International Journal of Scientific Engineering and Science
Volume 10, Issue 4, pp. 77-81, 2026. ISSN (Online): 2456-7361

Neural Matrix Factorization for Movie Recommendation: A Comparative Study on MovieLens 1M

Nguyễn Thị Quỳnh Trâm
Faculty of Mathematical Economics, Thuongmai University
Email address: tram.ntq@tmu.edu.vn

Abstract— *The strong development of streaming platforms, typically Netflix, has significantly increased the phenomenon of information and choice overload, thereby creating an urgent need for highly reliable and scalable movie recommender systems. This paper focuses on researching and evaluating a movie recommender system based on the Neural Matrix Factorization (NeuMF) model, in which a hybrid architecture between Generalized Matrix Factorization (GMF) and Multi-Layer Perceptron (MLP) is used to simultaneously model both linear and non-linear interactions between users and movies. Experiments were implemented on the standard MovieLens 1M dataset, including 1,000,209 explicit ratings from 6,040 users for 3,706 movies, with a rating scale from 1 to 5, based on a unified preprocessing process and an approximately 80/10/10 train/validation/test data split for all compared models. The results show that NeuMF achieved MAE = 0.6900, HR@10 = 0.7800 and NDCG@10 = 0.4746, thereby outperforming the two traditional recommender models SVD and item-based k-NN in both the rating prediction task and Top-10 recommendation. At the same time, SVD still demonstrates its role as a matrix factorization-based model with stable performance and relatively low computational cost, while item-based k-NN clearly reveals its limitations in the face of the sparsity of the interaction matrix. The experimental results obtained allow us to affirm the potential of deep learning models, specifically NeuMF, in improving the personalization level of digital content recommender systems using explicit rating data, while also providing additional empirical evidence for this approach on the MovieLens 1M dataset.*

Keywords— *Recommender System, Neural Matrix Factorization, NeuMF, MovieLens 1M, Collaborative Filtering, Deep Learning, Top-K Recommendation.*

I. INTRODUCTION

In the context of the digital era, users increasingly interact with a huge "sea of information" on digital content platforms and streaming services such as Netflix. The excessively large number of choices easily leads to information overload and "choice overload" states, making the entertainment decision-making process more difficult and time-consuming. In this situation, recommender systems (Recommender Systems – RS) have gradually become an indispensable infrastructure component, with the role of narrowing the choice space, personalizing content, and enhancing user engagement as well as experience.

Over the past decade, collaborative filtering (Collaborative Filtering – CF) and matrix factorization (Matrix Factorization, SVD) methods have proven their practical effectiveness on many large-scale datasets, typically the Netflix Prize and MovieLens. These models utilize the user–item interaction matrix to infer latent factors, thereby handling the data sparsity problem relatively well and providing acceptable recommendation quality in many application scenarios. However, traditional approaches mostly rely on the assumption of linear interactions between users and items, leading to limitations in capturing complex, non-linear behavioral patterns that are quite common in the current digital content environment.

The rapid development of deep learning has opened a new direction for recommender systems, with the emergence of architectures such as Neural Collaborative Filtering (NCF) and especially Neural Matrix Factorization (NeuMF). These models replace the traditional inner product with multi-layer neural networks, allowing them to directly learn a non-linear interaction function between users and items. Specifically, NeuMF is designed as a hybrid architecture between the Generalized Matrix Factorization (GMF) branch and the Multi-Layer Perceptron (MLP), aiming to simultaneously inherit the advantages of matrix factorization and enhance representation capability through non-linear layers.

Although there have been many positive results on implicit feedback data and Top-K ranking metrics, international studies on NeuMF generally have not systematically considered adapting this architecture for explicit rating data, as well as jointly evaluating both rating prediction performance (MAE) and Top-K recommendation quality (HR@K, NDCG@K) in the same experimental setting. In Vietnam, recommender system studies mainly focus on content-based models, k-NN, MF/SVD on small and medium-sized datasets; the implementation and detailed analysis of NeuMF on the MovieLens 1M dataset with explicit data has hardly been widely recorded.

Starting from the above research gaps, this paper aims at three main objectives. First, to build a movie recommender system based on NeuMF using explicit ratings data from the MovieLens 1M dataset, thereby simulating a Netflix-style recommender scenario. Second, to quantitatively compare NeuMF with two strong baseline models: SVD and item-based k-NN on both groups of metrics: rating prediction (MAE) and Top-10 recommendation (HR@10, NDCG@10). Third, to analyze and discuss the theoretical implications as well as practical applications of using the deep learning NeuMF model

in movie recommender systems operating on explicit rating data.

