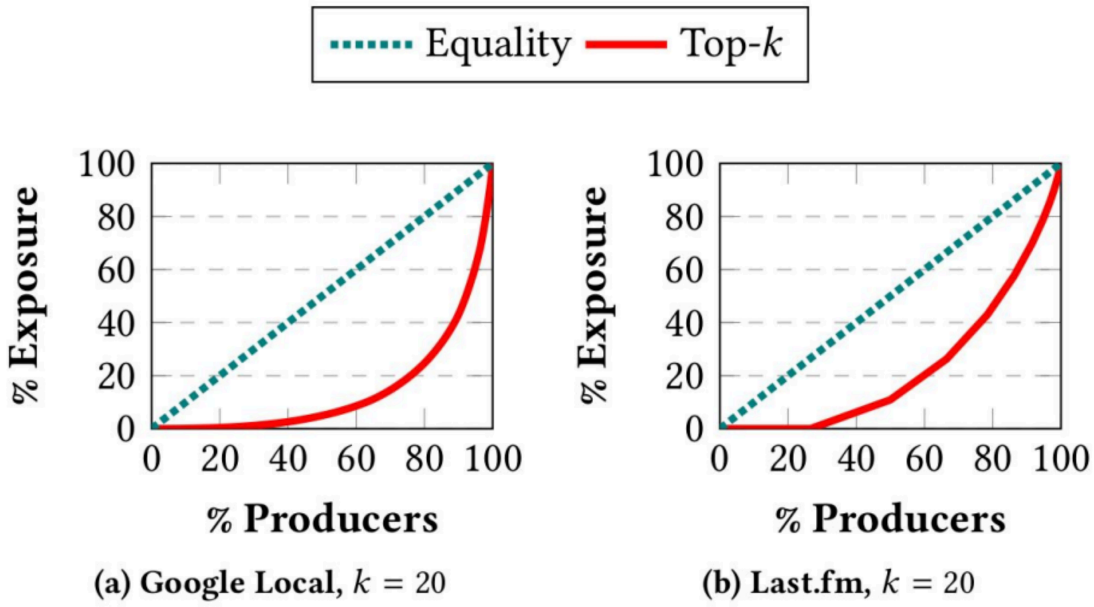


図 1.4: Google Local と Lastfm のデータセットから学習した推薦システムが提示するアイテム (Top-k) とユーザーへの露出の関係性．縦軸は露出割合の累積，横軸はアイテムを露出頻度の少ない順に並び替えた場合の割合．Patro et al. の論文の図 1 より [18].

同様のプレゼンテーションバイアスは，推薦システムにおいても生じる．Patro et al. [18] は，全体のアイテムのうち，推薦システムが提示するアイテムによってどれだけユーザーに露出しているかについて検討している．図 1.4 にて Patro et al. が Matrix Factorization を用いて Google Local と Lastfm のデータセットから学習した推薦システムを用いた検証結果を示す．その結果 Google Local を用いたデータセットからの学習では，推薦頻度上位 20%のアイテムが 80%のユーザーと接触しており，Lastfm を用いて学習させた事例では推薦頻度下位 60%のアイテムが全ユーザーの 20%未満の露出しか得られていないといった結果が得られた．

1.1.4 バイアスが与えるプラットフォームへの弊害

二次バイアス

二次バイアスとは，バイアスがかかったアルゴリズムによって提示された情報とユーザーが接触することで，バイアスの影響を受けた行動をユーザーが行った結果，システムのバイアスを強めてしまう現象を指す．プレゼンテーションバイアスを例にとると，上位に表示されるコンテンツはクリックされやすく，結果より多くのクリックを集めるといった循環が生まれる．

多様性の減少

二次バイアスが与えるオンラインプラットフォームへの悪影響として「多様性の減少」が挙げられる．集合知に基づくレビューサイトでは，投稿日時が古いものほど投票を受ける機会が増えた結果，新しい投稿についてほとんど票が集まらないといったことが知られており，ユーザーに表示される多様性が非常に少なくなっている (図 1.5)．

多様性の減少はユーザーのクリック情報に基づき最適化される推薦システムにおいても生じる．システムはユーザーに興味のありそうなコンテンツを提示し続けることで，ユーザーが触れる情報が既存の興味に限定され，好きになる可能性のある新しい情報に触れる機会が少なくなることで知られる．このように閉じた世界の中に閉じ込められることを例えてこれら弊害は “フィルターバブル” という名前で知られる [19]．

推薦システムとユーザーのコンテンツ消費の多様性との関連性について Anderson et al. [20] は音楽ストリーミングサービス Spotify[19]のデータを用いて分析を行った．これら推薦技術はユーザーの短期的なエンゲージメントを高める一方，中長期目線ではユーザーのコンテンツ消費の多様性を減少させ，サービスの離脱との相関関係が認められることが示されている．

また SNS においても意見の極化は近年問題になっている．ソーシャルメディアを使うユーザは繋がっているユーザーの意見に影響を受け，自分の意見とは異なるユーザとの繋がりを断つことで，閉鎖的で偏った意見がより強化される．これらは “エコーチェンバー” という名前で知られている [21]．

人気と質の乖離

これらのバイアスに伴う多様性の減少に起因して，人気と質との乖離が起きているとの報告が多数挙げられる．例えば上記図 1.3 において，上位のコンテンツが下位のコンテンツよりも 100 倍近く有用というのは感覚的にも反する．実際に Amazon のレビューの役に立った度 (Y 人中 X 人役に立ったと回答した場合 X/Y) は実際のレビューの品質と強く相関していないという報告が多くなされている [16, 22, 23, 24, 8]．

また Cheng et al. [25] は Yahoo! News の記事に対するコメントを用いて，それらに投票された高評価・低評価の集合と，有識者によるコメントの品質のアノテーションとがほぼ無相関であることを実験的に示した．また Qiu et al. [26] は情報拡散モデルを用い，SNS では情報過多の状況では人気と質には相関が弱くなることを

[19] http://spotify.com/

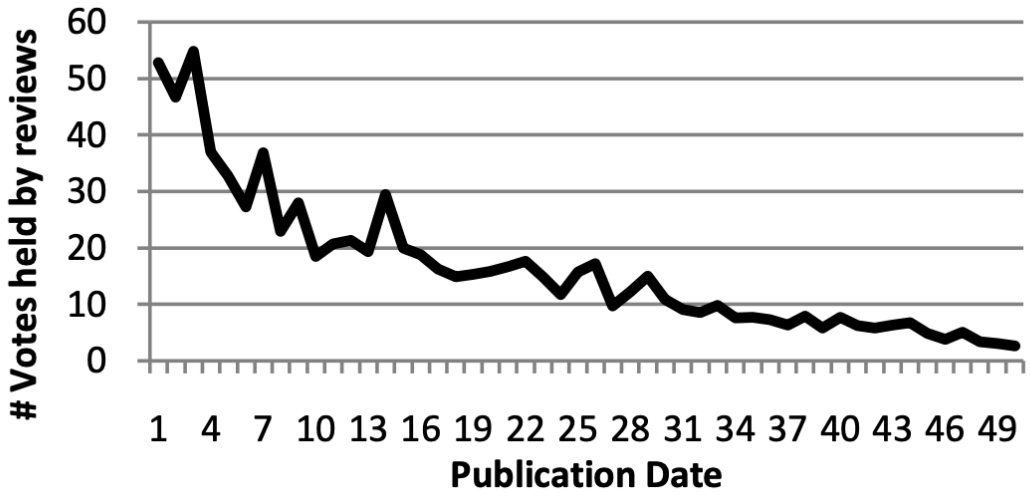


図 1.5: 投稿日に基づく「役に立った」の投票の累積数．Liu et al. の論文の図 4 より [15].

示した．これらの事例よりバイアスによる影響を考慮せずコンテンツの自動評価システムを実装することは，有用または求めている情報に迅速にアクセスできるオンラインプラットフォームの構築に必ずしも貢献しないことが分かる．

1.2 本研究の研究課題と構成

1.2.1 概要

前述の通り，オンラインプラットフォームには様々なバイアスが内在しており，既存のコンテンツの質の自動評価技術をナイーブに適用させるだけでは，真にコンテンツの質を表現または推定できているとは言い難い．そこで本研究では，オンラインプラットフォームにおけるコンテンツの質を，それらに対する様々な意見 (e.g., クリック，星) から，バイアスの影響を取り除いた形で，より正確に推定することを目的とする．

本研究では自動評価技術に関して二つに大別する．一つ目は “万人向け” に一元的にコンテンツの質を推定する技術である．これらはコンテンツを最上位から最下位まで質に基づき順位付けするように行われる．例として集合知に基づく試みが挙げられる．二つ目は “個人に特化した” 多元的なコンテンツの質の推定技術である．言い換えると人によってコンテンツの質は異なるという考えに基づき，ユーザー毎にコンテンツに関するランキングを嗜好度に基づき作成する．代表的な取り組みとし

ては推薦システムが挙げられる．

本論文では前者における既存の課題として二つを取り上げる．一つ目は認知バイアス情報 (e.g., コンテンツの位置,「いいね」数といったコンテンツに関する他者の評価) によって個人の評価傾向が歪められ，ユーザーが公平に評価を行えていない点である．二つ目は平均といった個人の意見を集約してユーザーに提示する際，やらせやアンチといった意見によって評価が歪められる点である．

同様に後者においても二つ取り上げる．一つ目は履歴数が少ないユーザーやアイテムに関してはる嗜好度の予測性能が大幅に低下する点である．二つ目は個人とアイテムとの嗜好度のモデリングに関し，予測に不確実性を持たすことができない点である．これらの課題の詳細に関しては第 2 章にてまとめる．

1.2.2 研究課題

本論文は上記課題に対し，以下の Research Question (RQ) を設定する．

Research Question 1： 複数の認知バイアス情報の影響を受けているユーザーの行動データから,「認知バイアス情報が無ければどのように行動したのか」という反実仮想のもと，それら影響を取り除いた真のコンテンツの質を測定できるか？

本研究では,「いいね」や「役に立った」といった投票形式の評価行動に焦点を当て，認知バイアス情報が及ぼす影響について明らかにする．例えば他者による事前の評価といった認知バイアス情報はユーザーの意思決定を歪めることが知られる [27]. もしそれら情報がなければ人はどう評価していたのかといった反実仮想な現象に対する推定技術の提案について述べる．

提案手法では，コンテンツの評判スコアのような認知バイアス情報の有無に基づく，ランダム化比較試験を用いて収集したデータセットについて，人々の投票行動の違いに着目する．続いて新たに人間の投票行動のモデリングを提案し，投票行動の違いについて機械学習モデルを用いて説明する．具体的には，ニューラルネットワークを用いて，周囲の認知バイアス情報によって「そのコンテンツが何倍有益に見えているか」を定量化する．さらに，認知バイアスの影響を割り引くことで，認知バイアスの影響を取り除いた形でも有用性を推定する手法を提案する．最後に，大手 Q&A サイトである Stack Exchange の行動ログを用いて，提案手法の有効性を実験的に示し，モデルが学習した知識について解釈を行う．提案手法は，オンラインプラットフォームにおける中心的な問題である，バイアスの影響を取り除いたコン

テンツの有用性の測定を実現する．

Research Question 2：　ユーザーのレビュー投票行動から不良ユーザー・レビューを発見し，信頼度に基づいた意見の集約によって従来よりも有識者に近い集合知を獲得できるか？

近年では，やらせレビューやアンチレビューのような，対象の評価を意図的に操作するために行われる信頼性の低いレビューが存在する．言い換えると既存のレビューの信頼度は一律ではなく偏りがあると言える．そのため星などの数で評価されたレビューを商品ごとに単に平均といった集約方法は，必ずしも良い集合知に繋がるとは限らない．

そこで，リンク構造に基づくレビューの信頼性評価アルゴリズムによって得られた「信頼度」によって意見を集約する．これによって得られた集合知を既存の平均といった集約方法と，有識者の意見との類似度という観点から比較を行う．レビューの信頼性評価には Kumar et al. [28] によって提案された REV2 を用いる．レビューのデータには東京都内のラーメン屋のレビューを独自に収集し利用する．これによって信頼性に基づいて集約することが有識者による集合知に近い形で獲得されるかどうかを確認する．また信頼性に基づいた評価によって平均と比べて評価が大きく変化した店舗についての特徴を分析する．さらに REV2 によって判断される信頼性とは何かといったモデルの解釈性についても検証を行う．

Research Question 3：　推薦システムは履歴の少ないユーザーやアイテムの組み合わせについても正しくそれぞれの間の嗜好度をモデリングすることは可能か？

Research Question 4：　推薦システムは予測の不確実性をモデリングすることは可能か？

推薦システムは，ユーザーとアイテムのインタラクションの履歴データを元に，未知のユーザーとアイテムのインタラクションを予測する．それらは個人ごとに嗜好度または質を推論している点において，多元的なコンテンツの質の評価を行っていると本研究では捉える．

従来の代表的な協調フィルタリング手法は，十分にユーザー・アイテムの嗜好度に関する履歴が手に入る場合には上手くモデリングができることが知られている．一方で，疎なデータ，つまりユーザー・アイテムの嗜好度に関する履歴がほとんど得ら

れていないケースでは十分にそれらの関連性をモデリングできない．しかし一般的に学習に用いられる履歴データは，プレゼンテーションバイアスの影響により，ごく一部のユーザーやアイテムのみ履歴データが豊富にあり，残りの殆どには履歴が僅かといった頻度バイアスを含むデータである．

また推薦システムの欠点として紹介したフィルターバブルの解消のためには，今後探索と活用を能動的に繰り返しながら，ユーザーの嗜好度を能動的に学習しながら推薦を行うといったアプローチが求められる．それらはつまり既存のオフラインデータに対してのみ，性能を向上させるようなモデルの開発と評価を行うことでは，それらの解消が根本的に困難であることを意味する．ユーザーのアイテムに対する嗜好度の予測とその予測の信頼度は，今後の能動的学習における探索と活用のバランスをとる上で非常に重要と考えられるが，既存の深層学習に基づく協調フィルタリング手法はそれらが行えないという欠点がある．

そこで本研究では，既存の潜在因子に基づく協調フィルタリング手法について，疎なデータからの学習と不確実性のモデリングを同時に可能にする汎用的な学習フレームワーク MetaCF を提案する．提案手法は推薦システムをメタラーニングの一種である Neural Processes の観点から定義し直すことで，従来のモデルにほとんど手を加えることなく，それらの予測を実現する．

1.2.3 本論文の構成

本稿の構成として，まず第 2 章において背景知識となるコンテンツの価値推定の試みについての既存研究を紹介し，それら手法に内在するバイアスの観点から既存研究の限界について述べる．第 3 章から第 5 章は大きく二つのパートに分けられる．第 1 パートである第 3 章，第 4 章は集合知ベースのようなモノの価値を一元的に評価する手法についてそれぞれ RQ1，RQ2 に答える形で検討する．第 2 パートである第 5 章は推薦システムに代表される，モノの価値を個人ベースで推定する手法について RQ3，RQ4 に答える形で検討する．最後に第 6 章で，本稿のまとめと今後の課題についてまとめる．

1.3 表記法

本論文で一般的に使用する表記法について説明を行う．

表記例	説明
x, y, z	小文字のイタリック体はスカラーを表す.
$\mathbf{x}, \mathbf{y}, \mathbf{z}$	小文字の太字はベクトルを表す.
$\mathbf{X}, \mathbf{Y}, \mathbf{Z}$	大文字の太字は行列を表す.
$\mathcal{X}, \mathcal{Y}, \mathcal{Z}$	カリグラフィーフォントの大文字は集合を表す．例外として $\mathcal{L}$ はスカラー値を持つ損失関数を表すのに用いられる.
$\mathbb{R}^N$	N 次元の実空間を表す.
$\mathbf{x}_i$	$\mathbf{x}$ ベクトルにおける i 番目の要素を表す.
$\mathbf{X}_{i,j}$	行列 $\mathbf{X}$ における i, j 番目の要素を表す.
$\mathbf{X}_{i,*}, \mathbf{X}_{*,j}$	行列 $\mathbf{X}$ における i 行目のベクトル，または j 列目のベクトルを表す.
$f(\mathbf{x}), f(\mathbf{X})$	$\mathbf{x}$, $\mathbf{X}$ を入力とする関数を表す.
$f_\theta(.)$	θ というパラメータによって制御された関数であることを表す.

第2章　既存研究とその限界

2.1　はじめに

本章では，本論文で議論するトピックと関わりの深い背景知識について説明を行う．それぞれのプラットフォームごとのコンテンツの質の既存自動評価技術について，定式化，既存の試みに潜むバイアス，バイアスを取り除くための試み，既存研究の限界について述べる．第 1 章では既存の試みを，集合知に基づく試みと機械学習に基づく試みに分けて説明した．本章では機械学習に基づく試みの対象として，検索エンジン，レビューサイト，Q&A サイト，推薦システムを取り上げる．

2.2　集合知に関する試み

2.2.1　既存の試み

レビューサイトや Q&A サイトでは，投稿されたコンテンツの有用性をユーザーが投票するシステムが導入されている．例としてはレビューに対する「役に立った」(第 1 章 図 1.1) 機能が挙げられる．ユーザーは他ユーザーによる評価の合計の多い順に並べ替えることで，容易に有用性の高いコンテンツに到達できるよう導線が準備されている．

これらは多くの個人による独立した推定値の集合が，単一の専門家による判断の性能を上回るといった「群衆の知恵」の考えに基づいている [6, 7]．集合知の研究はがんの診断 [29] から財務予測 [30] に至るまで，様々な問題に応用されている．これらは実装の容易さから多くのオンラインプラットフォームで広く用いられている．

2.2.2　既存手法に潜むバイアスと弊害

Burghardt et al. [27] によると，以下 5 項目の要因によって，ユーザーは投票形式の評価行動に影響を受けることを示している．

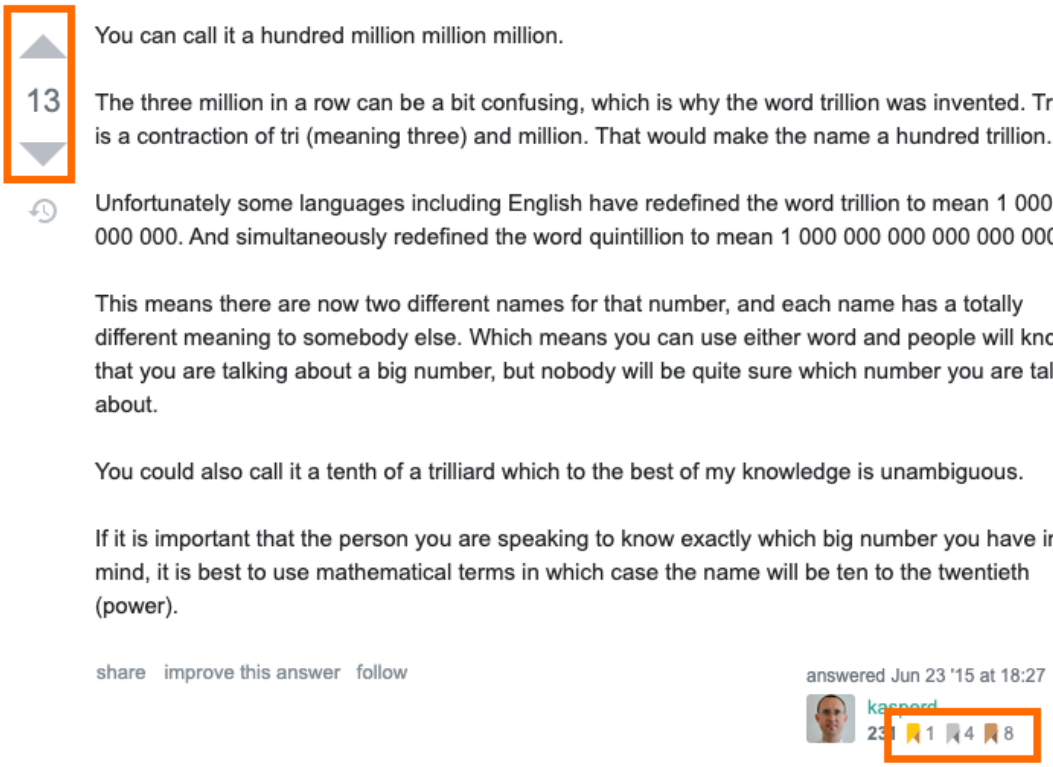


図 2.1: Stack Exchange からの回答のサンプル．左上には他ユーザーによる投票の合計で，通常はこのスコアの大きい順に並び替えられる．右下は，投稿者が過去の貢献具合に則り獲得した 3 種類のメダルである．

- コンテンツのトピック
- 内容の質
- 投稿者の評判 (社会的影響力バイアス)
- 他者による評判 (評判バイアス)
- 表示順番 (位置バイアス)

これらのうち，前者二つは投稿コンテンツ内容に依存する．後者三つは投稿以外の要因に由来し，人は認知ヒューリスティックを使用する傾向にある．このような投票行動に影響を与えるコンテンツ以外のバイアス情報を本論文では「認知バイアス情報」と呼称する．それぞれ三つの認知バイアス情報について下記で解説を行い，図 2.1 には Stack Exchange における例を示す．

社会的影響力バイアス

集合知の重要な前提として，個々人が他者の意見を知らない状態での意思決定を行う「意見の独立性」が挙げられる [31]．近年の研究によると，他者の意見を知った上で意思決定を行うと，意見の多様性が低下し，集合知の品質が落ちることが知られている [32, 33].

既存のプラットフォームにおける，他者が同一コンテンツにどれだけ「いいね」を押したかといった情報が観測できる状態は，集合知の前提である意思決定の独立性を歪める．その点，1.1.4 項で説明した人気と質の解離といった現象が生じる．

また多くの他者の意見が参照できる環境下では，群集的行動を示すことが先行研究から知られている [34]．同現象は最初の票がついていない環境下では，ユーザーは投票を行いにくく，また最初数件の高評価や低評価といった情報によって後続の意見が類似した傾向になることを指す.

評判バイアス

図 2.1 の右下に示すように Stack Exchange では投稿された回答に，投稿者の過去の貢献に基づいたスコアやバッジといった情報が付与されていることがある．他にも Amazon では，一部のレビューに Amazon が独自に指定したトップレビュワーである表示が付与されている場合がある．このようにユーザーの事前の評判に基づいてコンテンツの質を推測することは，社会的影響力バイアス同様，意見の多様性を減少させる．

位置バイアス

多くのオンラインプラットフォームでは，上から順番にコンテンツを並べられる形で表示される．ユーザーは下位に表示されたコンテンツは読みにくいといった傾向や，上位に表示されているものは他のコンテンツよりも有益だという認知バイアスを用いる傾向にある．これらがもたらす副次的影響については，1.1.3 項を再度参照されたい．

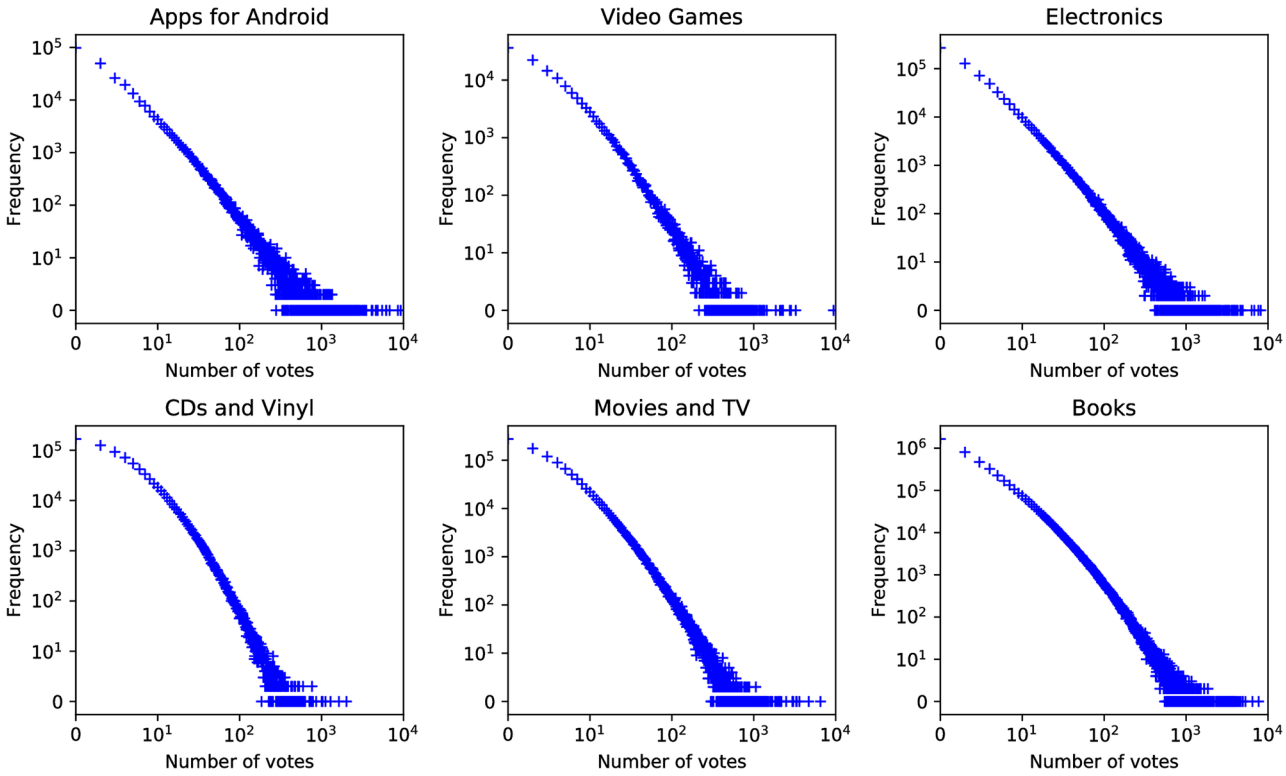


図 2.2: Amazon レビューにおける，カテゴリごとのレビューが持つ「役に立った」の投票数．Du et al. の図 1 より [35]．

2.2.3 既存手法の限界

前述した集合知の前提である意見の独立性が侵害されることに伴う，人気と質の乖離以外に加えて，本項ではさらに二つの既存の限界を取り上げる．

一つ目に，多くのコンテンツがそもそも投票を得られていないという点が挙げられる．1.1.3 項におけるプレゼンテーションバイアスや，社会的影響力バイアスで取り上げた集団の群衆的心理などに起因して，図 2.2 に示すように，各レビューが持っている投票の数には大きな偏りがあることが知られている．同結果よりほとんどのコンテンツはほぼ票が得られておらず，ごく一部のコンテンツのみ評価が行われている．それら偏った評価のみを元に全コンテンツの質を判断するシステムは公平とは言い難い．これら問題への取り組みとしては，後述 2.4 節で取り上げるコンテンツ内容やメタデータを元に推論する機械学習による推論が行われている．

二つ目に，これら投票の中には意図的に高評価，または低評価に見せようとするスパムユーザーが混在している可能性がある点が挙げられる．レビューサイトなどでは商品を使いもせずに意図的に評価を上げるために最上位の星や高評価のコメントをつけるなどの「やらせレビュー」，手当たり次第すべての商品に低評価を書き込む「アンチレビュー」，説明書きの文章をコピペしただけのものや翻訳ソフトにかけただけのような不自然な日本語のレビューなど数多く存在する．これら意見は特に図 2.2 のように得られている投票の数が少ない場合，影響が顕著に現れる．それらの怪しい意見を集合知に組み込んでいいのかという Research Question のもと，第 4 章ではこれら信頼性に基づいた集約方法により，より良い集合知が得られるかについて検討を行う．

2.3 情報検索システムにおける試み

2.3.1 リンク解析に基づく試み

既存の試み

ウェブページの重要度を測る試みとして，ページの内容ではなくハイパーリンク情報に着目した，リンク構造に基づく試みが知られる．代表的な手法としては PageRank が挙げられる．PageRank は「重要なページにリンクされているページは重要である」という発想に基づいている．同手法はランダムサーファーモデルを仮定し，あるウェブページの訪問者がリンクに基づきランダムに別のウェブページに遷移する状況が繰り返されていると想定した際，ランダムサーファーの各ウェブページにお

ける存在確率を重要度として用いている．同手法は様々な改良が提案されており，パーソナライズドサーチに応用した Personalized PageRank [36] や，トピックごとに PageRank を重み付けした Topic-Sensitive PageRank [37] など今日でも研究が続いており，文書要約といったウェブドメイン以外での応用も研究されている [38].

