Optimizing Collaborative Filtering in Very Sparse Datasets: a Confidence-based Approach

Vassiliki Stouraiti
Department of Informatics and Telecommunications
University of the Peloponnese
Tripoli, Greece
0009-0005-1160-7427

Giorgos Mpardis
Department of Digital Systems
University of the Peloponnese
Sparta, Greece
0009-0003-4209-317X

Konstantinos I. Roumeliotis
Department of Informatics and Telecommunications
University of the Peloponnese
Tripoli, Greece
0000-0002-8098-1616

Costas Vassilakis
Department of Informatics and Telecommunications
University of the Peloponnese
Tripoli, Greece
0000-0001-9940-1821

Christos Tryfonopoulos
Department of Informatics and Telecommunications
University of the Peloponnese
Tripoli, Greece
0000-0003-0640-9088

Dimitris Spiliotopoulos
Department of Management Science and Technology
University of the Peloponnese
Tripoli, Greece
0000-0003-3646-1362

***Abstract*— Collaborative filtering systems are a key tool for personalized recommendations, predicting user preferences based on historical data. A major challenge in collaborative filtering is the reliability of predictions, particularly in highly sparse datasets, where the limited overlap of common ratings between users impedes the identification of reliable neighbors, reducing both the accuracy and coverage of the predictions. So far, the literature has proposed confidence factors to quantify prediction reliability, typically using the Pearson and Cosine similarities with threshold-based neighbor selection. However, a recent study demonstrated that using the Sigmoid Cosine similarity metric in combination with the KNN neighborhood selection method and adapting the value of K to each level of dataset sparsity, significantly improves prediction performance. In this paper, we revisit the calculation of confidence factors under these improved settings, assessing their validity and effectiveness across datasets with varying sparsity levels. Continuous assessment of collaborative filtering methods and adjustment of their parameters according to dataset characteristics is particularly crucial in highly sparse datasets, where the reliability of neighbors strongly influences prediction quality.**

Keywords*— *Collaborative Filtering, Recommender Systems, Rating Predictions, Sparse Datasets, Optimization, Confidence Factors

I. Introduction

Recommender systems (RecSys) are nowadays an essential tool for delivering personalized content and services to users by leveraging preference and behavioral data. They are widely used on platforms, such as online stores and streaming services, with the aim to improve user experience and increasing engagement [1], [2], [3].

Collaborative Filtering (CF) is one of the most widely used techniques in RecSys. CF is based on the rationale that users with similar past preferences will also make similar choices in the future [4], [5], [6]. This method utilizes user ratings or interactions with items, without requiring information about the content itself.

In this context, CF systems aim to predict a user's preference for an item, either in the form of an estimated rating or the probability of interaction. These predictions are based on historical data, in the form of user ratings and/or reviews, and the similarity between users, enabling the generation of personalized recommendations.

The accuracy of predictions is a critical factor for the system's effectiveness, as it directly influences user satisfaction and the reliability of recommendations. A highly accurate CF system can more successfully identify users' actual preferences, strengthening their trust and increasing platform usage. Conversely, low accuracy can lead to irrelevant suggestions, demoting user experience, as well as the system's effectiveness and therefore its reliability [7], [8], [9].

The prediction process in CF is based on identifying "nearest neighbors" (NNs), i.e. users that share a high degree of similarity with the target/active user. Firstly, the similarity between users is calculated using similarity metrics, with the most widely used in the literature being the Pearson correlation and the Cosine similarity, which both consider ratings given to common items by users. Subsequently, the set of NNs (i.e., the users having the largest similarity value) is selected, and the real ratings of the selected neighbors are then used to estimate the unknown rating. The final prediction is typically calculated as the weighted average of the neighbors' ratings, where the weights are determined by the degree of similarity calculated in the previous step, boosting thus the weight of more relevant users [10], [11], [12].

In cases of very sparse datasets (i.e., datasets in which the number of available ratings constitutes a very small percentage of the total number of possible ratings), the process of generating predictions becomes significantly more difficult, as the available ratings are limited and unevenly distributed. This hinders the identification of sufficient and reliable NNs, on which the CF procedure relies. In particular, the scarcity of commonly rated items reduces the reliability of similarity metrics, leading to inaccurate predictions. Consequently, both the prediction accuracy and the coverage (i.e., the percentage of cases for which the system is able to calculate a prediction) decline [13], [14], [15].

The issue of rating prediction reliability in very sparse CF datasets is a research area that has been studied in recent literature. In this context, the work in [8] proposes the introduction of confidence factors, which take into account the quality and adequacy of the available data. Specifically, three key factors are proposed: (a) the number of NNs used to generate the prediction, with better performance when their number is sufficiently large, (b) the user's average rating, where higher accuracy is observed when this average is close

to the limits of the rating scale, and (c) the average rating of the item, with a corresponding improvement in accuracy when this average is also close to the limits of the scale. Incorporating these factors into the prediction generation process helps improve both the accuracy and stability of the results, particularly in conditions of severe data sparsity.

To draw the above conclusions, the experimental procedure of that study relied on the Pearson and Cosine metrics, to assess user similarity. Regarding the selection of NNs, the classical k-nearest neighbors (KNN) approach was not adopted; instead, a similarity threshold was applied, whereby all users with a positive similarity value (similarity > 0.0) were considered as NNs.

However, recent research [14] focusing on very sparse datasets has shown that the use of the Sigmoid Cosine similarity, in combination with the mean-centered rating prediction formula and with different optimal values of K for each level of sparsity, lead to increased accuracy and coverage. Specifically, for sparse datasets (with density ranging from 0.1% to 1%), the optimal value of K is estimated to be 250, for very sparse datasets is estimated to be 1000 (with density ranging from 0.01% to 0.1%), and for extremely sparse datasets (with density ranging from 0.001% to 0.01%) is estimated to be 1500. These findings suggest that the previous confidence factors proposed in [8] , may need to be recalculated or adjusted to remain valid under these new parameters and methods.

