推薦システムにおける人気バイアスを考慮したランキング評価指標

飯塚洸二郎† 久保 光証† 森田 一†

† 株式会社 Gunosy 〒 150-6139 東京都渋谷区渋谷 2-24-12 渋谷スクランブルスクエア 39 階
E-mail: †{kojiro.iizuka,hajime.morita,hajime.morita}@gunosy.com

あらまし 本論文では，推薦システムにおける人気バイアスを考慮したランキング評価指標を提案する．既存の推薦システムの評価ではユーザーのクリックの有無に応じて評価を行うことが一般的である一方で，提案する評価指標は，クリックしたアイテムの人気度合いに応じて重み付けを行う．これによって，精度と人気バイアスを同時に考慮しながら推薦モデルの学習を行うことを目指した．実験では，様々な推薦モデルで学習を行い，提案指標と既存指標を比較した．その結果，提案指標は，既存の精度指標に対して強い相関を持ち，かつ既存の人気バイアス指標に対しても強い相関を持つことを確認した．さらに，提案指標を学習過程に組み込むことで，既存の精度指標の毀損を抑えながら，人気バイアスを軽減できることを確認した．

キーワード 推薦システム，評価指標，人気バイアス

1 はじめに

推薦システムは多くのウェブサービスにおいて，ユーザーが求めるアイテムを発見するための重要な役割を担っている．この推薦システムに対して，オフラインでの評価やユーザーを介したオンラインでの評価が日々行われている．その中で，ユーザーの満足度やサービスの売上に影響を及ぼす要因として，推薦するアイテムの人気度合いが関連することが分かってきた．例えば，他のアイテムに比べて人気の低いロングテールアイテムは，ユーザーがそのアイテムをまだ知らない可能性が高いという点で，価値が高いと考えられている [1] [2]．また，ロングテールアイテムは，ユーザーの趣向を深く理解するために役立ち [3]，売上の増加というサービス側の利益につながるといった報告がある [4]．

このアイテムの人気度合いの偏りとして定義される人気バイアスの研究としては，人気バイアスを軽減させる研究と人気バイアスを評価する研究がある．人気バイアスを軽減させる手法としては，人気バイアスに対する正則化項を導入する手法 [5] [6] [7] [8] やリランキングを行う手法 [9] などがある．人気バイアスを評価する指標としては，推薦したアイテムの平均人気度である Average Recommendation Popularity (ARP) [10]，推薦したロングテールアイテムの割合である Average Percentage of Long Tail Items (APLT) [5]，人気アイテムの偏り具合を示すジニ係数，カタログアイテムカバレッジ [11] がある．しかしながら，これらの既存研究では，ユーザーにとって提示されたアイテムがユーザーの趣向にあうアイテムかという推薦精度に関する視点が欠けている課題があった．その逆に，既存の推薦精度を測る指標では，人気バイアスが考慮できない課題があった．

そこで本研究では，人気バイアスを考慮した上で，推薦精度を測定するためのランキング評価指標を提案する．推薦精度を評価するランキング評価指標は様々あるが，本研究ではランキング評価指標を，ユーザーがランキングから得られる利得の期待値の形で扱う．推薦システムの精度に関する評価を行う際には，利得はユーザーがアイテムをクリックしたか否かのまたは 1 で表現することが一般的である．本研究では，ランキング評価指標において，利得について適切に重み付けを行うことで，人気バイアスと推薦精度を同時に考慮する指標を提案する．

具体的には，我々はユーザーのクリックの有無に加えて，アイテムの人気度合いに応じた重みを考慮した利得を導入する．ここで，あるアイテムの人気度合いは，そのアイテムがクリックされた回数を全アイテムのクリック総数で割った値で定義する．このように人気バイアスと推薦精度を同時に一つの指標として表現することは，人気バイアスと精度を分けて個別に最適化を行い結果を一つにまとめる場合に比べて，両者の精度を高く保つことにつながることがある．

