PERFORMANCE EVALUATION OF COLLABORATIVE FILTERING TECHNIQUES ON AMAZON FASHION DATA

Lam Thanh Toan[1*], Tran Thi Thuy Oanh[2], Nguyen Bao Thai[2], Hoang Trung Hieu[2] and Nguyen Xuan Ha Giang[1]

[1]Can Tho University of Technology
[2]Student, Faculty of Information Technology, Can Tho University of Technology
Email: lttoan@ctuet.edu.vn

ARTICLE INFO

Received:
28/5/2025
Accepted:
11/9/2025
Published:
30/01/2026

Keywords:
Collaborative Filtering,
Deep Learning,
Matrix Factorization,
Recommendation System

ABSTRACT

Recommender systems play a crucial role in enhancing user experience by delivering personalized suggestions on e-commerce platforms. Among various techniques, collaborative filtering is widely adopted due to its ability to capture implicit relationships between users and items. However, the wide variety of available algorithms makes model selection challenging. This study conducts a comprehensive empirical evaluation of 18 collaborative filtering models categorized into three major groups: matrix factorization, heuristic/statistical methods, and deep learning, using the Amazon Fashion 2023 dataset. These models are assessed using AUC, NDCG@20, Precision@20, Recall@20, and training time. The results reveal that GMF achieves the highest AUC (0.7883), while VAECF and EASER demonstrate superior top-k recommendation quality, with Recall@20 reaching 0.1429. Conversely, models like BiVAECF and COE suffer from high training costs and limited effectiveness, especially in sparse data settings. The study highlights the strengths and weaknesses of each model group and recommends developing hybrid models that combine performance and efficiency for scalable recommender system deployment.

1. INTRODUCTION

In the era of digital transformation and information explosion, recommender systems have become an essential tool for enhancing user experience and optimizing business performance, especially in the e-commerce sector. Collaborative Filtering (CF) is one of the most widely adopted approaches, thanks to its ability to harness collective intelligence from user behaviors and preferences. By analyzing past interactions - such as purchase history, product ratings, or browsing activity - CF can predict individual user preferences and deliver personalized recommendations. This not only improves user satisfaction but also promotes product discovery and drives sales growth.

Moreover, the continuous growth in both the scale and complexity of user data, along with the emergence of numerous advanced machine learning models, has posed significant challenges in evaluating, comparing, and selecting the most suitable collaborative filtering (CF) models for specific application contexts. Previous studies have proposed various CF architectures, ranging from traditional matrix factorization methods such as Singular Value Decomposition (SVD), SVD++, and Non-negative Matrix Factorization (NMF), which are foundational and effective models in dense data environments but struggling with implicit feedback and sparse data [1] - to representative deep learning models like

AutoRec [2] and Neural Collaborative Filtering (NCF) [3], which open new directions by learning nonlinear interactions between users and items, achieving superior performance on complex datasets. Graph-based approaches, such as LightGCN [4] and Graph Convolutional Matrix Completion (GCMC) [5], exploit the user-item relationship structure in the form of graphs, enabling multi-step information propagation within interaction networks. Although each model offers certain advantages in specific scenarios, the absence of a unified and comprehensive evaluation framework makes it difficult to conduct objective comparisons and identify the strengths and weaknesses of each method. Existing research often focuses on proposing new models rather than deeply analyzing the performance of current models under varying data conditions. The lack of in-depth analyses regarding the impact of data characteristics, such as sparsity, long-tail distribution, or seasonality, on the performance of CF models represents a significant gap. This makes it challenging to make informed decisions when selecting the appropriate model for a given recommendation problem.

In this study, three groups of collaborative filtering techniques - Matrix Factorization, Heuristic/Statistical Models, and Deep Learning - are implemented and empirically evaluated on the Amazon Fashion 2023 e-commerce dataset. Through quantitative analysis, the research aims to provide a comprehensive insight into the effectiveness of each method, thereby offering suitable recommendations for selecting collaborative filtering techniques in recommender systems within the e-commerce domain

2. RELATED WORK

Collaborative Filtering (CF)-based recommender systems have been extensively studied for decades, starting with traditional methods such as User-based CF (UBCF) and Item-based CF (IBCF) [6]. These methods rely on the similarity between users or items to generate recommendations. However, they face challenges such as data sparsity and the cold start problem - when new users or items have limited interactions - resulting in poor recommendation quality and low accuracy.