In this paper, we revisit the calculation of the confidence factors proposed in [8], however using the settings that the more recent study [14] identified as optimal. More specifically, the sigmoid cosine similarity metric is employed alongside density-adaptive K values for neighborhood selection. Beyond a simple recalculation, this work reexamines established assumptions in CF reliability pipelines. While traditional confidence factors were evaluated under legacy similarity metrics, we demonstrate that modern, sparsity-optimized configurations have significant repercussions on prediction error behavior, introducing non-linear reliability dynamics. This insight is crucial for designing confidence-aware recommendation pipelines in highly sparse environments.

To this end, we conduct and present a comprehensive series of experiments to empirically validate these dynamics, evaluating the behavior of the recalculated confidence factors under the updated settings. Specifically, we conduct experiments on six widely used sparse CF datasets with varying levels of sparsity, including two sparse, two very sparse, and two extremely sparse datasets (as defined above), in order to cover the full range of sparsity levels described in [14]). All experiments are conducted using 5-fold cross-validation, for more reliable results, and the prediction accuracy is assessed using MAE and RMSE, the most commonly applied prediction error measures.

More generally, the need for continuous evaluation of CF methods and for readjusting their parameters according to the characteristics of the dataset is evident and becomes particularly critical in datasets with a high degree of sparsity, where the limited availability of shared evaluations hinders the identification of reliable neighbors and, consequently, reduces both the accuracy and coverage of the predictions, as we noted above.

The remainder of this work is organized as follows: Section II presents the relevant literature, while Section III briefly presents the foundations of the present work, including both the CF process and the confidence factors, to ensure self-containment. Section IV presents the experimental process and summarizes the results, and finally Section V concludes the paper and outlines for future work.

II. Related Work

CF is one of the most widely used methods in RecSys, leveraging historical user-item interactions to predict preferences. A major research focus in recent years has been improving prediction accuracy, especially in datasets with high levels of sparsity. Various contemporary studies have examined different similarity metrics, prediction calculation methods, and neighbor selection strategies, demonstrating that careful tuning of CF parameters can significantly impact performance, particularly in sparse datasets.

In this context, the work in [16] proposes an integrated CF-based recommendation framework that leverages item ontological semantics, users' demographic information and rating credibility. Users' credibility scores, computed from their rating behavior, are employed to identify reliable neighbors, enhancing recommendation accuracy. Additionally, demographic and ontological information are incorporated into similarity calculations for users and items, mitigating the issues of cold-start and sparsity in CF algorithms.

The work in [17] introduces a deep neural recommendation framework that includes reliability of ratings, generated from explicit feedback, using techniques such as intuitionistic fuzzy sets for noise detection and K-means clustering for threshold selection, aiming to mitigate noisy ratings and improve the accuracy, robustness, and overall reliability of predictions for active users.

The work in [18] examines how to improve CF in environments with limited or no explicit ratings by leveraging user reviews. It proposes methods for extracting implicit ratings using sentiment analysis, taking into account sentiment scores and view-sentiment pairs, and blending them with explicit user ratings for integration into various recommendation prediction algorithms.

The work in [19] proposes a new hybrid similarity model for CF, based on the Wasserstein distance, with the aim of improving recommendations under sparse data and cold-start conditions. It also introduces a new type of user similarity that considers both non-common and common products, as well as additional heuristic factors to mitigate the influence of popular items and users.

The work in [20] proposes a multi-factor similarity and fusion approach within a probabilistic matrix factorization framework, capturing both linear and nonlinear user correlations and integrating local relations into global ratings, thereby enhancing the accuracy and robustness of rating prediction, particularly CF datasets with very low density, as demonstrated through experiments on four widely used public datasets.

The study in [21] investigates user similarity measures in low-density CF datasets, considering multiple neighbor selection strategies, rating prediction formulas, and accuracy measures. The study emphasizes the importance of selecting effective similarity metrics to identify reliable users, a task

that is particularly challenging in sparse datasets and critical for the performance of CF RecSys.

The work in [22] introduces a user-based CF method, namely BinRec, to identify NNs within subgroups of users and items. By focusing on shared biclusters, the method efficiently addresses challenges of sparsity and cold-start in large-scale rating datasets, improving scalability and recommendation accuracy.

The work in [23] presents an enhanced K-means-based CF algorithm, namely KUR-CF, that integrates user attribute ratings and common ratings to improve similarity computation. By leveraging variance-weighted features and combining clustering with user-item similarities, the model effectively addresses sparsity and cold-start challenges, enhancing recommendation accuracy, precision, and recall compared to traditional CF approaches.

The study in [24] proposes a matrix reconstruction approach to improve CF on sparse datasets. By leveraging item features and user preferences to regenerate a denser user–item matrix, the method enhances prediction accuracy and reduces sparsity, demonstrating improved performance over conventional CF approaches using both the KNN method and singular value decomposition techniques.

The study in [25] introduces a CF algorithm, namely CFR-FD, which combines a neighborhood awareness attention mechanism with fine-grained mining. The method dynamically leverages neighborhood information and analyzes project attributes and user behaviors. As a result, it effectively addresses cold-start and sparsity issues, improving recommendation accuracy, personalization, and robustness across multiple real-world network datasets.

The work in [26] proposes a ranking-based CF technique, namely SVD-GSetRank, that integrates Singular Value Decomposition with normalized rating values and Gower similarity scores. The method addresses limitations of traditional memory-based approaches, improving Top-N recommendation accuracy, including measures like MRR, Hit Rate, and NDCG, and while maintaining efficient processing across multiple benchmark datasets.

In this context, the literature has also proposed confidence factors for quantifying the reliability of CF predictions. More specifically, the work in [8] proposes the introduction of confidence factors, which take into account the quality and adequacy of the available data. Specifically, three key factors are proposed: (a) the number of NNs used to generate the prediction, with better performance when their number is sufficiently large, (b) the user's average rating, where higher accuracy is observed when it is near either end of the rating range, and (c) the average rating of the item, with a corresponding improvement in accuracy when the values are also near either end of the rating range. Incorporating these factors into the prediction calculation procedure helps enhance both the accuracy and stability of the results, particularly under conditions of severe data sparsity. To draw the above conclusions, the experimental procedure of this study relied on the Pearson and Cosine metrics, to compute user similarity. Regarding the selection of NNs, a similarity threshold was applied, whereby all users with a positive similarity (>0.0) were considered as NNs.