II. Related Works

Research on recommender systems (Recommender Systems – RS) has developed through many stages, closely linked to the evolution of methodology and technology.

First, the initial stage (1990s) was associated with memory-based collaborative filtering methods, typically the Tapestry system, in which users' opinions and ratings were used directly to filter information. User-based and item-based k-NN algorithms became popular due to their simplicity and ease of implementation, but they revealed limitations in scalability and sensitivity to data sparsity when the number of users and items increased significantly.

Second, the model development stage (2000–2010) witnessed the explosion of model-based methods, especially matrix factorization (Matrix Factorization – MF) and SVD. The Netflix Prize was an important milestone that demonstrated the effectiveness of SVD in mapping users and items into a latent factor space, thereby handling sparse rating matrices better and significantly improving prediction accuracy compared to neighborhood-based CF. However, MF/SVD still relied on the assumption of linear interactions among latent factors, making it difficult to capture complex non-linear user behaviors.

Third, the hybridization and context-awareness stage (2010–2015) saw the emergence of hybrid recommender systems (Hybrid RS), combining collaborative filtering with content-based filtering to overcome the cold-start problem, along with context-aware recommender systems (CARS) that incorporated additional factors such as time, location, and device into the model.

Fourth, from around 2015 to the present, the development of deep learning has opened a new generation of recommender systems. Architectures based on autoencoders, CNN, RNN, attention, and Transformer have been applied to learn hidden representations and model sequential user behavior. Within this trend, the Neural Collaborative Filtering (NCF) line, typically Neural Matrix Factorization (NeuMF), directly "non-linearizes" matrix factorization models. NeuMF combines the GMF branch (modeling linear interactions through element-wise product of user–item embeddings) and the MLP branch (learning higher-order non-linear interactions), then concatenates the two branches at the top layer to create a hybrid interaction function. Many international studies have shown that NeuMF achieves superior performance compared to MF/SVD and k-NN on various recommender datasets, mainly in implicit feedback settings and Top-K ranking evaluation. Recently, some studies have integrated metaheuristics to fine-tune hyperparameters and explored federated learning for distributed recommender systems.

In Vietnam, recommender system research generally follows international trends but is implemented on problems and datasets of more modest scale. The works of Nguyễn Hùng Dũng and Nguyễn Thái Nghe applied collaborative filtering for electronic product recommendation, while Trần Nguyễn Minh Thư and Huỳnh Quang Nghi proposed the RecoLRC system — a hybrid of collaborative filtering and keyword indexing to support library document retrieval, helping personalize document lists according to users. In recent years, deep learning architectures have begun to appear, typically the study by Nguyễn Hoàng Anh (2024) which used Transformer for sequential news recommendation, emphasizing the ability to capture shifts in user preferences within a session. However, most studies still focus on k-NN, MF/SVD or neural models on implicit interaction data, while the detailed implementation and evaluation of NeuMF on explicit rating data at the scale of MovieLens 1M is still largely absent.

From the above overview, several research gaps can be identified. First, many international studies on NeuMF focus either on rating prediction (MAE, RMSE) or on Top-K ranking, but have not systematically analyzed the relationship between these two objectives on explicit rating data. Second, in the Vietnamese context, there is a lack of standardized experiments implementing NeuMF on MovieLens 1M with explicit ratings and directly comparing it with strong linear models such as SVD and item-based k-NN under the same data setup and metrics.

The current study was conducted to fill these gaps by implementing NeuMF on MovieLens 1M, establishing a unified experimental pipeline for NeuMF, SVD and item-based k-NN, and jointly evaluating them on MAE, HR@10, and NDCG@10, thereby providing a clearer view of the improvements that the GMF–MLP architecture brings in the context of movie recommender systems using explicit ratings.

III. Methodology

Dataset

The study uses the MovieLens 1M dataset, consisting of 1,000,209 ratings from 6,040 users for 3,706 movies, with a rating scale of 1–5. The data is kept in the form of explicit ratings and solves the rating prediction task, after which the predicted scores are used for ranking in the Top-K recommendation scenario.

The preprocessing process includes the following steps: (i) remapping UserID/MovieID to continuous integer indices for use in embedding layers; (ii) removing invalid records; (iii) optionally normalizing ratings to a suitable range during NeuMF training to stabilize gradients and performing inverse mapping when evaluating MAE on the 1–5 scale. The data is split into train/validation/test ≈ 80/10/10, with the most recent interactions (by timestamp) used for the test set to simulate future prediction scenarios.