既存手法の限界

重要な視点として，リンク解析に基づく試みにおける重要度の概念は開発者のヒューリスティックに基づいているという点が挙げられる．それらの利点として，ユーザーのクリックなどのフィードバックがなくても動作するため，サービスローンチ直後でもある程度の性能が保証される．一方でデメリットとして，さらなるユーザー体験の向上を目指す場合，アルゴリズムの改修をその都度ヒューリスティックを用いて，プラットフォームごとに行う必要があるといった，アドホックな側面が挙げられる．

2.3.2 Learning-to-Rank

そこでアドホックな改修を避けるために，ユーザーのフィードバックを学習データとして用い，プラットフォームに依存しない汎用的な教師あり機械学習手法を用いた，ウェブコンテンツの質を推定する試みが行われている．それらは多くの場合ランク付けモデルを構築するランク付け学習 (Learning-to-Rank) の枠組みで学習される．

定式化

検索システムの目的は検索クエリ $\mathbf{q}$ に対して，関連性 $r \in \mathbb{R}$ が大きい文書 $\mathbf{d}$ を提示することである．ここでは $\mathcal{D} = \{(\mathbf{q}, \mathbf{d}, r)\}$ を各 $\mathbf{q}$ と $\mathbf{d}$ に対応する r によって構成されたデータ集合を仮定する．ここでの r は人間のアノテーションによって明示的に作成された理想的なデータセットとする．

ここで, f_θ を学習させるランキングを作成する関数とすると，以下式を用いて最適化される．

$$\theta^* = \arg\min_{\theta} \quad \mathcal{L} \tag{2.1}$$

$$\mathcal{L} = \sum_{(\mathbf{q},\mathbf{d},r)\in\mathcal{D}} r \cdot f_\theta\left(\mathbf{q}, \mathbf{d}, \mathcal{X}_q\right) \tag{2.2}$$

ここでの $\mathcal{X}_q = \{\mathbf{d}|(\mathbf{q},\mathbf{d}) \in \mathcal{D}\}$ はデータ集合におけるクエリと共起する文書の集合を表す．損失関数には ARP (Average Relevance Position) や DCG (Discounted Cumulative Gain) が用いられる．

$$\mathcal{L}_{ARP} = \sum_{(\mathbf{q},\mathbf{d},r)\in\mathcal{D}} r \cdot \mathrm{rank}\,(\mathbf{q},\mathbf{d},\mathcal{X}_q) \tag{2.3}$$

$$\mathcal{L}_{DCG} = \sum_{(\mathbf{q},\mathbf{d},r)\in\mathcal{D}} r \cdot \frac{1}{\mathrm{rank}\,(\mathbf{q},\mathbf{d},\mathcal{X}_q)} \tag{2.4}$$

ここでは $\mathrm{rank}\,(\mathbf{q},\mathbf{d},\mathcal{X}_q)$ を $\mathbf{d}$ の $\mathcal{X}_q$ における順位とする．$\mathcal{L}_{ARP}$ は関連する文書の順位の和が小さくなるよう促し，$\mathcal{L}_{DCG}$ は関連する文書の順位の逆数の和が小さくなるよう促すことを表す．

これらランキング学習を行う手法としては，Ranking SVM [39] や LambdaMART [40] などが広く知られている．

既存手法に潜むバイアス

しかし現実的にはウェブコンテンツの有用性の人間によるアノテーションはコストが高く大量の収集が困難である．多くの研究では安価に収集が可能な，ユーザーのクエリに対するクリック情報といった暗黙的なフィードバック (implicit feedback) に基づいたデータ集合 $\mathcal{D} = \{(\mathbf{q},\mathbf{d},k,c)\}$ を用いて学習が行われる．ここでの k は $\mathbf{q}$ に対して $\mathbf{d}$ を提示した時の表示順位を表し, c はその時にクリックが発生したか否かを表す (式 2.5)．

$$c = \begin{cases} 1, & \text{フィードバックが観測された場合} \\ 0, & \text{それ以外} \end{cases} \tag{2.5}$$

しかしナイーブに r を c に置き換えて学習を行うことは，表示順位 k がバイアスとして学習に影響を与える．ここでは Joachims et al. [41] によって提案された Position-Based Model (PBM) の観点から解説を行う．

PBM では以下の関係性を仮定する．

$$\begin{aligned} \mathrm{P}(C=1 \mid \mathbf{q},\mathbf{d},k) &= \mathrm{P}(E=1 \mid k) \cdot \mathrm{P}(R=1 \mid \mathbf{q},\mathbf{d}) \\ &= \theta_k \cdot \gamma_{q,d} \end{aligned} \tag{2.6}$$

ここで C はクリックを表す二値確率変数，E はユーザーが観測したか (Examination) を表す二値確率変数とする． このモデル化により クエリ $\mathbf{q}$ と文書 $\mathbf{d}$ が関連性がある $(R=1)$ かつ観測された $(E=1)$ のときにクリックが発生する $(C=1)$ といった

仮定を置く．ナイーブに既存のアルゴリズムにおける r を c に置き換えて学習を行うことは，PBM の観点に基づくと，バイアスを含むことを以下に示す [41]：

$$\begin{aligned}\mathbb{E}\left[\mathcal{L}_{naive}\right] &= \sum_{(\mathbf{q},\mathbf{d},k,c)\in\mathcal{D}} \mathbb{E}[C]\cdot f\left(\mathbf{q},\mathbf{d},\mathcal{X}_{\mathbf{q}}\right) \\ &= \sum_{(\mathbf{q},\mathbf{d},k,c)\in\mathcal{D}} \theta_k\cdot\gamma_{q,d}\cdot f\left(\mathbf{q},\mathbf{d},\mathcal{X}_{\mathbf{q}}\right) \\ &= \sum_{(\mathbf{q},\mathbf{d},k,c)\in\mathcal{D}} \mathrm{P}(E=1\mid k)\cdot\mathrm{P}(R=1\mid\mathbf{q},\mathbf{d})\cdot f\left(\mathbf{q},\mathbf{d},\mathcal{X}_{\mathbf{q}}\right)\end{aligned} \tag{2.7}$$

ここで, $\mathrm{P}(E=1\mid k)$ の部分は本来不要な重みである．これよりナイーブな推定量は真の評価値に対して比例関係になく，バイアスを含むことが示された．ナイーブに r と c を置き換えて学習させることは，すなわち過去のランキングシステムが提示した高いポジションをより重要視するように学習させてしまっていると言える．言い換えると，上位に提示されたアイテムが不要に有用度が高く学習されることに繋がり，その結果常に上位に提示され続けることで最終的に多様性の減少に繋がる．

バイアスを取り除く試み

位置バイアスの影響を取り除いてランキングモデルの構築する試みは Unbiased -Learning-to-Rank という名前で知られる．PBM を提案した Joachims et al. [41] は，位置バイアスを取り除く試みとして，因果推論分野で広く用いられている IPW (Inverse Propensity Weighting) を活用して位置バイアスを取り除く試みをしている．

そこで事前に Result Randomization と呼ばれる提示結果をランダムに表示させる試みを行うことによって，θ_k を推定する．なお上位 N 個の文書がユーザに表示される前にランダムにシャッフルされたデータセット $\mathcal{R}$ に対して，位置 k から収集されたログのサブセットを $\mathcal{R}_k$ とする．

$$\begin{aligned}\mathbb{E}[C\mid k] &= \sum_{(\boldsymbol{q},\boldsymbol{d})\in\mathcal{R}_k} \mathbb{E}[C\mid\boldsymbol{q},\boldsymbol{d},k]P(\boldsymbol{q},\boldsymbol{d}) \\ &= \sum_{(\boldsymbol{q},\boldsymbol{d})\in\mathcal{R}_k} \theta_k\gamma_{\boldsymbol{q},\boldsymbol{d}}P(\boldsymbol{q},\boldsymbol{d}) \\ &= \theta_k\sum_{(\boldsymbol{q},\boldsymbol{d})\in\mathcal{R}_k} \gamma_{\boldsymbol{q},\boldsymbol{d}}P(\boldsymbol{q},\boldsymbol{d}) \\ &\propto \theta_k\end{aligned} \tag{2.8}$$

得られた θ_k を元にナイーブな損失関数に逆数を乗ずることで Unbiased な損失関数が得られていることが分かる：

$$\begin{aligned}
\mathbb{E}\left[\mathcal{L}_{IPW}\right] &= \sum_{(\mathbf{q},\mathbf{d},k,c)\in\mathcal{D}} \frac{\mathbb{E}[C]}{\theta_k} \cdot f\left(\mathbf{q},\mathbf{d},\mathcal{X}_q\right) \\
&= \sum_{(\mathbf{q},\mathbf{d},k,c)\in\mathcal{D}} \frac{\theta_k \cdot \gamma_{\mathbf{q},\mathbf{d}}}{\theta_k} \cdot f\left(\mathbf{q},\mathbf{d},\mathcal{X}_q\right) \\
&= \sum_{(\mathbf{q},\mathbf{d},k,c)\in\mathcal{D}} \gamma_{\mathbf{q},\mathbf{d}} \cdot f\left(\mathbf{q},\mathbf{d},\mathcal{X}_q\right) \\
&= \sum_{(\mathbf{q},\mathbf{d},k,c)\in\mathcal{D}} \mathrm{P}(R=1 \mid \mathbf{q},\mathbf{d}) \cdot f\left(\mathbf{q},\mathbf{d},\mathcal{X}_q\right)
\end{aligned} \tag{2.9}$$

直感的な理解として，表示順位が下位なことによって本来一番上のページと比べて 5 倍 Examination が行われないページがクリックを得た場合，一番上のページと比較して 5 倍大きく重み付けされる形で学習が行われる．

既存手法の限界

既存の PBM には大きく二つの課題が挙げられる．一つ目は IPW を行うための θ_k の推定に関するハードルの高さが挙げられる．一般的に Result Randomization はユーザー体験を著しく害したり，KPI に大きな打撃をもたらすためオンラインプラットフォーム運営の立場からは行うことが困難である．既存のその他の研究では θ_k の推定ハードルを下げるために回帰ベースの EM アルゴリズムを採用している [42]．

二つ目は PBM が置く仮定が非常に厳しい点である．PBM を拡張した試みとして Wang et al. [43] は誤クリック率をモデルに組み込んだ TrustPBM を提案している．しかし 2.2.2 項で示したように，ユーザーのクリックは位置以外の情報によってもバイアスを受けることが知られており，それら影響の考慮は PBM だけでは不十分と言える．式 2.9 は一見位置バイアスの影響を除去できているように見えるが，これらは PBM が成り立つ時場合のみ有効であり，それ以外のモデリングを仮定する場合は，同モデリングに応じた Unbiased な損失関数の設定が必要な点に留意する．

第 3 章では主にこれらの課題意識から，2.2 節で示した複数の認知バイアスにユーザーが晒されている場合におけるユーザーの投票形式の評価行動におけるモデリングと，同モデリングについて有効な Unbiased Learning-to-Rank 手法の提案を行う．

2.4 レビューサイト・Q&A サイトにおける試み

レビューサイトや Q&A サイトにおいて，2.2 節の既存の限界で取り上げた，投票を未だ得ていないコンテンツに対する質の推定は，公平性の高いシステム構築のために重要である．そこで近年では，投票形式の集合知に基づく試み以外に，機械学習を用いてそれらコンテンツの内容やリンク情報から質を推定する試みが行われている．本章ではそれらを二つに大別し解説する．

2.4.1 リンク解析に基づく悪質コンテンツ検知

既存の試み

一つ目は誰が何にどのような評価をしたかという評価ネットワークのリンク構造に基づく手法である．これらの多くの手法は不良レビュー・レビュワーの検出に焦点が置かれてることが多い．第 4 章で用いる REV2 [28] も本リンク解析に基づく手法に該当し，詳細は第 4 章で解説する．

反復学習を利用したアルゴリズムとしては，評価のコンセンサスに基づいて，評価ネットワーク内のスコアを共同で割り当てる手法が提案されている [44, 45, 46]．Sun et al. [44] は，ユーザー，レビュー，製品ごとにスコアを付け，Minnich et al. [45] は，ユーザーと製品ごとにスコアを付けている．FraudEagle [47] は，ユーザーをランク付けするための信念伝播モデルであり，詐欺師は良い製品を悪く評価し，悪い製品を肯定的に評価し，正直なユーザーはその逆であると仮定している．ランダムウォークベースのアルゴリズムは，ネット上の荒らし [48] や Twitter 上の談合から検索エンジンスパムであるリンクファームの検出に用いられている [49]．Jiang et al. [50]，Wang et al. [51]，Li et al. [46] は，ユーザーのローカルな近傍性に基づいて悪質ユーザーのグループを特定している．ネットワークベースの不正検知に関するサーベイは Akoglu et al. [52] によってまとめられている．

二つ目に時間 [53, 54, 55] やテキスト内容 [56, 57, 44] といったコンテンツに基づく手法もある．それらアルゴリズムは多くの場合，ドメイン知識の基づく特徴量をベースにしている．コンセンサスに基づく特徴量は [58, 59] で提案されており，REV2 における良否判定基準もこれら「群衆の知恵」の考えを用いている．一般的に使用されている特徴は，タイムスタンプ [60, 61, 62] やレビューテキスト [44, 63, 57] などから得られる．SpEagle [64] は，FraudEagle [47] を拡張して行動特徴を組み込んでいる．BIRDNEST [65] は，ベイズモデルを作成して，期待される行動からの各ユーザーの評価行動の偏差を用いて確信度を推定する．Jiang et al. [50]，Chen et al. [66]，

Viswanath et al. [67] は複数のユーザーの協調的なスパム行動を研究している．これらの特徴量ベースのアルゴリズムに関するサーベイは Jiang et al. [68] によってまとめられている．

既存手法の限界

これらの試みの利点と限界は 2.3.1 項のリンク解析に基づく試みの既存手法の限界と同じことが言える．すなわち開発者のヒューリスティックに基づいているため，ユーザーのクリックなどのフィードバックがなくても動作し，サービスローンチ直後でもある程度の性能が保証されるという利点が挙げられる．一方デメリットとして，さらなるユーザー体験の向上を目指す場合，アルゴリズムの改修をその都度ヒューリスティックを用いて，プラットフォームごとに行う必要があるといった，アドホックな側面が挙げられる．

2.4.2 Helpfulness Prediction

EC サイトにおける商品レビューのようなユーザー生成型コンテンツの品質を教師あり学習のフレームワークで推定することは，Helpfulness Prediction と呼ばれる分野で活発に研究されてきた．

定式化

Helpfulness Prediction は文書 $\mathbf{d}$ ごとに，有用度 $h \in \mathbb{R}$ を推定する．ここでは $\mathcal{D} = \{(\mathbf{d}, h)\}$ を各 $\mathbf{d}$ に対応する h によって構成された理想的なデータ集合と仮定する．このような場合 h は人間によるアノテーションによって明示的に作成された理想的なデータセットとする．

ここで, f_θ は有用度 h を推定するための関数とすると，以下式を用いて最適化される．

$$\theta^* = \arg\min_{\theta} \sum_{(\mathbf{d},h)\in\mathcal{D}} \delta(h, f_\theta(\mathbf{d})) \tag{2.10}$$

多くの場合，h を直接予測する回帰モデルを構築する，または閾値を設けて有用か有用でないかの二値分類モデルを作成する場合に分けられる．ここでの δ は損失関数を表し，タスクに応じて設計され，回帰の場合は式 2.11 のように SSE (Sum of

Squared Error) を，分類問題の場合は式 2.12 のように負の対数尤度 (Negative Binary Cross Entorpy) が用いられる．

$$\mathcal{L}_{SSE} = \sum_{(\mathbf{d},h)\in\mathcal{D}} (h - f_\theta(\mathbf{d}))^2 \tag{2.11}$$

$$\mathcal{L}_{BCE} = -\sum_{(\mathbf{d},h)\in\mathcal{D}} (h \cdot \log(f_\theta(\mathbf{d})) + (1-h) \cdot \log(1 - f_\theta(\mathbf{d})) \tag{2.12}$$

既存研究の多くは，これら $\mathbf{d}$ をいかに構築するかに重きが置かれていた．例えば TF-IDF や文書の読みやすさを測る ARI，文書の感情極性，文ごとの単語数の平均などが用いられてきた．これら特徴量のサーベイは Diaz et al. [69] によってまとめられている．一度選択された特徴量は連結され，線形モデルやロジスティック回帰，RandomForest といった広く用いられている機械学習アルゴリズムに入力される．また近年では深層学習を用いてこれら特徴量エンジニアリングを自動で行う試みが広がっており，従来の機械学習モデルよりも予測性能面で向上が確認されている [70, 71].

昨今の Helpfulness Prediction の新たな方向性としては，文書以外の周辺コンテキスト情報に基づく特徴量を取り入れる試み挙げられる．代表的なコンテキスト情報としては，レビューメタデータ [72, 73]，レビュワーの特徴 [74, 75] など，対象のレビューが持つ周囲と比較した情報量の増分 [76] などが検討されている．ある商品 $\mathbf{i}$ における周辺のコンテキスト情報を $\mathcal{X}_i = \{\mathbf{d} \mid (\mathbf{i},\mathbf{d}) \in D\}$ で表すと，それらは下記のように定式化される．

$$\theta^* = \arg\min_\theta \sum_{(\mathbf{i},\mathbf{d},h,\mathbf{X})\in\mathcal{D}} \delta(h, f_\theta(\mathbf{d}, \mathcal{X}_i)) \tag{2.13}$$

既存手法に潜むバイアス

ナイーブにユーザーのクリック情報を正解データとして用いることは，複数の認知バイアスが投票行動に影響を与える点において，真に価値 r を表しているとは言えない．この点に関しては集合知と情報検索エンジンに潜むバイアスと同様のことが言える．

バイアスを取り除く試み

近年の Helpfulness Prediction では，単にユーザーのクリックの合計を用いるのではなく，Y 人のユーザーのうち X 人がレビューが有用であると投票した場合，そのレビューの有用性を X/Y と定義し用いられることが多い．

これら試みは，情報検索エンジンにおける IPW を用いたバイアスを取り除く試みと関連性が高い．なぜなら投稿にユーザーが接触した回数を Y を用いて近似しているとみなすことができ，ユーザーのコンテンツ接触頻度によるバイアスを取り除こうとしていると解釈ができる．

既存手法の限界

既存手法の限界としては明確な教師シグナルの不足である．現状の多くが X/Y といったコンテンツとのユーザーの接触頻度による影響を緩和しようとしている一方で，しかしそれらは人間のアノテーションとの相関が低いという報告が多く行われている [16, 22, 23, 24, 8]．考えられる理由としては大きく二つが考えられる．一つ目はユーザーとの頻度以外に取り除くべき認知バイアスの影響が存在するということ，二つ目はユーザーによるフィードバックが数件しかない場合には X/Y という指標は不確実性を含みやすい，例えば 1000 人中 999 人が高評価したものと 2 人中 2 人が高評価したものでは後者の方が評価が高く計算されてしまう．

そのため，今後の Helpdulness Prediction を押し進めるためには，複数の認知バイアスの影響を取り除いた真の有用性 r を元に従来行われていたような機械学習モデルの構築を行うことが考えられる．Helpful Prediction の分野では前述 PBM のようなモデリングは我々の知る限りでは行われていない．本研究の第 3 章は新たにユーザーの投票形式の評価行動におけるモデリングと，同モデリングについて有効な Unbiased Learning-to-Rank 手法の提案を行う．提案手法によって得られたコンテンツの有用度を r として既存の Helpfulness Prediction の枠組みで学習させるといった応用展開が考えられる．

2.5 推薦システムにおける試み

推薦システムは近年の情報過多の問題を解決する上で非常に重要な役割を果たす．推薦システムは個人ごとに嗜好度，またはコンテンツの質を推論しているという観点で，本論文の主題と関連性が高い．

2.5.1 コンテンツベースフィルタリング

コンテンツベースフィルタリングは「過去にユーザーが高評価したアイテムと類似したアイテムを推薦する」という思考に基づいている [77]．例として映画の推薦では過去に試聴した映画の出演俳優やジャンルが近いものが推薦されるといった具合である．

手法の流れとして，ユーザーとアイテムはコンテンツ情報に基づき同一の特徴空間に写像され，同空間上での類似度を元に推薦が行われる．一般的に特徴量はドメイン知識を基づきジャンルといったメタデータや，TF-IDF などによってベクトル化されたテキスト情報が使用される．類似度指標にはコサイン類似度が広く用いられている．

コンテンツベースフィルタリングの利点として以下の二点が挙げられる．一つ目に新規のユーザーやアイテムといった履歴データが少ないアイテムにも特徴量が適切に選択されていれば推薦をすることができる点が挙げられる．二つ目にどの特徴量が推薦に寄与したのかといった推薦理由の可読性が高い点が挙げられる．

既存手法に潜むバイアスと限界

同手法は開発者の特徴量の選択にドメイン知識が不可欠になる．その結果，これまでのリンク解析に基づく手法で取り上げた例と同じく，開発者のヒューリスティックがシステムに内在し，アドホックな改修にはさらなるヒューリスティックの活用が必要になる．

欠点として，コンテンツベースフィルタリングの仮定上，ユーザーが過去に好んだアイテムと類似したアイテムばかりが推薦され，推薦にセレンディピティが欠けてしまうことが知られる．例えばプリンターを購入したユーザーには本来インクを推薦したいところだが，類似度に基づくことで類似のプリンターを推薦してしまうことに繋がりやすい．そのような問題を解決するために次項の協調フィルタリングに基づく推薦の研究が進められている．

2.5.2 協調フィルタリング

協調フィルタリングは,「ユーザーのアイテム利用履歴から，類似するユーザーの利用行動を元に推薦をする」という思考に基づいている．これらはコンテンツベースの手法と異なり，ドメイン知識に左右されずに学習を行える点において優れており，広く用いられている．

定式化

システム内にはM人のユーザーとN個のアイテムが存在すると仮定する．インタラクション行列$\mathbf{Y}_{u,i}$をユーザーuがアイテムiに対して有する真の嗜好度または価値とする．推薦システムは以下の誤差関数を最小化するような，ユーザーのアイテムに対する評価値の予測集合$\hat{\mathbf{Y}}$を得ることである．

$$\mathcal{L} = \frac{1}{M \cdot N} \sum_{M,N} \delta_{u,i} \left(\mathbf{Y}, \hat{\mathbf{Y}} \right) \tag{2.14}$$

ここでの$\delta_{u,i} \left(\mathbf{Y}, \hat{\mathbf{Y}} \right)$をユーザー$u$のアイテム$i$に対する予測値の局所損失を表す．評価値がレビューの星のような明示的なフィードバックを用いて回帰問題として扱う場合はMSE (式2.15)や，ユーザーのクリックといった二値変数を扱う場合負の対数尤度 (式2.16)などが用いられる．

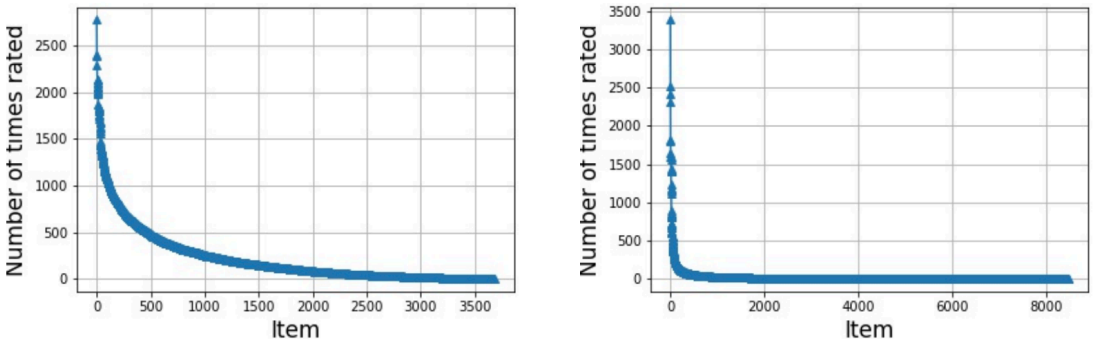
$$\mathcal{L}_{MSE} = \frac{1}{M \cdot N} \sum_{u,i} (\mathbf{Y}_{ui} - \hat{\mathbf{Y}}_{\mathbf{ui}})^2 \tag{2.15}$$

$$\mathcal{L}_{BCE} = -\frac{1}{M \cdot N} \sum_{u,i} (\mathbf{Y}_{ui} * \log(\hat{\mathbf{Y}}_{ui}) + (1 - \mathbf{Y}_{ui}) * \log(1 - \hat{\mathbf{Y}}_{ui})) \tag{2.16}$$

一般的にこれらインタラクション行列$\mathbf{Y}$は部分的に観測された欠損データになっている．欠損値はユーザーがアイテムを閲覧していない，またはアイテムを閲覧したが評価を行っていないことにより生じる．協調フィルタリングはこれら欠損値を埋める行列補間タスクとみなされる．

協調フィルタリングの手法として最も広く用いられている手法としてはMatrix Factorization (MF) [12]が挙げられる．MFはユーザーとアイテムをそれぞれ潜在的なベクトルで表現し同一空間上にマッピングを行い，内積といった類似度を図る指標に基づき，ユーザーのアイテムへの嗜好度をモデリングする手法である．

近年の研究ではユーザーとアイテムの潜在的なベクトルのインタラクションを深層学習を用いて，複雑なモデリングを学習し，性能が向上することが知られている．例えばDeepMF [78]はMFに深層学習を適応させており，加えてNCF [79]はgeneralized matrix factorization model (GMF)と多層パーセプトロン (MLP)の二つの機構を組み合わせた機構を持っている．CFNet [80]はユーザーとアイテムのembeddingのマッチングを複雑な非線形性を考慮した部分と，低ランクのマッチングを行う部分とに明示的にネットワークを分けて学習を行っている．



(a) MovieLens　　(b) Yahoo Movies

図 2.3: MovieLens と Yahoo Movies のデータセットにおけるアイテムごとのインタラクション数．Abdollahpouri et al. の図 2 より [81]．アイテムによって得られているインタラクション数に偏りがあることが確認される．

既存手法に潜むバイアス

ユーザーは一般的に自分の興味のありそうなアイテムに対してインタラクションを行うため，人気の高いアイテムはインタラクションが観測されやすく，人気でないアイテムはインタラクションが観測されにくい．そのためインタラクション行列 $\mathbf{Y}$ の欠損値の有無は，Missing Not At Random (MNAR) なデータと呼ばれ，アイテムの人気度に依存した自己選択バイアスのかかったデータであることが知られる．図 2.3 に推薦タスクのベンチマークで広く用いられているデータセット MovieLens 1M と Yahoo movies において，学習データに含まれるアイテムごと (横軸) のインタラクション数 (縦軸) を示す．アイテムによって得られているインタラクション数に偏りがあるといった頻度バイアスが確認される．

多くの場合観測された評価に関する予測を平均することで損失を計算する．ここで，ユーザー u のアイテム i に対する嗜好度 $\mathbf{Y}_{u,i}$ が観測できたかどうかを表す確率変数 $\mathbf{O}_{u,i}$ を導入するとナイーブに計算した損失関数は以下のように表される．

$$\mathcal{L}_{naive} = \frac{1}{|\{(u,i) : \mathbf{O}_{u,i} = 1\}|} \sum_{(u,i):\mathbf{O}_{u,i}=1} \delta_{u,i}(\mathbf{Y}, \hat{\mathbf{Y}}) \tag{2.17}$$

それゆえナイーブな推定を用いることは，バイアスを持つことを以下示す [82]．証明のために $\sum_{u,i} \mathbf{O}_{u,i}$ を任意の正の定数 C で置換した上で $\mathcal{L}_{naive}$ の確率変数 $\mathbf{O}_{u,i}$ に

ついての期待値をとると

$$\begin{aligned}\mathbb{E}\left[\mathcal{L}_{naive}\left(\hat{\mathbf{Y}}\right)\right] &= \mathbb{E}_{\mathbf{O}_{u,i}\left[\frac{1}{C}\sum_{(u,i):\mathbf{O}_{u,i}=1}\delta_{u,i}\left(\mathbf{Y},\hat{\mathbf{Y}}\right)\right]} \\ &= \mathbb{E}_{\mathbf{O}_{u,i}\left[\frac{1}{C}\sum_{u,i}\mathbf{O}_{u,i}\cdot\delta_{u,i}\left(\mathbf{Y},\hat{\mathbf{Y}}\right)\right]} \\ &= \sum_{u,i}\frac{\mathbb{E}_{\mathbf{O}_{u,i}}\left[\mathbf{O}_{u,i}\right]}{C}\cdot\delta_{u,i}\left(\mathbf{Y},\hat{\mathbf{Y}}\right)\end{aligned} \tag{2.18}$$

ここで $\mathcal{L}_{naive}$ と $\mathcal{L}$ が比例関係であるためには, 任意の u と i について，下記式が成り立つ必要がある：

$$\frac{\mathbb{E}\left[\mathbf{O}_{u,i}\right]}{C} \propto \frac{1}{M\cdot N} \tag{2.19}$$

しかし図 2.3 に示すように観測確率が一様ではない場合, この条件を満たす定数 N は存在しない．よってナイーブな推定量は真の評価値に対して比例関係になく，バイアスを含むことが示された. これらは推薦システムの学習に用いられる履歴データが MNAR なデータというバイアスを含んでいる特性に依るものである．

バイアスを取り除く試み

これまでの情報検索エンジンにおける位置バイアスの除去のために IPW が用いられていたのと同様，Schnabel et al. [82] によって推薦システムのバイアス除去のためにも IPW を用いる手法が提案されている．同手法で扱う傾向スコア $\mathbf{P}_{u,i} = P(\mathbf{O}_{u,i} = 1)$ は $\mathbf{Y}_{u,i}$ が観測される確率を表す．

$$\mathcal{L}_{IPS} = \frac{1}{M\cdot N}\sum_{(u,i):\mathbf{O}_{u,i}=1}\frac{\delta_{u,i}\left(\mathbf{Y},\hat{\mathbf{Y}}\right)}{\mathbf{P}_{u,i}} \tag{2.20}$$

以下 $\mathbf{P}_{u,i}$ で重み付けることで Unbiased な損失関数が得られていることが確認された：

$$
\begin{aligned}
\mathbb{E}\left[\mathcal{L}_{IPW}\right] &= \mathbb{E}_O\left[\frac{1}{M\cdot N}\sum_{(u,i):\mathbf{O}_{u,i}=1}\frac{\delta_{u,i}\left(\mathbf{Y},\hat{\mathbf{Y}}\right)}{\mathbf{P}_{u,i}}\right] \\
&= \mathbb{E}_O\left[\frac{1}{M\cdot N}\sum_{u,i}\mathbf{O}_{u,i}\cdot\frac{\delta_{u,i}\left(\mathbf{Y},\hat{\mathbf{Y}}\right)}{\mathbf{P}_{u,i}}\right] \\
&= \frac{1}{M\cdot N}\sum_{u,i}\frac{\mathbb{E}_{\mathbf{O}_{u,i}}\left[\mathbf{O}_{u,i}\right]}{\mathbf{P}_{u,i}}\cdot\delta_{u,i}\left(\mathbf{Y},\hat{\mathbf{Y}}\right) \\
&= \frac{1}{M\cdot N}\sum_{u,i}\delta_{u,i}\left(\mathbf{Y},\hat{\mathbf{Y}}\right) \\
&= \mathcal{L}
\end{aligned}
\tag{2.21}
$$

既存手法の限界

疎なデータからの学習

協調フィルタリング手法の欠点として，新規のユーザーやアイテムといった履歴データが少ない高度に疎なデータの場合に性能が極端に減少することが知られる．図 2.3 で示したように一般的な学習データは，人気度に基づいて観測される頻度に差があり，ほとんどのアイテムが少ない履歴しか保持していない．なお IPW に基づく損失関数は，アイテムごとの相対的な観測頻度の是正に有効ではあるが，少ないデータからの学習が可能になるわけではないことに留意する．多くのこれら履歴データが十分に利用できない場合にはコンテンツベースの手法との組み合わせたハイブリッドモデルを用いる試みが多く行われている [83, 84]．

少ない学習データからの学習は Few-shot Learning と呼ばれ，あらゆる機械学習分野で活発に研究が行われている．近年ではメタラーニングと呼ばれる学習フレームワークが着目を集めており，関連研究についての説明は第 5 章にて行う．

第 5 章では近年のメタラーニングの枠組みを協調フィルタリング手法に拡張することで少ない学習データからの学習が可能になるかといった Research Question の元，手法の提案と検証を行う．提案手法により，従来予測が難しかった履歴数の少ないユーザーやアイテムへの正確な予測が可能になる．

不確実性の推定

既存の多くの協調フィルタリング提案手法の欠点としては，予測に関する不確実性のモデリングの不足が挙げられる．つまり推薦システムが予測するユーザーのア

イテムに対する予測について，それらがどれほど自信のある推論かが分からないといった具合である．何故これら予測の不確実性が今後の推薦システムにおいて重要と我々が考えるかについては後述 2.5.3 項にて解説する．

2.5.3 能動学習に基づく興味の推定

前述 2.5.1，2.5.2 項で説明した推薦システムの研究は，ユーザーのアイテムに対する履歴データを元に，モデルの学習と評価が行われてきた．これらは実装と評価の簡便さから広く用いられてきた．しかしこれらはユーザーが過去にインタラクションしたものを予測している点において，本当は興味があったかもしれないが見られなかったアイテムに対する嗜好度の評価が行えないといった欠点がある．

そこで近年の新しい試みとして，推薦システムを実際にサービスで運用することで，長期的にユーザーの反応を確認するといったオンライン評価を行う試みが挙げられる．例えば動画投稿サイト YouTube では，ユーザーの動画視聴時間を報酬とした強化学習モデルによる推薦手法の開発を行っている [85]．また Netflix は Contextual Bandits アルゴリズムを用いて，ユーザーに提示するアートワークをユーザーに応じて最適化させる試みを行っている [86]．

これらオンライン評価は，オフライン評価の課題である未観測のユーザーとアイテムの嗜好度の評価を，実際に推薦することで評価できる点について，オフライン評価よりも説得力が高い．加えて適切な報酬関数を設定することでフィルターバブルが起きないような推薦システムの構築も可能になるといった点において注目を集めている．しかし欠点として実際に推薦システムをオンラインサービスに実装し，一定期間運用する必要があるといった点で，検証や再現実験がオフライン評価よりも遥かに難しい点が挙げられる．

推薦システムを能動学習させる場合は実サービスに実装を行う必要があると言った点において，低品質な推薦を長時間行うことはユーザー体験を損なう．そこで学習効率よくシステムが学習することが望ましい．一般的に能動学習では，多くの行動空間から良い行動を探す「探索」と，既存の知識を利用し最良の行動を取り続ける「活用」という，二つの相反する行動をバランスを取りながら行われる．それらバランスをとる試みとしては Greedy 法やソフトマックス法，Thomson Sampling などが知られている．これら手法は共通して早期の段階では多く探索を行い，時間が経つに連れ活用を増やす傾向にある．

第 5 章で提案する協調フィルタリング手法の予測に不確実性を持たせるというアイディアは，能動学習の初期フェーズにおいて早期の探索をより効率的にし，短期間

での学習を可能にすると考えられる．例えばユーザーとアイテムのインタラクションの予測の不確実さが高いペアから探索するといった具合である．多くの既存の能動学習は学習初期では嗜好度がわかっていないケースが多く，それらは特に初期の性能が安定しない．オフラインデータを用いて嗜好度の不確実性を事前にモデリングし，それらを能動学習に応用する試みは我々の知る限りでは行われていない．これらの仮説検証については推薦システムのオンライン評価のハードルの観点から今後の課題とする．

2.6 まとめ

本章では，本論文で議論するトピックと関わりの深い背景知識について説明を行った．それぞれのプラットフォームごとのコンテンツの質の既存自動評価技術について，定式化，既存の試みに潜むバイアス，バイアスを取り除くための試み，既存研究の限界について述べた．

集合知に基づく試みのように，多くの場合コンテンツの有用度としてユーザーのクリックの数が用いられてきた．しかし 2.2.2 項で述べたような認知バイアス情報 (既存のいいね，投稿者の評判，表示順位など) に起因して，それらクリックは人間のアノテーションによる質とは乖離していることが知られる．

機械学習を用いてコンテンツの質を測る手法は大きく二種類に分けられた．一つ目は自己教師あり学習に基づく手法，二つ目は教師あり学習に基づく手法である．自己教師あり学習のメリットとしては，クリック情報がない場合でもアルゴリズム開発者の置いた仮説が正しければ，一定程度動作するというというである．デメリットとしては，全オンラインプラットフォームにおいて汎用的な手法はなく，目的に応じたドメイン知識に基づく目的関数の開発と，アドホックな改修が必要になる点が挙げられる．

教師あり学習のメリットとしては一般的な教師あり学習の枠組みで学習を行うことができ，プラットフォームやドメイン知識に依存しない点が挙げられる．一方デメリットとしては，ナイーブにクリック情報を正解データとして用いることは前述のバイアスを含むことでアルゴリズムにバイアスが内在してしまう点である．教師あり学習における既存のバイアスを取り除く試みとしては IPW によるバイアスを取り除く手法が各ドメインで行われてきた．既存のそれら手法は多くの場合ユーザーとの接触頻度の影響のみを扱っていた．