However, a more recent study [14] showed that using the Sigmoid Cosine similarity metric, in combination with the KNN method and different values of K for each sparsity level, leads to improved prediction performance on sparse, very sparse, and extremely sparse datasets. More specifically, for sparse datasets (with density between 0.1% and 1%) the optimal value of K was found to be 250, for very sparse datasets was approximately 1000 (with density between 0.01% and 0.1%), and for extremely sparse datasets (with density between 0.001% and 0.01%) was approximately 1500.

The present work aims to recalculate the confidence factors proposed in [8], however using the settings that the more recent study [14] identified as optimal, specifically using the Sigmoid Cosine similarity metric and different values of K for each level of data sparsity, regarding the NN selection. In this way, we evaluate whether the confidence factors remain valid and effective when applied to sparse datasets.

III. Foundations

In this Section, we briefly present the fundamental concepts underlying the present work, to ensure self-containment. Specifically, we outline the complete process of CF prediction formulation, comprising user similarity measures, the KNN selection technique, the rating prediction formula, as well as the concept of prediction confidence factors in sparse CF datasets.

The process of generating rating predictions in CF is initially based on calculating user similarity. This is achieved using similarity metrics like the Cosine, the Pearson and the Sigmoid Cosine.

Cosine similarity (COS) assesses user similarity by calculating the angle between the vectors of their ratings. It utilizes the ratings that a pair of users have provided to jointly rated items and reflects how similar their preference patterns are, regardless of the absolute rating value [21], [27].

Pearson correlation (PC) can be viewed as a form of COS applied to normalized (mean-centered) data, where the mean score of each user is subtracted from their respective ratings. In this way, it compares patterns of preference rather than absolute rating levels (as COS does) [28], [29].

Sigmoid Cosine similarity (SIGM) is another variant of COS that applies a sigmoid transformation to emphasize the similarity between users who have many co-rated items and reduce the influence of cases with fewer ones [14], [30].

Table I presents the definitions of the three metrics mentioned above.

TABLE I. CF User Similarity Metrics

Cosine	$COS(u_1,u_2) = \frac{\sum_{k \in CR_{u_1,u_2}} r_{u_1,k} \bullet r_{u_2,k}}{\sqrt{\sum_{k \in CR_{u_1,u_2}} (r_{u_1,k})^2} \bullet \sqrt{\sum_{k \in CR_{u_1,u_2}} (r_{u_2,k})^2}}$
Pearson	$PC(U_1,U_2) = \frac{\sum_{k \in CR_{u_1,u_2}} (r_{u_1,k} - \overline{r_{u_1}}) \bullet (r_{U_2,k} - \overline{r_{u_2}})}{\sqrt{\sum_{k \in CR_{u_1,u_2}} (r_{u_1,k} - \overline{r_{u_1}})^2} \bullet \sqrt{\sum_{k \in CR_{u_1,u_2}} (r_{U_2,k} - \overline{r_{u_2}})^2}}$
Sigmoid	$SIGM(u_1,u_2) = COS(u_1,u_2) \bullet \frac{1}{1 + e^{\frac{-\lvert CR_{u_1,u_2} \rvert}{2}}}$
$I(u_a)$ denotes the set of items that user u_a has evaluated $CR_{a,b}$ denotes the set of items evaluated by both users u_a and u_b, i.e. $CR_{a,b} = I(u_a) \cap I(u_b)$ $r_{u,k}$ denotes the rating that user u has assigned to item k. $\overline{r_u}$ denotes the mean of all ratings user u has assigned to items	

As derived from the definition of PC and also noted in [14], when all of a user's ratings are identical, the similarity

between that user and every other user in the dataset cannot be calculated, since the denominator is equal to zero. In practice, this implies that a CF system can neither generate personalized predictions for that specific user (due to the inability to identify their NNs) nor utilize them as an NN to produce predictions for other users. Furthermore, the same study showed that in very sparse datasets, the proportion of these users (having identical ratings) is considerable (in the range of 10–20%), resulting in a significant loss in prediction coverage when the PC is used.

Once the similarities between all pairs of users have been calculated, the most common strategy for selecting each user's NNs is the KNN method. In this approach, the top K users having the largest similarity values with that user are considered their NNs (K is typically defined by the CF system's administrator). In this context, the work in [14] has experimentally found the optimal values of K, ensuring high accuracy and coverage. Specifically, for sparse datasets (with density ranging from 0.1% to 1%), the optimal value of K is found to be approximately 250, for very sparse datasets (with density ranging from 0.01% to 0.1%) is found to be approximately 1000 and for extremely sparse datasets (with density ranging from 0.001% to 0.01%) is found to be approximately 1500.

Lastly, rating predictions on items are generated using a prediction formula that combines the ratings given by the user's NNs for the same item. The similarity of each NN to the user, as previously calculated, acts as a weighting factor for that NN's rating. The work in [14] experimentally demonstrated that the best results are obtained using the mean-centered formula (MCF), whose definition is shown in equation (1). In equation (1), the similarity between the user and each NN is denoted as "sim(u_1, u_2)".

$$MCF_{u_1,i}=\overline{r_{u_1}}+\frac{\sum_{u_2\in NN_{u_1}} sim(u_1,u_2)\bullet(r_{u_2,i}-\overline{r_{u_2}})}{\sum_{u_2\in NN_{u_1}} sim(u_1,u_2)} \quad (1)$$

Regarding the concept of prediction confidence factors, the study in [8] demonstrated a correlation between basic rating prediction attributes and prediction accuracy in (very) sparse CF datasets. Specifically, the study demonstrated that (1) the number of NNs involved in the prediction calculation, (2) the user's average rating, and (3) the average rating of the item being predicted, are linked to higher prediction accuracy. Regarding the first factor, a rating prediction is considered as "highly accurate" when the NNs participating are ≥4. Regarding the other two factors, a rating prediction is considered as "highly accurate" when the average rating of either the user or the item is near either end of the rating range (either ⩽1.5/5 or ⩾4.5/5).

To draw the aforementioned conclusions, the experimental procedure of this study relied on the PC and COS similarity metrics for calculating similarity among users and, furthermore, regarding the selection of NNs, the KNN approach was not adopted (as suggested in [14]).