Baseline Models Two traditional models were selected as baselines: SVD and item-based k-NN, implemented using the Surprise library:

- *SVD*: Represents users and movies in a latent factor space, predicts the rating using the formula

$$\hat{r}_{ui} = \mu + b_u + b_i + q_i^T p_u$$

and optimizes the regularized squared error using SGD

- *Item-based k-NN*: Computes cosine similarity between movie rating vectors, selects the k nearest neighbors for each movie, and predicts the rating as a weighted average of ratings from similar movies that the user has rated.

Both models were trained on explicit 1–5 rating data and used for rating prediction, then applied to rank movies for Top-10 recommendation evaluation.

Proposed Model:

NeuMF The Neural Matrix Factorization (NeuMF) model is implemented with a hybrid architecture consisting of two branches: GMF and MLP.

- Embedding layer: Each user and item is mapped to a fixed-dimensional latent vector; different embedding sizes can be used for GMF and MLP or they can share embeddings depending on the configuration.
- GMF branch: Performs element-wise product between user and item embeddings to model linear interactions, similar to generalized matrix factorization.
- MLP branch: Concatenates user and item embeddings as input, passes through multiple hidden layers with a decreasing number of neurons (tower architecture, for example 64–32–16) and ReLU activation functions to learn non-linear interactions.
- Combination layer: The outputs of GMF and MLP are concatenated and fed into the final layer to predict the rating; the loss function uses Mean Squared Error (MSE), and evaluation is performed using MAE on the validation/test sets.

The model is trained using Adam optimizer, with hyperparameters such as embedding size, number of MLP layers and neurons, learning rate, and batch size tuned based on the validation set to balance prediction quality, computational cost, and the risk of overfitting.

Evaluation Protocol

The study uses two groups of metrics:

- Rating prediction: ◦ Mean Absolute Error (MAE) on the test set for all three models: NeuMF, SVD, and item-based k-NN.
- Top-K recommendation: ◦ Hit Ratio@10 (HR@10) and Normalized Discounted Cumulative Gain@10 (NDCG@10), in which the predicted rating scores are used to rank the list of candidate movies for each user.

The same train/validation/test split is used for all three models to ensure a fair comparison.

IV. Results and Discussion

- *Rating Prediction Results (MAE)* The main objective of the first experiment is to evaluate the accuracy of the models in predicting users' explicit ratings on the 1–5 scale. The Mean Absolute Error (MAE) of the three models on the MovieLens 1M test set is presented in Table 1.

TABLE 1. Mean Absolute Error (MAE) on the MovieLens 1M Test Set

Model	MAE
Item-based k-NN	0.7865
SVD (baseline)	0.7342
NeuMF (proposed)	0.6900

The results in Table 1 show that NeuMF achieved the lowest MAE value (0.6900), corresponding to an improvement of approximately 6.02% compared to SVD and 12.27% compared to item-based k-NN. The significant reduction in MAE demonstrates that the hybrid GMF–MLP architecture is more effective at learning latent feature representations from explicit rating data than traditional linear models.

While SVD only models linear interactions between users and movies in the latent factor space, NeuMF additionally utilizes the power of the multi-layer perceptron (MLP) to capture complex non-linear relationships in rating behavior, thereby producing predictions closer to the actual ratings. Although MAE is an important measure reflecting the stability of the model in point-wise prediction, evaluating the ability to personalize recommendation lists requires considering additional ranking metrics such as HR@10 and NDCG@10, which are presented in the next section.

- *Top-10 Recommendation Results* To comprehensively evaluate recommendation and personalization capability, the performance of the three models was analyzed through the Top-10 ranking protocol. Based on the predicted rating scores for each user–movie pair, each model generates a list of the 10 highest-scoring candidate movies. The detailed comparison results on Hit Ratio@10 (HR@10) and NDCG@10 are summarized in Table 2.

TABLE 2. HR@10 and NDCG@10 Results on the MovieLens 1M Test Set (K = 10)

Model	HR@10	NDCG@10
Item-based k-NN	0.4915	0.2830
SVD (baseline)	0.6782	0.3941
NeuMF (proposed)	0.7800	0.4746

The experiments show that item-based k-NN is the model with the lowest performance on both metrics. The main reason stems from the high sparsity of the interaction matrix in MovieLens 1M, making it difficult for the model to find a sufficient number of common users between pairs of movies to estimate stable similarity, thereby reducing recommendation quality. In contrast, SVD shows a clear improvement thanks to mapping the data into a latent factor space, contributing to solving the data sparsity problem and enhancing the quality of the suggested lists.