第I部

一元的な尺度によるコンテンツの質の自動評価手法について

第3章　認知バイアス情報の影響の除去による投票形式でのコンテンツの価値の推定

3.1　はじめに

本章では，第1章の **Research Question 1：　複数の認知バイアス情報の影響を受けているユーザーの行動データから,「認知バイアス情報が無ければどのように行動したのか」という反実仮想のもと，それら影響を取り除いた真のコンテンツの質を測定できるか？**に答える形で，事前の認知バイアス情報が無ければユーザーはどのように評価をしたのかという反実仮想的推定を行う．前述2.2節，2.4節で説明したように，既存検索システムなどの最適化に用いられたIPWに代表されるバイアスの影響を取り除く試みでは，ユーザーとコンテンツの接触頻度に関するバイアスのみを取り扱っていた．しかし，ユーザーの投票メカニズムは，2.2節で説明したように投稿ユーザーの事前評価 (評判バイアス)，これまでの投票の集計 (社会的影響力バイアス)，コンテンツの位置 (位置バイアス) など，様々なバイアスが絡み合って影響を受けることが知られている．これらのバイアスの結果，既存の人気とコンテンツの質が強く相関していないという知見が得られている [16, 22, 23, 24, 8]．

オンラインプラットフォームにおけるユーザーの投票行動に潜むバイアスの検出と定量化については，多くの研究が行われてきた．主なアプローチとしては，1) 参加者の投票条件を変えてランダム比較化試験を行う方法，2) 過去の投票データを解析するための統計モデルを開発する方法，3) 因果関係を説明するための反事実モデルを開発する方法，の三つがある．これら研究のサーベイに関してはDev et al. [87]によってまとめられている．既存の統計モデルを用いた手法は多くが2.4.2項で取り上げたHelpfulness Predictionの枠組みで行われ，認知バイアス情報などを表現した説明変数間の関連性の大きさの測定を行っている．しかし変数を取り除いた場合にどのような結果になったかといった反実仮想の推定の妥当性は担保されない点には留意が必要である．理由として，投票者の群衆行動はコンテンツの品質と相関する

[88] といったように認知バイアス情報に関する変数とコンテンツの質には相関が起きているためである．一方で因果推論を用いたアプローチ [87] ではランダム比較化試験を行わなくても各バイアスによる影響を除去することができるが，一つの変数の影響しか見られないといった欠点がある．これら先行研究の限界が，本章で提案するアプローチの動機となっている．

そこで本章では，複数の認知バイアス情報が存在するサイトにおいて，ユーザーのクリックデータからコンテンツの有用性をどのように推定するか，また，それらの情報がどのように相互作用し，ユーザーの投票行動に影響を及ぼすかを明らかにする手法を提案する．提案手法では従来の統計モデルを用いた試みのように直接コンテンツの有用度を予測するのではなく，認知バイアス情報によってどれだけ魅力的に見えていたかを統計モデルを用いて定量化し，それら影響を差し引くことでバイアスの影響を取り除くというアプローチを取る．

提案手法の流れとして，まずクラウドソーシングに基づく被験者を認知バイアス情報を与えた集団と与えなかった集団に分け，それら集団の投票行動の違いに着目する．続いて，それら集団の投票行動の違いを説明するためにグラフニューラルネットワーク (GNN) を採用し，対象のバイアスによってどれだけ投稿が魅力的に見えたのかを定量化する．続いて上記アプローチに基づき学習された GNN モデルを，大手 Q&A サイトである Stack Exchange のログデータに対して適用することで，認知バイアスを除去した形でのコンテンツの有用性の検証を行った．最後に，認知バイアス情報の相互関係を定量化し，学習したモデルの挙動についての解釈を行う．本研究は，オンラインプラットフォームにおける根本的な疑問である，コンテンツの有用性をどのようにして測定するかを明らかにする．

3.2 提案アプローチ

本研究では，ユーザーが複数の認知バイアス情報に晒されている際のユーザーの「いいね」などの投票形式のフィードバックから，バイアスによる影響を除いたコンテンツの有用性や品質を推定することを主な目的とする．ここでは QA サイトを例に挙げるが，レビューサイトの場合も同様である．

3.2.1 表記

ある質問 q に対して，回答文書 $\mathcal{D}_q = \{\mathbf{d}_1, \mathbf{d}_2, ..., \mathbf{d}_{N_q}\}$ が得られているとする．ここで N_q は質問 $\mathbf{q}$ について書かれた回答の数を表す．ユーザーが $\mathbf{d}_i \in \mathcal{D}_q$ に「役に立った」や「いいね」といったクリック c をすると，それはすなわちその回答が有用であることを意味し，下記式を用いて表現する

$$c_i = \begin{cases} 1 & \mathbf{d}_i \text{がクリックされた場合} \\ 0 & \text{それ以外} \end{cases} \tag{3.1}$$

提案手法ではそれぞれの文書 $\mathbf{d}_i$ は，固有の有用度 $h_i \in \mathbb{R}+$ を持つと仮定する ($\mathcal{H}_q = \{h_1, h_2, ..., h_{N_q}\}$)．ここでの h_i はテキストなどのコンテンツ情報からのみ得られる．逆に， $\mathcal{X}_q = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_{N_q}\}$ は，認知バイアス情報，例えば，ランキング位置，投稿者の評価，これまでの投票総数などを表す特徴ベクトルの集合と定義する．これらは 2.2 節で説明した Burghardt et al. [27] による評価行動に大きく影響を与える内在的な二つの要因と外来的三つの要因に由来する．

3.2.2 Competitive Voting Behavior Modeling

第 2 章で説明した多くの研究が示唆しているように，ユーザは投票時に認知ヒューリスティックを利用することが知られている．そこで，Q&A サイトにおけるコンテンツの価値と認知バイアスが人間の投票行動に影響を与える関係性を新たに以下のように定式化することを提案する:

$$\mathrm{P}(c_i = 1 \mid \mathbf{d}_i, \mathcal{D}_q, \mathcal{X}_q) = \frac{\exp(f(\mathbf{d}_i, \mathcal{X}_q))}{\sum_k \exp(f(\mathbf{d}_k, \mathcal{X}_q))} \tag{3.2}$$

$$f(\mathbf{d}_i, \mathcal{X}_q) = \alpha_\theta(\mathbf{d}_i, \mathcal{X}_q) \cdot h_i \tag{3.3}$$

ここでの $\mathrm{P}(c_i = 1 \mid \mathbf{d}_i, \mathcal{D}_q, \mathcal{X}_q)$ は $\mathcal{X}_q$ が与えられた際，$\mathbf{d}_i \in \mathcal{D}_q$ にユーザーが「いいね」を投票する確率を意味する．また認知バイアス情報 $\mathcal{X}_q$ によって何倍 $\mathbf{d}_i$ が有益に錯覚されるかをモデリングする関数 $\alpha_\theta(\mathbf{d}_i, \mathcal{X}_q)$ を考える．それゆえ $\alpha_\theta(\mathbf{d}_i, \mathcal{X}_q)$ が

1 より大きい際，$\mathbf{d}_i$ は本来よりも投票を得やすくなり，1 より小さい場合は本来より投票が得にくくなる．

直感的に $f(\mathbf{d}_i, \mathcal{X}_q)$ の出力は，ユーザーにとって観測される “見せかけの価値” を表しており，回答文書 $\mathcal{D}_q = \{\mathbf{d}_1, \mathbf{d}_2, ..., \mathbf{d}_{N_q}\}$ におけるそれら見せかけの価値をソフトマックス関数に入力することで，ユーザーのクリック率をモデリングしている．

3.2.3 バイアスの影響を取り除いた有用性の推定

コンテンツの価値 $\mathcal{H}_q$ を推定するために，本研究は二つのアプローチを提案する．一つ目は Result Randomization を用いる方法，二つ目は学習済みの α_θ を用いて実際の観測データからバイアスによる影響を割り引いて推定する方法である．

Result Randomization

Result Randomization においては，文書がランダムにシャッフルされ，スコアが表示されず，ユーザに表示される前にユーザーの評判が隠されていた状態でのクリックのログ情報に基づくデータセット $\mathcal{R}$ を収集する．質問 $\mathbf{q}$ に関するログのサブセットを $\mathcal{R}_q$ とする．h_i は下記式に基づき求められる：

$$
\begin{aligned}
\log(\mathbb{E}(c_i \mid \mathbf{d}_i, \mathcal{D}_q)) &= \sum_{\mathcal{X}_q \in \mathcal{R}_q} \log(\mathbb{E}(c_i \mid \mathbf{d}_i, \mathcal{D}_q, \mathcal{X}_q)) P(\mathcal{X}_q) \\
&= \sum_{\mathcal{X}_q \in \mathcal{R}_q} \log(\mathrm{P}(c_i = 1 \mid \mathbf{d}_i, \mathcal{D}_q, \mathcal{X}_q)) P(\mathcal{X}_q) \\
&= \sum_{\mathcal{X}_q \in \mathcal{R}_q} \log(\frac{\exp(f(\mathbf{d}_i, \mathcal{X}_q))}{\sum_k \exp(f(\mathbf{d}_k, \mathcal{X}_q))}) P(\mathcal{X}_q) \\
&= \sum_{\mathcal{X}_q \in \mathcal{R}_q} \alpha_\theta(\mathbf{d}_i, \mathcal{X}_q) \cdot h_i P(\mathcal{X}_q) \\
&\quad - \sum_{\mathcal{X}_q \in \mathcal{R}_q} \log(\sum_k \exp(\alpha_\theta(\mathbf{d}_k, \mathcal{X}_q) \cdot h_k)) P(\mathcal{X}_q) \\
&= h_i - C_q
\end{aligned}
\tag{3.4}
$$

これは，収集されたデータ $\mathcal{R}$ において，$\sum_{\mathcal{X}_q \in \mathcal{R}_q} \alpha_\theta(\mathbf{d}_i, \mathcal{X}_q) P(\mathcal{X}_q) = \mathbb{E}\{\alpha_\theta(\mathbf{d}_i, \mathcal{X}_q)\} = 1$ であり，第 2 項は特定の q については定数であるため，C_q と置くことができる．したがって，$h_i = \log(\mathbb{E}(c_i \mid \mathbf{d}_i, \mathcal{D}_q)) + C_q$ と導くことができる．

なお $\alpha_\theta \equiv 1$ という unbiased な状況では任意の実数 C_q に対して

$$\frac{\exp(h_i + C_q)}{\sum_k \exp(h_k + C_q)} = \frac{\exp(h_i)}{\sum_k \exp(h_k)} \tag{3.5}$$

が成り立つが，α_θ の出力が各 i について異なる値を取ると

$$\frac{\exp(\alpha_\theta(\mathbf{d}_i, \mathcal{X}_q) \cdot (h_i + C_q))}{\sum_k \exp(\alpha_\theta(\mathbf{d}_k, \mathcal{X}_q) \cdot (h_k + C_q))} \neq \frac{\exp(h_i)}{\sum_k \exp(h_k)} \tag{3.6}$$

となる．そのため実験を行う際には，事前に後述 α_θ の学習の際に事前に定めるハイパーパラメータである固定の C_q を定める必要がある．C_q に応じて何倍魅力的に見えているか (α_θ の出力) が決定されるため，事前にヒューリスティックに決める必要がある．なおクリックが未観測の $\mathbf{d}_i$ については $\log(\mathbb{E}(c_i \mid \mathbf{d}_i, \mathcal{D}_q))$ が計算できないため，十分小さな値を一律に加えた後 $h_i = 0$ になるよう C_q を足すことで調整を行う．そこで本研究の実装では十分小さな値として e^{-6}，それに伴い各 q について一律に $C_q = 6$ を用いた．

学習済み α_θ の活用

当然のことながら，Result Randomization に基づくデータの収集はコストが高い．そのため第二の方法は，認知バイアスによってどれだけ魅力的に見えているかを説明する学習済みの α_θ の出力を見せかけの価値から割り引くことで観測データから認知バイアス情報の影響を取り除いた価値を推定する方法が考えられる．そこで下記では α_θ の学習目的関数と，その後の h_i の推定方法について説明する．

まずは α_θ の目的関数について説明する．α_θ の学習段階には，式 3.4 で得られる ground-truth の h_i が必要となる．そしてログデータ $\{(\mathbf{q}^j, \mathbf{d}_i^j, h_i^j, \mathcal{X}_q^j, c_i^j)\}_{j=1}^M$ と組み合わせることで，θ を学習させる．

$$\theta^* = \arg\max_\theta \frac{1}{N_{q^j}} \sum_{j=1}^{M} \sum_{i=1}^{N_{q^j}} c_i^j \log(\mathrm{P}(c_i = 1 \mid \mathbf{d}_i, h_i, \mathcal{D}_q, \mathcal{X}_q, \theta)) \tag{3.7}$$

続いて未観測の h_i について式 3.7 にて得られた学習済みの α_{θ^*} を用いて推定を行う．ログデータ $\{(\mathbf{q}^j, \mathbf{d}_i^j, \mathcal{X}_q^j, c_i^j)\}_{j=1}^{M'}$ について，以下のように推定できる．

$$\hat{h}_i = \arg\max_{h_i} \frac{1}{N_{q^j}} \sum_{j=i}^{M'} \sum_{i=1}^{N_{q^j}} c_i^j \log(\mathrm{P}(c_i = 1 \mid \mathbf{d}_i, h_i, \mathcal{D}_q, \mathcal{X}_q, \theta^*)) \tag{3.8}$$

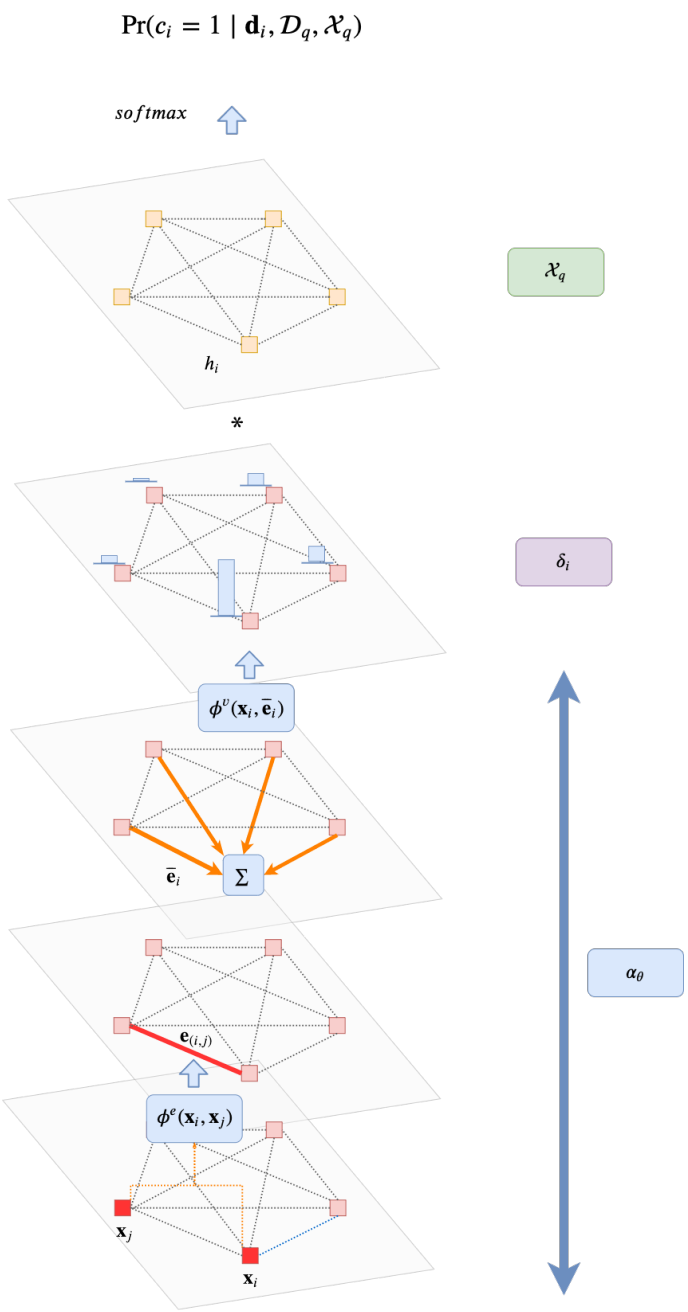


図 3.1: GNN モデルは共同学習された二つの部分 ϕ^e と ϕ^v から構成されている．GNN モデルの出力と h_i を乗算し，ソフトマックス関数を適用して確率 $\Pr(c_i = 1 \mid \mathbf{d}_i, \mathcal{D}_q, \mathcal{X}_q)$ を求める．

最適化には $\mathcal{H}_q$ をナイーブに測定した見せかけの有用度 $\log(\mathbb{E}(c_i \mid \mathbf{d}_i, \mathcal{D}_q)) + C_q$ を初期値とし，Adam [89] を最適化手法として，ミニバッチ勾配降下法を用いて学習を行う．

3.3 α_θ のモデリング

本節では，α_θ のネットワーク構造について詳述する．本論文では，$\mathcal{X}_q$ 間の相互作用を学習するために α_θ のモデリングにグラフニューラルネットワーク（GNN）を採用する．GNN はネットワーク構造情報のモデリングに優れていることが実証されており，多くの分野で最先端技術に貢献している．

3.3.1 概要

回答文書の認知バイアス情報 $\mathcal{X}_q$ をグラフ $G_q = (V_q, E_q)$ として表現する．ノードは各文書の認知バイアス情報 $\mathbf{x}_i$ が該当し，エッジは全てのノード同士繋がった完全グラフを想定する．一般的な Node Classification の枠組みにおいて，どの回答が次の投票を得るのかを予測する．

GNN の内部構造は，Edge model と Node model の二つの要素で構成されている．Edge model は，グラフ内のエッジで結ばれたノードの組 $\left(\mathbf{x}_i, \mathbf{x}_j \in \mathbb{R}^{L^v}\right)$ から，そのエッジのベクトル情報をメッセージベクトルとして写像する．これらのメッセージベクトルは，受信ノードごとに要素和をとり，それら和が Node model に入力される．Node model は，受信ノードと合計されたメッセージベクトルを受け取り，$\mathcal{X}_q$ によって $\mathbf{d}_i$ が何倍魅力的に見えさせているかのスカラー値 δ_i を計算する．最後に，全ノードにソフトマックス関数を適用して，次の投票がどの投稿に行われるかの確率を求める．

3.3.2 GNN の構造

Edge model: ϕ^e

あるノードから別のノードへのメッセージを計算するために，Edge model ($\phi^e : \mathbb{R}^{L^v} \times \mathbb{R}^{L^v} \to \mathbb{R}^{L^e}$) を作成する． ここで，$L^e$ はメッセージ特徴量の次元数である．本実験の実装では，$L^e = 1$ とする．本実験の実装では，20 個のユニットを持つ 3

つの隠れ層を持つ多層パーセプトロン，活性化関数では ReLU を用いる．全ての辺 (i, j) に対して下記式で表現される．

$$\mathbf{e}_{(i,j)} = \phi^e\left(\mathbf{x}_i, \mathbf{x}_j\right) \tag{3.9}$$

Aggregation

これらのメッセージは，各受信ノードの要素毎の合計を経て，ノード毎に加算プーリングされる ($\overline{\mathbf{e}}_i \in \mathbb{R}^{L^e}$)．この処理は下記式で表現される

$$\overline{\mathbf{e}}_i = \sum_{(i \neq j)} \mathbf{e}_{(i,j)} \tag{3.10}$$

Node Model: ϕ^v

ここでは，入力ノードと対応するメッセージベクトルを入力し，ノード毎に出力 y_i を行う Node model ($\phi^v : \mathbb{R}^{L^v} \times \mathbb{R}^{L^e} \to \mathbb{R}$) を定める．

$$\delta_i = \phi^v\left(\mathbf{x}_i, \overline{\mathbf{e}}_i\right) \tag{3.11}$$

本研究の実装では，ϕ^e 同様，20 個のユニットを持つ 3 つの隠れ層を持つ多層パーセプトロン，活性化関数では ReLU を用いる．

最終的に下記式に基づいて各回答 d_i が次の投票を得る確率 y_i を得る．

$$\begin{aligned} y_i &= \mathrm{P}(c_i = 1 \mid \mathbf{d}_i, \mathcal{D}_q, \mathcal{X}_q) \\ &= \frac{\exp(\delta_i)}{\sum_{j=1}^{N_q} \exp(\delta_j)} \end{aligned} \tag{3.12}$$

3.3.3 損失関数

最適化を行う過程で ϕ^v と ϕ^e はミニバッチ勾配降下法を用いて同時に学習される．最適化手法には Adam [89] を用い，ログデータ $\{(\mathbf{q}^j, \mathbf{d}_i^j, \mathcal{X}_q^j, c_i^j)\}_{j=1}^{M'}$ について最終的な損失関数は:

$$\mathcal{L} = -\frac{1}{N_{q^j}} \sum_{j=1}^{M'} \sum_{i=1}^{N_{q^j}} c_i^j \log(y_i^j) + \lambda \frac{1}{N_{q^j}} \sum_{j=1}^{M'} \sum_{i=1}^{N_{q^j}} ||\delta_i^j - 1||^2 \tag{3.13}$$

上記式にて λ は α_θ の出力が 1 に近づくよう学習することを促す L2 正則化項における重みを表す．

3.4 データセット

実験には 3 つのデータセット D1，D2，D3 を使用した．

3.4.1 Amazon Mechanical Turk を用いたデータ (D1, D2)

実験設定

3.2.3 項で説明したように，本研究では ground-truth の h_i を計算するために Result Randomization が施されたデータセットが必要になる．D1，D2 と呼称するデータセットは，Burghardt et al. [90] が行ったランダム実験に由来するデータを用いる．Burghardt et al. は Q&A プラットフォームである Stack Exchange を基にした実験環境を作成し，Amazon Mechanical Turk (MT) を利用してユーザーの投票データを収集した．

同実験では登録された被験者に一つの質問に対して，最も役に立ったと思った回答を選択させる (図 3.2 参照)．質問は英語学習者（ELL）フォーラムから選択された．それぞれの質問には，ELL のウェブサイトに掲載されている 8 つの最も古い回答が掲載されており，各被験者には合計 10 個の質問について回答を行う．被験者の 9 割は IP アドレスを元に，英語が一般的に話されているアメリカ，カナダ，イギリスの出身者がサンプリングされた．

被験者は回答の選択を完了するために最大 1 時間まで与えられ，各被験者は課題を完了するために 0.50 ドル支払われる．被験者が答えを選ぶと，ボタンをクリックして次の質問に進む．

D1

一つ目のデータセット D1 は認知バイアス情報を用いさせない形で収集されており，被験者は質問につきにランダムな順序で表示される回答について最も有用だと感じた回答について投票を行う．D1 の実験条件下ではテキスト以外の情報は表示されない．ログは全部で 1787 件収集された．

[1] https://ell.stackexchange.com/questions/30/what-is-the-difference-between-nope-and-no

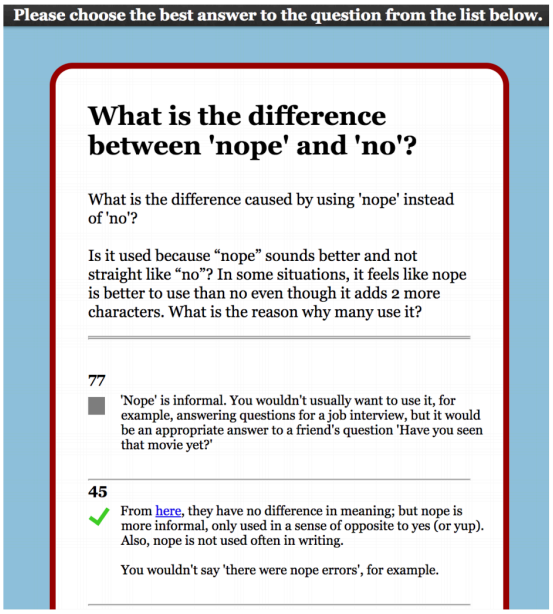


図 3.2: Stack Exchange からの質問を表示した実験で使用した画面の例[1]．例では投票数が表示されており，社会的影響力バイアスを与えた条件（Result Randomization では，投票数は非表示）を表す．回答が選択された後，ユーザーは「承諾」ボタンをクリックして次の質問に進む．Burghardt et al. の図 1 より [90]．

D2

二つ目のデータセット D2 は，回答の横にスコアが表示されており，回答はそのスコア順に並べられている．各回答の横に表示されているスコアは,「過去にこの回答を選んだ人の数を表している」と被験者に伝えられる．しかし，実際には「スコア」は 0 から 25 までのランダムに生成された数字であり，これらの数字は，各被験者ごとに独立して生成されていて内容とは一切関係ない．これは意図的に社会的影響力バイアスを発生させるためである．ログは全部で 1668 件収集された．

3.4.2 実観測データの利用 (D3)

三つ目のデータセット D3 は Stack Exchange の観測ログを用いる．Stack Exchange の観測データは CC BY-SA 3.0 ライセンスの下で公開されている[2]．これらのデータセットには，ユーザーの投票行動のログやタイムスタンプが完全に記録されており，投票行動時の $\mathcal{X}_q$ を再現することができる．用いる特徴量は 3.4.4 項にて解説を行う．ログは全部で 738 件収集された．

[2] `https://archive.org/details/stackexchange`

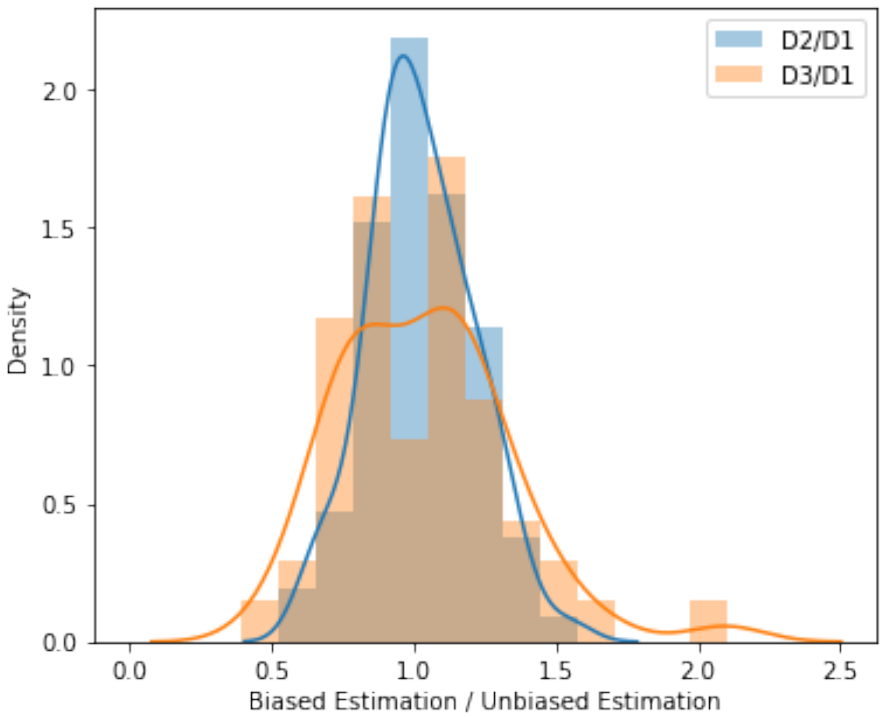


図 3.3: Ground-truth の有用度と D2，D3 についてナイーブに推定した有用度との比に基づくヒストグラムとカーネル密度推定を用いた結果.

3.4.3 ナイーブな推定における有用度の偏り

まずナイーブに推定した有用度 $\log(\mathbb{E}(c_i \mid \mathbf{d}_i, \mathcal{D}_q))) + C_q$ がデータセットごとにどれだけ異なっているかを確認する．図 3.3 に，D2 または D3 からナイーブに計算された有用度が ground-truth である D1 から推定された有用度と比較して何倍違うかについてヒストグラム，またカーネル密度推定を用いた結果を示した．カーネル密度推定のバンド幅については Scott のルールを用いて決定した [91]．これらは ground-truth の h_i が何倍歪められているかを表しており，これら違いを α_θ を用いて $\mathcal{X}_q$ から予測することが学習の目的であることに留意する．言い換えると図 3.3 において 1 付近に集まっている回答は，バイアスのかかった推定値 (D2， D3) と偏っていない推定値 (D1) がほぼ同じ値を持っていることを表す.

まずそれぞれの分布について Maurus et al. [92] の手法を用い，単峰性の検証を行った．分布は単峰であるという帰無仮説は，D2/D1 で P 値が 0.99, D3/D1 で P 値が 0.30 となり，それぞれ有意水準 5% で棄却されなかった．続いてそれぞれの分布の分散について検証を行ったところ，D2/D1 で 3.4×10^{-2}，D3/D1 で 8.7×10^{-2} となった．ルビーン検定 [93] を行った結果 P 値が 0.001 となり分布間の分散が有意に異なることが示された.

3.4.4 認知バイアス情報

以下では，本稿で扱う $\mathcal{X}_q$ を構成する特徴量について説明を行う．第 2 章で示した図 2.1 は，Stack Exchange のサンプルページで，本研究で考慮した特徴量についての注釈を加えている．

他ユーザーによる投票の合計 (D2，D3)

ここまでの各回答における他ユーザーによる投票の合計投票を $\mathcal{X}_q$ の特徴量として用いる．これを利用した動機としては，総投票数の多いコンテンツほど投票を受けやすいという社会的影響力バイアスを学習するためである．D2 の実験では，コンテンツの質に関係なく，付与される 0 から 25 までの数字を使用する．D3 の実験では，ユーザーが新たに投票を行った際に，それまで既に獲得した投票の数を使用する．本実験では α_θ に入力するための前処理として Box-Cox power transformation [94] を施した後，標準化を行った g_λ を用いる．

$$g_\lambda(x) = \begin{cases} \left(x^\lambda - 1\right)/\lambda & \text{if } \lambda \neq 0 \\ \log(x) & \text{if } \lambda = 0 \end{cases} \tag{3.14}$$

本処理では，変数の変換後の分布が最も正規分布に近づくように，λ の値は決定される．

コンテンツの位置 (D2,D3)

$\mathcal{X}_q$ の特徴として，回答の位置を用いる．これを利用した動機としては，位置が高い方がより多くの投票を受けやすい (位置バイアス) を学習するためである．同特徴量も α_θ への入力が 0，標準偏差が 1 になるように標準化を行う．

投稿者の事前評価 (D3)

投稿者が過去の貢献に基づき獲得したメダルの数 (金，銀，銅) を $\mathcal{X}_q$ の特徴量として用いる．これを利用した動機としては，より高い評価を得ている投稿者のコンテンツほど，より多くの評価を得られる (評判バイアス) を学習させるためである．同特徴量も α_θ への入力前に Box-Cox power transformation を行ったのち，標準化を行う．なお本特徴量は D3 データセットのみ利用可能であることに留意する．

3.5 実験

ここでは，以下の研究課題に答えることを目的として実験を行う．

RQ1.1 提案モデリングはユーザーの投票行動を正しくモデル化できているか？

RQ1.2 提案手法による有用性の推定はバイアスの影響を取り除けているか？

RQ1.3 α_θ はどのような関数を学習したのか？

RQ1.4 見せかけの価値に左右されずに有用な回答を高評価するユーザーはどのようなユーザーか?