実験では，様々な推薦アルゴリズムを用いて，既存の評価指標と提案指標の比較実験を 2 つ行った．1 つ目の実験は，提案指標が既存の評価指標についてどれほど相関があるかについて評価した．結果として，提案指標は，既存のランキング指標に対して強い相関を持ち，かつ既存の人気バイアス指標に対しても強い相関を持つことがわかった．2 つ目の実験は，推薦アルゴリズムの学習において，提案手法を組み込むことで，どれほど精度と人気バイアスに影響するかを評価した．結果としては，推薦精度の毀損を抑えながら，人気バイアスが軽減できることを確認した．

本研究の貢献は以下のとおりである．

- 人気バイアスと推薦精度を同時に評価するための指標を提案した．
- 数値実験の結果，提案指標は，既存の精度指標と人気バイアス指標について強い相関を持つことがわかった．
- 提案指標を学習に組み込むことで，推薦精度の毀損を抑えながら，人気バイアスが軽減できることを確認した．

本論文の構成は下記のとおりである．第二章では関連研究について述べる．第三章では，指標の提案を行う．第四章では，

数値実験の設定と実験結果を述べる．最後に，第五章で本論文のまとめを行う．

2 関 連 研 究

2.1 推薦システムの評価指標

推薦システムでの評価に関連して，古くから情報検索の分野で数多くの評価に関する研究が行われてきた．特に，ランキング上でのユーザーのクリック行動のモデルは，クリックモデル [12] と呼ばれ研究されている．代表的な評価指標としては，Mean Average Precision (MAP)，Mean Reciprocal Rank (MRR) [13]，Normalized Discounted Cumulative Gainn (nDCG) [14] などがある．これらの指標は指標の内部に，クエリに対する文書の適合度や，ユーザーがランキングから得られる利益である利得に関する項が含まれている．多様性や複数の検索意図を考慮した利得を用いる指標には α-nDCG [15] や D-nDCG [16] などが知られている．一方で，推薦システムの精度の評価を行う場合は，クリックの有無によってユーザーの利得を定義することが一般的である．本研究では，ランキング評価指標における利得について，従来のように単にクリックの有無の 2 値で扱うのではなく，アイテムの人気度合いを考慮した評価指標を提案する．

2.2 人気バイアス

人気バイアスに関する研究には，人気バイアスを軽減させる研究と人気バイアスを評価する研究がある．人気バイアスを軽減させる手法としては，人気バイアスに対する正則化項を導入する手法 [5] [6] [7] [8] や敵対的学習を行う手法 [17]，因果推論を活用する手法 [18] [19] [20] [21]，リランキングを行う手法 [9] がある．人気バイアスを評価する指標としては，推薦したアイテムの平均人気度である Average Recommendation Popularity (ARP) [10]，推薦したロングテールアイテムの割合である Average Percentage of Long Tail Items (APLT) [5]，人気アイテムの偏り具合を示すジニ係数，カタログアイテムカバレッジ [11] がある．しかしながら，これらの既存研究では，ユーザーにとって提示されたアイテムがユーザーの趣向にあうアイテムかという推薦精度に関する視点が欠けている課題があった．本研究では，人気バイアスと精度の両方考慮する評価指標を提案し，エンドツーエンドで両側面の最適化を行うことを目指す．

3 指標の設計

推薦システムの精度を測定するためのランキング評価指標は MAP， MRR [13]， nDCG [14] など様々ある．これらの指標について，nDCG を除く代表的なランキング評価指標は下記のように Normalised Cumulative Utility (NCU) [22] としてユーザーの利得の期待値の形として一般化できる．

$$NCU = \sum_{r=1}^{n} P(r)G(r)$$

表 1　ml-100k の統計情報

ユーザー数	944
アイテム数	1683
ユーザーとアイテムの接触数	100,000

前提として，ユーザーはランキングの上位から下位に向かってアイテムを閲覧していき，アイテムに満足した時点でランキング上でのアイテムの閲覧をやめるユーザーの母集団を考える．ここで，$P(r)$ は，ユーザーがランキングの r 番目のアイテムに満足して停止する確率を表し，$G(r)$ は r 番目のアイテムから得られる利得を表している．