In the next Section, the three confidence factors proposed in [8] will be recalculated, however using the settings that the more recent study [14] identified as optimal, specifically using the SIGM similarity metric and different values of K for each level of data sparsity.

IV. Experimental Evaluation

In this section, the three confidence factors discussed above will be recalculated using the SIGM similarity, with K=250, 1000, and 1500, for sparse, very sparse, and extremely sparse datasets, respectively.

Table II summarizes the datasets employed in the present study. As shown, we include datasets representing all sparsity levels (i.e., sparse, very sparse, and extremely sparse), thereby covering all three cases identified in [14].

Furthermore, datasets from different item categories are included (movies, music, office products, etc.), to mitigate potential biases associated with the characteristics of a single type of data.

TABLE II. Dataset Information

Name	***Attributes***	***Optimal K [14]***
Digital Music [31]	0.3% density, 65K ratings	250
Yahoo [32]	0.2% density, 221K ratings	250
Musical Instruments [33]	0.04% density, 512 K ratings	1000
Videogames [33]	0.03% density, 815K ratings	1000
Movies and TV [33]	0.006% density, 7.4M ratings	1500
Office Products [33]	0.009% density, 1.8M ratings	1500

For more reliable results, all experiments are conducted using the 5-fold cross-validation, where each dataset is divided into five subsets of equal size, with four of them used for training and the remaining one for testing. The CF system then calculates the similarities between users using the training data and attempts to predict the ratings contained in the test subset. Notably, the datasets are utilized in their original forms, as sourced from their respective repositories, with the five Amazon datasets (all except the Yahoo Movies dataset) being standard 5-core subsets to ensure a baseline of at least 5 ratings per user and item, while grouping of confidence-factor intervals was performed programmatically, based only on training set history. Prediction accuracy is assessed using two widely recognized error metrics, MAE and RMSE. The first measures the average magnitude of prediction errors without considering whether each error overestimates or underestimates the real value, while the latter assigns greater weight to larger errors.

A. Number of NNs Confidence Factor

Fig. 1 depicts the experimental results for the first confidence factor (the number of NNs involved in the rating prediction), using the MAE error metric. The baseline for comparison is the overall MAE of all rating predictions in each dataset, which corresponds to the 0% change on the vertical axis. Values above 0% indicate higher MAE than the overall MAE, while values below 0% indicate lower MAE (and therefore better prediction accuracy).

When examining the change in MAE, in relation to the overall MAE of each dataset, we observe that predictions based on up to 4 NNs exhibit higher MAE, i.e. demoted accuracy, as compared to the average performance. For instance, predictions using only 1 NN result, on average, in a 19% higher MAE than the overall MAE of the dataset (across all predictions). Beyond this point, the MAE is lower than the average performance and gradually improves. More

specifically, in cases where 5-8 NNs are involved, small reductions are observed (in the range 3%-5%).

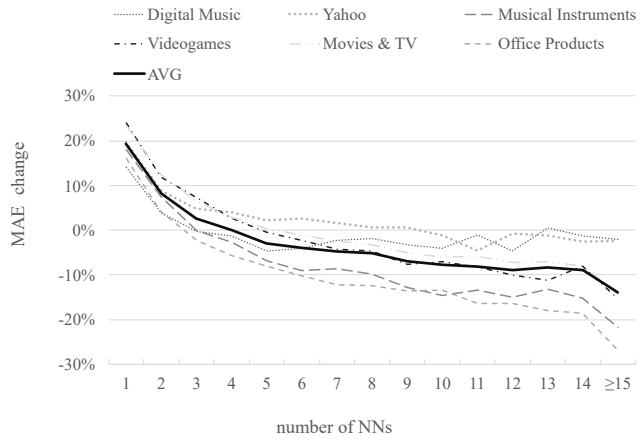

Fig. 1. Effect of the number of NNs on the MAE of rating predictions

When the number of NNs increases to 9-14 slightly larger reductions (7%-9%, on average) occur, while for NNs⩾15, the improvements become more significant, with an average 14% decrease compared to the overall MAE. At the dataset level, when the number of NNs involved in the predictions is ⩾10, the average MAE is reduced in all cases.

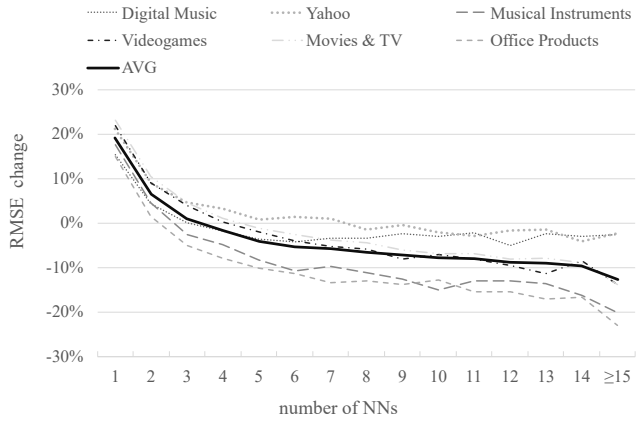

Fig. 2. Effect of the number of NNs on the RMSE of rating predictions

Similar behavior can be observed when analyzing the change in RMSE in relation to the overall RMSE of each dataset, as depicted in Fig. 2. When considering rating predictions where NN⩽3, the RMSE is higher than the global RMSE average. At NNs=4, a slight reduction is observed, averaging a 1.7% decrease. When the number of NNs ranges from 5 to 8, small reductions of 4-7% are noted, on average, while for predictions based on 9 to 14 NNs, slightly larger reductions of 7-9% occur. For NNs⩾15, the improvements become more substantial, averaging a 13% decrease compared to the overall RMSE. At the dataset level, all datasets exhibit a reduction in RMSE once the number of NNs reaches 8.

Overall, we do observe a positive correlation between the number of NNs and prediction accuracy, with both MAE and RMSE decreasing as the number of NNs increases. However, while the work in [8] found that predictions became very highly accurate once the number of NNs exceeded 4 (using different CF settings), our findings (with the updated settings) suggest that at NN=4, there is only a slight improvement in MAE. High accuracy, according to our results, is first observed for NN values between 9 and 14, while very high accuracy is achieved for NNs⩾15.