3.5.1 投票行動のモデリングに関する性能比較 (RQ1.1)

まずコンテンツの有用度と認知バイアスが人間の投票行動に影響を与える関係性のモデル化 (CVBM) の妥当性を検証するために，投票行動の予測という観点から本手法の有効性を示す．

評価方法

D1 から計算した h_i を ground-truth として用いて，D2 と D3 について投票行動の予測性能を評価する．実験設定として D2,D3 を学習/テスト (9 問，1 問) に分割し，10-Fold Cross Validation を行う．テストセットの各回答が投票される予測の確信度に応じてランク付けを行う．

ランク付けを評価するための二つの代表的な指標である HR (Hit Ratio) と NDCG (Normalized Discounted Cumulative Gain) [95] を用いて性能を検証する．以下それぞれの指標について説明を行う．

$$HR@k = \begin{cases} 1, \text{ if } c_i \in R_k \\ 0, \quad \text{otherwise} \end{cases} \tag{3.15}$$

R_k は確信度に基づく上位 k 件のリストである．直感的に HR は上位 k 番目までに正解であるクリック c_i が存在するかを表す．

$$NDCG@n = \frac{DCG@n}{IDCG@n} \tag{3.16}$$

$$DCG@k = \sum_{j=1}^{k} \frac{2^{rel_j} - 1}{\log_2(j+1)} \tag{3.17}$$

ここでの rel_j は関連性スコアで，本検証では j 番目の予測が c_i であれば 1，そうでなければ 0 とする．NDCG@k は，完全順位を測定する理想的な iDCG@k を用いて DCG@k を正規化することで 0 から 1 の間で求められる．直感的に NDCG は実際のクリック c_i がどれだけその中でも上位にあるかを表す．

ベースラインモデルとの比較

GNN ベースの提案手法は，回答ごとの認知バイアス情報の依存関係をモデル化することに重きを置いているため，以下のように GNN モデルの構造のそれぞれの構成要素に着目して性能比較を行う．

- **Naive**： 推定には α_θ を用いずに ground-truth の h_i が大きい順にソートして予測リストを作成する
- **Node Only (Linear)**: GNN における Edge Model ϕ^e を用いずに Node Model ϕ^v のみを用いてユーザーの投票行動を予測する．ここで ϕ^v は 1 層の線形回帰を用いて実装する．
- **Node Only (Deep)**： Node Only (Linear) と同じく ϕ^v のみを用いて予測を行う．ϕ^v には 3 層の隠れ層を持つ MLP を用いて実装する．
- **GNN**: 3.3 章で説明した方法で，同時に学学習させた ϕ^v と ϕ^e を元に投票行動を予測する．

図 3.4 (a)(b) は D2 での結果を示し，(c)(d) は D3 での結果を示す．D2 を用いた実験では，GNN, Node Only (Linear) と Node Only (Deep) がほぼ同じ性能を発揮し，Naive 推定よりも良い結果が得られていることがわかる．この結果は，ユーザーの投票行動を予測するためには，認知バイアス情報を取り入れることが有効であることを示している．一方，D2 における実験では，α_θ の構造を変化させても違いは見られなかった．この結果は，認知バイアスの特徴量が 2 個と少なく，複雑な非線形モデリングが必要なかったためと考えられる．

D3 を用いた実験では，GNN と Node Only (Deep) の方が Node Only (Linear) よりもわずかに優れているが，有意な差は見られなかった．この結果は，D2 実験でも同じように示された．興味深い発見としては，D2 実験よりも D3 実験の方が Naive 推定からの性能の改善幅が大きいことが挙げられる．これには二つの理由が考えられる．一つ目に MT を用いた実験環境 (D2) に比べて，実際のサービス運用環境 (D3) では認知バイアスの影響が強い傾向があることが考えられる．二つ目に D3 に含ま

れない他の認知バイアス (メダル数など) がユーザーの予測に支配的であることが考えられる．これら特徴の重要性については，後の 3.5.3 項で検証する．

3.5.2 バイアスの影響を取り除いた有用性の推定 (RQ1.2)

ここでは，バイアスの影響を取り除いた有用性の推定に関してモデルごとの性能検証を行う．

評価方法

3.5.1 項と同じ 10-Fold Cross Validation を行い，事前に訓練した α_θ を用いて，3.2.3 項で説明した提案手法に基づき，テストセットにおける h_i を推定する．

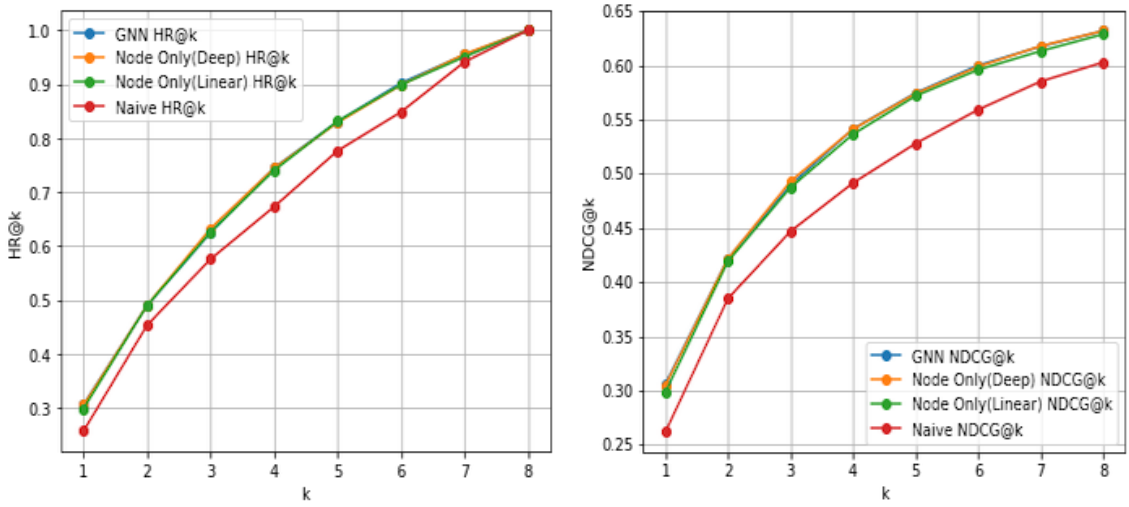
性能は二つの指標，RMSE (Root Mean Squared Error) と KL Divergence を用いて性能を検証する．RMSE は推定された h_i と ground-truth との差を測定する．KL Divergence の計算はソフトマックス関数を用いて h_i を投票確率に変換し，$p_i = P(c_i = 1 \mid \mathbf{d}_i, \mathcal{D}_q)$ を得て，$\mathcal{P}'_q = \{p'_1, p'_2, ..., p'_{N_q}\}$ と正解の分布 $\mathcal{P}_q = \{p_1, p_2, ..., p_{N_q}\}$ との間について求める．

$$D_{KL}(\mathcal{P}_q \| \mathcal{P}'_q) = \sum_{i=1}^{N_q} p_i \cdot \log_2 \frac{p_i}{p'_i} \tag{3.18}$$

ベースラインモデルとの比較

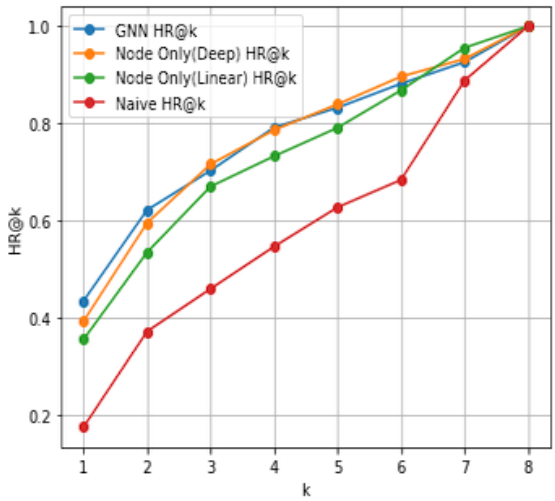
本項では以下ベースラインモデルとの比較を行う．

- **Naive** : $\log(\mathbb{E}(c_i \mid \mathbf{d}_i, \mathcal{D}_q))) + C_q$ を h_i として用いる．学習済みの α_θ を用いた補正を行わない．
- **Node Only (Linear)** : 3.5.1 項で説明した GNN のうち，ϕ^e を使わずに学習済みの ϕ^v のみを用いて h_i を得る． ϕ^v は線形回帰モデルで実装を行う．
- **Node Only (Deep)** : 3.5.1 項で説明した GNN のうち，ϕ^e を使わずに学習済みの ϕ^v を用いて h_i を得る．ϕ^v は隠れ層 3 層の MLP を用いて実装を行う．
- **GNN** : 3.5.1 項で説明した ϕ^v と ϕ^e を同時に学習させた GNN を用いて h_i を得る .

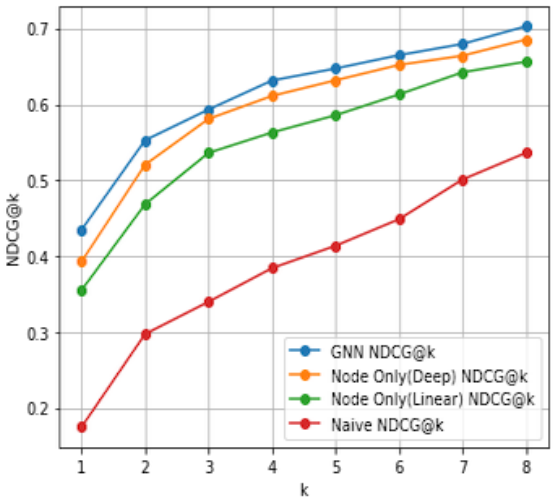


(a) D2 — HR@k

(b) D2 — NDCG@k

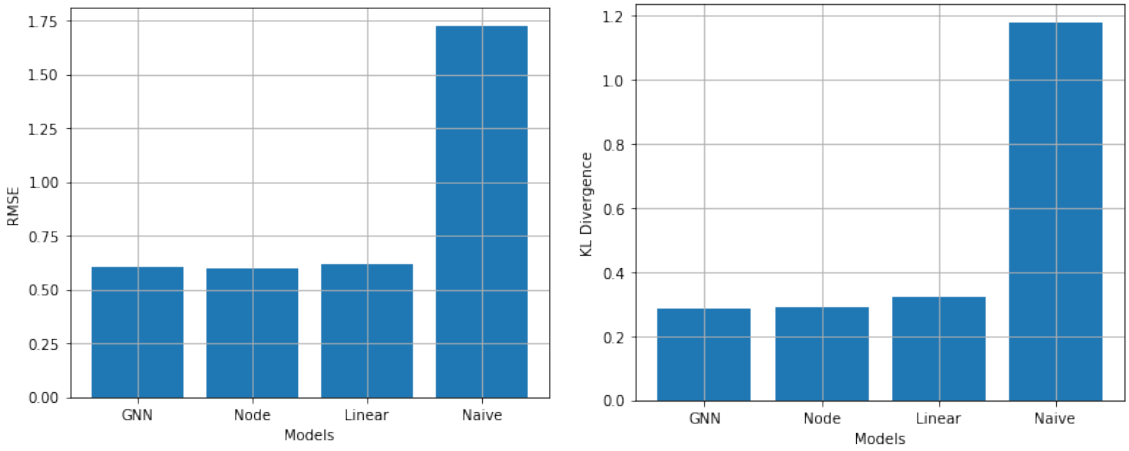


(c) D3 — HR@k



(d) D3 — NDCG@k

図 3.4: D2 で k が 1 から 8 まで，D3 で N の方法分類が 2 から 8 までの場合のユーザーの投票行動予測の評価．(a), (b) は GNN, Node Only (Linear) と Node Only (Deep) の違いがなく，ナイーブ推定よりも優れていることを示している．



(a) D3 — RMSE

(b) D3 — KL Divergence

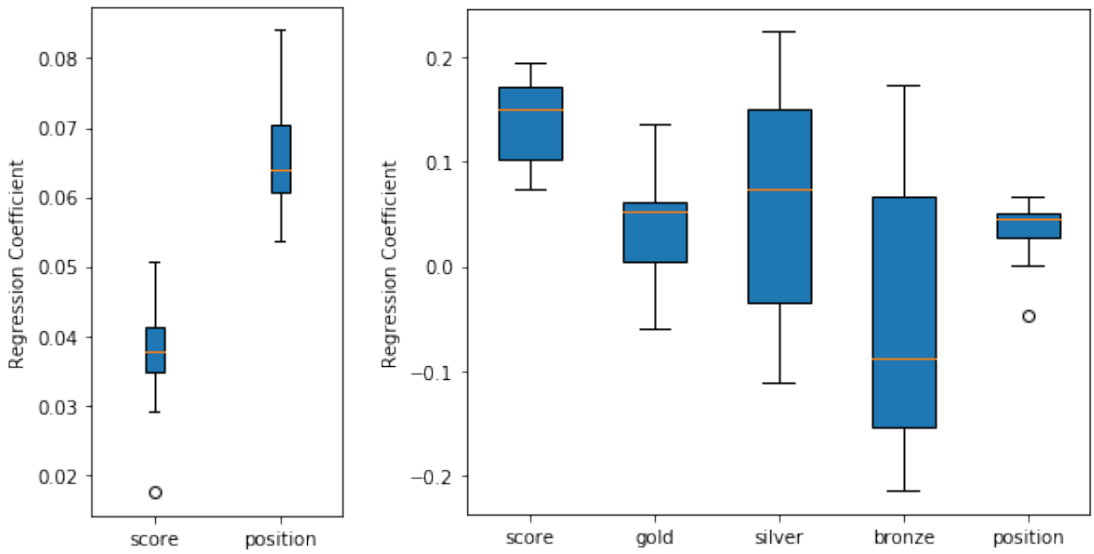
図 3.5: ベースラインモデル間のバイアスの影響を取り除いた有用性推定の評価

なお，D2 では，ランキングとスコアがランダムに割り当てられているため，観測される期待値（ナイーブ推定）の計算は D1 で推定した期待値とほぼ同じであることに留意する．一方ユーザーが継続的にバイアスにさらされている D3 では，ナイーブな推定が偏ってしまう．そこで提案手法の妥当性の検証は D3 データセットを用いて行う．

図 3.5(a) は RMSE，(b) は KL Divergence での結果を示す．いずれの結果からも推定した有用性 h_i がバイアスの影響を取り除けていることがわかる．また，すべてのケースにおいて Naive 推定は ground-truth の有用度に比べて大きくバイアスがかかっていることがわかり，認知バイアス情報を取り入れることが有用性予測に必要であることがわかる．また前項同様，GNN，Node Only (Linar)，Node Only (Deep) はほぼ同じ性能を示しており，単純な線形モデルで複数の認知バイアスによる影響は分析可能であることがわかった．

3.5.3 モデルの解釈 (RQ1.3)

3.5.1，3.5.2 項の二つの実験により，単純な線形回帰モデルが深層学習モデルとほぼ同等の性能を示すことがわかった．モデルが学習した知識を解釈するために，線形回帰モデルの重みに着目する．図 3.6(a)(b) は，D2 と D3 実験で 10-Fold Cross Validation から得られた 10 モデルにおける回帰パラメーターのボックスプロットを示す．また表 3.1 にはそれらの統計量を示す．説明変数は全て正規化されており，重みが大きな値をとるほど出力に寄与していることを示す．以下に同 10 個のモデルの重みの平均を用いた関数を示す．



(a) D2　　(b) D3

図 3.6: 10-Fold Cross Validation から得られた 10 個の α_θ(Node Only (Linear)) の回帰係数のボックスプロット．より高い値を示す属性は，コンテンツがより投票を集めやすくなることに貢献していることを示す．

D2 と D3 の score と position の係数について Welch の t 検定を用いて平均値の差を比較した結果，有意水準 1% で score については有意な差が ($p \approx 0.0003$)，position については有意な差が確認されなかった ($p \approx 0.33$)．これら結果より D3 のような実際のサービス上では，ユーザーは score といった社会的影響力バイアスの影響をより受けていることがわかった．もう一つの知見として，D3 実験では，他社スコアの集計 (score) が最終アウトプットと最も強く関連していたが，位置 (position) による影響は予想外に低くなった．これは，D3 の収集が行われる実環境下では位置がスコアでソートされているために，二つの変数が相関しすぎていること，または訓練サンプルが限られていたことが原因と考えられる．

3.5.4　個人単位でのバイアスによる影響度と属性の関連性 (RQ1.4)

本項では，各ユーザーがどれだけ認知バイアスを用いて投票行動を行っていたのかに着目し，ユーザーの回答の傾向と個人の属性との関連性について検討する．具体的には「見せかけの有用度は高かったが，真の有用度は低かったコンテンツを高評価した人 (公平度の低いユーザー)」と「見せかけの有用度は低かったが，真の有用度は高かったコンテンツを高評価した人 (公平度の高いユーザー)」について傾向を調べる．

表 3.1: 10-Fold Cross Validation から得られた 10 個の α_θ(Node Only (Linear)) の回帰係数に関する統計量．より高い値を示す属性は，コンテンツがより投票を集めやすくなることに貢献していることを示す．

データセット	特徴量	平均	標準偏差	95%信頼区間	
				上限	下限
D2	score	0.037	0.009	0.031	0.044
	position	0.066	0.009	0.059	0.072
	切片	0.990	0.003	0.988	0.992
D3	score	0.150	0.064	0.105	0.196
	gold	0.045	0.059	0.002	0.087
	silver	0.070	0.117	-0.014	0.154
	bronze	-0.055	0.147	-0.161	0.050
	position	0.026	0.034	0.002	0.050
	切片	0.260	0.177	0.134	0.387

データの取得

本項の実験を行う上では，どのようなユーザーがどの回答を高評価したかが必要になる．しかし前項の実験で用いた Stack Exchange のログデータにはどの投稿に誰が高評価したかといった情報は保存されていなかった．そこで代替として，質問投稿者がどの回答をベストアンサーとして選んだかの情報は取得可能であったため，今回はこちらを回答傾向と個人の属性を紐づける上で用いる．

また前項の実験では質問数が 10 問と数が限られており，上記実験を行うには少ないと判断した．そこで本項では 3.5.3 項で得られた学習済みの α_θ を用いて，前項で扱った質問以外に，同一の英語学習者 (ELL) フォーラム内に投稿された質問 39542 件のうち，質問投稿者が 5 択以上の中からベストアンサーを選択した 448 件について，それらに投稿された回答それぞれの有用度 h_i を擬似的に作成した．

ベストアンサー選択の傾向

図 3.7 に質問投稿者の選んだベストアンサーについて，その他投稿されていた回答との有用度との相対関係を表す散布図を示す．横軸にナイーブに推定した有用度における NDCG@1 (Naive NDCG@1)，縦軸に学習済みの α_θ を用いて推定した h_i を元に計算した NDCG@1 (Unbiased NDCG@1) を表す．NDCG@1 は選択した回答の有用

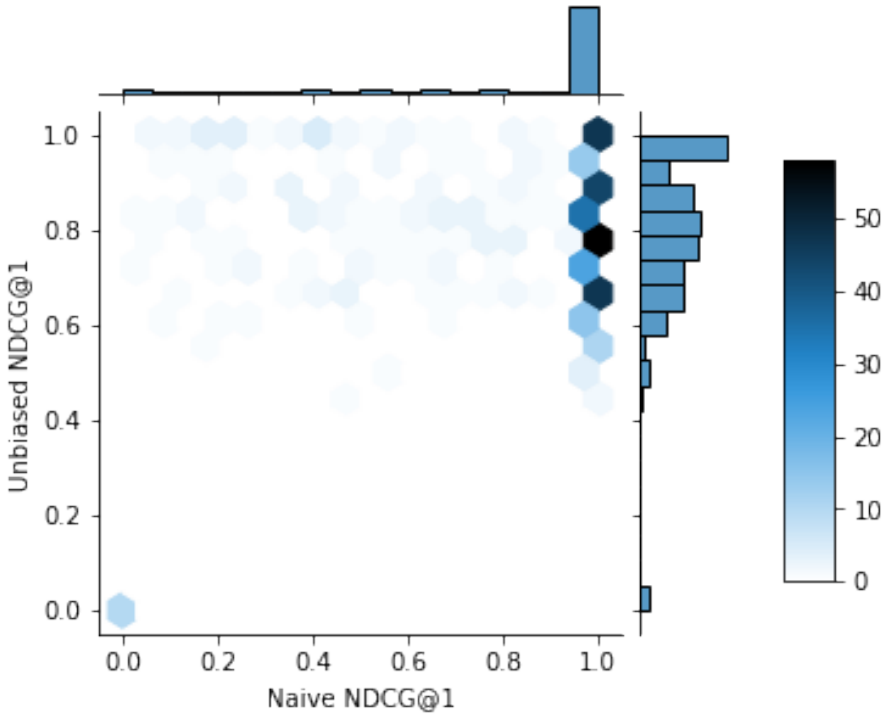


図 3.7: 質問投稿者の選んだベストアンサーについて，その他投稿されていた回答との有用度との相対関係を表す散布図

度を，候補となる回答における有用度の最大値で割った値を示す．つまり NDCG@1 が 1 の場合，有用度が最大の回答を選んだことを表す．

図 3.7 の直感的解釈として，右上のユーザーは，見せかけの有用度も真の有用度も最も高い回答を選択したことを表す．右下のユーザーは見せかけの有用度が高い回答を選択しているが，実際の真の有用度は高くなかった回答を選択したことを表す．同様に左上のユーザーは見せかけの有用度が低いが，真の有用度は高い回答を選択したしたことを表す．最後に左下のユーザーは見せかけの有用度も真の有用度も低い回答を選択したことを表す．

結果として，Naive NDCG@1 が 1 付近，Unbiased NDCG@1 が 0.5 から 1 のユーザーが多く確認される．これら結果より，多くのユーザーが見せかけの価値が高い回答，つまり多くの人が高評価している回答をベストアンサーとして選択していることがわかる．しかし，実際にはより有用度の高い回答が存在しており，選べていない場合がある．逆に見せかけの有用度は最大ではないが，真の有用度が高い回答が最大の回答を選択できているユーザーは，比較的少ないことがわかる．

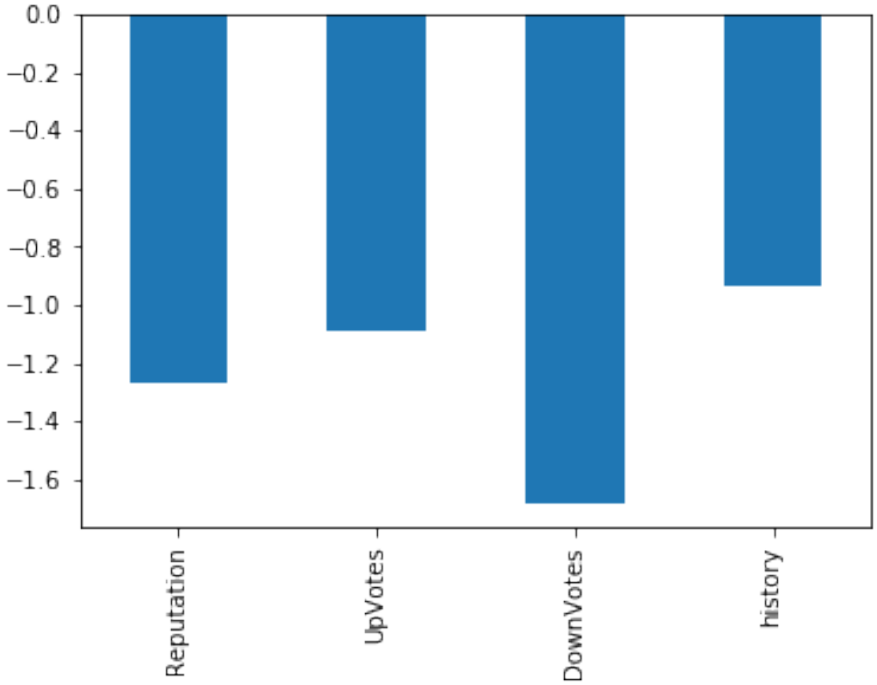


図 3.8: 横軸にはそれぞれの属性，縦軸には公平度の高い集団と公平度の低い集団の各属性の集団ごとの平均に関する増減を対数オッズ比を用いて表す．

属性との比較

続いてどのようなユーザーが，見せかけの価値に左右されずに有用な回答を高評価したのかについて分析を行う．そこで図 3.7 において左上と右下のユーザーについて着目し，それぞれの属性を比較する．

分析の流れとして，Unbiased NDCG@1 と Naive NDCG@1 の差が大きいユーザー上位 15% (左上 : 公平度の高いユーザー) と，小さいユーザー下位 15% (右下 : 公平度の低いユーザー) の二つの集団に分割する[3]．そこで二つの集団を質問投稿者の「評判スコア (Reputaion)」と「高評価 (Upvotes)」,「低評価 (Downvotes)」,「これまでに投稿した質問の数 (history)」といった属性に基づき検証を行う．

結果を図 3.8 に示す．横軸にはそれぞれの属性，縦軸には公平度の高い集団と公平度の低い集団の各属性の集団ごとの平均に関する増減を対数オッズ比を用いて表す．結果として公平度の高いユーザーは全ての属性において負の値を示した．つまり公平度の高いユーザーは低いユーザーに比べてそれぞれの属性において値が小さくなったことを表す．DownVotes が一番大きな負の値を示したことから公平度が高いユーザーは低評価が付けられにくい傾向にあることが分かったが，その他の指標

[3]閾値の変化に基づく結果の堅牢さの調査を行なったが，結果の傾向は同じであることを確認した．

についても低くなることが観測された．これらことより，公平度の高いユーザーだからといって他ユーザーから必ずしも認められているわけではないことが確認された．

今後はこれら回答の価値に基づく回答傾向から公平度または，エキスパートの発見を行うといったユーザーのスコアリングへの応用先が考えられる．

3.6 まとめ

本章では，オンラインプラットフォームにおけるコンテンツの有用度を，認知バイアス情報によってどれだけ魅力的に見えていたかを回帰モデルを用いて定量化し，それら影響を差し引くことでバイアスの影響を取り除いた形で推定するといった学習フレームワークの提案を行った．

本提案手法の主な貢献は以下の通りである．

- 本研究では，ユーザーの投票行動は周囲のコンテンツの有用度に影響を受けるというアイディアのもと，コンテンツの価値と認知バイアスが人間の投票行動に影響を与える関係性のモデリングを新たに定式化 (Competitive Voting Behavior Modeling) した．
- CVBM の条件下におけるバイアスの影響を取り除いた上でのコンテンツの有用度の推定手法を提案した．
- 学習済みのモデルから，認知バイアス情報によってコンテンツが何倍有益に見えているかを，またそれぞれの変数の関連性について定量化を行い，Stack Exchange のデータセットを用いて実験的に有用性を示した．
- ユーザーが投票行動を行う際，複数の認知バイアスの影響は単純な線形モデルで表現可能であることがわかった．
- 本提案手法によって推定された有用性の高い回答を，見せかけの価値に左右されず高評価するユーザーは必ずしも周囲からの評価が高くないことが得られた．

3.7 今後の方向性

今後の方向性として三つ取り上げる．

一つ目は提案手法の汎用性を確認するために他ドメインへの転用が考えられる．提案手法で用いた Q&A サイト以外への応用事例として，1) 認知バイアスが及ぼすレビューサイトなどの星といった明示的なフィードバックへの転用，2) 既存の検索エンジンの最適化を，位置バイアスだけでなく，URL のドメインやメディア名といった評判バイアスになるうる情報の影響を加味することで改善を行う試み，3)Twitter などにおける情報拡散を，コンテンツや投稿者のプロフィール，既存のシェア数などが情報拡散に与える影響分析，などが考えられる．

二つ目は，本章で扱ったバイアスによる影響の受けやすさを個人単位に拡張する方向性が考えられる．本研究では仮定として，一元的に有用度やバイアスの影響の受けやすさをモデリングしたが，認知バイアスによる影響の受けやすさは個人によって異なることが考えられる．そこで個人単位でそれらバイアスの影響度の推定を行うことは，よりユーザーにとって有用な情報の提供に繋がると考えられる．

三つ目に本研究で得られた有用度 h_i を元に，コンテンツ情報などから有用度を予測する Helpfulness Prediction モデルの構築が今後応用先として考えられる．これまでの Helpfulness Prediction はバイアスの影響を明示的に取り除いていなかったため，人間のアノテーションと異なるケースが多数報告されていた．そこで本研究で得られた有用度 h_i をコンテンツ情報だけから予測するモデルを構築することで，どのような文章の特徴が有用と人間に判断されやすいのかが明らかになると考えられる．

第4章　レビューの信頼性に基づく集合知の定量的評価と信頼性の解釈性の検討

4.1　はじめに

近年，不正に操作されたスパムレビューが多く横行しており社会問題と化している [96]．商品を使いもせずに意図的に評価を上げるために最上位の星や高評価のコメントをつけるなどの「やらせレビュー」，手当たり次第すべての商品に低評価を書き込む「アンチレビュー」，説明書きの文章をコピペしただけのものや翻訳ソフトにかけただけのような不自然な日本語のレビューなど数多く存在する．それらレビューによって消費者はもっと良いものが買えたはずという機会損失と，出品者は不当なアンチレビューによる，もっと売れたはずという機会損失が両者に生じている．Mukherjee et al. [58] は EC サイト (e.g. Amazon, Yelp, Alibaba) における悪質ユーザーの動機について分析し，多くの場合製品を宣伝する目的や消費者を誤解させるために偽のレビューを書くことが多いと報告している．

第 3 章では，集合知の前提となる意見の独立性について，認知バイアス情報の影響を受けたクリックデータから推定を試みた．しかし公平に評価を行っていないスパムレビューが存在する点において，バイアスの影響を取り除いたとしても集合知に組み込んで良いのかという疑問が生じる．特に平均などの既存集約方法は，レビュー履歴数が少ない場合に外れ値の影響を受けやすい．また第 2 章 図 2.3 が示すように，基本的にほとんどのアイテムがレビュー履歴数は得られていない点において，懸念はほとんどのアイテムについて当てはまる．そこで本章では第 1 章 **Research Question 2：　ユーザーのレビュー投票行動から不良ユーザー・レビューを発見し，信頼度に基づいた意見の集約によって有識者に近い集合知を獲得できるか？**の元検証を行う．

第 2 章で紹介したように，近年では機械学習を用いた低品質なレビューを自動で取り除く試みが行われている．本研究では，2018 年 2 月時点で不良レビュワー検出を，自己教師あり学習の枠組みで行った研究で state-of-the-art の性能を示した，Kumar

et al. [28] による REV2 に着目した．同手法は評価者と被評価者と評価に基づく，評価ネットワークのリンク構造に着目し，各評価の「信頼性」，評価者の「公平度」，被評価者の「モノの価値」をスコアリングし，悪質ユーザーの発見に応用される．同手法は事前にラベル付けされた悪質ユーザーの検知性能を評価することで有効性を検証しており，インドの最大手 EC サイト Flipkart では既に実運用されている．

しかし，悪質ユーザーの検知性能が示された一方，本章の主眼であるレビューの信頼性に基づく「モノの価値」が有識者による感覚と一致しているかや，何故悪質と判断されてしまったかというモデルの説明可能性の観点での検証は行われていない．これらの観点による分析は更なる実運用を想定した際に重要であると言える．レビューの質が高いプラットフォームの構築は，消費者の購買体験の向上，さらに出品者間での競争を活性化させ，潜在的に利益向上への貢献が考えられる．

以上を受け，本章では下記の検証を行った．

- REV2 を用いて算出したレビューの信頼度に基づく集合知が平均などのベースライン手法と比較して，より有識者に近い形で得られていることを実験的に示した．評価には東京都内のラーメン店に関するレビューを独自に収集し，手法毎に有識者が作成した評価の高い店舗との比較を行った．
- 信頼度に基づく集約のスコアが元々の平均といった集計方法から大きく異なる店舗について調査を行った．その結果，信頼度に基づく集約のスコアと単純平均スコアの差が正の値を持つ店舗は，信頼性の高い高評価と信頼性の低い低評価 (アンチレビュー) が多かったことが示された．反対に負の値を持つ店舗は，信頼性の低い高評価 (やらせレビュー) と信頼性の高い低評価が多かったことが示された．
- REV2 により，信頼性が低いレビューを行うレビュワーについて分析を行ったところ，「周りの人との意見のずれが大きい」，「高評価か低評価ばかりを付けやすい」といった傾向が確認された．

4.2 背景技術：REV2

本節ではレビューの信頼性を判定するために，Kumar et al. による REV2 [28] の説明を行う．

4.2.1 表記

ここでは評価者と被評価店舗の二部評価ネットワークを考える．評価者を u，被評価者を p とする．レビュワー $u \in \mathcal{U}$ が被評価店舗 $p \in \mathcal{P}$ に対し評価 $(u,p) \in \mathcal{R}$ を与えるといった，有向重み付き二部グラフ $G = (\mathcal{U}, \mathcal{R}, \mathcal{P})$ を考える．またここではレビュワーが店舗に対して与えた評価の点数を score (u,p) と表現する．score (u,p) は-1 から 1 の値に正規化されて入力される．また，$\mathcal{U}$,$\mathcal{R}$,$\mathcal{P}$ はそれぞれ，全てのレビュワー，一対一で与えた評価，被評価店舗の集合を表す．さらに，$\mathcal{O}(u)$ はレビュワー u による評価の集合，$\mathcal{I}(p)$ は被評価店舗の受けた評価の集合を表し，$|\mathcal{O}(u)|$,$|\mathcal{I}(p)|$ はそれぞれの集合の要素数を示す．

Goodness：被評価店舗の真の価値

被評価店舗 p は Goodness という指標を用いてモノの価値を定義する．Goodness のスコアは仮定として公平な評価者が与えるであろう評価として最も妥当であると考えられる一つの数値で表される．直感的に，良い店舗は公平な評価者から高評価を受け，逆に良くない店舗は公平な評価者から低い評価を受けていると考えられる．全ての被評価者 p について，その Goodness のスコア $G(p)$ は-1 から+1 までの範囲で算出される．