推薦システムの評価においては，ユーザーがランキングのアイテムをクリックするか否かの 0,1 によって利得を表現することが一般的である．本研究では，利得 $G(r)$ を下記によって定義する．

$$G(r) = C(r)/B(r),$$

ここで，$C(r) \in \{0, 1\}$ は r 番目のアイテムがクリックされるか否かの 0,1 を表し，$B(r)$ は r 番目のアイテムの人気バイアスを表す．本研究では，人気バイアス項である $B(r)$ は，データセット内の全アイテムの総クリック回数を T,r 番目のアイテムの総クリック回数を $N(r)$ として，$B(r) = N(r)/T$ で定義する．本研究では，このように定義した利得についての NCU を特に Popularity-aware Ranking Metrics (PRM) と呼ぶ．

4 実　　験

4.1 データセット

実験には，Movielens のデータセットである ml-100k を用いた．ml-100k の統計情報は表 1 の通りである．Movielens のデータセットはユーザーがアイテムに対して評価値を与えるデータセットである．本研究ではユーザーがアイテムに評価値を与えた場合，ユーザーがアイテムに対してクリックを行ったとみなす．学習データとテストデータの分割に関しては，学習: 検証: テスト=8:1:1 になるようにランダムに分割した．

4.2 評 価 指 標

評価指標には，精度を測る指標と人気バイアスを測る指標を用いた．ここでの各指標は，ランキング上位 k 番目までのアイテムから定義されたものを用いる．

4.2.1 精 度 指 標

精度指標の定義はそれぞれ下記の通りである．

$$MRR@k = \frac{1}{|U|} \sum_{u \in U} \frac{1}{\mathrm{rank}_u^*},$$

ここで，$rank_u^*$ はユーザー u がランキングの上位 k 位までで，クリックしたアイテムの位置が最上位の位置を表す．

$$HIT@k = \frac{1}{|U|} \sum_{u \in U} \min(\sum_{r=1}^{k} G_u(r), 1),$$

表 2　各モデルに対する推薦精度指標と人気バイアスの指標の値

モデル名	PRM@10	MRR@10	HIT@10	nDCG@10	ItemCoverage@10	AvgPop@10	Gini@10	TailPercent@10
Pop	46.90	0.1951	0.4698	0.1034	0.0362	315.3	**0.9886**	0.0000
BPR	189.2	0.4970	0.7911	0.2956	0.3381	222.1	0.8968	0.2183
LightGCN	196.6	0.5030	0.8049	0.3049	0.3518	215.2	0.8875	**0.2449**
RecVAE	**200.4**	**0.5303**	**0.8165**	**0.3260**	0.3446	215.4	0.8826	0.2281
SimpleX	162.6	0.4756	0.7667	0.2835	0.2644	228.8	0.9253	0.1334
ItemKNN	165.5	0.4623	0.7847	0.2834	0.2513	217.7	0.9135	0.1551
SpectralCF	57.44	0.2674	0.4995	0.1334	0.0392	**338.7**	0.9875	0.0000
GCMC	189.6	0.4720	0.7943	0.2922	**0.3832**	220.8	0.8836	0.2401

ここで，$G_u(r)$ は評価対象のランキングの r 番目のアイテムがユーザー u にクリックされるかの 0,1 を表す．

$$nDCG@k = \frac{1}{|U|}\sum_{u\in U}\frac{\sum_{r=1}^{k} G_u(r)/\log(r+1)}{\sum_{r=1}^{k} G_u^*(r)/\log(r+1)},$$

ここで，$G_u(r)$,$G_u^*(r)$ は評価対象のランキングおよび理想的なランキングの r 番目のアイテムがユーザー u にクリックされるかの 0,1 を表す．