To overcome these limitations, matrix factorization techniques such as Singular Value Decomposition (SVD), SVD++, and Non-negative Matrix Factorization (NMF) have been developed [1]. These methods represent users and items in a latent factor space, effectively capturing hidden relationships within the data. SVD++ improves upon SVD by incorporating information about unobserved user-item interactions, thereby optimizing recommendation quality in sparse data scenarios. However, matrix factorization methods still face challenges when handling nonlinear features or complex contextual information.

In recent years, with the rapid advancement of deep learning, models such as AutoRec [2] and Neural Collaborative Filtering (NCF) [3] have been proposed to overcome the limitations of matrix factorization methods. These models utilize neural networks to learn nonlinear representations of users and items, enhancing recommendation quality, especially in large and diverse datasets like Amazon Fashion. Specifically, NCF models nonlinear interactions between users and items, delivering superior performance in complex scenarios.

Moreover, integrating contextual data (such as time, location, device, user behavior, etc.) into collaborative filtering models has been shown to significantly improve the effectiveness of recommender systems [7]. Contextual data helps mitigate the cold start problem and optimizes recommendations based on additional information. Hybrid models that combine collaborative filtering with contextual or content information are becoming a research trend to address this issue. These hybrid models improve recommendation accuracy and increase the system's adaptability to changes in user behavior and the expansion of product catalogs.

Although many advanced models have been developed, the lack of a comprehensive empirical evaluation framework to compare and assess collaborative filtering techniques under real-world data conditions remains a notable gap in current research. Previous studies have mainly focused on developing new models, paying less attention to thoroughly evaluating and analyzing the performance of existing methods in practical data scenarios. This issue is particularly important in e-commerce environments like Amazon Fashion, where user and product data are abundant and diverse, presenting challenges such as sparsity, long-tail distributions, and imbalanced numbers between users and products.

Therefore, this study aims to evaluate and compare the performance of current collaborative filtering methods under real - world data conditions, specifically using the Amazon Fashion 2023 dataset. Through this approach, we hope to provide a unified evaluation framework that assists researchers and developers in selecting the most suitable model for specific recommendation tasks.

3. RESEARCH METHODOLOGY

The experimental procedure of this study is illustrated in Figure 1, beginning with the preprocessing of the Amazon 2023 dataset, which includes fields such as UserID, ItemID, Rating, and Timestamp. After preprocessing, the data is split into 80% for training and 20% for testing. Three groups of models Matrix Factorization, Heuristic/ Statistical and Deep learning are trained and evaluated under a unified experimental framework. Model performance is measured using metrics such as AUC, NDCG@K, Precision@K, Recall@K, along with training and testing times.

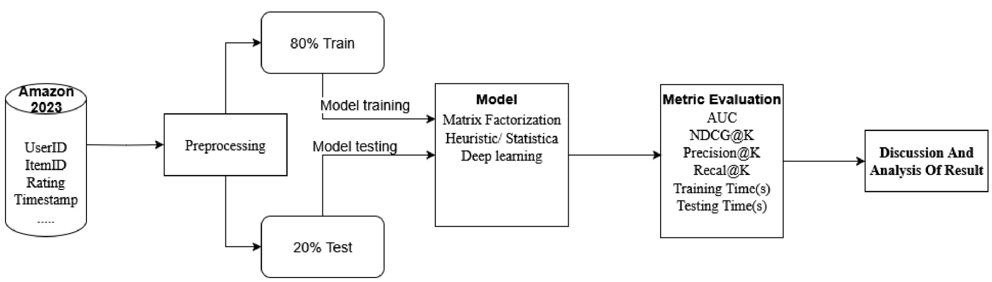

Figure 1 Performance evaluation process of collaborative filteringtechniques on the Amazon Fashion 2023 dataset

3.1. Dataset

In this study, the dataset used is Amazon Fashion, a large dataset containing user reviews of fashion products on the Amazon Fashion platform. The dataset includes important fields such as: UserID (user identifier), ItemID (product identifier), Title (review title), GroupItem (product group), Rating (review score), Description (review content), Helpful_vote (number of helpful votes for the review), Verified_purchase (purchase verification status), Item_helpful_vote (number of helpful votes for the product), Average_rating (average rating score), Rating_number (total number of reviews), and Details (additional information related to the product or review). In total, this dataset comprises approximately 2 million users, 825,900 products, and 2.5 million reviews

3.2. Data Preprocessing

From the Amazon Fashion 2023 dataset, the study randomly selects 10,000 samples to build and evaluate the models. Among these, three data columns are extracted for training and testing purposes, including ItemID and Rating. After preprocessing, the data is randomly split into two separate sets for training and evaluation. Specifically, 80% of the data is used as the training set, while the remaining 20% is reserved for the testing set. This split ensures the model's generalization capability and provides an objective evaluation of the proposed algorithms' performance.