Fairness：レビュワーの公平度

またレビュワー u は Fairness という指標を用いて，その公平性を判断される．公平な評価者は偏見を持たずに評価を行う．つまり良い店舗に高評価，逆の場合には低評価をつける．そのため Goodness から逸脱したレビュワーを「不公平」と判断する．例えば，Goodness が高い店舗に低い評価をつける者は Fairness が低いといえる．全てのレビュワー u について，その Fairness のスコア $F(u)$ は 0 から 1 までの範囲で算出される．REV2 ではこの Fairness の指標を用いて不良ユーザーの検出を行い，それら妥当性をアノテーションが施された Amazon などのレビューデータセットを用いること評価している．

Reliability：レビューの信頼性

最後に一対一での評価 $score(u,p)$ は Reliability という指標を用いてそのレビューがどれだけ信頼できるかを判断される．Reliability のスコア $R(u,p)$ は 0 から 1 までの範囲で算出される．

定式化

式 4.1 にて Goodness を, 式 4.2 にて Fairness を, 式 4.3 にて Reliability を定義する．これら式より，それぞれが依存しあった関係を仮定していることがわかる．最もこれら関係を満たす Goodness, Fairness, Reliability を得るために，アルゴリズム 1 を用いて学習される．

$$G(p) = \frac{\sum_{(u,p)\in\mathcal{I}(p)} R(u,p)\cdot \mathrm{score}(u,p)}{|\mathcal{I}(p)|} \tag{4.1}$$

$$F(u) = \frac{\sum_{(u,p)\in\mathcal{O}(u)} R(u,p)}{|\mathcal{O}(u)|} \tag{4.2}$$

$$R(u,p) = \frac{\left(F(u) + \left(1 - \frac{|\mathrm{score}(u,p) - G(p)|}{2}\right)\right)}{2} \tag{4.3}$$

Algorithm 1 REV2 アルゴリズム

Input: 評価ネットワーク $(\mathcal{U}, \mathcal{R}, \mathcal{P})$

Output: Fairness, Reliability, Goodness

1: 初期化：$F^0(u) = 1$, $R^0(u,p) = 1$, and $G^0(p) = 1$ $\forall u \in \mathcal{U}, (u,p) \in \mathcal{R}, p \in \mathcal{P}$

2: $t = 0$

3: **do**

4: $t = t + 1$

5: Goodness の更新:
$\forall p \in \mathcal{P}$, $G^t(p) = \frac{\sum_{(u,p)\in\mathcal{I}(p)} R^{t-1}(u,p)\cdot \mathrm{score}(u,p)}{|\mathcal{I}(p)|}$

6: Reliability の更新:
$\forall (u,p) \in \mathcal{R}$, $R^t(u,p) = \frac{F^{t-1}(u) + \left(1 - \frac{\left|\mathrm{score}(u,p) - G^t(p)\right|}{2}\right)}{2}$

7: Fairness の更新:
$\forall u \in \mathcal{U}$, $F^t(u) = \frac{\sum_{(u,p)\in\mathcal{O}(u)} R^t(u,p)}{|\mathcal{O}(u)|}$

8: error $=\max\Big(\sum_{u\in\mathcal{U}} |F^t(u) - F^{t-1}(u)|,$
$\sum_{(u,p)\in\mathcal{R}} |R^t(u,p) - R^{t-1}(u,p)|, \sum_{p\in\mathcal{P}} |G^t(p) - G^{t-1}(p)|\Big)$

9: **while** $error > \epsilon$

10: **return** $F^t(u), R^t(u,p), G^t(p), \forall u \in \mathcal{U}, (u,p) \in \mathcal{R}, p \in \mathcal{P}$

4.2.2 Goodnessの入次数に基づく補正

Kumar et al. [28] は受けた評価の数に基づいてGoodnessの値を補正する手法についても提案を行っている．これは少数の評価しか受けていない被評価店舗の評価について，情報が不十分なことによる不確実性を排除するためである．具体的には，少数の評価しか受けていない場合，それが適正であるのかの判断をするための比較材料がないといった例が挙げられる．この問題に対処するためにここでは受けた評価の総数によってGoodnessの値を修正する．具体的には上で定義した$G(p)$の中央値を$median(G(p))$とし，その後ネットワークにおいてGoodnessののノードに入ってくるエッジの総数に1を加えた数値の自然対数をとったもの$G(p) - median(G(p))$との積を取る．定式化すると式4.4によってGoodnessの値を再定義する．

$$G'(p) = (G(p) - median(G(p))) \cdot \log(|\mathcal{I}(p)| + 1) \tag{4.4}$$

式4.4ではアルゴリズム1で計算されたGoodnessが同じ複数店舗について，それらが中央値よりも大きい場合，レビュー数が多い店舗ほど評価が高い店であるといったヒューリスティックが用いられている．また逆に中央値よりも低い場合，レビュー数が多いほどより評価が低い店であると判断される．そしてこれらレビュー数による影響は対数関数を用いることで，数レビューの差による影響が，レビュー数が比較的少ない場合は顕著に，比較的多い場合は軽微になるような工夫が施されている．なおこれら設定は2.4.1項で解説した，教師なし学習アルゴリズム特有の開発者が事前に設定するヒューリスティックに基づいており，アノテーション付きの学習データの追加などによりチューニングされうる箇所であることに注意する．

4.3 データセット

本稿では独自に収集した以下3つの東京都のラーメン店に関するデータセットを用いる．

Google Places Review

Google PlaceはGoogle社が提供する店や場所に関する情報をまとめたサービスである[1]．データの収集期間は2019年1月1日から10月12日の間で，同期間に東

[1] https://cloud.google.com/maps-platform/places

表 4.1: Google Places Review データセット概要

評価人数	被評価店舗数	評価数	評価者毎の平均評価数	被評価店毎の平均評価数	スパース度
33748 人	1208 店	108762 件	5.18 件	72.60 件	99.73%

京都内のラーメン店に対して投稿されたレビューを対象に分析を行う．概要を表 4.1 に載せる．

食べログ ラーメン 百名店 TOKYO 2019

株式会社カカクコムは同社が運営する口コミサービス食べログ[2]にて，ユーザーから高い評価を集めた 100 店を，様々なカテゴリーで一年ごとに掲載している[3]．集計方法については公開されていないが，一般に食べログは有識者によるレビューほど星に反映されやすい集計方法とされている．そのためそれら 100 店は真に価値が高いという仮定の元，本稿では 2019 年における東京都内のラーメン百名店[4] (以下，百名店と呼称) を検証用データとして以降の分析にて用いる．

ミシュランガイド東京 2019

「ミシュランガイド東京 2019」[97] は日本ミシュランタイヤが行う，東京都におけるレストランとホテルを厳選した格付けガイドブックである．それらは同社の覆面調査員が複数人で匿名で取材・評価を行い，広く信頼と人気を獲得している．2016 年より追加された「ラーメン」のカテゴリでは，2019 年度版に 24 店舗が選出された．本稿では前データセット同様，同 24 店舗を検証用データ (以降，ミシュランと呼称) として以降の分析にて用いる．

4.4 実験

本章では以下三つのリサーチクエスチョンに答えることを主眼に実験を行う．

- **RQ2.1** ： REV2 により算出された信頼度に基づく意見の集合 (Goodness) がベースライン手法と比較してより正しくモノの価値を推論できているか

[2] https://tabelog.com/
[3] https://award.tabelog.com/hyakumeiten
[4] https://award.tabelog.com/hyakumeiten/ramen_tokyo(最終閲覧日 2021 年 1 月 27 日)

表 4.2: 各手法におけるスコア上位 k 店舗が有識者による名店とどれだけ一致しているかを表す．評価には Precision と Recall を Oracle の性能で割ることで，最大性能の何%性能を発揮しているかを用いる．最も性能が良い結果を太字で表現する．全体を通して REV2 が全手法について同等かそれ以上の結果を示した．

データ	指標	Average	Base(1, 1)	Base(5, 5)	Base(10, 10)	Base(15, 15)	REV2
百名店	Precision@10	0.75	0.75	0.80	0.80	0.80	**1.00**
	Precision@50	0.72	0.70	0.76	0.74	0.74	**0.90**
	Precision@100	0.55	0.57	0.59	0.64	0.65	**0.71**
	Precision@200	0.78	0.81	0.83	0.86	0.88	**0.89**
	Recall@10	0.80	0.90	0.90	0.90	0.90	**1.00**
	Recall@50	0.72	0.70	0.76	0.74	0.74	**0.90**
	Recall@100	0.55	0.57	0.59	0.64	0.65	**0.71**
	Recall@200	0.78	0.81	0.83	0.86	0.88	**0.89**
ミシュラン	Precision@20	0.20	0.20	0.20	0.20	0.25	**0.30**
	Precision@50	0.38	0.38	0.42	0.38	0.38	**0.46**
	Precision@100	0.58	0.58	0.62	0.67	0.67	**0.71**
	Precision@200	**0.79**	**0.79**	**0.79**	**0.79**	**0.79**	**0.79**
	Recall@20	0.20	0.20	0.20	0.20	0.25	**0.30**
	Recall@50	0.38	0.38	0.42	0.38	0.38	**0.46**
	Recall@100	0.58	0.58	0.62	0.67	0.67	**0.71**
	Recall@200	**0.79**	**0.79**	**0.79**	**0.79**	**0.79**	**0.79**

- **RQ2.2**：Goodness が従来の平均による集計と比較して大きく異なる店にはどのような特徴があったか
- **RQ2.3**：REV2 によりレビュワーの公平性 (Fairness) が低いと判断されたレビュワーにはどのような特徴があったのか

4.4.1 性能の比較 (RQ2.1)

比較手法

Goodness の比較手法として，以下のベースライン手法を用いる．

Average : 単純に被評価店舗のレビューをすべて平均したスコアを用いる．レビューサイトでは最も広く使われている手法である．

Base(l, h):下位 $l\%$ と上位 $h\%$ の星のレビューを取り除いて平均したスコアを用いる．著しい外れ値が含まれていた際に影響を過度に受けないようにするためである．

Oracle:百名店やミシュラン登録店舗を事前に知っていて，それらを上位に並べたランキング．つまり比較手法として理論上の最大性能を表す．

本項では Goodness が上位の店舗がベースライン手法と比較してどれだけ検証用データと一致しているかについて定量的な評価を行う．

Gooodness とベースライン手法の比較

図 4.1 に，Goodness と Average の上位 k 件 (横軸) が，百名店の中に何件含まれているか (縦軸) を示す．また図 4.2 に，Goodness と Average の上位 k 件 (横軸) が，ミシュランの中に何件含まれているか (縦軸) を示す．なお破線に Oracle の結果を示す．

結果として，両データセットについて，Goodness は Average と比較して，いずれの上位 k 件についても同等かそれ以上の性能を示すことが確認できた．これにより，Goodness は Average と比較して，ランキング作成の観点でより有識者の感覚と近い結果が得られていることが示された．

続いて REV2 とその他ベースライン手法 (Average, Base(1, 1), Base(5, 5), Base(10, 10), Base(15, 15)) との性能比較を行う．評価指標として，推論結果上位 k 件について何件検証用データセットに含まれていたか (Precision@k)，検証用データセットの内何件が推論結果上位 k 件に含まれていたか (Recall@k) を用いて評価を行う．検証

のための k の値は [20, 50, 100, 200] を採用した．表 4.2 に各手法が Precision@k と Recall@k が Oracle の何%達成できているかを示す．例えば百名店において $k = 10$ を採用した場合に，Goodness 上位 10 件に 9 件百名店が含まれていた場合，Recall@10 は 0.09 となり，理論上の最大値は 0.1 なため表 4.2 では 0.9 と記される．そのため，同表は 0.0 から 1.0 の値を取り 1 に近いほど良いことを示す．

結果として Goodness は両データセットについてベースライン手法よりも Precision・Recall 共に同等かそれ以上の性能を示した．また Base(15, 15) のように極端な意見を除くことで Average よりも性能が向上することも確認された．以上の結果より REV2 を用いた評価の集約である Goodness は平均以外のベースライン手法と比較しても，より有識者の感覚と近く判断できていると言える．

なおミシュランの場合 $k = 200$ では比較手法全てで同じ結果が確認され，それら店舗について比較を行なったところ，同一のミシュラン登録店舗が選択されていることを確認した．それに伴い，各手法による評価の類似性を検証する．

図 4.3 に手法ごとの各店舗への評価点数に基づく相関係数のヒートマップを示す．その結果それぞれの手法ごとの相関係数が最低でも 0.98 という非常に高い相関を示していることが確認された．また各ベースライン手法により選ばれた店舗が REV2 とどれだけ類似しているかについて図 4.4 に示す．同図はベースライン手法と REV2 の上位 k 件の集合がどれだけ一致しているかを jaccard 係数を用いて表す．同図は k が上昇するにつれ右肩上がりの図になるため，比較のために店舗の順位をランダムにシャッフルした 100 個のランキングと REV2 との類似度の平均も示すその結果ベースライン手法はランダムな結果と比べて REV2 の結果に遥かに類似していることが確認された．

4.4.2 評価の変動に関する考察 (RQ2.2)

Goodness と Average の比較

本項では，Goodness と Average を比較した際，スコアが変動著しく変動した店舗に着目し，それぞれの特徴を分析する．そこで両者の比較のために両指標を平均 0，分散 1 になるよう標準化を施し，Goodness と Average の差を表すヒストグラムを図 4.5 に示す．釣鐘型の分布が確認されたことから，多くの場合両指標は近しい値を指し示す一方，部分的に乖離している点も散見された．

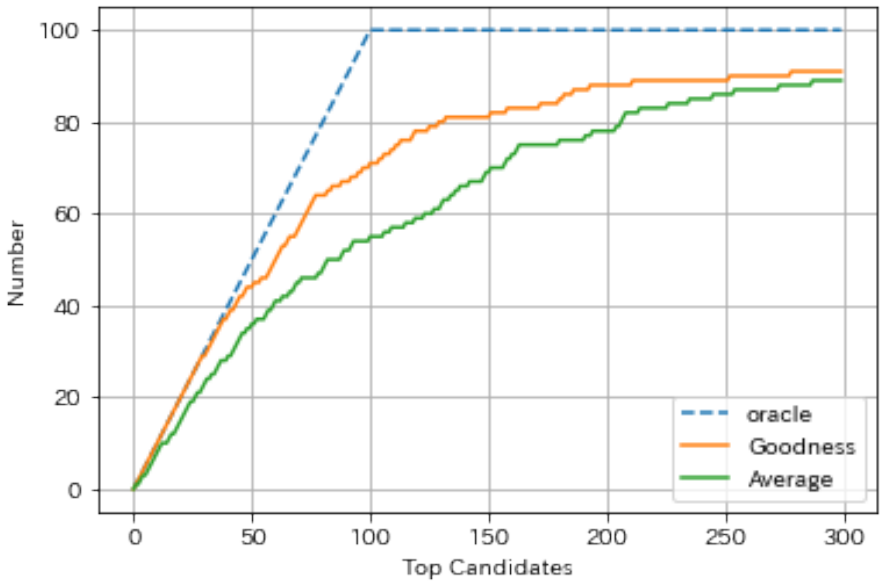


図 4.1: Goodness，Average の上位 k 件 (横軸) のうち, 百名店に含まれている店が何件含まれているか (縦軸).

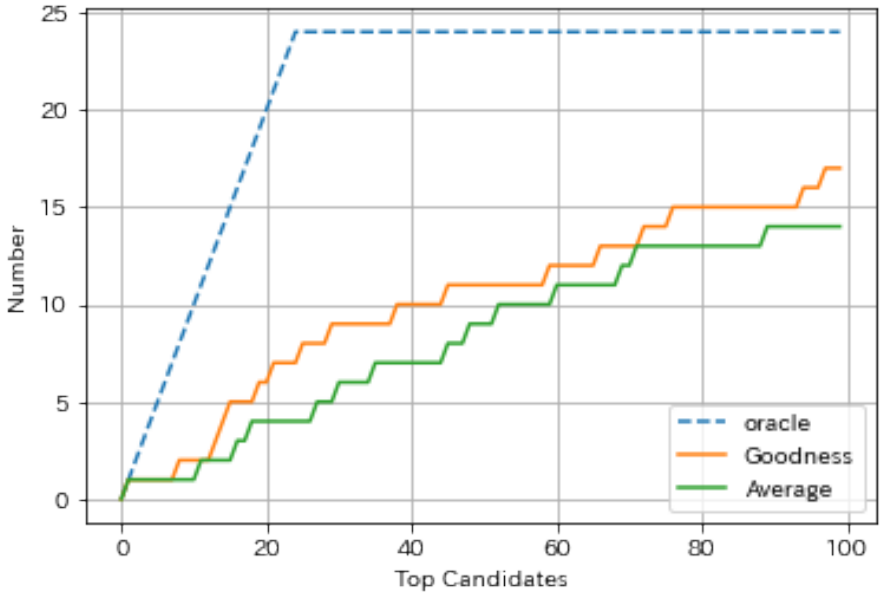


図 4.2: Goodness，Average の上位 k 件 (横軸) と，それらがミシュランの中に幾つ含まれているか (縦軸).

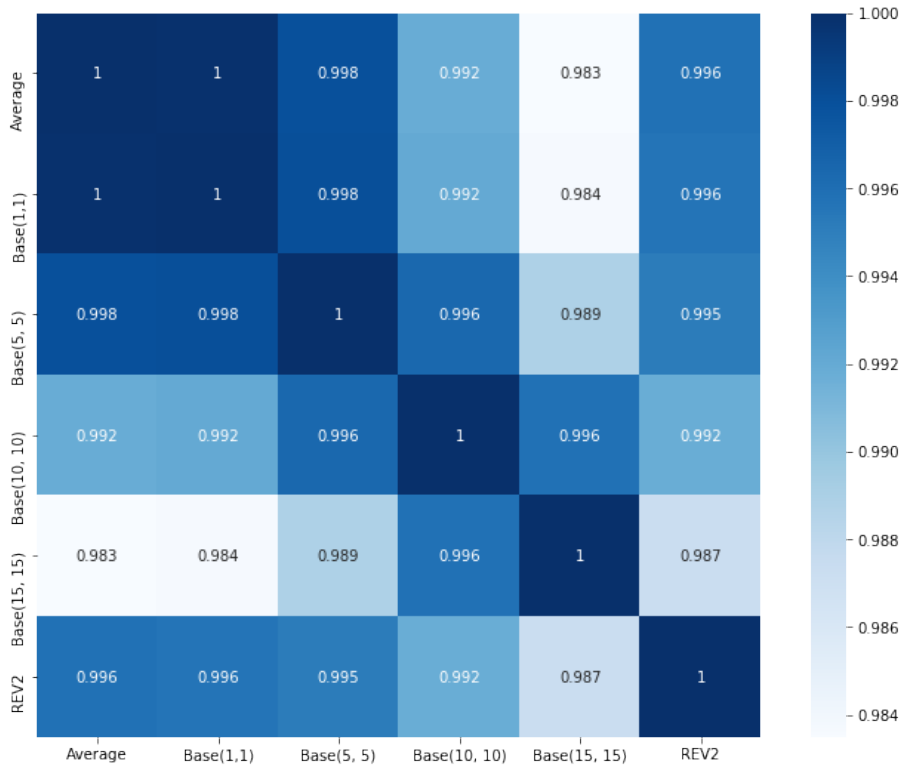


図 4.3: 手法ごとの各店舗への評価点数に基づく相関係数のヒートマップ

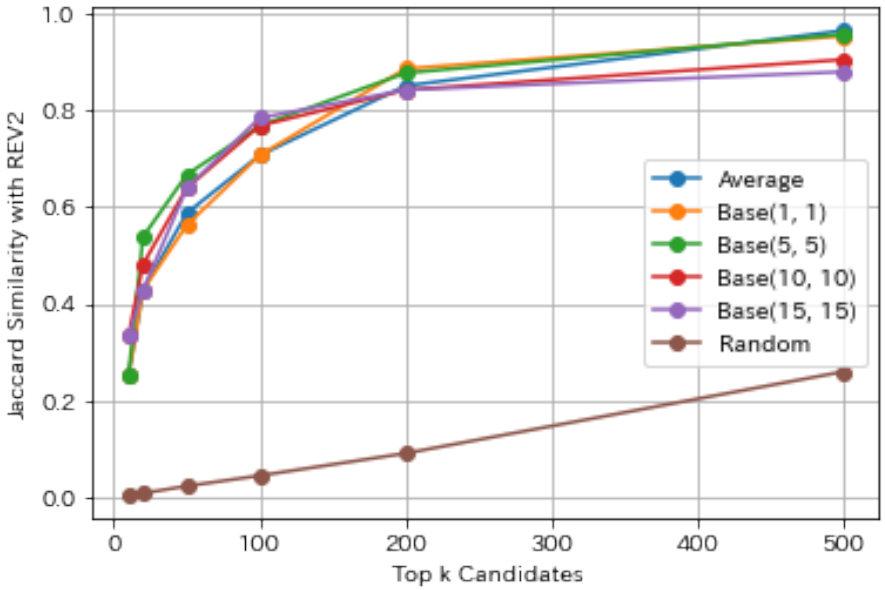


図 4.4: ベースライン手法とランダムシャッフルによる上位 k 件の集合が REV2 の上位 k 件の集合との一致度.

評価の変動に関する考察

続いてどのような場合に Goodness と Average が異なる値を示すのかについて分析を行う．各店舗を Goodness が Average と比較して大きい上位 N%と，少ない店舗の N%の二つの集団に分割する．そこで二つの集団を投稿された「レビューの信頼性」と「評価の高さ」の観点に基づき検証を行う．レビューの信頼性について上位 50%を信頼性の高いレビューと定義し，下位 50%を信頼性の低いレビューと定義する．またレビューの星が 3 以上をポジティブな評価，3 未満をネガテイブな評価として定義する ($score(u, p)$ としての表現ではスケーリングされているため 0 以上以下で評価を分ける)．以上のようにレビューを四種類に分類し，それぞれの集団について，これら四種類のレビューの発生確率を計算し，対数オッズ比を計算することで，どのように分布が変化したかを確認する．

結果を表 4.3 に示す．同表は Goodness が Average と比較して大きい集団と少ない集団について，信頼性が高い高評価，信頼性が低い高評価，信頼性の高い低評価，信頼性の低い低評価のレビューの割合がどれだけ増加したかを対数オッズ比を用いて N が 5, 10, 15, 20 の際の結果を示す．値が大きいほど Goodness が Average と比較して増加した集団にて観測される割合が増えたことを意味する．その結果 Goodness が Average と比較して増加した集団は，減少した集団と比較して，信頼性の高い高評価が多く，信頼性の高い低評価が少ない．また信頼性の低い低評価の割合が多く，信頼性の高い低評価が少ないといった傾向が確認された．この結果より，Goodness が Average と比較して減少した集団は信頼性が低い高評価が多いことから「やらせレビュー」の割合が多かったと解釈でき，また Goodness が Average と比較して増加した集団は信頼性が低い低評価が多いことから「アンチレビュー」の割合が多かったと解釈できる．

4.4.3 Fairness の解釈可能性 (RQ2.3)

本項では，レビュワーの Fairness について「レビュワーの投稿したスコアと他の評価者との差の平均」,「レビュワーが過去に投稿したスコアの平均」,「レビュワーのレビュー履歴数」の三つの観点から比較を分析することで，モデルの説明可能性を検証する．また結果の比較として，今回分析に使用した Google Place Review のデータセットをリンク情報を維持したまま評価値のみシャッフルしたサロゲートネットワークを用いる． 図 4.6 にレビュワーの Fairness の分布を示す．Google Place Review データセット (図 4.6 (a)) では Fairness が高い人が多い右側に偏った分布であることが確認された．またサロゲートネットワーク (図 4.6 (b)) でも同様の結果が確認され

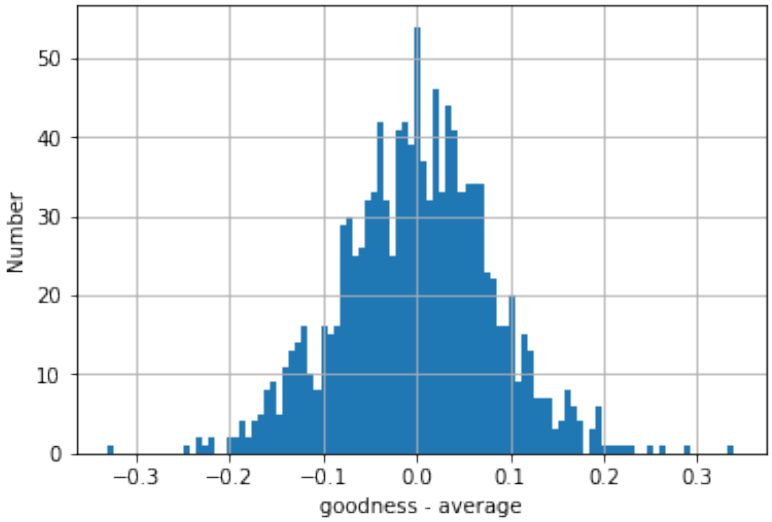


図 4.5: Goodness と Average との差のヒストグラム

表 4.3: Goodness が Average と比較して増加した集団と減少した集団について，信頼性が高い高評価，信頼性が低い高評価，信頼性の高い低評価，信頼性の低い低評価のレビューの割合がどれだけ増加したかを対数オッズ比を用いて表す．それぞれの集団は図 4.5 の上位下位 N%の閾値を用いて分けられる．値が大きいほど Goodness が Average と比較して増加した集団にて観測される割合が増えたことを意味する．

閾値	信頼性の高い		信頼性の低い	
	高評価	低評価	高評価	低評価
5 %	0.43	-0.43	-0.22	0.25
10 %	0.37	-0.41	-0.16	0.24
15 %	0.36	-0.45	-0.1	0.23
20 %	0.35	-0.46	-0.08	0.22

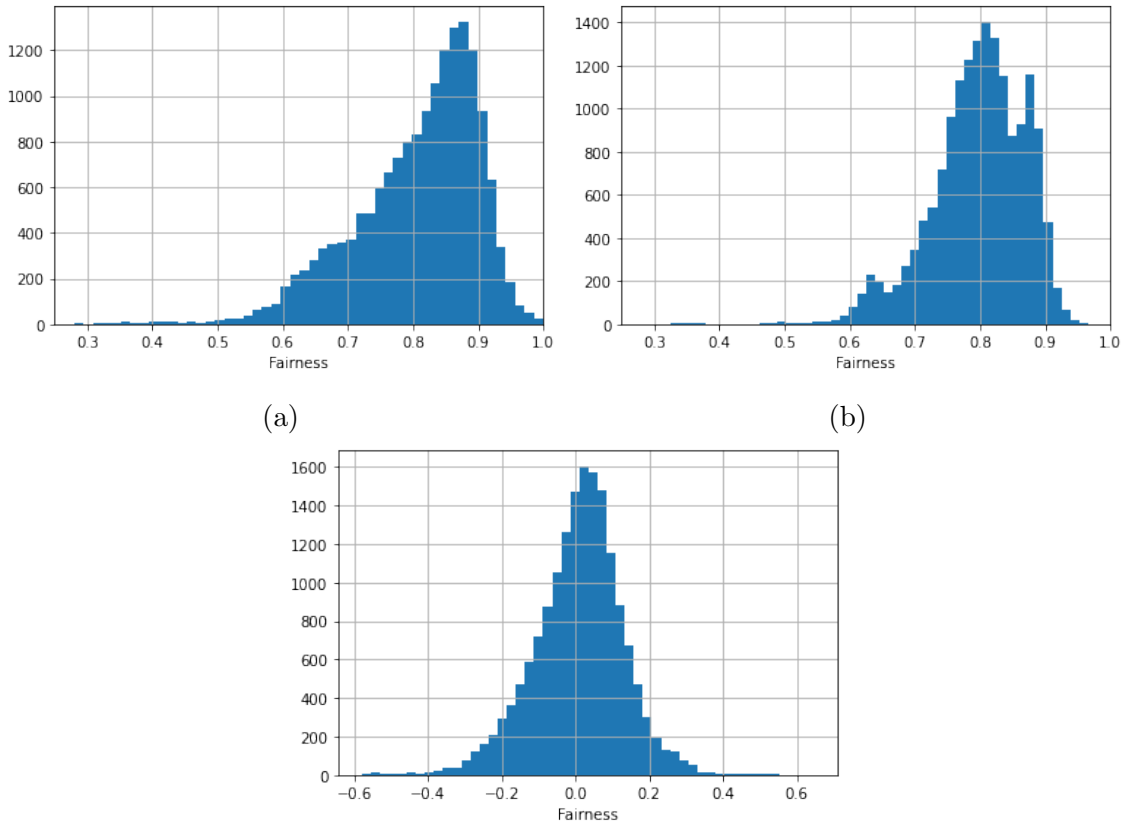


図 4.6: Fairness の分布．(a)Google Places Review (b) ランダムネットワーク (c) (a)-(b) に基づく Faieness の増減のヒストグラム

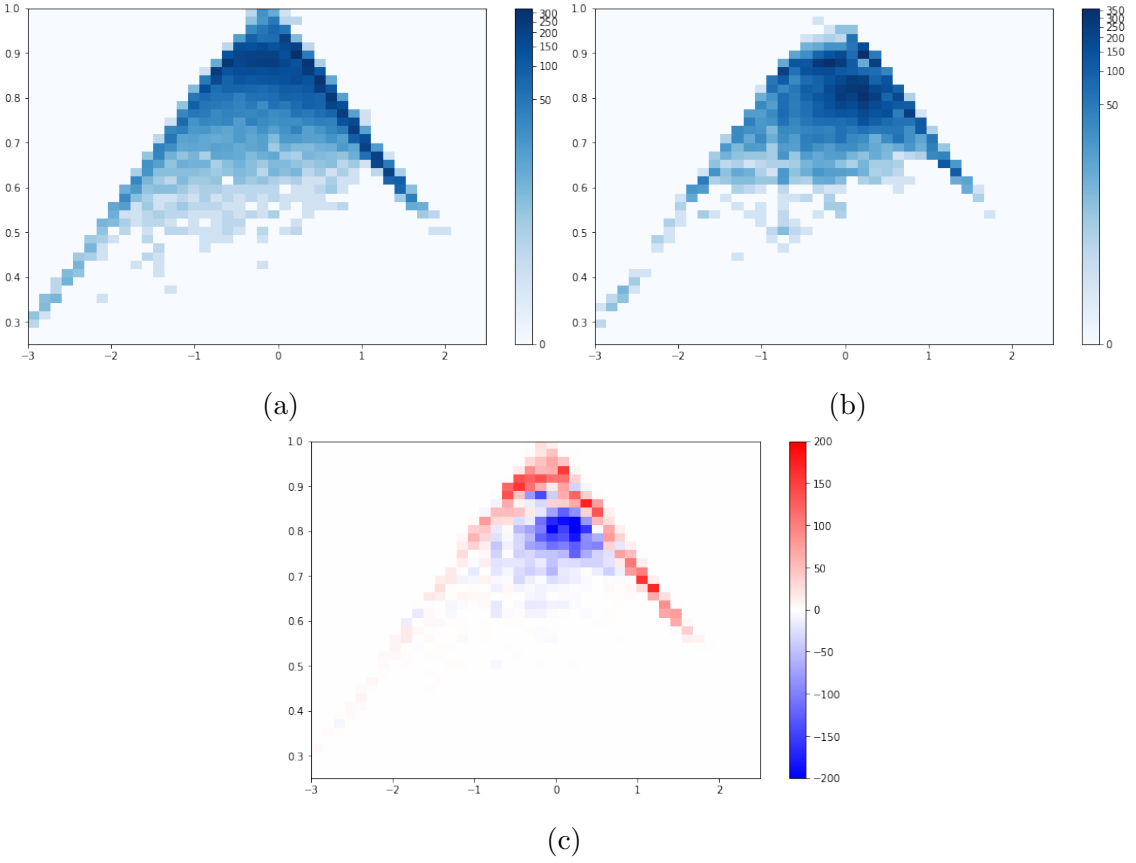


図 4.7: 「レビュワーの投稿したスコアと他の評価者との差の平均」(横軸) と Fairness(縦軸) の関係．対数スケールで可視化を行った．Fairness の分布．(a)Google Places Review (b) ランダムネットワーク (c) (a)-(b) による各ピクセルの人数の増減に基づくヒートマップ