4.2.2　人気バイアス指標

人気バイアス指標の定義はそれぞれ下記の通りである．

$$ItemCoverage@k = \frac{\left|\bigcup_{u\in U} R_u\right|}{|I|},$$

ここで，R_u はユーザー u に提示するランキングのアイテム集合を表す．

$$AvgPop@K = \frac{1}{|U|}\sum_{u\in U}\frac{\sum_{i\in R_u} N(i)}{k},$$

ここで $N(i)$ は，学習データ中に含まれるアイテム i の総クリック回数である．

$$TailPercentage@k = \frac{1}{|U|}\sum_{u\in U}\frac{\sum_{i\in R_u}\delta(i\in T)}{k},$$

ここで，R_u はユーザー u に提示するランキング，δ は指示関数，T はロングテールアイテムの集合を表す．本研究でのロングテールアイテムは，クリック回数の多い順にアイテムを並べたときの下位 80%のアイテムとした．

$$Gini@k = \frac{\sum_{i=1}^{|I|}(2i-|I|-1)P(i)}{|I|\sum_{i=1}^{|I|}P(i)},$$

ここで，I は全アイテムの集合を表し，$P(i)$ はアイテム i が推薦ランキングに含まれる回数を表し，$P(i)\leqq P(i+1)$ となるように並び替えられているものとする．

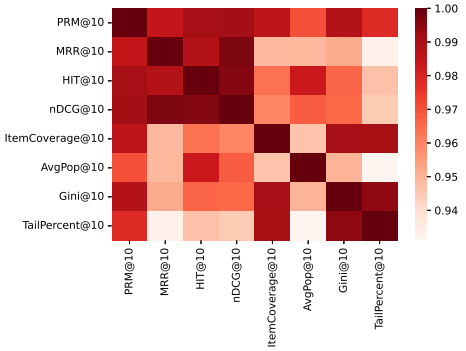

図 1　各指標間の相関値ヒートマップ

4.3　推薦モデル

実験には，ユーザーとアイテムの接触のみから構築できるモデルを用いた．クリック回数の多い人気順に推薦する Pop，過去に接触したアイテムの近傍アイテムを推薦する ItemKNN [23]，Bayesian Personalized Ranking (BPR) [24] を始めとして，近年開発された推薦モデルである LightGCN [25]，RecVAE [26]，SimpleX [27]，SpectralCF [28]，GCMC [29] を実験に用いた．実装は Recbole [30] を利用した．本研究は，各推薦モデルについての精度を各々詳細に議論することが目的ではないため，モデルのパラメータは全てデフォルトのものを用いた．

4.4　結　　果

表 2 に各モデルに対する評価指標の値を示した．単に人気順に推薦を行う Pop に比べて，その他の各種法は精度が高くなっており，正しく学習が行えていることが確認できる．

図 1 に各指標間の相関値をヒートマップとして示した．なおここでは，AvgPop@k と Gini@k については ItemCoverage@k と TailPercent@k と同様に値が大きいほうがバイアスが小さくなるように，符号を反転させた結果を載せている．提案指標である PRM は，全ての指標に対して 0.97 以上の値を取っている．この結果に対して，精度指標は人気バイアス指標に対して相関値が相対的に低くなっており，人気バイアス指標は精度指標に対して相関値が相対的に低くなっている．このように，提案指標は，精度指標と人気バイアス指標両方に対して最も強く相関しているといえる．

表 3 に PRM を学習に活用した場合の各指標の増減値を示す．ここで，PRM を学習に活用するとは，検証データに対し

表 3　PRM を学習に活用した場合の各指標の増減値

モデル名	PRM@10	MRR@10	HIT@10	nDCG@10	ItemCoverage@10	AvgPop@10	Gini@10	TailPercentage@10
BPR	8.589	-0.0159	-0.0085	-0.0095	**0.0653**	-14.99	-0.0278	**0.0645**
LightGCN	**9.346**	-0.0010	-0.0064	-0.0029	0.0392	-7.309	-0.0157	0.0330
RecVAE	1.243	-0.0021	0.0053	-0.0011	0.0089	**0.2152**	0.0004	0.0056
SimpleX	6.587	0.0030	**0.0223**	**0.0067**	0.0125	-2.448	-0.0040	0.0199
GCMC	-4.166	**0.0000**	-0.0064	-0.0094	-0.0255	-0.0012	**0.0078**	-0.0091