3.3. Models

The study evaluates three groups of collaborative filtering techniques - Matrix Factorization, Heuristic/Statistical, and Deep Learning - which are presented in Table 1.

Table 1. Collaborative Filtering Technique Groups: Matrix Factorization, Heuristic/Statistical, and Deep Learning

Model Name	Technique Group	Short Description	Reference
MF	Matrix Factorization	Decomposes the user–item interaction matrix into latent feature vectors.	[1]
PMF	Matrix Factorization	A probabilistic version of MF assuming Gaussian distribution for latent factors.	[8]
NMF	Matrix Factorization	Adds non-negativity constraints to MF for better interpretability.	[9]
WMF	Matrix Factorization	Assigns weights to interactions to reflect data confidence.	[10]
MMMF	Matrix Factorization	Matrix factorization based on a margin-optimization learning framework.	[11]
SVD	Matrix Factorization	Decomposes matrix into three orthogonal matrices; commonly used in RS.	[6]
EASE	Heuristic / Statistical	A simple linear model that learns the item-item reconstruction matrix.	[12]
COE	Heuristic / Statistical	Embeds ordinal preferences to better handle ranking data.	[13]
SKMEANS	Heuristic / Statistical	Combines spectral clustering with K-means for user/item grouping.	[14]
HPF	Heuristic / Statistical	Hierarchical probabilistic factorization using Poisson distribution.	[15]
WBPR	Heuristic / Statistical	Extends BPR by incorporating weights to handle data imbalance.	[16]
Online_IBPR	Heuristic / Statistical	An online learning version of IBPR for dynamic data.	[17]
GMF	Deep Learning	Extends MF with a neural layer for learning non-linear interactions.	[3]
MLP	Deep	A deep neural network to model user–item	[3]

Model Name	Technique Group	Short Description	Reference
	Learning	interactions.	
NeuMF	Deep Learning	Combines GMF and MLP to leverage both linear and non-linear effects.	[3]
VAECF	Deep Learning	Uses probabilistic autoencoders to learn the latent distribution of user preferences.	[18]
BiVAECF	Deep Learning	An improvement on VAECF by modeling both user and item perspectives.	[19]
RecVAE	Deep Learning	A variational autoencoder with strong regularization for top-N recommendation.	[19]

3.4. Model Evaluation

To evaluate whether a recommendation system is effective, commonly used metrics include AUC, NDCG, Top-K, Precision, and Recall.

Additionally, train (s) and test (s) refer to the time the model takes for training and prediction, respectively. The technical performance of the two model groups is presented in Table 2.

Table 2. Performance evaluation methods of the models used in the study.

Z	Citation
AUC, Precision, Recall	[20]
NDCG	[21]
Training time, Testing time	[22]

4. DISCUSSION AND RESULT ANALYSIS

In this study, we evaluate the performance of three popular collaborative filtering algorithm groups on the Amazon Fashion 2023 dataset, including Matrix Factorization, Heuristic/Statistical Models, and Deep Learning, based on metrics such as AUC, NDCG@20, Precision@20, Recall@20, along with training and testing time.

The experimental results of matrix factorization models on the Amazon Fashion 2023 dataset are presented in Table 3, showing significant differences in performance among the models, both in overall accuracy and top-k recommendation effectiveness. Overall, the models achieved average AUC scores ranging from 0.53 to 0.68, with the highest being NMF (0.6802) and MMMF (0.6788), reflecting a fairly good ability to distinguish between interaction and non-interaction pairs. However, the NDCG@20, Precision@20, and Recall@20 metrics are all very low; for example, MF and SVD only achieved Precision@20 of 0.0001 and Recall@20 around 0.0030, indicating that these models are ineffective in recommending items that users are truly interested in. WMF is the only model showing a clear improvement, with Recall@20 reaching 0.0938 and NDCG@20 at 0.0816, but this comes at a very high training cost (nearly 2000 seconds). Some models like PMF and NMF have moderate training times but only marginal improvements in recommendation effectiveness.

