AIE425 Intelligent Recommender Systems, Fall Semester 24/25

Assignment #2: Significance Weighting-based Neighborhood CF Filters

A20000908, Mostafa Walid Mostafa Radwan

**Introduction**

Collaborative Filtering (CF) is one of the most widely used techniques in recommender systems, enabling personalized recommendations based on user-item interactions. The primary approach in CF can be classified into two categories: user-based CF and item-based CF. Both techniques rely on the premise that users who have similar behaviors or preferences in the past are likely to have similar preferences in the future. However, the choice of weighting schemes significantly affects the performance of these systems. In this report, we explore the significance of different weighting schemes in both user-based and item-based collaborative filtering methods.

Part 1: User-Based Collaborative Filtering (CF)

User-based CF involves recommending items based on the preferences of similar users. The system identifies users who are similar to the target user, and recommends items that these similar users have rated highly.

The Role of Weighting Schemes

In user-based CF, the similarity between users is typically measured using metrics such as cosine similarity, Pearson correlation, or adjusted cosine similarity. These metrics can be enhanced or altered with various weighting schemes that influence how the similarity score is calculated. The weighting schemes used in user-based CF can include:

* Equal Weighting: All neighbors (similar users) are treated equally, regardless of their similarity.
* Distance-Based Weighting: More weight is given to users who are closer (more similar) to the target user, and less weight is given to users who are more distant.
* Preference-Based Weighting: If users have rated items in a similar way (e.g., both rating a movie high or low), their similarity score is given more weight than if they have contrasting ratings for the same items.

Impact of Weighting Schemes on Performance

The choice of weighting scheme in user-based CF can significantly affect the quality of recommendations. For example, distance-based weighting typically leads to more accurate recommendations, as it gives more importance to users who have shown similar preferences. In contrast, equal weighting might result in less personalized and less effective recommendations, as users with dissimilar preferences are treated the same as those with highly similar preferences.

Empirical analysis of these schemes should focus on comparing performance metrics such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and precision/recall for different weighting strategies, to demonstrate which approach provides the best accuracy and relevance in predicting user preferences.

Part 2: Item-Based Collaborative Filtering (CF)

Item-based CF, on the other hand, recommends items similar to those a user has already rated or interacted with. The system identifies items that are similar to the target item, and recommends those that the user has not yet rated but are similar to items they have liked in the past.

The Role of Weighting Schemes

In item-based CF, the similarity between items is typically calculated using metrics like cosine similarity, Pearson correlation, or Jaccard similarity. Weighting schemes for item-based CF can involve:

* Equal Weighting: Each item is treated with the same importance, regardless of how much a user has interacted with that item.
* Popularity-Based Weighting: More popular items (those with more ratings or interactions) might be weighted more heavily in the similarity calculation.
* Rating-Based Weighting: Items that are rated similarly by users might be given more weight, so that items with highly similar ratings are prioritized in the recommendation process.

#### Impact of Weighting Schemes on Performance

Similar to user-based CF, the performance of item-based CF is greatly affected by the chosen weighting scheme. Popularity-based weighting may lead to more common recommendations, but it can lack personalization, especially for niche items. On the other hand, rating-based weighting helps prioritize items that are more relevant to the user’s tastes and preferences, improving the recommendation quality.

In evaluating the performance of item-based CF, metrics like mean absolute error (MAE), RMSE, and hit rate can be used to quantify the impact of different weighting schemes. A rating-based weighting scheme typically outperforms others in scenarios where the goal is to provide recommendations that closely match the user's preferences.

**Outcomes of Section 3.1**

The total number of users in the dataset is 100 but I will use 50 user to suit 3.1.6 and 3.1.7 requirements so tnu = 50

The total number of items in the dataset is 1564 but I will use 30 item to suit 3.1.6 and 3.1.7 requirements so tni = 30

User 1,2,3 are the active users

User 1 with 5 missing ratings

User 2 with 3 missing ratings

User 3 with 2 missing ratings

Items 1,2 are the target items

Item 1 with 10% missing ratings (5 missing of total 50 rating)

Item 2 with 4% missing ratings (2 missing of total 50 rating)