Most notably, NeuMF records superior performance on both HR@10 and NDCG@10. Specifically, with HR@10 = 0.7800, NeuMF improves by approximately 15.01% compared to SVD and 58.7% compared to k-NN. At the same time, NDCG@10 = 0.4746 is 20.43% higher than SVD, showing that NeuMF not only increases the probability of "hitting" the target movie in the Top-10 but also tends to rank the target movies in higher priority positions within the recommendation list.

This superiority can be explained by the hybrid GMF–MLP architecture:

- The GMF branch inherits the generalization ability of linear matrix factorization models, helping to effectively model the latent factor structure in the data.

- The MLP branch with non-linear layers allows NeuMF to learn complex, higher-order relationships between users and movies that MF/SVD finds difficult to capture.
- The combination layer at the output helps integrate information from both branches, while retaining the advantages of MF and exploiting the strong representation power of deep learning.

Combined with the MAE optimization results analyzed in the previous section, it can be affirmed that the deep learning NeuMF architecture simultaneously meets two objectives: (i) accurately predicting users' explicit ratings and (ii) providing high-quality Top-10 recommendation lists, overcoming the inherent limitations of traditional linear collaborative filtering methods such as SVD and item-based k-NN.

Figure 1 aims to compare the Top-10 recommendation performance between item-based k-NN, SVD and NeuMF on the MovieLens 1M dataset according to the two metrics HR@10 (hit accuracy) and NDCG@10 (ranking quality). The chart shows that NeuMF achieves HR@10 = 0.7800 and NDCG@10 = 0.4746, outperforming SVD and item-based k-NN, thereby confirming the advantage of the hybrid GMF–MLP architecture in improving the quality of Top-10 recommendation lists.

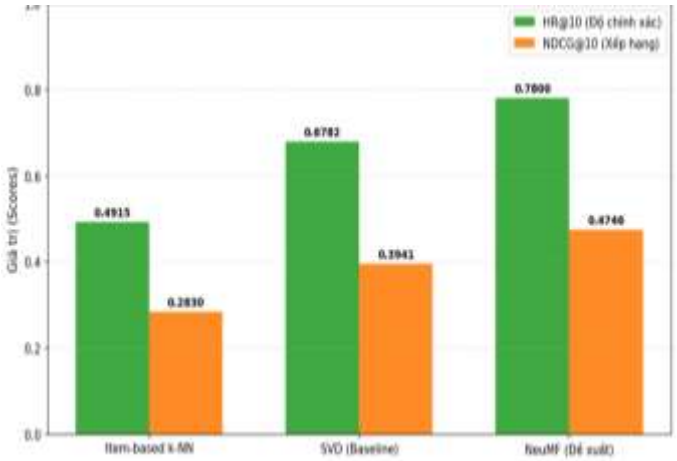

Figure 1. Performance Comparison between NeuMF and Baseline Models

V. Analysis and Discussion

First, compared to SVD, NeuMF retains the latent representation idea of matrix factorization but replaces the linear inner product with a multi-layer MLP, allowing it to learn non-linear relationships between users and movies. This helps NeuMF achieve lower MAE and higher HR@10 and NDCG@10, especially in the context where rating behavior depends on multiple factors simultaneously (genre, style, actors, viewing context…).

Second, item-based k-NN operates directly on the rating matrix and depends heavily on the set of common users who have rated both movies, therefore it is very sensitive to data sparsity. This is reflected in the experimental results: the model has the highest MAE and the lowest Top-10 metrics among the three models.

Third, the experiments also show that NeuMF performance depends significantly on the choice of hyperparameters such as embedding size, MLP structure, learning rate, and batch size; appropriate tuning helps substantially reduce MAE and improve HR@10 and NDCG@10.

From an application perspective, if a real-world movie recommender system uses explicit rating data similar to MovieLens 1M and has suitable computational infrastructure, NeuMF is the preferred choice because it delivers superior recommendation quality. In resource-constrained environments, SVD remains a reliable baseline thanks to its lower training and inference costs, while item-based k-NN can be used when the need for intuitive explainability ("recommended because it is similar to movies X, Y, Z") is more important than absolute accuracy.