Table 3. Experimental Results of Models in the Matrix Factorization Group

Model	AUC	NDCG@20	Precision@20	Recall@20	Train(s)	Test(s)
MF	0.5885	0.0010	0.0001	0.0030	0.0713	1.7617
PMF	0.5713	0.0356	0.0024	0.0476	3.1176	1.6793
NMF	0.6802	0.0050	0.0007	0.0134	0.5812	1.7572
WMF	0.5374	0.0816	0.0048	0.0938	1999.32	2.4691
MMMF	0.6788	0.0033	0.0004	0.0089	9.3549	2.2501
SVD	0.5879	0.0010	0.0001	0.0030	0.0722	2.1743

The experimental results of models in the Heuristic/Statistical group are presented in Table 4, showing that HPF achieved the highest AUC (0.7242), demonstrating strong ability to distinguish interactions, while SKMEANS had the lowest performance (AUC = 0.3431). EASER showed good results in top-k metrics such as Recall@20 (0.1429) and Precision@20 (0.0071), along with very low training time, indicating a good capability to recommend relevant products in the list. Meanwhile, COE, despite having a high training cost (over 3500 seconds), yielded limited top-k results, which is not commensurate with the model's complexity. OnlineIBPR and WBPR performed at an average level, with OnlineIBPR having a better Recall@20 (0.0474) compared to WBPR (0.0047)

Table 4. Experimental results of models in the Heuristic/Statistical group

Model	AUC	NDCG@20	Precision@20	Recall@20	Train(s)	Test(s)
EASER	0.4880	0.0553	0.0071	0.1429	0.0671	0.0061
COE	0.5209	0.015	0.0001	0.0030	3546.66	1.9619
Skmeans	0.3431	0.0171	0.0019	0.0379	1.8918	0.3061
HPF	0.7242	0.0166	0.0019	0.0379	35.7935	0.2736
WBPR	0.5147	0.0047	0.0002	0.0047	1.0271	0.2899
Online_Ibpr	0.5611	0.0253	0.0024	0.0474	60.7203	0.5315

Table 5. Experimental Results of Models in the Deep Learning Group

Model	AUC	NDCG@20	Precision@20	Recall@20	Train(s)	Test(s)
GMF	0.79	0.0148	0.0012	0.0241	250.536	0.4287
MLP	0.7519	0.0079	0.0008	0.0161	304.158	2.0719
NeuMF	0.6327	0.0565	0.0040	0.0803	322.603	2.7818
VAECF	0.558	0.0331	0.0071	0.1429	1.2195	0.004
BiVAECF	0.0122	0.0005	0.0001	0.0017	15926.04	30.4047

The experimental results in Table 5 show that the GMF model achieves the highest AUC value (0.7883), indicating a strong ability to distinguish between interactions and non-interactions in sparse data. The BiVAECF model has a very low AUC

(0.0122) and a very long training time, suggesting it is unsuitable for the current dataset. Although VAECF has a low AUC (0.558), it attains the highest Recall@20 (0.1429), demonstrating accurate top-k product recommendations. NeuMF outperforms GMF and MLP in terms of NDCG@20 (0.0565) and Recall@20 (0.0803), but still falls short of VAECF in recommendation coverage. MLP shows average performance and relatively long training time. Overall, GMF stands out for its discrimination capability, VAECF is effective in top-k recommendations, while BiVAECF is unsuitable due to its high training cost and low effectiveness.