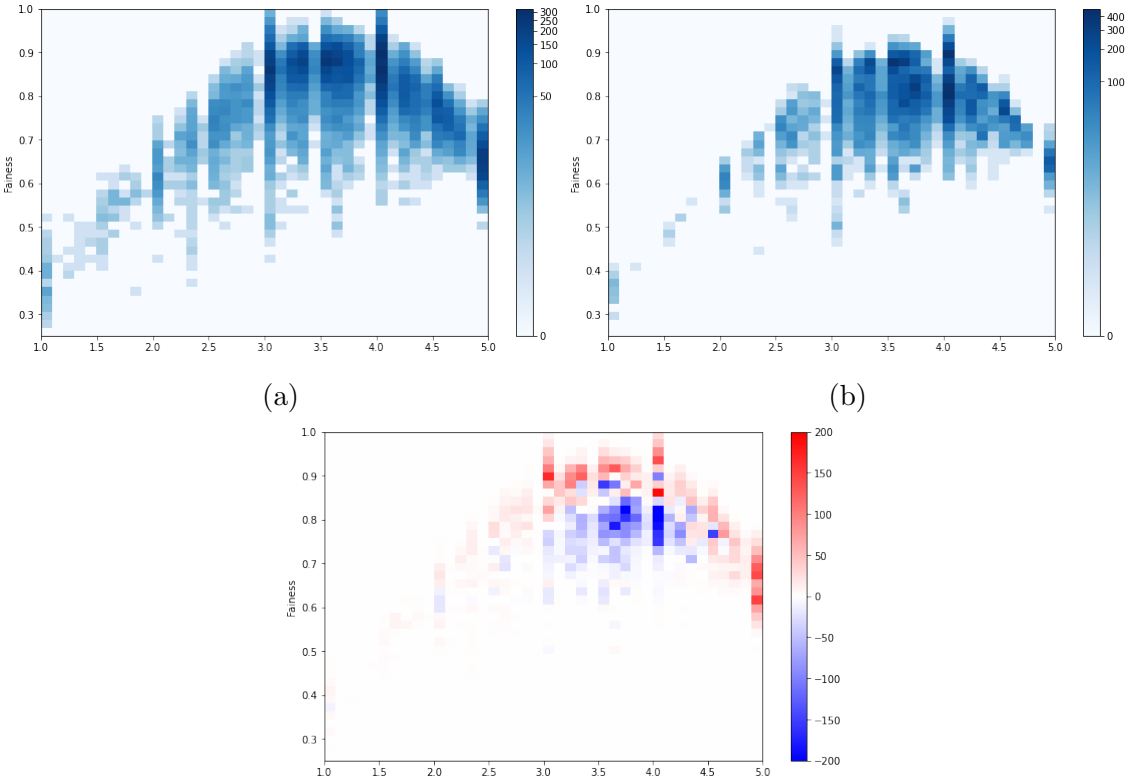


図 4.8: 「レビュワーが過去に投稿したスコアの平均」(横軸) と Fairness(縦軸) の関係．対数スケールで可視化を行った．Fairness の分布．(a)Google Places Review (b) ランダムネットワーク　(c) (a)-(b) による各ピクセルの人数の増減に基づくヒートマップ

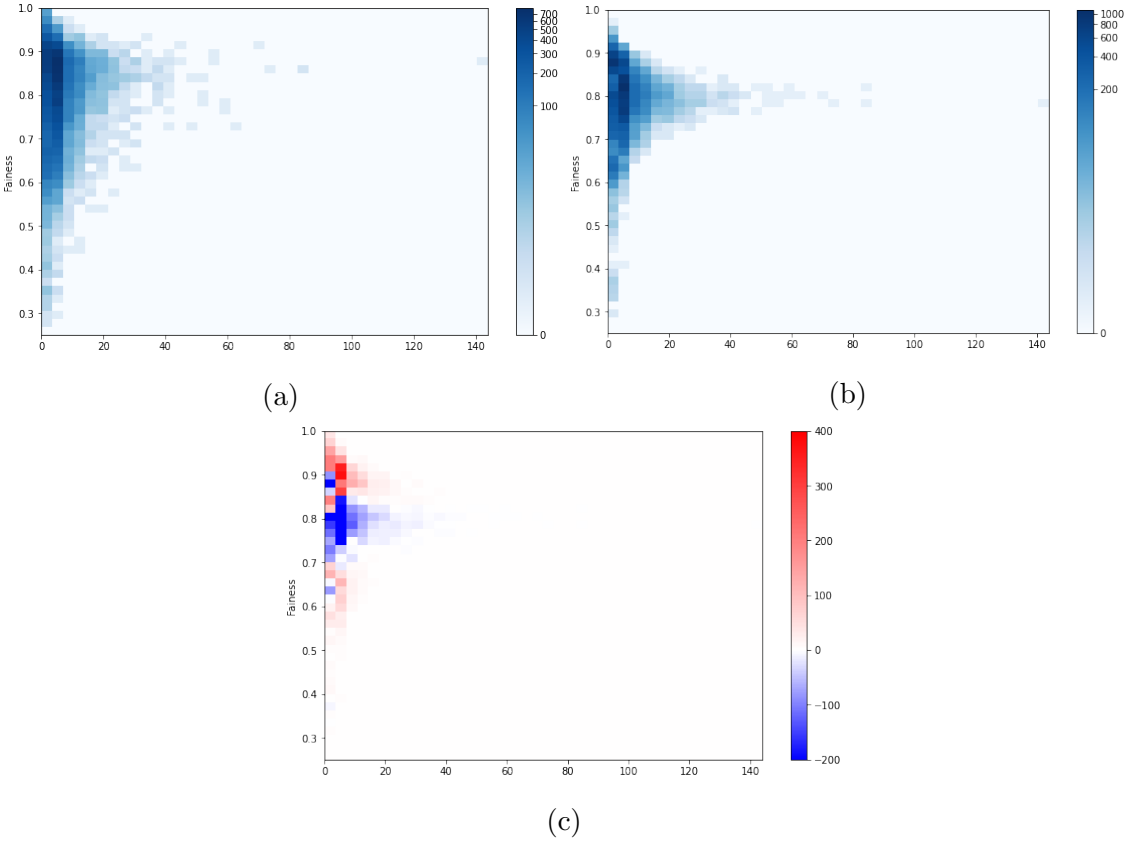


図 4.9: レビュワーのレビュー履歴数」(横軸) と Fairness(縦軸) の関係．対数スケールで可視化を行った．Fairness の分布．(a)Google Places Review (b) サロゲートネットワーク (c) (a)-(b) による各ピクセルの人数の増減に基づくヒートマップ

た．マン・ホイットニーの U 検定を用いて分布が異なるかを検定したところ有意水準 0.01% で，それら二つの集団で有意な差が確認された．データセットごとのレビュワーの Fairness の差 (Google Place Review - サロゲート) に基づくヒストグラムを図 4.6 (c) に示す．その結果中央値が 0.018，また増加した集団が 57%となり Google Place Review データセットはサロゲートネットワークに比べて Fairness が高い傾向にあることが確認された．

レビュワーの投稿したスコアと他の評価者との差の平均

図 4.7 に「レビュワーがある店舗に投稿したスコアと他の評価者が同一店舗に行った評価の平均との差の平均 (Goodness - Average)」と Fairness の関係を示す．図 4.7 (a)(b) は縦軸が Fairness，横軸がレビュワーの投稿したスコアと他の評価者との差の平均となっており，対数スケールで可視化を行った．

結果として Google Place Review データセット (図 4.7 (a)) ではレビュワーの投稿したスコアと他の評価者との差の平均が 0 を中心として，大きくなればなるほど Fairness が下がっていくという傾向が強く確認された．またサロゲートネットワークの結果 (図 4.7 (b)) についても類似した結果が確認された．そこで図 4.7 (c) では (a)-(b) による各ピクセルの人数の増減に基づくヒートマップを示す．その結果 Google Place Review の結果はサロゲートを用いた場合と比べて．レビュワーの投稿したスコアと他の評価者との差の平均が 0 を中心に Fairness が低下していくレビュワーが多く確認された．

レビュワーが過去に投稿したスコアの平均

図 4.8 に「レビュワーが過去に投稿したスコアの平均」と Fairness の関係を示す．図 4.8 (a)(b) は縦軸が Fairness，横軸がレビュワーが過去に投稿したスコアの平均となっており，対数スケールで可視化を行った．

結果として Google Place Review データセット (図 4.8 (a)) では，レビュワーが過去に投稿したスコアの平均が総じて低い場合，Fairness が低くなる傾向が確認された．また評価が高すぎる場合も Fairness が低くなることが確認される．またサロゲートネットワークの結果 (図 4.8 (b)) についても同様の結果が確認された．そこで図 4.7 (c) では (a)-(b) による各ピクセルの人数の増減に基づくヒートマップを示す．その結果過去に投稿したスコアの平均が 3 から 4 点付近の場合，Google Place Review の結果では Fairness が高くでており，また同時に 3 未満のような比較的低評価ばかりつけるレビュワーと 4 より大きい高評価ばかりつけるレビュワーでは Fairness が比

較的低くなることが確認された．これらは図 4.7 で確認されたように，Average との差分による Fairness の低下と強い相関があると考えられる．レビュワーが過去に投稿したスコアの平均が低いということは，低評価のみをつけているレビュワーである可能性が高い．公平性の高い低評価のみをつける辛口レビュワーは 4.4.3 項より，周囲の人も低くつけていれば Fairness が高くなるはずだが，その様なユーザー群は確認されず，多くの場合は公平性の低い「アンチレビュワー」が確認された．レビュワーが過去に投稿したスコアの平均も，モデルの出力を解釈する上で重要な指標であることが確認された．

レビュワーのレビュー履歴数

図 4.9 に「レビュワーのレビュー履歴数」と Fairness の関係を示す．縦軸が Fairness，横軸がレビュワーのレビュー履歴数の平均となっており，対数スケールで可視化を行った．

結果として Google Place Review データセット (図 4.9 (a))，レビュワーのレビュー履歴数が総じて低い場合 Fairness の分散は大きく．レビュー履歴数が多くなるにつれ分散が小さくなっていくことが確認された．またサロゲートネットワークの結果 (図 4.8 (b)) についても同様の結果が確認された．そこで図 4.7 (c) では (a)-(b) による各ピクセルの人数の増減に基づくヒートマップを示す．その結果レビュー数が多く Fairness が低いといった，やらせレビューを専門で行っていそうな特徴的なレビュワーはデータセット固有な特徴として確認されなかった．

両データセットについてレビュー履歴数と Fairness の関連性について検証を行う．ここではそれぞれ履歴数の中央値である 4 件以下のレビュワー集団と 4 件より多いレビュワー集団について，マン・ホイットニーの U 検定を用いて分布が異なるかを検定した．その結果帰無仮説は棄却されず，それぞれそれら二つの集団で有意な差があるとは言えなかった．

4.5 まとめ

本章では，レビューサイトなどにおける不良ユーザー検知アルゴリズムである REV2 [28] を用いて，三つの点について検証を行った．一つ目にレビューの信頼性に基づく意見の集約が平均などのベースライン手法と比較してより有識者の感覚と近い結果が得られることを確認した．これらの結果は，第 1 章 **Research Question 2： ユーザーのレビュー投票行動から不良ユーザー・レビューを発見し，信頼度に**

基づいた意見の集約によって有識者に近い集合知を獲得できるか？に対する答えとなっており，信頼度が低いユーザーのフィードバックは集合知に含めないほうが有識者の集合知に近い結果が得られることが示された．二つ目に信頼性に基づく集約と平均とのランキングに基づく比較を行った．その結果評価が高まった店舗は信頼性の高い高評価と，信頼性の低い低評価 (アンチレビュー) が多かったことが示された．反対に低くなった店舗は信頼性の低い高評価 (やらせレビュー) と，信頼性の高い低評価が多かったことが示された．三つ目に公平性が低いと REV2 アルゴリズムで示されるたユーザーについての解釈性の検証を実データとサロゲートネットワークの比較を用いて行った．結果「周りの人との意見のずれが大きい」,「低評価を付けやすい」といった傾向が確認された．

4.6 今後の方向性

本章で行った試みの今後の方向性として以下三つが考えられる．一つ目は，第 3 章で提案した認知バイアスの情報を取り除いた有用度を正解データとして，信頼度に基づく集合知の評価を行うことが考えられる．これまで一般的に集合知の評価は，答えが明確なものに限られていた．集合知を用いた，オンラインプラットフォームのコンテンツの評価を一元的に行う試みは，我々が調べた範囲では見当たらなかった．そこで 3.7 節で述べた派生手法により，本研究における有識者による評価部分を置き換えて，信頼性に基づく信頼性のさらなる方向性が考えられる．

二つ目は，本章で用いた通常の星の評価から，第 3 章で述べた事前バイアスの影響を取り除いた上で，同様の信頼度に基づく集合知の再評価が考えられる．第 2 章で行った議論から，本章で用いたデータセットも，有名店だからという評判バイアスによって高評価をつけているレビューが存在する可能性が高い．そこでそれらバイアスの影響を取り除いて再分析することは新たな知見が生まれると考えられる．しかし，3.7 節で述べたように，ユーザーがレビューした時点での認知バイアス情報などは，一般的に手に入らないことが多く，それら情報を管理しているプラットフォーム提供者からのデータ提供が必須になる点で難しさが残る．

三つ目は REV2 以外の信頼性評価アルゴリズムの検討が考えられる．REV2 以外に提案された不良ユーザー検出のアルゴリズムを用いることは容易に考えられる一方，第 3 章で述べた事前バイアスの影響の受けやすさといった観点が新たな，信頼性評価になるのではと考える．その点において，新たな信頼性評価アルゴリズムを提案することにより，再検証を行う方向性が考えられる．

第II部

個々人に応じた多元的なコンテンツの質の自動評価手法について

第5章　協調フィルタリングにおけるメタラーニングの適用による疎なデータからの学習と不確実性の推論

5.1　はじめに

近年の情報過多に伴う負担を軽減するために，あらゆる方面で推薦システムの研究が広く行われている．推薦システムは近年の情報過多の問題を解決する上で非常に重要な役割を果たす．推薦システムは個人ごとに嗜好度，またはコンテンツの質を推論しているという観点で，本論文の主題と関連性が高い．代表的なアプローチとして協調フィルタリング (Collaborative Filtering (CF)) は，過去の履歴のみから学習を行い，推薦するアイテムのコンテンツ情報に非依存といった手軽な点において，広く用いられている [98, 99, 100].

近年の深層学習を活用した協調フィルタリング手法には二つの欠点が大きく挙げられる．第一に，これら手法はユーザーやアイテムの過去の履歴データが豊富に利用できる場合には有効だが，それらが高度に疎な場合には性能が極端に減少することが知られる．疎なデータの事例としては，新しいアイテムや全くの新規ユーザーに対しての推薦といったコールドスタートな場面が挙げられる．それらへの対処としてはコンテンツベースフィルタリングか，ハイブリッドモデルを用いることが一般的だが，それは協調フィルタリングの簡便性を損なう．

第二の欠点としては，モデルの予測の不確実性のモデリングが行えない点が挙げられる．推薦システムにおける不確実性のモデリングには，能動学習への応用が考えられる．能動学習のように，能動的かつ適用的にユーザーのアイテムへの趣向を学習していくことはフィルターバブルの解消に貢献できると考えられる．これまで不確実性のモデリングは探索と活用のバランスを取る上で活用されてきた [101].

そこで本章では上記を踏まえ第 1 章 **Research Question 3：　推薦システムは**

履歴の少ないユーザーやアイテムの組み合わせについても正しくそれぞれの間の嗜好度をモデリングすることは可能か？ Research Question 4： 推薦システムは予測の不確実性をモデリングすることは可能か？に答える形で研究を行う．本章では取り上げた二点の問題を同時に解決する学習フレームワークを提案する．提案手法の着想としては，上記二つの問題は互いに関連し合っているという仮説に基づく．推薦システムにおいてとりわけ新規に近いユーザーは識別情報の欠如により不確実性が生じうることが知られており [102]，そのためより正確な推薦を可能にするために不確実性のモデリングが重要と考えたためである．

我々はメタラーニングの派生手法の一つである Conditional Neural Processes (CNP)[103, 104] を，既存の協調フィルタリングに応用した学習方法 (MetaCF) を提案する [105]．本提案手法は学習のさせ方に焦点を置いており，その点あらゆる既存の深層学習ベースの手法に，モデルの機構に実質手を加えることなく適用が可能である．MetaCF を適用させることで，実験的に推薦システムのタスクの一つである暗黙的フィードバックを用いた Top-N 推薦タスクにおいて，既存の state-of-the-art の手法に対し大きく上回る有効性を示す．推論結果を分析することで MetaCF はとりわけ過去履歴の少ないユーザー，もしくはアイテムに対し. 高い不確実性を示し，またそれらが最終的な推薦タスクのスコアに貢献していることを示す．

5.2 背景知識

5.2.1 メタラーニング

大規模なデータセットにおける機械学習の性能の向上の傍ら，より少ないデータからの効率的な学習への需要が高まっている．そこで近年ではメタラーニングを用いた，複数のタスクで学習することで未知のタスクにも少ないデータや学習ステップで適応できるといった学習手法に注目が集まっている．

一般的な機械学習の学習設定では，学習データ $\mathbf{x}$ と予測対象 $\mathbf{y}$ の集合であるデータセット $\mathcal{D} = \{(\mathbf{x}_1, \mathbf{y}_1), \ldots, (\mathbf{x}_N, \mathbf{y}_N)\}$ が与えられ，損失関数 $\mathcal{L}$ を最小化するように予測モデル $\hat{\mathbf{y}} = f_\theta(\mathbf{x})$ を構築する．

$$\theta^* = \arg\min_\theta \mathcal{L}(\mathcal{D}; \theta, \omega) \tag{5.1}$$

ここでの ω は，例えば θ の最適化手法や f のアーキテクチャ構造といった事前に開発者が決定する “how to learn” に関する情報で，メタ知識として知られる．一般

的な機械学習の特徴としてそれらは事前に決定され，θ の最適化は各タスクについて一から行われる点が挙げられる．

一方メタラーニングの特徴として上記 $\mathcal{D}$ のような最適化を行うタスクを複数同時に扱い ω を学習する点が挙げられる．データセット $\mathcal{D}$ と損失関数 $\mathcal{L}$ を持つタスクを $\mathcal{T} = \{\mathcal{D}, \mathcal{L}\}$ で表すと，学習させるタスクの分布 $p(\mathcal{T})$ について下記 ω の性能を評価する：

$$\min_{\omega} \mathop{\mathbb{E}}_{\mathcal{T}\sim p(\mathcal{T})} \mathcal{L}(\mathcal{D}; \omega) \tag{5.2}$$

ここで $\mathcal{L}(\mathcal{D}; \omega)$ は，データセット $\mathcal{D}$ 上で ω を用いて訓練されたモデルの性能を評価する．これら学習の仕方 ω を学習している観点からメタラーニングは “learning to learn” などと呼ばれる．

Hospedales et al. [106] はこれら既存のメタラーニング手法を Meta-Representation (“What?”)，Meta-Optimizer (“How?”)， Meta-Objective (“Why?”) によって分類することを提唱している．Meta Representation は上記 ω の選択を意味し，パラメータ初期値，最適化手法，ネットワーク構造，ロス関数などが挙げられる．Meta-Optimizer はそれら学習方法を意味し，勾配ベース，強化学習，進化学習などが挙げられる．Meta-Ojective はそれら学習の目的を意味し，Few-Shot Learning や，転移学習，Domain Adaptation，継続学習などが挙げられる．例えば代表的メタラーニング手法の MAML [107] は ω としてニューラルネットの初期値が該当し，学習方法として勾配ベースを，目的として Few-Shot Learning が挙げられる．

5.2.2 Conditional Neural Processes

Garnelo et al. [103] は，少ないデータでの学習や不確実性のモデリングを行うメタラーニングの派生手法の一つである Conditional Neural Processes (CNP) を提案した．CNP は階層ベイズと確率過程の観点からメタラーニングを捉え直した手法である．通常の教師あり学習との相違点として，Gaussian Processes [108] と類似するように確信度に基づく分布を出力に持つ．データ数を N とした際，Gaussian Process は計算時間はオーダー $\mathcal{O}(N^3)$ であるのに対し，CNP は $\mathcal{O}(N)$ という点で大規模データに対してスケーラブルな点で優れた手法と言える．

定式化

CNP は確率過程に基づくデータから DNN を用いることで直接予測分布を学習する．未知の確率過程 $f : \mathcal{X} \to \mathcal{Y}$ 由来のいくつかの観測が与えられた際，CNP は条

件付き表現 $\mathbf{r}$ に基づき f の未観測の値を推論するように学習を行う．

定式化すると $\mathbf{x_i} \in \mathcal{X},\ \mathbf{y_i} \in \mathcal{Y}$ における $(\mathbf{x_i}, \mathbf{y_i})$ を f のそれぞれの関数に対して二つの集合に分割する，つまり n 個データが観測された際， m 個の文脈データ $\mathcal{C} = \{(\mathbf{x_i}, \mathbf{y_i})\}_{i=1}^{m}$ と $n-m$ 個の目標データ $\mathcal{D} = \{(\mathbf{x_j}, \mathbf{y_j})\}_{j=m+1}^{n}$ に分けられる．これらデータは以下のように処理される．

$$\mathbf{r_i} = h_\theta\left(\mathbf{x_i}, \mathbf{y_i}\right), \quad \forall\left(\mathbf{x_i}, \mathbf{y_i}\right) \in \mathcal{C} \tag{5.3}$$

$$\mathbf{r} = \mathbf{r_1} \oplus \mathbf{r_2} \oplus \ldots \mathbf{r_{m-1}} \oplus \mathbf{r_m} \tag{5.4}$$

$$\boldsymbol{\phi_j} = g_\theta\left(\mathbf{x_j}, \mathbf{r}\right), \quad \forall\left(\mathbf{x_j}\right) \in \mathcal{D} \tag{5.5}$$

最初にエンコーダー h_θ を用いてすべての文脈データ $(\mathbf{x_i}, \mathbf{y_i})$ の集合を条件付き表現 $\mathbf{r_i}$ に変換を行う $(h_\theta : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^d)$．続いて $\mathbf{r_i}$ を順列非依存な集約計算（平均や和）などを用いて固定長な表現である $\mathbf{r}$ へ集約を行う．最終的にデコーダー g_θ を用いて予測分布の推論を目標点 $\mathbf{x_j} \in \mathcal{D}$ に対して行う $(g_\theta : \mathcal{X} \times \mathbb{R}^d \to \mathbb{R}^e)$． 回帰タスクにおいては $\boldsymbol{\phi_j}$ は正規分布 $\mathcal{N}\left(\mu_j, \sigma_j^2\right)$ における平均と分散 $\boldsymbol{\phi_j} = \left(\mu_j, \sigma_j^2\right)$ をモデリングし，また分類タスクにおいて $\boldsymbol{\phi_j}$ はカテゴリカル分布におけるクラス c における生起確率をモデリングする．これらを式で表現すると以下の形で表現できる：

$$P\left(\mathcal{Y} \mid \mathcal{X}, \mathcal{C}\right) = \int P(\mathcal{Y} \mid \mathcal{X}, \mathbf{r}) P\left(\mathbf{r} \mid \mathcal{C}\right) d\mathbf{r} \tag{5.6}$$

式 5.6 は，文脈データ $\mathcal{C}$ を固定次元のパラメータ $\mathbf{r}$ に埋め込むという点で，メタ知識を獲得しているとみなせる点においてメタラーニングとして捉えられる．

通常のノンパラメトリックなベイジアン手法では単純な事前確率を仮定し，事後分布の追跡を行うが，CNP は代わりにニューラルネットを用いることで直接確率過程に基づくデータから予測分布の学習を行う．この手法によってデータから複数の関数の生成や出力における分散のモデリングが可能になる．CNP は Few-shot 分類・回帰や画像補完などへの適用が行われている [103, 104]．本研究では CNP を協調フィルタリング手法に拡張し，省データからの学習のみならず不確実性の考慮を可能にした．

5.3 提案手法

本章ではまず全体的な提案手法の枠組みについて，続いて提案手法を適用させるネットワークの機構について説明を行う．上記で確認したようにメタラーニングを推薦システムに応用した研究は過去にも存在し，Vartak et al. [109] は Twitter

や Facebook における新規投稿をコールドスタート問題として捉えメタラーニングの枠組みを適用している．また Chen et al. [110] はメタ連合学習 (federated meta learning) の枠組みを用いて分散学習とプライバシー保護の観点から提案を行っている． 本研究では CNP の発想に帰着を得ており，深層学習を用いた協調フィルタリングの手法と組み合わせることで上記で取り上げた問題の解決を図る．深層学習を用いた協調フィルタリングの手法の多くは，ネットワークの構造によりいかにユーザーとアイテムの embedding を合成させるかに焦点が置かれており，どのようにしてモデルを学習させるかという点における研究はあまり行われていない．

5.3.1 MetaCF サンプリング

ユーザーとアイテムのインタラクション行列 $\mathbf{Y}$ は確率過程に基づいた一部の観測結果とみなすことができる．なぜなら $\mathbf{Y}$ における $\mathbf{Y}_{ui} = 0$ は必ずしも u が i のことを好んでいないだけではなく，ただ気づいていないだけかもしれないという，真の状態の一部のみを確率的に観測したデータであるためである．CNP に則り，提案手法の主なアイディアは観測された点のうち確率的にサンプリングした任意の数の点を用い，予測分布を学習させることにある．

行列における未観測の要素の予測というタスクは先行研究の CNP を画像補完タスクに応用した事例に近い部分がある [104]．同研究では文脈データの候補として任意の要素を用いていたが，提案手法では推論を行う目的の点 $\mathbf{Y}_{ui}$ に対して，対応する同行・同列における観測されている点のみを文脈データの候補として用いる点において異なる． これは目的の点における出力をモデリングする際，行も列も異なる点の情報よりも同一行・列内の情報との依存性が他の任意の点に比べ大きいと考えたためであり，この考え方は MF といった一般的な協調フィルタリングの手法とも考え方が共通している．

定式化

以下 $\mathbf{Y}_{u*}$ を $\mathbf{Y}$ におけるユーザー u と対応する行ベクトルとし，$\mathbf{Y}_{*i}$ をアイテム i と対応する列ベクトルとする．続いて予測対象の $\mathbf{Y}_{ui}$ に対し，$\mathbf{Y}_{u*}$ と $\mathbf{y}_{*\mathbf{i}}$ のうち観測された点を確率的に任意の数サンプリングし，文脈点として用いる． 本研究ではこの過程のことを MetaCF サンプリングと呼ぶこととする．MetaCF サンプリングは CNP と同様，学習時毎バッチごとに生成される．

これは CNP におけるランダムに文脈点を選ぶ過程と動機を共有する．直感的な理解としては同ユーザーは過去の履歴のうち一部が欠けていたとしても同様のフィー

ドバックを行っていたであろうとみなすことができる．MetaCF サンプリングは以下のように記述される:

$$p_u \sim \text{ Uniform } (0, 1) \tag{5.7}$$

$$\mathbf{m_{u*}} \sim \text{ Bernoulli } (p_u) \tag{5.8}$$

$$\mathbf{Y}'_{u*} = \mathbf{m_{u*}} \odot \mathbf{Y}_{u*} \tag{5.9}$$

ここでの $\odot$ は要素積を意味する．$\mathbf{m_{u*}}$ はそれぞれ独立なベルヌーイ分布からサンプリングされた 0 か 1 かを持つマスキングのためのベクトルであり，同ベルヌーイ分布のパラメータ p_u は $[0, 1]$ の一様分布からサンプリングされる． $\mathbf{m_{u*}}$ は対応する行 $\mathbf{Y}_{u*}$ との要素積を行い，$\mathbf{Y}'_{u*}$ を計算する．

同様の計算はアイテム i の列についても対応する列 $\mathbf{Y}_{*i}$ と同様の計算を行い，MetaCF サンプリング後の列 $\mathbf{Y}'_{*i}$ を計算する．

$$p_i \sim \text{ Uniform } (0, 1) \tag{5.10}$$

$$\mathbf{m_{*i}} \sim \text{ Bernoulli } (p_i) \tag{5.11}$$

$$\mathbf{Y}'_{*i} = \mathbf{m_{*i}} \odot \mathbf{y_{*i}} \tag{5.12}$$

本提案手法は Dropout [111] と類似しているが，Dropout は学習時の p_u などと対応するノードを落とす割合が事前に固定であり，出力に同割合の逆数でスケールアップさせるのに対し，本手法ではノードの落とす割合も変動し，スケールアップも行わない点において異なる．

これら $\mathbf{Y}'_{u*}$ と $\mathbf{Y}'_{*i}$ はニューラルネットワークを用いて以下のように処理される:

$$\mathbf{r_u} = h^U_\theta(\mathbf{y}'_{\mathbf{u}*}) \tag{5.13}$$

$$\mathbf{r_i} = h^I_\theta(\mathbf{Y}'_{*i}) \tag{5.14}$$

$$\boldsymbol{\phi_{ui}} = g_\theta(\mathbf{r_u}, \mathbf{r_i}) \tag{5.15}$$

最初にエンコーダー h^U_θ を用いて $\mathbf{Y}_{u*}$ を変換しユーザーの潜在表現 $\mathbf{r_u}$ を得る．そしてエンコーダー h^I_θ を用いて $\mathbf{y_{*i}}$ を変換しアイテムの潜在表現 $\mathbf{r_i}$ を得る．この過程は文脈データを潜在変数としてベクトル空間に埋め込み，固定次元への圧縮を行っている．すなわち NP における条件付きの潜在表現を得る式 5.3 と固定次元への集約を同時に行っている点において式 5.4 と対応している．その後デコーダー g_θ は $\mathbf{r_u}$ と $\mathbf{r_i}$ を入力とし式 5.15 に示す出力を行う．本研究の問題設定ではベルヌーイ分布のパラメータを出力し，対数尤度の最大化を目的関数に学習を行う．このようにして

MetaCF の学習の枠組みで学習するモデルは NP の枠組みに基づき出力に予測の分布を仮定した形で学習を行うことができる．

ネットワークの出力 $\boldsymbol{\phi_{ui}}$ はタスクに基づいて選択される．本研究では Top-N 推薦タスクのため $\boldsymbol{\phi_{ui}}$ はクリックが観測される確率 p_{ui} を出力する．また rating prediction を解く場合は $\boldsymbol{\phi_{ui}}$ は正規分布 $\mathcal{N}(\mu_{ui}, \sigma_{ui}^2)$ の平均と分散 $\boldsymbol{\phi_{ui}} = (\mu_{ui}, \sigma_{ui}^2)$ を出力する．もしくは多腕バンディット問題のように多クラス分類を扱う場合 ϕ_{ui} はカテゴリカル分布の各クラス c の生起確率 p_c と対応する． 以上これら MetaCF サンプリングと出力のモデリングを合わせて，提案手法を MetaCF と呼称する． MetaCF は Keras [112] を用いて数行での実装が行える点と従来の深層学習ベースの協調フィルタリングモデルのうちネットワーク機構を変えることなく実装できるという点において利便性が高い．

5.3.2 ネットワーク機構

本研究での提案手法である MetaCF はネットワークの構造に依存しないため，ここではその有効性の検証として従来手法の CFNet に適用させることで有効性を確認する．従来の深層学習を用いた協調フィルタリングの手法は大きく 2 種類の大別される．一つ目は表現学習 (representation learning) に重きを置いたもの，二つ目はマッチング学習 (matching learning) に重きを置いたものである． CFNet はその両方の機構を採用した形をしており，複雑なユーザーとアイテムの合成を行う関数や低ランクな関係をモデリングする．本章では本研究で用いた実装について述べる．また CFNet を MetaCF の枠組みで学習させたネットワークを MetaCFNet と呼ぶこととし，全体の構造について図 5.1 に示す．

表現学習 (Representation Learning)