B. User's Average Rating Confidence Factor

Fig. 3 depicts the experimental results for the second confidence factor, the user's average rating (avgU), based on the MAE metric and using the same baseline as in Fig. 1.

When examining changes in MAE, relative to the overall MAE of each dataset, we observe that predictions for users with average rating value (computed across all their ratings) ⩽4.4 exhibit lower accuracy than the average case. For instance, predictions for users with an average rating in the range [1.0, 4.0] have been found to exhibit a MAE that is 60% higher than the overall MAE across all datasets (including all predictions). Beyond the threshold of 4.4/5 for avgU, MAE decreases rapidly, reaching an average of 56% reduction for predictions for users with avgU≥4.9/5 (a small subset of users who rate almost every item as "excellent"). At the dataset level, when the average user rating is ≥4.5/5, the average MAE is reduced in all cases.

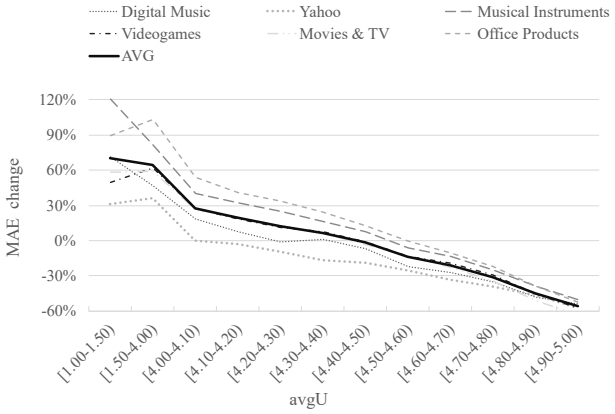

Fig. 3. Effect of the avgU on the MAE of rating predictions

Similar results are observed when analyzing the change in RMSE in relation to avgU and compared to the overall RMSE of each dataset, as depicted in Fig. 4. When avgU ranges from 1/5 to 4.3/5, the mean RMSE is higher than the average case. When avgU exceeds the threshold of 4.3/5, the RMSE decreases rapidly, reaching an average of 29% reduction for predictions concerning users with avgU≥4.9/5. At the dataset level, when the average rating is ≥4.5/5, the average RMSE is reduced across all datasets.

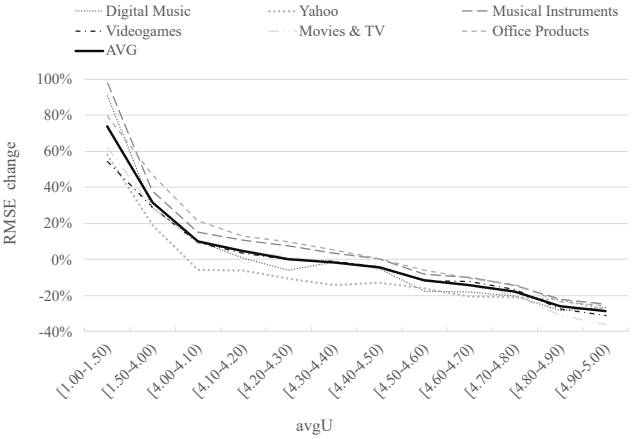

Fig. 4. Effect of the avgU on the RMSE of rating predictions

Overall, we do observe a positive correlation between the user's average rating and the improvement in prediction accuracy, in both MAE and RMSE. However, unlike the work in [8], which reported higher prediction accuracy at both ends of the rating scale (upper and lower), the experimental results of the present study indicate that this improvement occurs only when the user's average rating is near the upper bound.

C. Item's Average Rating Confidence Factor

Fig. 5 depicts the experimental results for the third confidence factor, the item's average rating (avgI), based on the MAE metric, with the same baseline as in Figs. 1 and 3.

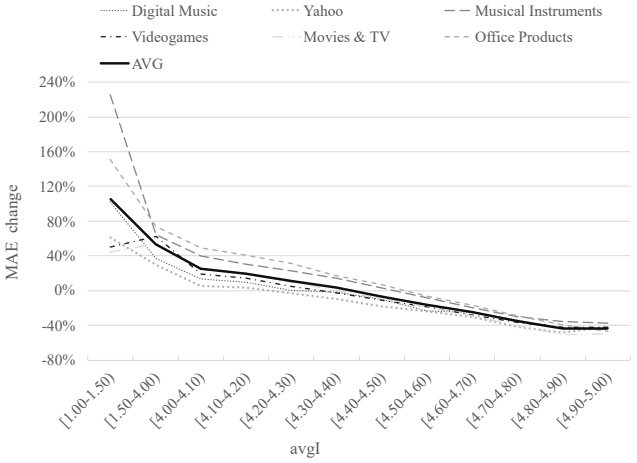

Fig. 5. Effect of the avgI on the MAE of rating predictions

When examining changes in MAE, relative to the overall MAE of each dataset, we observe that predictions for items with average rating (computed from all users who have rated them) ≤4.4 exhibit, on average, higher MAE -and therefore limited accuracy- compared to the mean MAE for all predictions across all datasets. For instance, predictions concerning users with average rating in the range [1.0/5, 4.0/5), on average, demonstrate a 50% higher MAE than the overall MAE of the dataset (across all predictions). Beyond the threshold of avgI=4.4/5 MAE decreases rapidly, reaching a 43% reduction for predictions concerning users with an average rating ≥4.8/5. Therefore, predictions for items rated as "excellent" by almost all users tend to be highly reliable. At the dataset level, when the average rating is ≥4.5/5, the average MAE decreases in all cases.

Similar results are observed when analyzing changes in RMSE relative to avgI, and comparing against the overall RMSE of each dataset, as depicted in Fig. 6. When avgI falls within the range [1.0/5, 4.4/5) the RMSE for predictions concerning the corresponding item is higher (on average) than the mean RMSE, indicating lower accuracy. Beyond the threshold of avgI=4.4/5, RMSE decreases rapidly, reaching an average reduction of 30% for predictions concerning items with average rating ≥4.8/5. At the dataset level, when the average item rating is ≥4.5/5, the average RMSE for predictions concerning these items is reduced across all datasets.