The experimental results on the Amazon Fashion 2023 dataset show clear differences among the three model groups in both discrimination accuracy and top-k (k=20) recommendation effectiveness. The deep learning group stands out with GMF achieving the highest AUC (0.7883), demonstrating a strong ability to distinguish between interactions and non-interactions, while VAECF, despite its low AUC (0.558), attains the highest Recall@20 (0.1429), reflecting precise top-k product recommendations. In the heuristic/statistical group, EASER delivers good recommendation performance (Recall@20 = 0.1429) with very low training cost, indicating it as an efficient and easy-to-deploy model. Meanwhile, the matrix factorization group generally achieves moderate average AUC values (0.53–0.68), with NMF and MMMF being the best models for interaction discrimination; however, their top-k metrics are very low, except WMF, which reaches Recall@20 = 0.0938 but with high training cost. Some models such as BiVAECF and COE have long training times but offer disproportionate effectiveness, suggesting they are not suitable for the current dataset. Therefore, if the goal is good interaction discrimination, GMF and HPF are appropriate choices; if priority is given to top-k recommendation effectiveness, VAECF and EASER are models worth considering; meanwhile, high-cost, low-performance models like BiVAECF, COE, or Skmeans are unsuitable for highly sparse datasets. It should be noted that this study employed only 10,000 randomly selected interactions from the original dataset instead of using the entire data. This choice was motivated by the very large scale of the dataset, which exceeds the available computational capacity and time constraints within the current scope. Random sampling (rather than selection by user or item) was adopted to minimize bias and preserve data diversity, while ensuring that the models could be compared under stable and feasible conditions. As a future direction, the experiments will be extended to the full dataset and complemented with alternative sampling strategies (e.g., user-based or item-based) in order to evaluate the generalizability and adaptability of different model groups across more diverse real-world contexts.

5. CONCLUSION

The study conducted a comprehensive evaluation of the performance of three groups of collaborative filtering techniques - matrix factorization, heuristic/statistical models, and deep learning - on the Amazon Fashion 2023 dataset. Experimental results indicate that each group of models has distinct strengths and weaknesses. Deep learning models such as GMF and VAECF demonstrate high interaction discrimination ability and good top-k (k=20) recommendation effectiveness, albeit with considerable training costs. The heuristic/statistical group, especially EASER, shows impressive top-k recommendation performance with very low computational cost, making it suitable for practical system deployment. Meanwhile, matrix factorization models like NMF, MMMF, and WMF achieve moderate discrimination accuracy but generally limited top-k recommendation effectiveness.

Based on these findings, future research directions should focus on combining the strengths of different model groups, such as developing hybrid models between EASER and VAECF to leverage both cost efficiency and recommendation quality. Additionally, optimizing the training cost for deep learning models (such as NeuMF and VAECF) is an important approach to enhance feasibility for large-scale recommendation systems. Finally, it is necessary to extend research to datasets with varying characteristics (such as sparsity and user interaction levels) to evaluate the generalizability and adaptability of each model group in different real-world contexts.

References

[1] Koren Y, Bell R, Volinsky C. Matrix factorization techniques for recommender systems. Computer. 2009; 42(8): 30-37.

[2] Sedhain S, Menon AK, Sanner S, et al. AutoRec: Autoencoders meet collaborative filtering. Proceedings of the 24th International Conference on World Wide Web (WWW 2015 Companion). Florence: International World Wide Web Conferences Steering Committee; 2015. 111-112.

[3] He X, Liao L, Zhang H, et al. Neural collaborative filtering. Proceedings of the 26th International Conference on World Wide Web (WWW 2017). Perth: International World Wide Web Conferences Steering Committee; 2017. 173-182.

[4] He X, Deng K, Wang X, et al. LightGCN: Simplifying and powering graph convolution network for recommendation. Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval. Xi'an: ACM; 2020. 639-648.

[5] Van den Berg R, Kipf TN, Welling M. Graph convolutional matrix completion. Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2018). London: ACM; 2018. 395-404.

[6] Sarwar B, Karypis G, Konstan J, et al. Item-based collaborative filtering recommendation algorithms. Proceedings of the 10th International Conference on World Wide Web (WWW 2001). Hong Kong: ACM; 2001. 285-295.

[7] Zhang Y, Cao B, Yu Y, et al. Incorporating contextual information in recommender systems via a neural network model. Proceedings of the 23rd ACM International Conference on Information and Knowledge Management (CIKM 2014). Shanghai: ACM; 2014. 1427-1436.

[8] Mnih A, Salakhutdinov R. Probabilistic matrix factorization. Advances in Neural Information Processing Systems (NeurIPS 2007). Vancouver: Curran Associates; 2007. 1257-1264.

[9] Lee DD, Seung HS. Algorithms for non-negative matrix factorization. Advances in Neural Information Processing Systems (NeurIPS 2001). Vancouver: MIT Press; 2001. 556-562.

[10] Hu Y, Koren Y, Volinsky C. Collaborative filtering for implicit feedback datasets. Proceedings of the 2008 IEEE International Conference on Data Mining (ICDM 2008). Pisa: IEEE; 2008. 263-272.