表現学習は DNN を用いて複雑な関数を近似しユーザーとアイテムの embedding を同一空間に写像し，直接同空間内での類似度を図ることで推論を行うことを目的とする．本研究では多層パーセプトロン (MLP) を用いてユーザーアイテムそれぞれの表現を学習する．同学習手法は以下のように定義される:

$$\mathbf{p}_u = \mathbf{P}^T \mathbf{y}'_{u*} \tag{5.16}$$

$$\mathbf{q}_i = \mathbf{Q}^T \mathbf{y}'_{*i} \tag{5.17}$$

$$\mathbf{r}_u = MLP_u(\mathbf{p}_u) \tag{5.18}$$

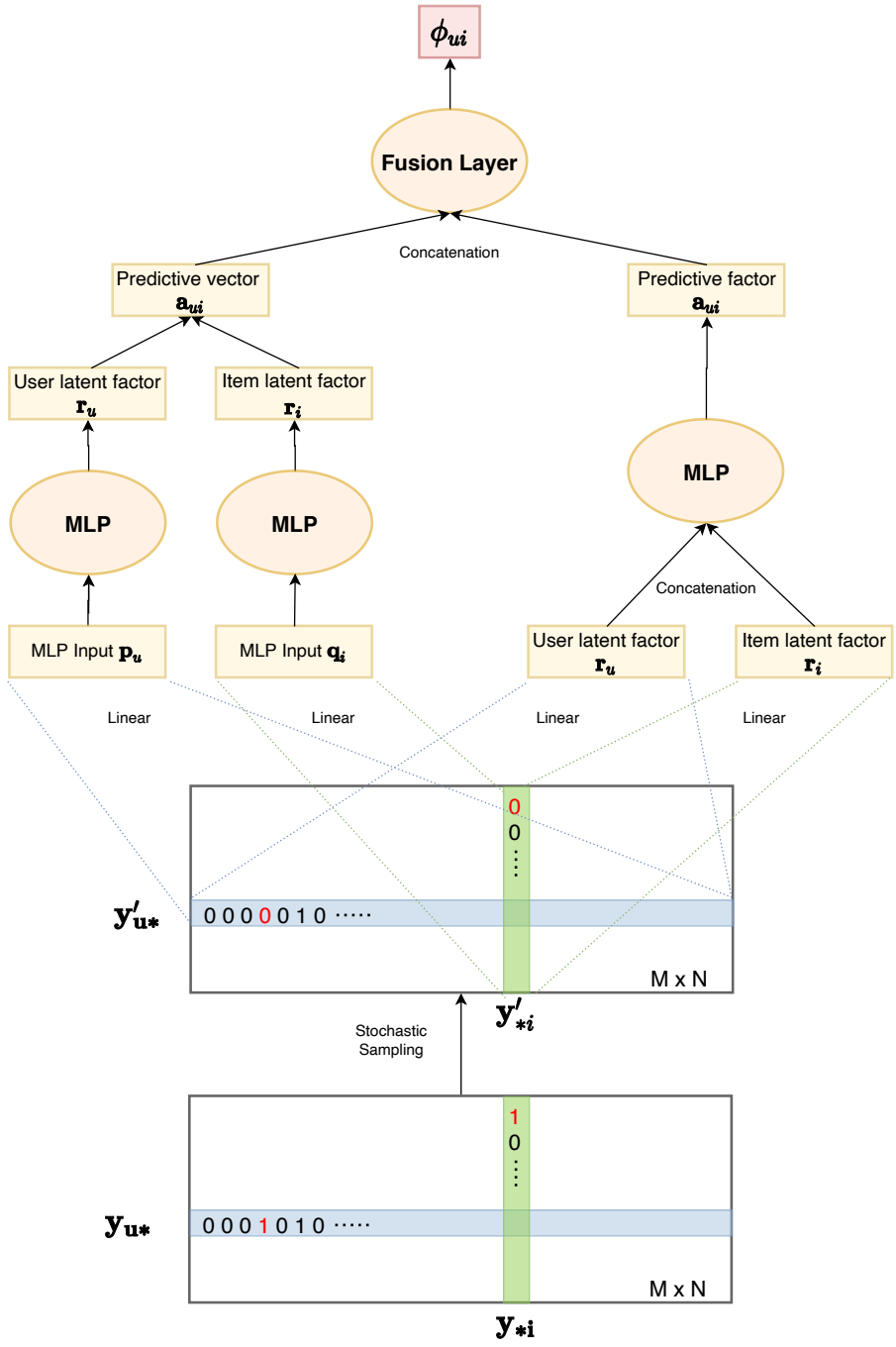


図 5.1: MetaCFNet の構造 (福馬らの論文図 1 より [105])

$$\mathbf{r}_i = MLP_i(\mathbf{q}_i) \tag{5.19}$$

$$\mathbf{a}_{ui} = \mathbf{r}_u \odot \mathbf{r}_i \tag{5.20}$$

$$\phi_{ui} = \sigma\left(\mathbf{W}_{out}^T \mathbf{a}_{ui}\right) \tag{5.21}$$

$\mathbf{P}$ と $\mathbf{Q}$ はインタラクション行列の行・列から embedding を得るための線形変換のパラメータを表す ($\mathbf{P} : \mathbb{R}^M \to \mathbb{R}^{M'}, \mathbf{Q} : \mathbb{R}^N \to \mathbb{R}^{N'}$)．$MLP_u(*)$ と $MLP_i(*)$ は MLP を用いた写像を意味する ($MLP_u : \mathbb{R}^{M'} \to \mathbb{R}^d, MLP_i : \mathbb{R}^{N'} \to \mathbb{R}^d$)．$MLP_u(*)$ と $MLP_i(*)$． 類似度計算にノンパラメトリックな内積を用いることでモデルは低ランクな関係を学習する．本研究で提案手法 MetaCF と組み合わせた表現学習のネットワークを MetaCF-rl と以下では呼称する．

マッチング学習 (Matching Learning)

マッチング学習では DNN の非線形性を活用し，ユーザーとアイテムの表現の合成を複雑な関数で近似することで推論を行う． 本研究では MLP を用いて関数の近似を行い，以下のように定式化される:

$$\mathbf{r}_u = \mathbf{P}^T \mathbf{y}'_{u*} \tag{5.22}$$

$$\mathbf{r}_i = \mathbf{Q}^T \mathbf{y}'_{*i} \tag{5.23}$$

$$\mathbf{a}_{ui} = MLP(\begin{bmatrix} \mathbf{r}_u \\ \mathbf{r}_i \end{bmatrix}) \tag{5.24}$$

$$\boldsymbol{\phi_{ui}} = \sigma\left(\mathbf{W}_{out}^T \mathbf{a}_{ui}\right) \tag{5.25}$$

潜在表現 $\mathbf{r_u}$ と $\mathbf{r_i}$ は連結させることで集約を行い，MLP を用いて推論を行う．本研究で提案手法 MetaCF と組み合わせたマッチング学習のネットワークを MetaCF-ml と以下では呼称する．

フュージョンレイヤー

先二つで提唱したモデルによって得られた表現 $\mathbf{a}_{ui}$ を組み合わせることで両者の利点を生かした表現が学習される．二つの表現を連結させ全結合層へと入力することで最終的な予測を行う．$\mathbf{a}_Y^{rl}$ と $\mathbf{a}_Y^{ml}$ をそれぞれ表現学習由来の表現とマッチング学習由来の表現とみなした場合，以下のように定式化される:

$$\boldsymbol{\phi_{ui}} = \sigma\left(\mathbf{W}_{out}^{T}\begin{bmatrix}\mathbf{a}_{ui}^{rl}\\ \mathbf{a}_{ui}^{ml}\end{bmatrix}\right) \tag{5.26}$$

式5.26を用いることで，MetaCF-rlとMetaCF-mlを組み合わせ最終的な提案ネットワーク機構MetaCFNetとなる．

5.4 実験

本章では下記のリサーチクエスチョン (RQ) に答える形で実験を行う:

RQ3.1 ネットワークの様々な学習設定のうち性能に寄与する部分はどこか？

RQ3.2 提案手法は暗黙的フィードバックにおけるTop-N推薦において既存のstate-of-the-artの手法を上回るか？

RQ3.3 提案手法は不確実性を扱えているか？

5.4.1 実験設定

データセット

検証には一般公開されている4つのデータセットを用いる: MovieLens 1M (ml-1m)[1], LastFM (lastfm)[2], Amazon music(AMusic)[3], and Amazon toys (AToy)[3]. ml-1mとlastfmは1ユーザーあたり最低20個の履歴があるユーザーのみが含まれるように事前に加工され公開されている．一方AMusicでは59%のユーザーが，ATоyでは49%のユーザーが20より少ない履歴しかもっていない．そのような点で後者二つはより疎なデータセットであるといえる．それぞれの統計情報について表5.1に載せる．

Top-N 推薦

最初に暗黙的フィードバックにおけるTop-N推薦の定式化を行う．先行研究 [79] の設定に基づき，M人のユーザーとN個のアイテムからインタラクション行列 $\mathbf{Y} \in \mathbb{R}^{M\times N}$ をユーザーの暗黙的フィードバック，例えばクリックや視聴，収集，購買の履歴から構築する．

[1] https://grouplens.org/datasets/movielens/
[2] http://ocelma.net/MusicRecommendationDataset/
[3] http://jmcauley.ucsd.edu/data/amazon/

表 5.1: データセットの統計情報

統計量	ml-1m	lastfm	AMusic	AToy
ユーザー数	6040	1741	1776	3137
アイテム数	3706	2665	12929	33953
評価数	1000209	69149	46087	84642
スパース度	0.9553	0.9851	0.9980	0.9992

$$y_{ui} = \begin{cases} 1, & \text{フィードバックが観測された場合} \\ 0, & \text{それ以外} \end{cases} \tag{5.27}$$

$y_{ui} = 1$ は，ユーザーからアイテムへの暗黙的フィードバックが観測されたことを意味する． 暗黙的フィードバックを用いた推薦は， $\mathbf{Y}$ における未観測の要素の値を推定し，アイテムのランキングに用いることで定式化される．

上記の問題設定は One-Class Collaborative Filtering (OCCF) 問題として知られており，本研究では [113, 114] らのアプローチに則る．すなわち観測されていないユーザーアイテムの組み合わせの一部をサンプリング (negative sampling) し， 二値分類として定式化する．しかしただ分類問題を解くだけではランキング化が行えず，各 y_{ui} にベルヌーイ分布を仮定し， 確率的にモデリングすることでその尤度を元にランキングを構築する．

$$\begin{aligned} P\left(y_{ui} = k | p_{ui}\right) &= \begin{cases} 1 - p_{ui}, & k = 0 \\ p_{ui}, & k = 1 \end{cases} \\ &= p_{ui}^{k} \left(1 - p_{ui}\right)^{1-k} \end{aligned} \tag{5.28}$$

p_{ui} は $y_{ui} = 1$ となる確率を意味する．そのため p_{ui} は同時にどれだけ u が i とマッチするかの確信度としても解釈することができる．言い換えると p_{ui} が 1 に近いということは完璧に u が i を高い確信度でマッチすると予測しており，また 0.5 は最もモデルの推論として確信を持てていないと解釈できる．

Netflix Prize [115] に端を発し，協調フィルタリングの初期の研究はユーザーの付与した星の数といった明示的なフィードバックに着目し，回帰タスク (rating prediction) として定式化されていた．しかし近年の研究により，同タスクで高性能なモデルで

あっても，ユーザーの趣向の高さ順にアイテムを並び替えて推薦する Top-N 推薦タスクでは必ずしも良い性能を示さないということが明らかになり [116]， よりランキングタスクでの評価を呼びかける機運が高まっている．また [117] によると，単純な Bayesian MF [118] などのベースラインモデルを注意深くチューニングした結果，近年提案されたモデルのいずれもそれのスコアを超えることができなかったと報告され，rating prediction における評価の難しさが近年挙げられている．本研究ではこれらの流れを踏まえ，Top-N 推薦タスクでの評価を用いた有効性の検証を行った．

評価指標

ランキングの評価には二つの代表的な指標 Hit Ratio (HR) と Normalized Discounted Cumulative Gain (NDCG) [95] を用いる．

$$HR@k = \begin{cases} 1, \text{ if } r_{\text{test}}(u,i) \in R_k \\ 0, \quad \text{otherwise} \end{cases} \tag{5.29}$$

$$NDCG@k = \frac{DCG@k}{IDCG@k} = \sum_{j=1}^{k} \frac{2^{r(u,j)} - 1}{\log(i+1)} \tag{5.30}$$

R_k は確信度に基づく上位 k 件のリストである．また $r(u,j)$ は j 番目の推薦アイテムが $r_{\text{test}}(u,i)$ と一致する場合 1, それ以外は 0 を表す．

直感的に HR は上位 k 番目までにインタラクションしたアイテムが存在するか，NDCG はどれだけその中でも上位にあるかを示した指標である．先行研究 [80] に則り，両指標における評価は上位 10 件で評価を行う．

先行研究 [80, 79] に則り，leave-one-out 評価を行う; 全部のアイテムでのランキング評価は時間的コストがとても大きいので，ユーザーごとに最新のインタラクションしたアイテム $r_{\text{test}}(u,i)$ とランダムに選択した観測されていないアイテム 100 件を用いランキングを構築する．

学習

モデルの最適化には Adam [89] を用いたミニバッチ学習で行う．バッチサイズは 256 に，学習率は 0.001 に固定した．モデルのパラメータは平均 0，分散 0.01 の正規分布からのサンプリングで初期化を行った．negative sampling の数は 4 とし，毎エポックごとに更新する．検証用データにはユーザー毎にランダムに 1 件のインタラクションを用い，エポック数のハイパーパラメータチューニングを行った．

表 5.2: 事前学習の有無に伴う MetaCFNet の性能比較.

Datasets	Without pre-training		With pre-training	
	HR	NDCG	HR	NDCG
ml-1m	0.6033	0.3474	0.7280	0.4353
lastfm	0.7025	0.4306	0.8949	0.6074
AMusic	0.2302	0.1193	0.5557	0.3083
AToy	0.3009	0.1614	0.6063	0.3405

事前学習

先行研究より学習済みのモデルの重みを用いて初期化を行うことで最終的なスコアの向上と収束が早くなることが知られている [79, 80]．本研究では事前学習した MetaCF-rl と MetaCF-ml を用いて MetaCFNet の重みの初期化を行った．MetaCF-rl と MetaCF-ml は Adam を用いて未学習の状態から学習を行い，MetaCFNet は確率的勾配降下法 (SGD) を用いて最適化を行った．Adam は以前のモメンタムといった情報が必要になってしまうため事前学習後には用いなかった．

比較手法

比較手法は以下である．

- **CFNet-rl**：表現学習の機構を用いた手法である．MetaCF-rl とネットワークの機構は同じであり，唯一の違いは MetaCF の枠組みの下で学習させたか否かである．
- **CFNet-ml**：マッチング学習の機構を用いた手法である．MetaCF-ml とネットワークの機構は同じであり，唯一の違いは MetaCF の枠組みの下で学習させたか否かである．
- **CFNet**： CFNet-rl と CFNet-ml の機構を組み合わせた state-of-the-art の手法である．MetaCFNet とネットワークの機構は同じであり，唯一の違いは MetaCF の枠組みの下で学習させたか否かである．

5.4.2 学習機構による感度性 (RQ3.1)

RQ3.1 に答える形で以下 3 つの設定に対する性能の変化を確認する．第一に事前学習の有無，第二に MetaCF サンプリングをユーザーのみアイテムのみもしくは両方から行った場合，第三に表現学習とマッチング学習の比較について行う．

最初に MetaCFNet における事前学習の有無を比較した結果を表 5.2 に載せる．事前学習を行った MetaCFNet はしなかった場合に比べて大きな性能の向上を果たした．しかし事前学習を行わなかった MetaCFNet は後述の MetaCF-rl と MetaCF-ml よりも結果が悪くなった．このことより事前学習が有効であるということだけでなく，ネットワークの機構によっては MetaCF サンプリングの効果にばらつきがあることが確認された．考察としては MetaCFNet は二股構造をしており，より明示的に条件付き表現 $\mathbf{r}$ を学習させる機構の方が望ましかった可能性がある．

第二に MetaCF サンプリングをユーザーのみから行った場合，アイテムのみから行った場合，両方から行った場合を比較してどれが最終的な結果に貢献したかを確認する．結果を表 5.3 に載せる．結果として両方から行った場合が最も結果が良くなりやすく，ユーザーからのみ行った場合がアイテムからのみ行うよりも効果的であることが確認された．それは AMusic や AToy データセットといったコールドなユーザーを多く含んでいる場合で顕著であった．

第三に表 5.3 と表 5.4 に表現学習とマッチング学習との比較を行った結果を載せる．すべての場合においてマッチング学習が上回る結果となった．しかし両者を事前学習して組み合わせることで性能は向上することも確認された．これらのことより条件付き表現を得る上で early ステージでのフュージョンが late ステージでのフュージョンよりも効果的であることが確認された．

5.4.3 総合評価 (RQ3.2)

RQ3.2 に答えるため， MetaCFNet をベースライン手法と比較した結果を表 5.5 に載せる．一番性能の良かった手法と二番目の手法を太字で表現している． 結果として MetaCFNet は ml-1m と lastfm においては CFNet とほぼ同等の結果を示し，AMusic と AToy では最大約 40%の大きな性能の向上を示した．前者はコールドユーザーを含まないような前処理が行われているが，後者は多くのコールドユーザーを含むという点において，本提案手法が疎なデータセットからの学習に強みを示しており，ml-1m や lastfm といった事前に履歴が豊富になるに従い CFNet と同等の結果を示すのは仮説どおりの結果といえる．

表 5.3: 学習機構と MetaCF サンプリングに基づく HR@10 の結果．太字は各行で最も性能のよかったものを表す．

Datasets	Models	MetaCF Sampling		
		Users	Items	Both
ml-1m	MetaCF-rl	**0.5692**	0.5510	0.5374
	MetaCF-ml	**0.7179**	0.6995	0.7142
	MetaCFNet	0.7206	0.7005	**0.7280**
lastfm	MetaCF-rl	**0.8569**	0.8530	0.8507
	MetaCF-ml	0.8845	0.8736	**0.8926**
	MetaCFNet	0.8943	0.8845	**0.8949**
AMusic	MetaCF-rl	0.4865	0.3778	**0.5051**
	MetaCF-ml	0.5462	0.4144	**0.5456**
	MetaCFNet	0.5546	0.4245	**0.5557**
AToy	MetaCF-rl	0.5202	0.3522	**0.5413**
	MetaCF-ml	0.5757	0.3828	**0.6025**
	MetaCFNet	0.5990	0.3848	**0.6063**

表 5.4: 学習機構と MetaCF サンプリングに基づく NDCG@10 の結果．各行ごとに最大性能を太字で表す．

Datasets	Models	MetaCF Sampling		
		Users	Items	Both
ml-1m	MetaCF-rl	**0.3155**	0.3062	0.2967
	MetaCF-ml	**0.4340**	0.4260	0.4311
	MetaCFNet	**0.4390**	0.4190	0.4353
lastfm	MetaCF-rl	0.5579	**0.5667**	0.5629
	MetaCF-ml	0.5994	0.5827	**0.6047**
	MetaCFNet	0.5996	0.5952	**0.6074**
AMusic	MetaCF-rl	0.2691	0.2310	**0.2753**
	MetaCF-ml	0.3022	0.2561	**0.3027**
	MetaCFNet	0.3049	0.2673	**0.3083**
AToy	MetaCF-rl	0.2885	0.2063	**0.2988**
	MetaCF-ml	0.3182	0.2343	**0.3346**
	MetaCFNet	0.3343	0.2482	**0.3405**

また MetaCF sampling の有用性について，疎なデータセットにおいては，どのネットワーク機構についても大幅に性能の向上に寄与することが確認された．さらに，ml-1m と lastfm では，マッチング学習の機構で性能の向上が確認されたが，表現学習の機構では性能が低下した．

5.4.4 不確実性の可視化と性能面との比較 (RQ3.3)

RQ3.3 に答えるために我々の提案手法を情報理論の観点から検証を行う．ϕ はベルヌーイ分布のパラメータをモデリングしており，それ自体が確信度を表していると解釈する．そこで不確実性の評価を式 5.31 に示す二値エントロピー関数 H で表す．同式は $\phi = 0.5$ のとき最大値 1.0 を出力し，$\phi = 0$ もしくは 1 のとき最低値の 0 を出力する．

$$\mathrm{H}(\phi) = -\phi \log_2 \phi - (1-\phi) \log_2 (1-\phi) \tag{5.31}$$

ここでは一例として最も性能が向上したデータセット AToy を用いて以下の可視化を行う．図 5.2 は MetaCF-ml と CFNet-ml それぞれにおける出力のヒストグラムである．両者ネットワーク機構で唯一の違いは提案手法である MetaCF サンプリングで学習を行ったか否かである．CFNet-ml はほとんどの出力が 0 か 1 に偏っているのに対し，MetaCF-ml はより広範囲に広がっていることが視覚的に確認できる．このことより本提案手法はより明示的に不確実性をモデリングできている手法といえる．

次に不確実性とユーザーとアイテムにおけるスパースさとの関連を調べる．図 5.3 (a) にユーザーの履歴数 (横軸) とアイテムの履歴数 (縦軸) ごとのエントロピーの変化のヒートマップを示す．それぞれのマスは CFNet-ml と比較したエントロピーの増大量の平均値に基づいて色分けを行なっている．図よりユーザー及びアイテムが持つ過去のインタラクションの履歴が少ない場合に，不確実性が高くなることが明らかとなった．

最後にスコアの向上とユーザー・アイテム間のスパースさとの関連を調べる．図 5.3 (b) のヒートマップは CFNet-ml と MetaCF-ml 間の HR@10 の向上に基づいて表現している．図より提案手法を用いることで，アイテムの履歴が少ない場合でも高いスコアを得ることができることが確認された．

表 5.5: NDCG@10 と HR@10 による結果の比較．ベースライン手法は [80] から引用．各データセットの評価指標ごとに，一番性能の良かった手法と二番目の手法を太字で表現している．

Datasets	Measures	Existing Methods			MetaCF			Improvement
		CFNet-rl	CFNet-ml	CFnet	MetaCF-rl	MetaCF-ml	MetaCFNet	from CFnet
ml-1m	HR	0.7127	0.7073	**0.7253**	0.5354	0.7142	**0.7280**	0.37%
	NDCG	0.4336	0.4264	**0.4416**	0.2960	0.4339	**0.4353**	-1.4%
lastfm	HR	0.8840	0.8834	**0.8995**	0.8489	0.8925	**0.8949**	-0.5 %
	NDCG	0.6001	0.5919	**0.6186**	0.5585	0.6046	**0.6074**	-1.8%
AMusic	HR	0.3947	0.4071	0.4116	0.5051	**0.5456**	**0.5557**	35.0%
	NDCG	0.2504	0.2420	0.2601	0.2753	**0.3027**	**0.3083**	18.5%
AToy	HR	0.3746	0.3931	0.4150	0.5413	**0.6025**	**0.6063**	44.9%
	NDCG	0.2271	0.2293	0.2513	0.2988	**0.3346**	**0.3405**	35.1%

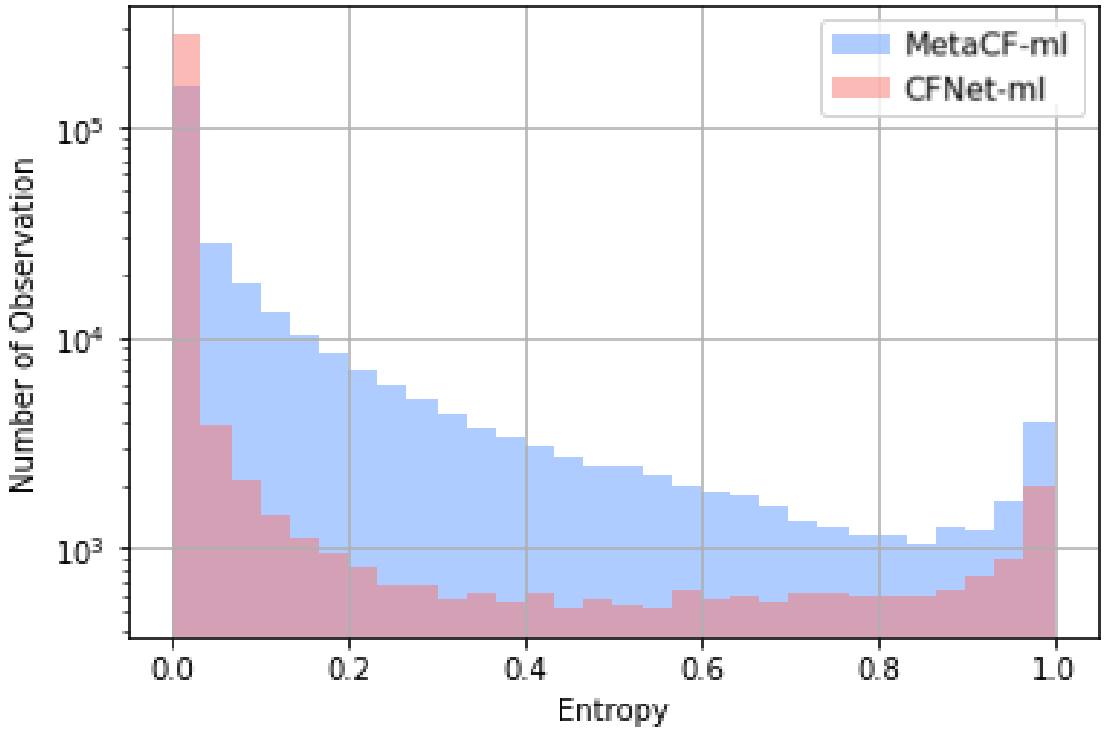
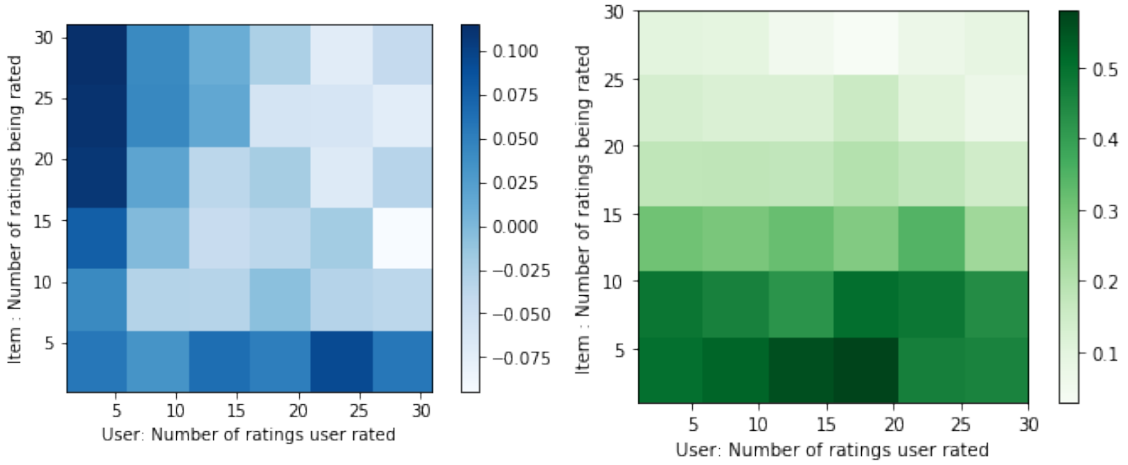


図 5.2: MetaCF-ml と CFNet-ml のおけるエントロピーのヒストグラム (福馬らの論文図 2(a) より [105]



(a) (b)

図 5.3: (a) CFNet-ml から MetaCF-ml へのエントロピーの増大に伴うヒートマップでの可視化 (b) HR@10 における性能の向上に基づくヒートマップ (福馬らの論文図 2(b)(c) より [105])

5.5 まとめ

本章のまとめとして，主な貢献は以下である．

- 本研究ではメタラーニングに基づいた発想より，協調フィルタリングの手法を拡張し，疎なデータからの学習を効率化し，モデルの出力における不確実性を考慮する MetaCF を提唱した．
- 近年提案された CFNet [80] に MetaCF を適用した MetaCFNet というネットワークを提案した．Top-N 推薦のタスクにおいて，4 つのデータセットにおいて優れた結果を出し，とりわけ疎なデータセットにおいては最大約 40%の大きな性能向上を示し，実験的に有用性を検証した．
- 実験的に過去履歴の少ないユーザーやアイテムに対する推論には高い不確実性を持って出力を示していることを確認し，またそれらが最終的なベンチマークにおけるスコアの向上に貢献していることを確認した．

5.6 今後の方向性

今後の研究の方針としては以下が考えられる．第一に MetaCF sampling とネットワーク機構の相性に関する理論的考察である．本稿ではマッチング学習についての適用が，表現学習に比べ性能の大幅な向上を示した．またそれらを両方用いた CFNet の機構については，事前学習を行わない場合，性能の劣化が確認された．これら二つの原因の解明は今後の課題として扱う．第二にコンテンツデータといった外部データを用いる場合の提案手法への組み込み方の検討がある．疎なデータに対する既存の試みとしてはコンテンツベースフィルタリングが挙げられる．それらとの組み合わせは今後の推薦技術の向上に重要であると考えられる．第三に提案手法によって得られた不確実性を能動学習における探索と活用のバランス調整に利用し，性能向上やフィルターバブルの解消への応用可能性の検証が考えられる．

第6章　結論

6.1　個々研究のまとめ

本研究ではオンラインプラットフォームにおけるコンテンツの質の評価を，それらに対する意見 (e.g., クリック，星) から，様々なバイアスの影響を考慮することで，より正確に推定する試みを行った．第一部では集合知に基づく試みのように，一元的な尺度でのコンテンツの質の推定を扱った．第二部は推薦システムのように，多元的なコンテンツの質の推定，つまり人によってコンテンツの質は異なるという状況下のもとで個々人に対するコンテンツの質の推定を扱った．

本論文は以下の Research Question に答える形で行われた．

Research Question 1：　複数の認知バイアス情報の影響を受けているユーザーの行動データから,「認知バイアス情報が無ければどのように行動したのか」という反実仮想のもと，それら影響を取り除いた真のコンテンツの質を測定できるか？

本研究では，まず「いいね」や「役に立った」などの投票形式の評価行動に焦点を当て，事前のバイアスが及ぼす影響について明らかにすることを目的に提案を行った．またそれら情報がなければ人はどのように評価していたのかといった反事仮想な現象に対する推定技術の提案について述べた．

これらの背景として，まず多くの既存研究ではユーザーによるクリックといったフィードバックをコンテンツの価値としてシステムを最適化していた．一方それらのデータは様々な認知バイアス (e.g. 評判バイアス, 社会的影響力バイアス, 位置バイアス) の影響を受けるため，実際に人気と質とは解離が起きていることが知られている．

既存の Unbiased Learning-to-Rank 手法は主にユーザーとコンテンツの接触頻度しか扱えていなかった．それら手法を拡張し，複数の認知バイアスにユーザーが晒されている場合におけるユーザーの投票形式の評価行動におけるモデリングの提案と，同モデリングについて有効な最適化手法の提案を行った．

提案手法では，まずコンテンツの評判スコアのような認知的要因の有無をユーザに提示した対照実験から収集したデータセットを用いて，投票行動の違いに着目した．次に，人間の投票行動の概念を定式化し，認知的要因を入力した場合の違いを説明するモデルを提案した．具体的には，グラフニューラルネットワークを用いて，周囲の情報によってコンテンツがどのように有益に見えるようになるかを定量化した．さらに，認知バイアスの影響を割り引いて，偏りのない有用性を推定する戦略を設計した．最後に，実世界の Stack Exchange のデータセットを用いて，我々の手法の有効性を実証的に示した．また学習したモデルについて解釈を行った結果，特に他者からの事前評価がユーザーの投票行動に影響を与える傾向が確認された

Research Question 2：　ユーザーのレビュー投票行動から不良ユーザー・レビューを発見し，信頼度に基づいた意見の集約によって有識者に近い集合知を獲得できるか？

もし仮にバイアスのかかっていない評価を推定できたとしても，それらを単に平均などを集約させることは必ずしも良い集合知が取れるとは限らない．近年では，やらせレビューやアンチレビューのようにそもそも公平に評価をつけていない人が存在し，それらは集合知の前提を満たさない．

そこでレビューの評価ネットワークにおけるリンク構造に基づき，レビューの信頼度を推論する REV2 を用いることで，それら信頼性に基づく集約方法がより有識者に近い集合知をもたらすかについて検証を行った．そこで我々は東京都内のラーメン屋のレビューを独自に収集し，有識者のレビューとの比較を行った．その結果レビューの信頼性に基づく意見の集約が平均などのベースライン手法と比較してより有識者の感覚と近い結果が得られることを確認した．これらの結果より信頼度が低いユーザーのフィードバックは集合知に含めないほうが良い結果が得られることが示唆された．また信頼性に基づく集約と平均とのランキングに基づく比較を行った．その結果評価が高まった店舗は信頼性の高い高評価と，信頼性の低い低評価 (アンチレビュー) が多かったことが示された．反対に低くなった店舗は信頼性の低い高評価 (やらせレビュー) と，信頼性の高い低評価が多かったことが示された．最後に公平性が低いと REV2 アルゴリズムで示されるたユーザーについての解釈性の検証を行った．その結果「周りの人との意見のずれが大きい」,「低評価を付けやすい」といった傾向が確認された．

Research Question 3：　推薦システムは履歴の少ないユーザーやアイテムの組

み合わせについても正しくそれぞれの間の嗜好度をモデリングすることは可能か？

Research Question 4： 推薦システムは予測の不確実性をモデリングすることは可能か？

推薦システムなどで用いられる Matrix Factorization に代表される行列因子分解モデルには大きく二つの欠点が存在した．

一つ目は十分にユーザー・アイテムの嗜好度に関する履歴が手に入る場合には上手くモデリングができる一方で，疎なデータ，つまりユーザー・アイテムの嗜好度に関する履歴がほとんど得られていないケースではそれらの関連性を十分にモデリングできないという問題がある．一般的な学習データは，人気度に基づいて観測される頻度に差があり，ほとんどのアイテムが少ない履歴しか保持していないため大きな問題となる．