Overall, we do observe a positive correlation between the item's average rating value and the improvement in prediction accuracy, in both MAE and RMSE. However, unlike the findings in [8], which reported higher prediction accuracy at both ends of the rating scale (upper and lower), the present study indicates that this improvement occurs only when an item's average rating is near the upper bound.

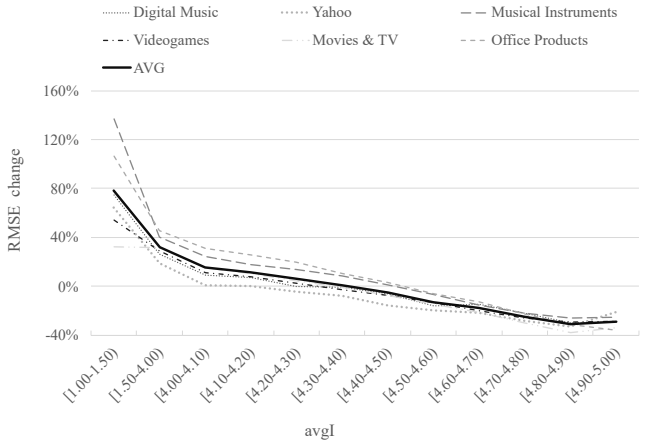

Fig. 6. Effect of the avgI on the RMSE of rating predictions

The statistical robustness of these findings is validated by the observation that measurements converge across all experiments (Figs. 1–6). Rather than relying solely on macro-averages, the individual curves and error metrics for each of the six diverse datasets exhibit almost identical behavior, with relatively low deviation. The uniformity observed across datasets demonstrates that the witnessed asymmetric error dynamics and threshold shifts are structurally rooted on the proposed configuration, rather than owing to dataset-specific features or anomalies.

D. Comparison with Previous Work: Key Differences in Confidence Factors

This subsection presents a comparative analysis of the three rating prediction confidence factors between the present study and the findings of the study in [8]. The comparison highlights how the updated settings proposed in [14] (the SIGM similarity and adaptive K-NN based on the exact density of each dataset), affect the behavior of the confidence factors and the corresponding prediction accuracy.

1) Confidence Factor 1 - Number of NNs

- Work in [8]: high accuracy was observed when NN≥4.
- Present work: prediction accuracy is worse compared to the dataset's overall accuracy when NNs<4. Accuracy begins to improve gradually for NNs⩾4, reaching high levels between 9 and 14 NNs, and optimal performance for NNs⩾15.

2) Confidence Factor 2 - avgU

- Work in [8]: high accuracy was observed for users with avgU at either end of the scale (specifically, avgU≤1.5/5 or avgU≥4.5/5).
- Present work: high accuracy is observed only for users with avgU near the upper end of the rating scale (avgU≥4.4/5). Accuracy increases significantly from this threshold and continues to improve as avgU increases, suggesting that higher avgU values correspond, on average, to more reliable rating predictions.

3) Confidence Factor 3 - avgI

- Work in [8]: high accuracy was observed for items with avgI at either end of the scale (specifically, avgI≤1.5/5 or avgI≥4.5/5).
- Present work: high accuracy is observed only for items with avgI near the upper end of the rating scale (avgI ≥4.4/5). Accuracy increases significantly from this threshold and continues to improve as avgI rises, suggesting that higher avgI values correspond, on average, to more reliable rating predictions.

The comparison mentioned above shows that the three prediction confidence factors identified in [8] do remain valid, confirming their importance for assessing the reliability of rating predictions in very sparse CF datasets. However, the thresholds-ranges at which these factors indicate high-accuracy predictions have shifted due to the use of the SIGM similarity and the NN selection strategy (as proposed in [14]), based on the density of each dataset. These findings highlight the need to reassess reliability parameters and adjust them according to the density of the dataset, particularly in sparse environments, in order to achieve optimal performance.

V. Conclusions and Future Work

In this study, we revisited the prediction confidence factors introduced in [8] for assessing the reliability of rating predictions in highly sparse CF datasets. By applying the updated settings proposed in [14] -including the SIGM similarity and an adaptive K-NN selection strategy based on the density of each dataset- we examined whether the confidence factors remain effective under the new CF configurations. To evaluate prediction accuracy, we employed two fundamental prediction error metrics in CF, MAE and RMSE. Additionally, we conducted experiments on six sparse datasets exhibiting varying sparsity degrees (from sparse to extremely sparse) and domains, using the 5-fold cross-validation technique to ensure more reliable results.

Our experimental findings indicate that, although the three reliability factors proposed in [8] remain valid, the thresholds-ranges defining high-accuracy rating predictions have shifted, with the adoption of the new settings described above.

Specifically, regarding the first factor, the number of NNs involved in the prediction, the work in [8] reported that the highest prediction accuracy occurs for NNs≥4. In contrast, the results of the present study show that predictions reach high accuracy only when NNs≥15. Predictions with 5-8 NNs show a modest increase of accuracy, while those with 9-14 NNs exhibit a larger improvement. Lastly, predictions involving ≤4 NNs are considered less accurate.

Regarding the second factor, the user's average rating, the work in [8] indicated that prediction accuracy is higher when the user's average rating is near either end of the rating scale (upper or lower). In contrast, the results of the present study show that high accuracy is achieved only for users with average ratings near the upper end of the scale (avgU ≥4.4/5). Prediction accuracy increases significantly from this point and continues to improve as avgU rises, indicating that higher user average ratings correspond, on average, to more reliable predictions.

Similarly, for the third factor, the item's average rating, [8] reported higher prediction accuracy when the item's average rating is near either end of the rating scale. In contrast, our results show that high accuracy is achieved only for items with avgI ≥4.4/5. Prediction accuracy increases significantly from this threshold and continues to improve as avgI rises, indicating that higher item average ratings correspond, on average, to more reliable predictions.

More generally, the findings of this study highlight the essential need for continuous evaluation of CF methods and the (re-)adjustment of their settings and parameters according to the characteristics of each dataset. Particularly in the case of very sparse datasets, the limited availability of commonly rated items hinders the identification of reliable neighbors, directly impacting the accuracy of predictions. The present study indicates that revising and adjusting prediction confidence factors is vital for improving the performance of CF systems under such challenging conditions.