て，PRM 値を算出し，パラメータの更新を行うか否かを決定することを意味する．この検証は各学習エポックごとに行った．この結果は，PRM に増減があったモデルの結果のみを表している．PRM が増加したモデルは，人気バイアス指標がバイアスを小さくする方向に改善していることから，人気バイアスを軽減できている．精度指標に関しては精度指標が多少毀損していることが見て取れる一方で，その減少割合は人気バイアスの改善具合に比べて小さい．このように，PRM を学習過程に組み込むことで，既存のランキング指標の毀損を抑えながら，人気バイアスを軽減できることを確認した

5 ま と め

本論文では，推薦システムにおける人気バイアスを考慮したランキング評価指標を提案した．実験では，様々な推薦モデルで学習を行い，提案指標と既存指標を比較した．その結果，提案指標は，既存のランキング指標に対して強い相関を持ち，かつ既存の人気バイアス指標に対しても強い相関を持つことを確認した．さらに，提案指標を学習過程に組み込むことで，既存のランキング指標の毀損を抑えながら，人気バイアスを軽減できることを確認した．

今後は，実験のデータの種類と規模を増やすこと，学習の損失項自体に提案手法を組みより直接的に最適化を行うことが課題である．さらに，バイアスを軽減する手法との比較実験を行い，提案指標が満たすべき性質をいくつかの公理 (axioms) に基づいて説明することを目指す．

文　献

[1] Chris Anderson. The long tail: Why the future of business is selling less of more. Hachette UK, 2006.

[2] Guy Shani and Asela Gunawardana. Evaluating recommendation systems. In Recommender systems handbook, pp. 257–297. Springer, 2011.

[3] Himan Abdollahpouri, Masoud Mansoury, Robin Burke, Bamshad Mobasher, and Edward Malthouse. User-centered evaluation of popularity bias in recommender systems. In Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization, pp. 119–129, 2021.

[4] Erik Brynjolfsson, Yu Jeffrey Hu, and Michael D Smith. From niches to riches: Anatomy of the long tail. Sloan management review, Vol. 47, No. 4, pp. 67–71, 2006.

[5] Himan Abdollahpouri, Robin Burke, and Bamshad Mobasher. Controlling popularity bias in learning-to-rank recommendation. In Proceedings of the eleventh ACM conference on recommender systems, pp. 42–46, 2017.

[6] Zhihong Chen, Rong Xiao, Chenliang Li, Gangfeng Ye, Haochuan Sun, and Hongbo Deng. Esam: Discriminative domain adaptation with non-displayed items to improve long-tail performance. In Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 579–588, 2020.

[7] Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. Correcting popularity bias by enhancing recommendation neutrality. In RecSys Posters, 2014.

[8] Ziwei Zhu, Yun He, Xing Zhao, Yin Zhang, Jianling Wang, and James Caverlee. Popularity-opportunity bias in collaborative filtering. In Proceedings of the 14th ACM International Conference on Web Search and Data Mining, pp. 85–93, 2021.

[9] Himan Abdollahpouri, Robin Burke, and Bamshad Mobasher. Managing popularity bias in recommender systems with personalized re-ranking. In The thirty-second international flairs conference, 2019.

[10] Hongzhi Yin, Bin Cui, Jing Li, Junjie Yao, and Chen Chen. Challenging the long tail recommendation. arXiv preprint arXiv:1205.6700, 2012.

[11] Mouzhi Ge, Carla Delgado-Battenfeld, and Dietmar Jannach. Beyond accuracy: evaluating recommender systems by coverage and serendipity. In Proceedings of the fourth ACM conference on Recommender systems, pp. 257–260, 2010.

[12] Aleksandr Chuklin, Ilya Markov, and Maarten de Rijke. Click models for web search. Synthesis lectures on information concepts, retrieval, and services, Vol. 7, No. 3, pp. 1–115, 2015.