もう一点は予測の不確実性が行えない点が挙げられる．つまり推薦システムが予測するユーザーのアイテムに対する予測について，それらがどれほど自信のある推論かが分からないという問題がある．

そこで本研究では，既存の潜在因子に基づく協調フィルタリング手法について，疎なデータからの学習と不確実性のモデリングを同時に可能にする汎用的な学習フレームワーク MetaCF を提案した．提案手法は推薦システムをメタラーニングの一種である Neural Processes の観点から定義し直すことで，従来のモデルにほとんど手を加えることなく，それらの予測を可能にした．極めて疎なデータからは提案手法を用いることで従来手法と比較して最大 40%近い性能向上を示した．また特に，ユーザーまたはアイテムの履歴数が少ないペアに対して，高い不確実性を予測するといった感覚的にも一致する結果を実験的に示した．本研究により，より高性能にユーザーとアイテムの嗜好度のモデリングが可能になった．加えて今後は不確実性に基づく能動的学習により学習効率の高い能動学習が可能になると期待される．

6.2 本論文のまとめ

本研究ではオンラインプラットフォームにおけるコンテンツの質の評価を，様々なバイアスの影響を考慮することで，より正確に推定するという試みを行った．コンテンツの質を正しく推定できることは，プラットフォーム運営者にとってユーザーに有益な情報を提供しやすくなり，情報過多の解決のために重要である．オンライ

ンプラットフォームには様々なバイアスが内在しており，既存のコンテンツの質の自動評価技術をナイーブに適用させるだけでは，真にコンテンツの質を表現または推定ができているとは言い難い．

我々の研究では，一元的なコンテンツの質の評価 (一位から最下位までコンテンツの質を並び替える) と多元的な評価 (個人ごとにそれぞれのコンテンツに興味があるかを予測) を対象に行った．

前者における既存の課題として二つを取り上げた．一つ目は認知バイアス情報によって個人の評価傾向が歪んでしまい，正しく評価を行えていないこと．二つ目は個人の意見を集約の際にやらせやアンチといった意見によって評価が歪められることを取り上げた．

後者における課題も二つ本論文では扱った．一つ目は履歴数が少ないユーザーやアイテムに関してはそれらに関するまたは対する嗜好度の予測性能が大幅に低下するということ．二つ目は個人とアイテムとの嗜好度のモデリングに関し，予測に不確実性を持たすことができない点を取り上げた．

我々の提案手法・分析によって，以下のことが解決・明らかになった．1) 事前バイアスがなければ人はどのように評価したか推定することで．コンテンツの質を認知バイアスの影響を取り除いた形で推定が可能になった．2) コンテンツの質を測るにあたり，個人の意見の集約は，評価者の過去の評価履歴などから判断した信頼性に基づいた形で行うことでより有識者に近い意見が得られるようになった．3) 推薦システムにおいて疎なデータからの学習が可能になった．4) 推薦システムの予測に対して確信度をモデリングすることが可能になった．

6.3 今後の課題

今後の課題と発展性として，各章で述べた今後の方向性を 1) 提案手法の融合，2) オンラインでの評価，3) 他ドメインの適用に分類する．

手法の融合

それぞれの提案手法は，今後組み合わせることでさらなる検証が考えられる．例えば第 3 章と第 4 章のアイディアのように，認知バイアスを取り除いた上での信頼度に基づく集合知の検証が考えられる．

他にも第 3 章のアイディアを第 5 章のような個々人ベースに適応することが挙げられる．すなわち認知バイアスによる影響は個人によって異なり，それらをモデリ

ングすることで，より高度な推薦システムの開発可能性が考えられる．

また第 4 章で信頼性の低いレビューの意見を集合知に用いないことで全体の質が向上したことから，第 5 章のように推薦システムにおける学習時の重みとして用いることが可能である．協調フィルタリングの損失関数として，フィードバックを信頼度で重み付けて学習することは，傾向スコアによって重み付け流ことと非常に類似しており，一部のスパムレビュワーの意見を軽減した形での推薦システムの構築が考えられる．

オンラインでの評価

ここでのオンラインでの評価とは実サービスに実装し，ユーザーの反応を見ながらモデルを評価することを指す．特に第 3，4 章にて提案した一元的なコンテンツの質の自動評価技術はオフラインデータでの評価には限界があり，説得力のある検証を行うには実サービスに実装し，オンライン評価を行うことが考えられる．

また第 5 章で述べた提案モデルによる予測における不確実性の推論について，これらが果たして能動学習を行う際，探索と活用について有効で，早期学習段階でのの性能向上に貢献できるかはオンライン評価を行う必要がある．

他ドメインへの転用

本研究における提案手法の汎用性を確認するために他ドメインへの転用が考えられる．例えば第 3 章で行った提案手法は Q&A サイトにおける投票形式のフィードバックを用いたが，今後はレビューサイトにおける星といった数的なフィードバックへの転用も考えられる．

しかしこれら他ドメインへの転用を考える上での大きなハードルとして，第 3 章で用いたようなユーザーの評価時の認知バイアス情報は一般にデータセットとして公開されていない点が挙げられる．多くのレビューデータセットでは，ある期間内で収集された結果のみを公開しており，完全に本実験を行えるような形で公開されているデータセットは我々が調べた限りでは Stack Exchange 以外見つけることはできなかった．そのため今後はオンラインプラットフォームを運営している企業や研究者との連携により，これら情報を含むデータを大量かつ様々なドメインについて収集し，提案手法の応用と検証が考えられる．

また他ドメインの転用という観点では，本研究で扱わなかった SNS などが考えられる．例えば第 3 章と第 5 章と関連しユーザーが好むコンテンツの推定を，既存の

いいね数やリツイート者の評判と行った認知バイアスを除いた形で推定するなどが考えられる．また昨今 SNS などではフェイクニュースが頻繁に拡散されており，コンテンツとバイアス情報，それらが情報拡散にあたえる影響の分析などはフェイクニュースの拡散を防ぐ上でも有意義である．

最後に本研究で分析されたバイアスを取り除いた形でのコンテンツの表示，推薦を行うオンラインプラットフォームを，現実のウェブサービスとして構築することが今後の課題である．

謝辞

本研究は東京大学大学院工学系研究科システム創成学専攻における博士論文研究として行ったものです．論文作成にあたりましては，先生方をはじめ様々な方のご協力を賜りました．指導教員の鳥海先生には，修士から含めて5年間にわたってご指導賜りました．研究の仕方，発表の仕方，論文の書き方から私生活に至るまで，様々な姿勢や考え方の指導を手厚くしていただきました．先生と出会った当初，何もできなかった自分が，博士論文の執筆に至る事が出来たのは，先生の熱心な指導の賜物だと思います．鳥海先生には深く感謝を申し上げます．また鳥海研究室のすべての方々からいつも刺激を受けて楽しく研究することができました．ここに深く御礼申し上げます．大澤先生，島田先生，合田先生，山崎先生には，論文の審査にあたり大変お世話になりました．様々な議論をする事によって，論文を深める事につながりました．大変ご多忙にもかかわらず論文の副査を担当していただき，本当に有難うございます．

また大学学部時代から続けている競技ダンスという芸術スポーツにおける経験は，本研究における課題意識を形成する最大の要因でした．第3章の背景動機としては，既に勝っている選手が常に勝ち続けている現状を見て感じたことに由来しています．第4章の背景動機は複数人での審査で，明らかにおかしな審査員が紛れ込んでいたことで涙を飲んだ先輩を見ていたためです．第5章の背景動機としては，自分が出たコンペティションで海外の国際的に有名な先生だけが自分に票をくれたという経験から，もし今回もあの人が審査していたらどうなっていただろうという思考が元になっています．これら経験は研究課題につながっており，これらを考えるきっかけになった方達にも感謝申し上げます．

博士課程に在籍しながらもアマチュア競技生活を続けられたのは数多くの人たちのおかげです．卒部してからも仲良くして励ましてくれる競技ダンス部時代の友人達，アマチュア全日本チームのメンバー達，いつも熱心に指導してくださる先生方，ここに深く感謝申し上げます．

並びにこれまでTDAI Labに携わってくれたメンバーにも感謝申し上げます．修士課程の時に起業を経験し，経営しながら博士課程を続けられたのは，メンバーが非常に優秀でいてくれたおかげです．日々の業務内での議論を通じて受ける刺激は

少なからず，自分の研究マインドに大きな影響を与えています．心より感謝申し上げます．毎週の勉強会での論文発表は自分も学ぶことが多く，論文や発表の仕方など非常に学ぶことが多かったです．業務を通じて得られたものは自分の成長にとって欠かせないものばかりでした．起業を勧めていただいた鳥海先生には改めて感謝申し上げます．

最後に，一番近くで常に自分を支えてくれた妻と，長い間サポートし続けてくれた両親にも深く感謝を申し上げます．

参考文献

[1] 将吾 石田 and 信夫 河口. 個人を見守るサーバシステムに基づくコンテキストアウェアな情報提示手法. **マルチメディア、分散、協調とモバイル** *(DICOMO)* **シンポジウム論文集**, pages 1162–1170, 2009.

[2] 経済産業省. 令和元年度電子商取引に関する市場調査, 2020. `https://www.meti.go.jp/policy/it_policy/statistics/outlook/r1_kohyoshiryo.pdf`.

[3] PowerReviews. The growing power of review. White Paper, 2018.

[4] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab, November 1999.

[5] Luís Borges, Bruno Martins, and Pável Calado. Combining similarity features and deep representation learning for stance detection in the context of checking fake news. *J. Data and Information Quality*, 11(3), June 2019.

[6] Nicolas De Condorcet et al. *Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix*. Cambridge University Press, 2014.

[7] F Galton. Vox populi [electronic version]. *URL: http://galton. org/cgi—bin/searchlmages/galton/search/essays/pages/galton—1907—vox—populi_l. htm (Stand: 19.06. 2012)*, 1907.

[8] Yinfei Yang, Yaowei Yan, Minghui Qiu, and Forrest Bao. Semantic analysis and helpfulness prediction of text for online product reviews. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 38–44, 2015.

[9] Brian Dolhansky, Russ Howes, Ben Pflaum, Nicole Baram, and Cristian Canton Ferrer. The deepfake detection challenge (dfdc) preview dataset, 2019.

[10] Malay Haldar, Mustafa Abdool, Prashant Ramanathan, Tao Xu, Shulin Yang, Huizhong Duan, Qing Zhang, Nick Barrow-Williams, Bradley C. Turnbull, Brendan M. Collins, and Thomas Legrand. Applying deep learning to airbnb search. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery Data Mining*, KDD 2019, page 1927–1935, 2019.

[11] Lionel Martin and Pearl Pu. Prediction of helpful reviews using emotions extraction. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*, AAAI 2014, page 1551–1557, 2014.

[12] Yehuda Koren, Robert Bell, and Chris Volinsky. Matrix factorization techniques for recommender systems. ieee, computer journal, 42(8), 30-37. *Computer*, 42:30 – 37, 09 2009.

[13] 神嶌 敏弘（産業技術総合研究所）. 公平性に配慮した学習とその理論的課題. 機械学習と公平性に関するシンポジウム, 2020.

[14] Eduardo Graells-Garrido, Mounia Lalmas, and Filippo Menczer. First women, second sex: Gender bias in wikipedia. In *Proceedings of the 26th ACM Conference on Hypertext amp; Social Media*, HT 2015, page 165–174, 2015.

[15] Ricardo Baeza-Yates. Bias on the web. *Communications of the ACM*, 61(6):54–61, May 2018.

[16] Jingjing Liu, Yunbo Cao, Chin-Yew Lin, Yalou Huang, and Ming Zhou. Low-quality product review detection in opinion summarization. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 334–342, 2007.

[17] Everdell I. Maynes, R. The evolution of google search results pages and their effects on user behaviour. Technical report, Mediative., 2014.

[18] Gourab K Patro, Arpita Biswas, Niloy Ganguly, Krishna P. Gummadi, and Abhijnan Chakraborty. Fairrec: Two-sided fairness for personalized recom-

mendations in two-sided platforms. In *Proceedings of The Web Conference 2020*, WWW 2020, page 1194–1204, 2020.

[19] Eli Pariser. *The filter bubble : what the Internet is hiding from you.* Penguin Press, New York, 2011.

[20] Ashton Anderson, Lucas Maystre, Ian Anderson, Rishabh Mehrotra, and Mounia Lalmas. Algorithmic effects on the diversity of consumption on spotify. In *Proceedings of The Web Conference 2020*, WWW 2020, page 2155–2165, 2020.

[21] Kathleen Hall Jamieson and Joseph N Cappella. *Echo chamber: Rush Limbaugh and the conservative media establishment.* Oxford University Press, 2008.

[22] Cristian Danescu-Niculescu-Mizil, Gueorgi Kossinets, Jon Kleinberg, and Lillian Lee. How opinions are received by online communities: A case study on amazon.com helpfulness votes. In *Proceedings of the 18th International Conference on World Wide Web*, 2009.

[23] Oren Tsur and Ari Rappoport. Revrank: A fully unsupervised algorithm for selecting the most helpful book reviews, 2009.

[24] A. Ghose and P. G. Ipeirotis. Estimating the helpfulness and economic impact of product reviews: Mining text and reviewer characteristics. *IEEE Transactions on Knowledge and Data Engineering*, 23(10):1498–1512, 2011.

[25] Bee-Chung Chen, Jian Guo, Belle Tseng, and Jie Yang. User reputation in a comment rating environment. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD 2011, page 159–167, 2011.

[26] Xiaoyan Qiu, Diego Oliveira, Alireza Shirazi, Alessandro Flammini, and Filippo Menczer. Limited individual attention and online virality of low-quality information. *Nature Human Behaviour*, 1:0132, 06 2017.

[27] Keith Burghardt, Emanuel F. Alsina, Michelle Girvan, William Rand, and Kristina Lerman. The myopia of crowds: Cognitive load and collective evaluation of answers on stack exchange. *PLOS ONE*, 12(3):1–19, 03 2017.

[28] Srijan Kumar, Bryan Hooi, Disha Makhija, Mohit Kumar, Christos Faloutsos, and V.S. Subrahmanian. Rev2: Fraudulent user prediction in rating platforms. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, WSDM 2018, page 333–341, 2018.

[29] Ralf HJM Kurvers, Stefan M Herzog, Ralph Hertwig, Jens Krause, Patricia A Carney, Andy Bogart, Giuseppe Argenziano, Iris Zalaudek, and Max Wolf. Boosting medical diagnostics by pooling independent judgments. *Proceedings of the National Academy of Sciences*, 113(31):8777–8782, 2016.

[30] Russ Ray. Prediction markets and the financial "wisdom of crowds". *The Journal of Behavioral Finance*, 7(1):2–4, 2006.

[31] James Surowiecki. *The Wisdom of Crowds*. Anchor, 2005.

[32] Jan Lorenz, Heiko Rauhut, Frank Schweitzer, and Dirk Helbing. How social influence can undermine the wisdom of crowd effect. *Proceedings of the National Academy of Sciences*, 108(22):9020–9025, 2011.

[33] Serguei Kaniovski and Alexander Zaigraev. Optimal jury design for homogeneous juries with correlated votes. *Theory and Decision*, 71(4):439–459, 2011.

[34] Maria Glenski and Tim Weninger. Rating effects on social news posts and comments. *ACM Trans. Intell. Syst. Technol.*, 8(6), 2017.

[35] Jiahua Du, Jia Rong, Sandra Michalska, Hua Wang, and Yanchun Zhang. Feature selection for helpfulness prediction of online product reviews: An empirical study. *PLOS ONE*, 14(12):1–26, 12 2019.

[36] Peter Lofgren, Siddhartha Banerjee, and Ashish Goel. Personalized pagerank estimation and search: A bidirectional approach. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*, WSDM 2016, page 163–172, 2016.

[37] Taher H. Haveliwala. Topic-sensitive pagerank. In *Proceedings of the 11th International Conference on World Wide Web*, WWW 2002, page 517–526, 2002.

[38] Rada Mihalcea. Graph-based ranking algorithms for sentence extraction, applied to text summarization. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 170–173, 2004.

[39] Yunbo Cao, Jun Xu, Tie-Yan Liu, Hang Li, Yalou Huang, and Hsiao-Wuen Hon. Adapting ranking svm to document retrieval. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR 2006, page 186–193, 2006.

[40] Christopher J. C. Burges. From RankNet to LambdaRank to LambdaMART: An overview. Technical report, Microsoft Research, 2010.

[41] Thorsten Joachims, Adith Swaminathan, and Tobias Schnabel. Unbiased learning-to-rank with biased feedback. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, WSDM 2017, page 781–789, 2017.

[42] Xuanhui Wang, Nadav Golbandi, Michael Bendersky, Donald Metzler, and Marc Najork. Position bias estimation for unbiased learning to rank in personal search. In *Proceedings of the 11th ACM International Conference on Web Search and Data Mining (WSDM)*, pages 610–618, 2018.

[43] Aman Agarwal, Xuanhui Wang, Cheng Li, Michael Bendersky, and Marc Najork. Addressing trust bias for unbiased learning-to-rank. In *The World Wide Web Conference*, WWW 2019, page 4–14, 2019.

[44] Huan Sun, Alex Morales, and Xifeng Yan. Synthetic review spamming and defense. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD 2013, page 1088–1096, 2013.

[45] Amanda J. Minnich, Nikan Chavoshi, Abdullah Mueen, Shuang Luan, and Michalis Faloutsos. Trueview: Harnessing the power of multiple review sites. In *Proceedings of the 24th International Conference on World Wide Web*, WWW 2015, page 787–797, 2015.

[46] Huayi Li, Geli Fei, Shuai Wang, Bing Liu, Weixiang Shao, Arjun Mukherjee, and Jidong Shao. Bimodal distribution and co-bursting in review spam detection. In *Proceedings of the 26th International Conference on World Wide Web*, WWW 2017, page 1063–1072, 2017.

[47] Leman Akoglu, Rishi Chandy, and Christos Faloutsos. Opinion fraud detection in online reviews by network effects, 2013.

[48] G. Wang, S. Xie, B. Liu, and P. S. Yu. Review graph based online store review spammer detection. In *2011 IEEE 11th International Conference on Data Mining*, pages 1242–1247, 2011.

[49] Saptarshi Ghosh, Bimal Viswanath, Farshad Kooti, Naveen Kumar Sharma, Gautam Korlam, Fabricio Benevenuto, Niloy Ganguly, and Krishna Phani Gummadi. Understanding and combating link farming in the twitter social network. In *Proceedings of the 21st International Conference on World Wide Web*, WWW 2012, page 61–70, 2012.

[50] Meng Jiang, Peng Cui, Alex Beutel, Christos Faloutsos, and Shiqiang Yang. Catchsync: Catching synchronized behavior in large directed graphs. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD 2014, page 941–950, 2014.

[51] Jing Wang, Anindya Ghose, and Panos Ipeirotis. Bonus, disclosure, and choice: What motivates the creation of high-quality paid reviews? In *Proceedings of the International Conference on Information Systems, ICIS 2012, Orlando, Florida, USA, December 16-19, 2012*, 2012.

[52] Leman Akoglu, Hanghang Tong, and Danai Koutra. Graph based anomaly detection and description: A survey. *Data Min. Knowl. Discov.*, 29(3):626–688, May 2015.

[53] Amanda J. Minnich, Nikan Chavoshi, Abdullah Mueen, Shuang Luan, and Michalis Faloutsos. Trueview: Harnessing the power of multiple review sites. In *Proceedings of the 24th International Conference on World Wide Web*, WWW 2015, page 787–797, Republic and Canton of Geneva, CHE, 2015.

[54] Guangyu Wu, Derek Greene, and Pádraig Cunningham. Merging multiple criteria to identify suspicious reviews. In *Proceedings of the Fourth ACM Conference on Recommender Systems*, RecSys 2010, page 241–244, 2010.

[55] Sihong Xie, Guan Wang, Shuyang Lin, and Philip Yu. Review spam detection via temporal pattern discovery. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 08 2012.

[56] Amir Fayazi, Kyumin Lee, James Caverlee, and Anna Squicciarini. Uncovering crowdsourced manipulation of online reviews. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR 2015, page 233–242, New York, NY, USA, 2015.

[57] Vlad Sandulescu and Martin Ester. Detecting singleton review spammers using semantic similarity. In *Proceedings of the 24th International Conference on World Wide Web*, WWW 2015 Companion, page 971–976, 2015.

[58] Arjun Mukherjee, Vivek Venkataraman, Bing Liu, and Natalie Glance. What yelp fake review filter might be doing? *Proceedings of the 7th International Conference on Weblogs and Social Media, ICWSM 2013*.

[59] Ee-Peng Lim, Viet-An Nguyen, Nitin Jindal, Bing Liu, and Hady Wirawan Lauw. Detecting product review spammers using rating behaviors. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, CIKM 2010, page 939–948, 2010.

[60] Sihong Xie, Guan Wang, Shuyang Lin, and Philip S. Yu. Review spam detection via temporal pattern discovery. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD 2012, page 823–831, 2012.

[61] Bimal Viswanath, Muhammad Ahmad Bashir, Muhammad Bilal Zafar, Simon Bouget, Saikat Guha, Krishna P. Gummadi, Aniket Kate, and Alan Mislove. Strength in numbers: Robust tamper detection in crowd computations. In *Proceedings of the 2015 ACM on Conference on Online Social Networks*, COSN 2015, page 113–124, 2015.

[62] Amanda J. Minnich, Nikan Chavoshi, Abdullah Mueen, Shuang Luan, and Michalis Faloutsos. Trueview: Harnessing the power of multiple review sites. In *Proceedings of the 24th International Conference on World Wide Web*, WWW 2015, page 787–797, 2015.

[63] Amir Fayazi, Kyumin Lee, James Caverlee, and Anna Squicciarini. Uncovering crowdsourced manipulation of online reviews. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR 2015, page 233–242, 2015.

[64] Shebuti Rayana and Leman Akoglu. Collective opinion spam detection: Bridging review networks and metadata. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD 2015, page 985–994, 2015.

[65] Bryan Hooi, Neil Shah, Alex Beutel, Stephan Günnemann, Leman Akoglu, Mohit Kumar, Disha Makhija, and Christos Faloutsos. Birdnest: Bayesian inference for ratings-fraud detection. pages 495–503, 2016.

[66] Cheng Chen, Kui Wu, Venkatesh Srinivasan, and Xudong Zhang. Battling the internet water army: Detection of hidden paid posters. In *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, ASONAM 2013, page 116–120, 2013.

[67] Bimal Viswanath, Muhammad Ahmad Bashir, Muhammad Bilal Zafar, Simon Bouget, Saikat Guha, Krishna P. Gummadi, Aniket Kate, and Alan Mislove. Strength in numbers: Robust tamper detection in crowd computations. In *Proceedings of the 2015 ACM on Conference on Online Social Networks*, COSN 2015, page 113–124, 2015.

[68] M. Jiang, P. Cui, and C. Faloutsos. Suspicious behavior detection: Current trends and future directions. *IEEE Intelligent Systems*, 31(1):31–39, 2016.

[69] Gerardo Ocampo Diaz and Vincent Ng. Modeling and prediction of online product review helpfulness: A survey. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 698–708, 2018.

[70] Cen Chen, Minghui Qiu, Yinfei Yang, Jun Zhou, Jun Huang, Xiaolong Li, and Forrest Sheng Bao. Multi-domain gated cnn for review helpfulness prediction. In *The World Wide Web Conference*, WWW 2019, page 2630–2636, 2019.

[71] Shuzhe Xu, Salvador Barbosa, and Don Hong. *BERT Feature Based Model for Predicting the Helpfulness Scores of Online Customers Reviews*, pages 270–281. 02 2020.

[72] Lotte M. Willemsen, Peter C. Neijens, Fred Bronner, and Jan A. de Ridder. “highly recommended!” the content characteristics and perceived usefulness

of online consumer reviews. *Journal of Computer-Mediated Communication*, 17(1):19–38, 2011.

[73] Kevin K. Y. Kuan, Kai Lung Hui, Pattarawan Prasarnphanich, and Hok-Yin Lai. What makes a review voted? an empirical investigation of review voting in online review systems. *Journal of the Association for Information Systems*, 16(1):1, 2015.

[74] Yi-Hsiu Cheng and Hui-Yi Ho. Social influence's impact on reader perceptions of online reviews. *Journal of Business Research*, 68, 04 2015.

[75] Ya-Han Hu and Kuanchin Chen. Predicting hotel review helpfulness: The impact of review visibility, and interaction between hotel stars and review ratings. *International Journal of Information Management*, 36(6, Part A):929 – 944, 2016.

[76] Jorge E. Fresneda and David Gefen. A semantic measure of online review helpfulness and the importance of message entropy. *Decision Support Systems*, 125:113117, 2019.

[77] Pasquale Lops, Marco de Gemmis, and Giovanni Semeraro. *Content-based Recommender Systems: State of the Art and Trends*, pages 73–105. 01 2011.

[78] Hong-Jian Xue, Xinyu Dai, Jianbing Zhang, Shujian Huang, and Jiajun Chen. Deep matrix factorization models for recommender systems. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pages 3203–3209, 2017.

[79] 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*, WWW 2017, pages 173–182, 2017.

[80] Zhi-Hong Deng, Ling Huang, Chang-Dong Wang, Jian-Huang Lai, and Philip S. Yu. Deepcf: A unified framework of representation learning and matching function learning in recommender system. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):61–68, Jul. 2019.

[81] Himan Abdollahpouri, Masoud Mansoury, Robin Burke, and Bamshad Mobasher. The impact of popularity bias on fairness and calibration in recommendation. *CoRR*, abs/1910.05755, 2019.

[82] Tobias Schnabel, Adith Swaminathan, Ashudeep Singh, Navin Chandak, and Thorsten Joachims. Recommendations as treatments: Debiasing learning and evaluation. volume 48 of *Proceedings of Machine Learning Research*, pages 1670–1679, 2016.

[83] David H. Stern, Ralf Herbrich, and Thore Graepel. Matchbox: Large scale online bayesian recommendations. In *Proceedings of the 18th International Conference on World Wide Web*, WWW 2009, pages 111–120, 2009.

[84] Xuan Nhat Lam, Thuc Vu, Trong Duc Le, and Anh Duc Duong. Addressing cold-start problem in recommendation systems. In *Proceedings of the 2Nd International Conference on Ubiquitous Information Management and Communication*, ICUIMC 2008, pages 208–211, 2008.

[85] Minmin Chen, Alex Beutel, Paul Covington, Sagar Jain, Francois Belletti, and Ed H. Chi. Top-k off-policy correction for a reinforce recommender system. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, WSDM 2019, page 456–464, 2019.

[86] Fernando Amat, Ashok Chandrashekar, Tony Jebara, and Justin Basilico. Artwork personalization at netflix. In *Proceedings of the 12th ACM Conference on Recommender Systems*, RecSys 2018, page 487–488, 2018.

[87] Himel Dev, Karrie Karahalios, and Hari Sundaram. Quantifying voter biases in online platforms: An instrumental variable approach. *Proc. ACM Hum.-Comput. Interact.*, 3(CSCW), November 2019.

[88] Pascal Van Hentenryck, Andrés Abeliuk, Franco Berbeglia, Felipe Maldonado, and Gerardo Berbeglia. Aligning popularity and quality in online cultural markets. *Proceedings of the International AAAI Conference on Web and Social Media*, 10(1), Mar. 2016.

[89] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference*

on Learning Representations, ICLR 2015, , Conference Track Proceedings, 2015.

[90] Keith Burghardt, Tad Hogg, and Kristina Lerman. Quantifying the impact of cognitive biases in question-answering systems. In *Proceedings of the Twelfth International Conference on Web and Social Media, ICWSM*, pages 568–571, 2018.

[91] D.W. Scott and John Wiley & Sons. *Multivariate Density Estimation: Theory, Practice, and Visualization*. A Wiley-interscience publication. Wiley, 1992.

[92] Samuel Maurus and Claudia Plant. Skinny-dip: Clustering in a sea of noise. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD 2016, page 1055–1064, 2016.

[93] Benoit Mandelbrot. Contributions to probability and statistics: Essays in honor of harold hotelling (ingram olkin, sudhist g. ghurye, wassily hoeffding, william g. madow, and henry b. mann, eds.). 3(1):278–292, 1961.

[94] G. E. P. Box and D. R. Cox. An analysis of transformations. *Journal of the Royal Statistical Society. Series B (Methodological)*, 26(2):211–252, 1964.

[95] Kalervo Järvelin and Jaana Kekäläinen. Cumulated gain-based evaluation of ir techniques. *ACM Trans. Inf. Syst.*, 20(4):422–446, October 2002.

[96] Richard J. Bolton and David J. H. Statistical fraud detection: A review. *Statistical Science*, 17:2002, 2002.

[97] ミシュランガイド東京 *2019*. 日本ミシュランタイヤ, 2018.

[98] Qi-Ying Hu, Zhi-Lin Zhao, Chang-Dong Wang, and Jian-Huang Lai. An item orientated recommendation algorithm from the multi-view perspective. *Neurocomput.*, 269(C):261–272, December 2017.

[99] Shuhui Jiang, Zhengming Ding, and Yun Fu. Deep low-rank sparse collective factorization for cross-domain recommendation. In *Proceedings of the 25th ACM International Conference on Multimedia*, pages 163–171, 2017.

[100] Xiang Wang, Xiangnan He, Fuli Feng, Liqiang Nie, and Tat-Seng Chua. Tem: Tree-enhanced embedding model for explainable recommendation. In *Proceedings of the 2018 World Wide Web Conference*, WWW 2018, pages 1543–1552, 2018.

[101] David D. Lewis and William A. Gale. A sequential algorithm for training text classifiers. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR 1994, page 3–12, 1994.

[102] Dingyuan Zhu, Peng Cui, Daixin Wang, and Wenwu Zhu. Deep variational network embedding in wasserstein space. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD 2018, pages 2827–2836, 2018.

[103] Marta Garnelo, Dan Rosenbaum, Christopher Maddison, Tiago Ramalho, David Saxton, Murray Shanahan, Yee Whye Teh, Danilo Rezende, and S. M. Ali Eslami. Conditional neural processes. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1704–1713, 2018.

[104] Hyunjik Kim, Andriy Mnih, Jonathan Schwarz, Marta Garnelo, S. M. Ali Eslami, Dan Rosenbaum, Oriol Vinyals, and Yee Whye Teh. Attentive neural processes. *ArXiv*, abs/1901.05761, 2019.

[105] 福馬 智生 and 鳥海 不二夫. 協調フィルタリングにおけるメタラーニングの適用による疎なデータからの学習と不確実性の推論. **人工知能学会論文誌**, 35(5):p. F–JC3_1–9, 2020.

[106] Timothy Hospedales, Antreas Antoniou, Paul Micaelli, and Amos Storkey. Meta-Learning in Neural Networks: A Survey. *arXiv e-prints*, page arXiv:2004.05439, April 2020.

[107] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks, 2017.

[108] Carl Edward Rasmussen. Gaussian processes for machine learning. 2006.

[109] Manasi Vartak, Arvind Thiagarajan, Conrado Miranda, Jeshua Bratman, and Hugo Larochelle. A meta-learning perspective on cold-start recommendations for items. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 6904–6914. 2017.

[110] Fei Chen, Zhenhua Dong, Zhenguo Li, and Xiuqiang He. Federated meta-learning for recommendation. *CoRR*, abs/1802.07876, 2018.

[111] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.*, 15(1):1929–1958, January 2014.

[112] François Chollet et al. Keras. `https://keras.io`, 2015.

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

[114] Rong Pan, Yunhong Zhou, Bin Cao, Nathan N. Liu, Rajan Lukose, Martin Scholz, and Qiang Yang. One-class collaborative filtering. In *Proceedings of the 2008 Eighth IEEE International Conference on Data Mining*, ICDM 2008, pages 502–511, 2008.

[115] James Bennett, Charles Elkan, Bing Liu, Padhraic Smyth, and Domonkos Tikk. Kdd cup and workshop 2007. *SIGKDD Explor. Newsl.*, 9(2):51–52, December 2007.

[116] Paolo Cremonesi, Yehuda Koren, and Roberto Turrin. Performance of recommender algorithms on top-n recommendation tasks. In *Proceedings of the Fourth ACM Conference on Recommender Systems*, RecSys 2010, pages 39–46, 2010.

[117] Steffen Rendle, Li Zhang, and Yehuda Koren. On the difficulty of evaluating baselines: A study on recommender systems. *CoRR*, abs/1905.01395, 2019.

[118] Ruslan Salakhutdinov and Andriy Mnih. Bayesian probabilistic matrix factorization using markov chain monte carlo. In *Proceedings of the 25th International Conference on Machine Learning*, ICML 2008, pages 880–887, 2008.