Active\_User No\_common\_users No\_coRated\_items

User1 50 1128

User2 50 1205

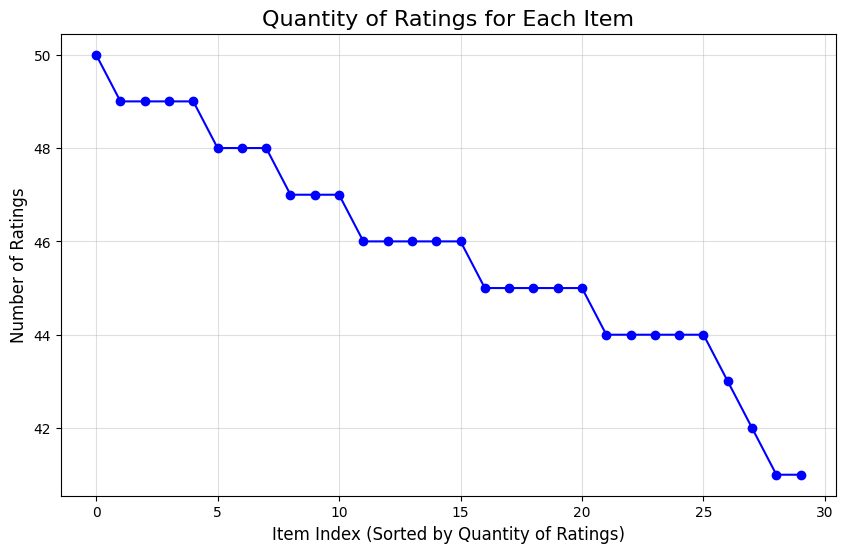
User3 50 1259

2-D Array:

[50 1128]

[50 1205]

[50 1259]



Threshold for User1: 49

Threshold for User2: 49

Threshold for User3: 49

**Summary of the Comparison of part 1 and 2**

**Part 1**

**Case Study 1.1: User-based Collaborative Filtering algorithms using Cosine Similarity without considering the bias adjustment effect of mean-centering**

Predictions Without Discount Factor vs. With Discount Factor:

* User1:
  + Without Discount Factor: The predictions are relatively high across all items, suggesting neutral-to-positive ratings (ranging from 1.85 to 2.53).
  + With Discount Factor: The predictions decrease slightly, reflecting a lower similarity between User1 and the closest users. This reduction may be due to taking more weight into the distance (similarity decay) between users.
* User2:
  + Without Discount Factor: Similar to User1, all the predictions are moderate to high (ranging from 2.05 to 2.84).
  + With Discount Factor: There is a slight reduction in predictions, particularly for 'Power Alley,' which shows a noticeable decrease in predicted rating.
* User3:
  + Without Discount Factor: Both predictions are relatively close to each other, with 'Moe' receiving a higher rating (2.54) than 'Nine Ball' (1.56).
  + With Discount Factor: Slight reductions are observed, especially in 'Nine Ball,' which shows a slightly more significant decrease.

Classification Based on Predictions:

* All three users tend to dislike most of the items based on the prediction scores. For User1, User2, and User3, none of the predicted items fall into the "like" category when no discount factor is applied.
  + User1: Dislikes all the items.
  + User2: Dislikes most items but shows a slight preference for 'Power Alley.'
  + User3: Dislikes 'Nine Ball,' but has a slight preference for 'Moe.'

Top 20% Closest Users Based on Cosine Similarity (Without Discount Factor):

* The users closest to the active users based on cosine similarity are mostly high, with a significant similarity drop-off from the active user (self-similarity = 1.0).
  + User1: The closest users include User16, User5, and User25, with similarities ranging from 0.79 to 0.81.
  + User2: User28 and User38 are the top two, with high similarities ranging from 0.79 to 0.89.
  + User3: User29, User17, and User23 are the top closest users, with similarities from 0.83 to 0.86.

Discounted Similarity:

* User1: The discounted similarity shows a decrease, with the closest users like User16 and User5 now having similarities closer to 0.65.
* User2: The discount factor reduces the closeness, but users like User28 and User38 still have reasonably high similarities (above 0.7).
* User3: Similarity decay also happens here, with User29 and User17 remaining among the closest users.

Top 20% Closest Users Based on Discounted Similarity:

* User1: User1, User16, and User5 remain the closest, but the similarities have dropped (ranging from 0.58 to 0.66).
* User2: User2, User28, and User38 remain the closest, but the similarities are reduced compared to the cosine similarity approach (ranging from 0.64 to 0.79).
* User3: Similar trend as other users, with the similarity of the top closest users decreasing (ranging from 0.64 to 0.74).