[13] Ellen M Voorhees, et al. Overview of the trec 2001 question answering track. In Trec, pp. 42–51, 2001.

[14] Kalervo Järvelin and Jaana Kekäläinen. Ir evaluation methods for retrieving highly relevant documents. In ACM SIGIR Forum, Vol. 51, pp. 243–250. ACM New York, NY, USA, 2017.

[15] Charles LA Clarke, Maheedhar Kolla, Gordon V Cormack, Olga Vechtomova, Azin Ashkan, Stefan Büttcher, and Ian MacKinnon. Novelty and diversity in information retrieval evaluation. In Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval, pp. 659–666, 2008.

[16] Tetsuya Sakai and Ruihua Song. Evaluating diversified search results using per-intent graded relevance. In Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval, pp. 1043–1052, 2011.

[17] Adit Krishnan, Ashish Sharma, Aravind Sankar, and Hari Sundaram. An adversarial approach to improve long-tail performance in neural collaborative filtering. In Proceedings of the 27th ACM International Conference on information and knowledge management, pp. 1491–1494, 2018.

[18] Wenjie Wang, Fuli Feng, Xiangnan He, Xiang Wang, and Tat-Seng Chua. Deconfounded recommendation for alleviating bias amplification. In Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining, pp. 1717–1725, 2021.

[19] Tianxin Wei, Fuli Feng, Jiawei Chen, Ziwei Wu, Jinfeng Yi, and Xiangnan He. Model-agnostic counterfactual reason-

ing for eliminating popularity bias in recommender system. In Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining, pp. 1791–1800, 2021.

[20] Yang Zhang, Fuli Feng, Xiangnan He, Tianxin Wei, Chonggang Song, Guohui Ling, and Yongdong Zhang. Causal intervention for leveraging popularity bias in recommendation. In Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 11–20, 2021.

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

[22] Tetsuya Sakai, Stephen Robertson, and I Newswatch. Modelling a user population for designing information retrieval metrics. In EVIA@ NTCIR, 2008.

[23] Mukund Deshpande and George Karypis. Item-based top-n recommendation algorithms. ACM Transactions on Information Systems (TOIS), Vol. 22, No. 1, pp. 143–177, 2004.

[24] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. Bpr: Bayesian personalized ranking from implicit feedback. In Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence, pp. 452–461, 2009.

[25] Xiangnan He, Kuan Deng, Xiang Wang, Yan Li, Yongdong Zhang, and Meng Wang. Lightgcn: Simplifying and powering graph convolution network for recommendation. In Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, pp. 639–648, 2020.

[26] Ilya Shenbin, Anton Alekseev, Elena Tutubalina, Valentin Malykh, and Sergey I Nikolenko. Recvae: A new variational autoencoder for top-n recommendations with implicit feedback. In Proceedings of the 13th International Conference on Web Search and Data Mining, pp. 528–536, 2020.

[27] Kelong Mao, Jieming Zhu, Jinpeng Wang, Quanyu Dai, Zhenhua Dong, Xi Xiao, and Xiuqiang He. Simplex: A simple and strong baseline for collaborative filtering. In Proceedings of the 30th ACM International Conference on Information & Knowledge Management, pp. 1243–1252, 2021.

[28] Lei Zheng, Chun-Ta Lu, Fei Jiang, Jiawei Zhang, and Philip S Yu. Spectral collaborative filtering. In Proceedings of the 12th ACM conference on recommender systems, pp. 311–319, 2018.

[29] Rianne van den Berg, Thomas N Kipf, and Max Welling. Graph convolutional matrix completion. arXiv preprint arXiv:1706.02263, 2017.

[30] Wayne Xin Zhao, Shanlei Mu, Yupeng Hou, Zihan Lin, Yushuo Chen, Xingyu Pan, Kaiyuan Li, Yujie Lu, Hui Wang, Changxin Tian, et al. Recbole: Towards a unified, comprehensive and efficient framework for recommendation algorithms. In Proceedings of the 30th ACM International Conference on Information & Knowledge Management, pp. 4653–4664, 2021.