Our future work will focus on incorporating the updated confidence factors derived in this study into the recommendation process, by combining the prediction score with its corresponding prediction confidence. In terms of practical deployment, this allows designers to improve recommendation decisions by dynamically down-ranking predictions with high uncertainty; for instance, effectively prioritizing a high-confidence 4.7/5 prediction over a low-confidence 4.9/5 prediction (as proposed in [34], [35]). The ultimate goal is to transition from traditional error metrics to directly quantifying these threshold limits on Top-N ranking metrics like Precision, Recall, and NDCG, ensuring items with a high likelihood of being accepted by users are recommended, rather than simply recommending items with the highest prediction score. Furthermore, we plan to investigate additional confidence factors and evaluate their correlation with prediction accuracy, aiming to further improve the reliability of recommendations. Finally, similar work focusing on dense datasets, such as MovieLens, is planned to assess whether the proposed or analogous factors and settings can also improve performance in these environments [36], [37], [38].

References

[1] Z. Xia, A. Sun, J. Xu, Y. Peng, R. Ma, and M. Cheng, "Contemporary Recommendation Systems on Big Data and Their Applications: A Survey," *IEEE Access*, vol. 12, pp. 196914–196928, 2024, doi: 10.1109/ACCESS.2024.3517492.

[2] X. Yang, "Research on personalized distance education recommendation system based on deep learning," *Sci Rep*, vol. 15, no. 1, p. 42158, Nov. 2025, doi: 10.1038/s41598-025-26020-1.

[3] E. Hasan, M. Rahman, C. Ding, J. X. Huang, and S. Raza, "Review-based Recommender Systems: A Survey of Approaches, Challenges and Future Perspectives," *ACM Comput. Surv.*, vol. 58, no. 1, pp. 1–41, Jan. 2026, doi: 10.1145/3742421.

[4] Z. Cui *et al.*, "Personalized Recommendation System Based on Collaborative Filtering for IoT Scenarios," *IEEE Trans. Serv. Comput.*, vol. 13, no. 4, pp. 685–695, Jul. 2020, doi: 10.1109/TSC.2020.2964552.

[5] O. Alshareet and A. Awasthi, "Collaborative filtering in the age of AI: foundations, innovations, and emerging trends," *Computing*, vol. 107, no. 11, p. 216, Nov. 2025, doi: 10.1007/s00607-025-01564-2.

[6] H. Khatter, S. Arif, U. Singh, S. Mathur, and S. Jain, "Product Recommendation System for E-Commerce using Collaborative Filtering and Textual Clustering," in *2021 Third International Conference on Inventive Research in Computing Applications (ICIRCA)*, Coimbatore, India: IEEE, Sep. 2021, pp. 612–618. doi: 10.1109/ICIRCA51532.2021.9544753.

[7] L. Wu, X. He, X. Wang, K. Zhang, and M. Wang, "A Survey on Accuracy-oriented Neural Recommendation: From Collaborative Filtering to Information-rich Recommendation," *IEEE Trans. Knowl. Data Eng.*, pp. 1–1, 2022, doi: 10.1109/TKDE.2022.3145690.

[8] D. Margaris, C. Vassilakis, and D. Spiliotopoulos, "On Producing Accurate Rating Predictions in Sparse Collaborative Filtering Datasets," *Information*, vol. 13, no. 6, p. 302, Jun. 2022, doi: 10.3390/info13060302.

[9] R. E. Veras De Sena Rosa, F. A. S. Guimaraes, R. D. S. Mendonca, and V. F. D. Lucena, "Improving Prediction Accuracy in Neighborhood-Based Collaborative Filtering by Using Local Similarity," *IEEE Access*, vol. 8, pp. 142795–142809, 2020, doi: 10.1109/ACCESS.2020.3013733.

[10] S. C. Mana and T. Sasipraba, "Research on Cosine Similarity and Pearson Correlation Based Recommendation Models," *J. Phys.: Conf. Ser.*, vol. 1770, no. 1, p. 012014, Mar. 2021, doi: 10.1088/1742-6596/1770/1/012014.

[11] G. Jain, T. Mahara, and S. C. Sharma, "Performance Evaluation of Time-based Recommendation System in Collaborative Filtering Technique," *Procedia Computer Science*, vol. 218, pp. 1834–1844, 2023, doi: 10.1016/j.procs.2023.01.161.

[12] S. Nudrat, H. U. Khan, S. Iqbal, M. M. Talha, F. K. Alarfaj, and N. Almusallam, "Users' Rating Predictions Using Collaborating Filtering Based on Users and Items Similarity Measures," *Computational Intelligence and Neuroscience*, vol. 2022, pp. 1–13, Jul. 2022, doi: 10.1155/2022/2347641.

[13] H. Koohi and K. Kiani, "Two new collaborative filtering approaches to solve the sparsity problem," *Cluster Comput*, vol. 24, no. 2, pp. 753–765, Jun. 2021, doi: 10.1007/s10586-020-03155-6.

[14] S.-A. Lapadaki, J. Nanos, D. Margaris, C. Vassilakis, and D. Spiliotopoulos, "Optimizing Collaborative Filtering for Accurate Rating Predictions in Very Sparse Datasets," *Future Internet*, vol. 18, no. 2, p. 114, Feb. 2026, doi: 10.3390/fi18020114.

[15] S. Natarajan, S. Vairavasundaram, S. Natarajan, and A. H. Gandomi, "Resolving data sparsity and cold start problem in collaborative filtering recommender system using Linked Open Data," *Expert Systems with Applications*, vol. 149, p. 113248, Jul. 2020, doi: 10.1016/j.eswa.2020.113248.

[16] N. R. Kermany, W. Zhao, T. Batsuuri, J. Yang, and J. Wu, "Incorporating user rating credibility in recommender systems," *Future Generation Computer Systems*, vol. 147, pp. 30–43, Oct. 2023, doi: 10.1016/j.future.2023.04.029.

[17] J. Deng, Q. Wu, S. Wang, J. Ye, P. Wang, and M. Du, "A novel joint neural collaborative filtering incorporating rating reliability," *Information Sciences*, vol. 665, p. 120406, Apr. 2024, doi: 10.1016/j.ins.2024.120406.