VI. Conclusion and Future Work

This paper presents the construction and evaluation of a movie recommender system based on NeuMF using the MovieLens 1M dataset with explicit ratings, while comparing it with two traditional models: SVD and item-based k-NN.

Main results:

- NeuMF is the best model among the three in terms of MAE, HR@10 and NDCG@10, affirming the advantage of the hybrid GMF–MLP architecture in modeling non-linear interactions between users and movies.
- SVD is a strong baseline, delivering stable results with lower computational cost than NeuMF, but still inferior in accuracy.
- Item-based k-NN suffers limitations due to data sparsity and yields the lowest results, although it has the advantage of explainability.

Limitations: (i) The study only experimented on one MovieLens 1M dataset and has not been validated on other datasets or real-world data; (ii) the model only uses user/item IDs and has not exploited metadata (genre, actors, release year) or context; (iii) it has not addressed the cold-start problem for new users/movies.

Future research directions

(i) expanding the evaluation of NeuMF on multiple datasets and different content domains; (ii) developing hybrid NeuMF variants that combine CF embeddings with content and contextual features; (iii) integrating cold-start handling mechanisms and explainable recommendation techniques based on embeddings or attention; (iv) applying automatic hyperparameter optimization techniques (Bayesian optimization, metaheuristics) for NeuMF to improve quality and reduce manual tuning effort.

References

1. Adomavicius, G., & Tuzhilin, A. (2005). Towards the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE Transactions on Knowledge and Data Engineering*, 17(6), 734–749.
2. Chen, R., Hua, Q., Chang, Y. S., Wang, B., Zhang, L., & Kong, X. (2018). A survey of collaborative filtering-based recommender systems: From traditional methods to hybrid methods based on social networks. *IEEE Access*, 6, 64301–64320.
3. Roy, D., & Dutta, M. (2022). *A systematic review and research perspective on recommender systems*. Journal of Big Data, 9(1), 59.

4. Zhang, S., Yao, L., Sun, A., & Tay, Y. (2019). Deep learning based recommender system: A survey and new perspectives. *ACM Computing Surveys*, 52(1), Article 5.
5. Ariyanto, Y., & Widiyaningtyas, T. (2024). A systematic review of movie recommender systems. *ITEGAM Journal of Engineering and Industrial Applications (JETIA)*, 10(47), 34–41.
6. Lops, P., de Gemmis, M., & Semeraro, G. (2011). *Content-based recommender systems: State of the art and trends.* In F. Ricci, L. Rokach, B. Shapira, & P. B. Kantor (Eds.), Recommender systems handbook (pp. 73–105). Springer.
7. Trần Nguyễn Minh Thư, & Huỳnh Quang Nghi. (2016). Hệ thống gợi ý hỗ trợ tra cứu tài liệu. *Tạp chí Khoa học Trường Đại học Cần Thơ*, (43), 126-134.
8. Koren, Y., Bell, R., & Volinsky, C. (2009). Matrix factorization techniques for recommender systems. *Computer*, 42(8), 30-37
9. Nguyễn Hùng Dũng, & Nguyễn Thái Nghe. (2014). Hệ thống gợi ý sản phẩm trong bán hàng trực tuyến sử dụng kỹ thuật lọc cộng tác. *Tạp chí Khoa học Trường Đại học Cần Thơ*, (31), 36-51.
10. He, X., Liao, L., Zhang, H., Nie, L., Hu, X., & Chua, T. S. (2017). Neural collaborative filtering. In *Proceedings of the 26th International Conference on World Wide Web* (pp. 173–182).
11. Sarwar, B., Karypis, G., Konstan, J., & Riedl, J. (2001). Item-based collaborative filtering recommendation algorithms. In *Proceedings of the 10th International Conference on World Wide Web* (pp. 285-295).
12. Nguyễn Hoàng Anh. (2024). Phương pháp khuyến nghị tin tức trên cổng thông tin điện tử dựa trên dữ liệu tuần tự sử dụng Transformer. *Tạp chí Khoa học Công nghệ Thông tin và Truyền thông*, (01), 63-71.
13. Jayalakshmi, S., Ganesh, N., Čep, R., & Senthil Murugan, J. (2022). Movie recommender systems: Concepts, methods, challenges, and future directions. *Sensors*, 22(13), 4904.
14. Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT press.
15. Herlocker, J. L., Konstan, J. A., Terveen, L. G., & Riedl, J. T. (2004). Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems*, 22(1), 5–53.