[11] Rennie JDM, Srebro N. Fast maximum margin matrix factorization for collaborative prediction. Proceedings of the 22nd International Conference on Machine Learning (ICML 2005). Bonn: ACM; 2005. 713-719.

[12] Steck H. Embarrassingly shallow autoencoders for sparse data. Proceedings of the World Wide Web Conference (WWW 2019). San Francisco: ACM; 2019. 3251-3257.

[13] Zeng J, Yu X, Tong X, et al. Enhancing automatic readability assessment with pre-training and soft labels for ordinal regression. Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL 2022). Dublin: Association for Computational Linguistics; 2022. 1234-1245.

[14] Wang X, Zhang Y. Spectral clustering with prior knowledge. Pattern Recognit Lett. 2015; 52: 1-7.

[15] Gopalan P, Hofman JM, Blei D. Scalable recommendation with hierarchical Poisson factorization. Proceedings of the 31st Conference on Uncertainty in Artificial Intelligence (UAI 2015). Amsterdam: AUAI Press; 2015. 326-335.

[16] Gantner Z, Drumond L, Freudenthaler C, et al. Personalized ranking for non-uniformly sampled items. Proceedings of the 2012 ACM Conference on Recommender Systems (RecSys 2012). Dublin: ACM; 2012. 269-276.

[17] Le D, Lauw HW. Indexable Bayesian personalized ranking for efficient top-k recommendation. Proceedings of the 26th International Conference on World Wide Web (WWW 2017). Perth: International World Wide Web Conferences Steering Committee; 2017. 1383-1392.

[18] Liang D, Krishnan RG, Hoffman MD, et al. Variational autoencoders for collaborative filtering. Proceedings of the World Wide Web Conference (WWW 2018). Lyon: ACM; 2018. 689-698.

[19] Shenbin I, Alekseev A, Tutubalina E, et al. RecVAE: A new variational autoencoder for top-N recommendations with implicit feedback. Proceedings of the 13th International Conference on Web Search and Data Mining (WSDM 2020). Houston: ACM; 2020. 528-536.

[20] Powers DMW. Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation. J Mach Learn Technol. 2020; 2(1): 37-63.

[21] Liu TY. Learning to rank for information retrieval. Found Trends Inf Retr. 2009; 3(3): 225-331.

[22] Mitchell TM. Machine learning. New York: McGraw-Hill; 1997.

ĐÁNH GIÁ HIỆU NĂNG CỦA CÁC KỸ THUẬT LỌC CỘNG TÁC TRÊN TẬP DỮ LIỆU AMAZON FASHION

TÓM TẮT

Hệ thống gợi ý đóng vai trò quan trọng trong việc nâng cao trải nghiệm người dùng thông qua việc cung cấp các đề xuất cá nhân hóa trên các nền tảng thương mại điện tử. Trong số các kỹ thuật hiện có, lọc cộng tác được sử dụng rộng rãi nhờ khả năng khai thác các mối quan hệ tiềm ẩn giữa người dùng và sản phẩm. Tuy nhiên, sự đa dạng của các thuật toán khiến việc lựa chọn mô hình phù hợp trở nên khó khăn. Nghiên cứu này thực hiện một đánh giá thực nghiệm toàn diện trên 18 mô hình lọc cộng tác, được phân thành ba nhóm chính: phân rã ma trận, các phương pháp heuristic/thống kê và các mô hình học sâu, sử dụng tập dữ liệu Amazon Fashion 2023. Các mô hình được đánh giá dựa trên các chỉ số AUC, NDCG@20, Precision@20, Recall@20 và thời gian huấn luyện. Kết quả cho thấy GMF đạt giá trị AUC cao nhất (0,7883), trong khi VAECF và EASER thể hiện chất lượng gợi ý top-k vượt trội, với Recall@20 đạt 0,1429. Ngược lại, các mô hình như BiVAECF và COE có chi phí huấn luyện cao và hiệu quả hạn chế, đặc biệt trong bối cảnh dữ liệu thưa. Từ các kết quả này, nghiên cứu làm rõ ưu điểm và hạn chế của từng nhóm mô hình, đồng thời đề xuất hướng phát triển các mô hình lai nhằm cân bằng giữa hiệu năng và hiệu quả tính toán cho việc triển khai hệ thống gợi ý ở quy mô lớn.

Từ khóa: *Collaborative filtering, deep learning, matrix factorization, recommendation system*