[18] S. AL-Ghuribi, S. A. Mohd Noah, and M. Mohammed, "An experimental study on the performance of collaborative filtering based on user reviews for large-scale datasets," *PeerJ Computer Science*, vol. 9, p. e1525, Aug. 2023, doi: 10.7717/peerj-cs.1525.

[19] J. Guan, B. Chen, and S. Yu, "A hybrid similarity model for mitigating the cold-start problem of collaborative filtering in sparse data," *Expert Systems with Applications*, vol. 249, p. 123700, Sep. 2024, doi: 10.1016/j.eswa.2024.123700.

[20] C. Feng, J. Liang, P. Song, and Z. Wang, "A fusion collaborative filtering method for sparse data in recommender systems," *Information Sciences*, vol. 521, pp. 365–379, Jun. 2020, doi: 10.1016/j.ins.2020.02.052.

[21] K. Sgardelis, D. Margaris, D. Spiliotopoulos, and C. Vassilakis, "An evaluation review of user similarity metrics in sparse collaborative filtering datasets," *Int J Data Sci Anal*, vol. 20, no. 7, pp. 6665–6693, Nov. 2025, doi: 10.1007/s41060-025-00846-4.

[22] D. Rodríguez-Baena, F. Gómez-Vela, A. Lopez-Fernandez, M. García-Torres, and F. Divina, "BinRec: addressing data sparsity and cold-start challenges in recommender systems with biclustering," *Appl Intell*, vol. 55, no. 12, p. 830, Aug. 2025, doi: 10.1007/s10489-025-06725-6.

[23] S. Zhang, S. Chen, X. Yu, and S. Mei, "Research on collaborative filtering algorithm based on improved K-means algorithm for user attribute rating and co-rating," *Sci Rep*, vol. 15, no. 1, p. 19600, Jun. 2025, doi: 10.1038/s41598-025-96705-0.

[24] S.-M. Choi, D. Lee, K. Jang, C. Park, and S. Lee, "Improving Data Sparsity in Recommender Systems Using Matrix Regeneration with Item Features," *Mathematics*, vol. 11, no. 2, p. 292, Jan. 2023, doi: 10.3390/math11020292.

[25] J. Jun and Y. Li, "Collaborative filtering recommendation algorithm based on fine-grained mining and neighborhood awareness attention," *Discov Artif Intell*, vol. 5, no. 1, p. 156, Jul. 2025, doi: 10.1007/s44163-025-00414-6.

[26] T. Widiyaningtyas, I. Saifudin, I. A. E. Zaeni, Moh. Z. N. Maulana, and W. Caesarendra, "Memory-based collaborative filtering based on matrix factorization and Gower's set rank," *J. King Saud Univ. Comput. Inf. Sci.*, vol. 37, no. 9, p. 289, Nov. 2025, doi: 10.1007/s44443-025-00261-6.

[27] Y. Adilaksa and A. Musdholifah, "Recommendation System for Elective Courses using Content-based Filtering and Weighted Cosine Similarity," in *2021 4th International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)*, Yogyakarta, Indonesia: IEEE, Dec. 2021, pp. 51–55. doi: 10.1109/ISRITI54043.2021.9702788.

[28] P. Chowdhury and B. B. Sinha, "Evaluating the Effectiveness of Collaborative Filtering Similarity Measures: A Comprehensive Review," *Procedia Computer Science*, vol. 235, pp. 2641–2650, 2024, doi: 10.1016/j.procs.2024.04.249.

[29] A. B. M. K. Hossain, Z. Tasnim, S. Hoque, and M. A. Rahman, "A Recommender System for Adaptive Examination Preparation using Pearson Correlation Collaborative Filtering," *Int J Auto AI Mach learn*, pp. 30–45, Mar. 2021, doi: 10.61797/ijaaiml.v2i1.55.

[30] F. Fkih, "Similarity measures for Collaborative Filtering-based Recommender Systems: Review and experimental comparison," *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 9, pp. 7645–7669, Oct. 2022, doi: 10.1016/j.jksuci.2021.09.014.

[31] J. Ni, J. Li, and J. McAuley, "Justifying Recommendations using Distantly-Labeled Reviews and Fine-Grained Aspects," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Hong Kong, China: Association for Computational Linguistics, 2019, pp. 188–197. doi: 10.18653/v1/D19-1018.

[32] Yahoo! Research, "Yahoo! Webscope Dataset YData-Set R4: Yahoo! Movies User Ratings," 2010. [Online]. Available: https://webscope.sandbox.yahoo.com/

[33] Y. Hou, J. Li, Z. He, A. Yan, X. Chen, and J. McAuley, "Bridging Language and Items for Retrieval and Recommendation," 2024, *arXiv*. doi: 10.48550/ARXIV.2403.03952.

[34] D. Margaris, C. Vassilakis, and D. Spiliotopoulos, "From Rating Predictions to Reliable Recommendations in Collaborative Filtering: The Concept of Recommendation Reliability Classes," *BDCC*, vol. 9, no. 4, p. 106, Apr. 2025, doi: 10.3390/bdcc9040106.

[35] D. Margaris, D. Spiliotopoulos, K. Sgardelis, and C. Vassilakis, "Using Prediction Confidence Factors to Enhance Collaborative Filtering Recommendation Quality," *Technologies*, vol. 13, no. 5, p. 181, May 2025, doi: 10.3390/technologies13050181.

[36] G. Behera and N. Nain, "Collaborative Filtering with Temporal Features for Movie Recommendation System," *Procedia Computer Science*, vol. 218, pp. 1366–1373, 2023, doi: 10.1016/j.procs.2023.01.115.

[37] D. Spiliotopoulos, D. Margaris, and C. Vassilakis, "On Exploiting Rating Prediction Accuracy Features in Dense Collaborative Filtering Datasets," *Information*, vol. 13, no. 9, p. 428, Sep. 2022, doi: 10.3390/info13090428.

[38] A. Gonzalez, F. Ortega, D. Perez-Lopez, and S. Alonso, "Bias and Unfairness of Collaborative Filtering Based Recommender Systems in MovieLens Dataset," *IEEE Access*, vol. 10, pp. 68429–68439, 2022, doi: 10.1109/ACCESS.2022.3186719.