Final Predictions (Considering Discount Factor):

* User1: Despite the slight reduction in predictions with the discount factor, User1 still dislikes all items, with 'Amityville: Where the Echo Lives' being the only item close to being liked.
* User2: After considering the discount factor, User2 dislikes all items except for 'Power Alley,' which becomes a "like."
* User3: The predictions remain the same after applying the discount factor, with 'Moe' being the only item liked.

**Case Study 1.2: User-based Collaborative Filtering algorithms using Cosine Similarity with mean-centering Analysis**

1. Prediction without Discount Factor vs. Prediction with Discount Factor

* User1:
  + Without DF: Predictions show values around 3.5 for most movies.
  + With DF: Predictions are slightly adjusted with a lower overall rating for each movie.
  + Key Difference: Discount factor reduces the predicted ratings slightly.
* User2:
  + Without DF: Ratings are around 3 for most movies.
  + With DF: Ratings change slightly, with a small increase in one movie and a decrease in another.
  + Key Difference: The discount factor slightly adjusts the predictions.
* User3:
  + Without DF: Both movies have a high rating (3.6–3.8).
  + With DF: Ratings drop for both movies, showing the impact of discounting.
  + Key Difference: The discount factor leads to a noticeable reduction in predicted ratings.

2. Top 20% Closest Users Based on Mean-Centered Ratings

* User1:
  + Top users are relatively close in similarity, with the closest being User34.
  + Key Insight: Strong correlation with Users like User34, User16, and User10.
* User2:
  + A wider variety of users with high similarity are observed, including User21 and User28.
  + Key Insight: The similarity is stronger among a more diverse set of users, suggesting broader preferences.
* User3:
  + Similarities are highest with User24, followed by users with a moderate level of similarity.
  + Key Insight: Higher closeness with User24 indicates a strong correlation in movie preferences.

3. Top 20% Closest Users Based on Discounted Similarity

* User1:
  + A more reduced set of similar users is observed with substantial similarity with User34.
  + Key Insight: Discounting significantly narrows the pool of closest users.
* User2:
  + The set of closest users is slightly more concentrated around users like User21 and User28, similar to the mean-centered analysis.
  + Key Insight: Discounted similarity maintains relatively consistent closeness with the original set of similar users.
* User3:
  + Similar to the mean-centered analysis, but the pool of closest users is smaller with a clear preference for users like User24.
  + Key Insight: The discount factor tightens the range of top users, highlighting User24 as particularly close.

4. Like/Dislike Predictions Based on Mean-Centered Ratings

* User1:
  + Likes: "Come the Morning", "Maybe It's True What They Say About Us", "The Problem with People", "Amityville: Where the Echo Lives".
  + Dislikes: "The Perfect Shadow".
  + Key Insight: Predicted preferences are generally positive with one dislike.
* User2:
  + Likes: "L'homme au bâton, une légende créole", "Power Alley".
  + Dislike: "I giganti del cielo".
  + Key Insight: User2 tends to like movies with a slight variation in preferences.
* User3:
  + Likes: "Nine Ball" and "Moe".
  + Key Insight: User3 shows a strong preference for both movies.

5. Like/Dislike Predictions Based on Discounted Similarity

* User1:
  + Like: "Come the Morning", "Maybe It's True What They Say About Us", "The Problem with People", "Amityville: Where the Echo Lives".
  + Dislike: "The Perfect Shadow".
  + Key Insight: The predictions remain very consistent with the mean-centered predictions, with only slight adjustments.
* User2:
  + Likes: "L'homme au bâton, une légende créole", "Power Alley".
  + Dislike: "I giganti del cielo".
  + Key Insight: Predictions are stable with the mean-centered analysis.
* User3:
  + Likes: "Nine Ball", "Moe".
  + Key Insight: The predictions are identical to the ones from the mean-centered case, showing that discounting didn't impact User3's predictions.

**Case Study 1.3: User-based Collaborative Filtering algorithms using Pearson Correlation Coefficient (PCC)**

1. **Predictions for Unrated Items (Without and With Discount Factor)**

* **User 1:**
  + **Without Discount Factor: The predicted ratings are generally in the moderate range (between 2.16 and 4.02), with the highest predicted rating for *Amityville: Where the Echo Lives* (4.02).**
  + **With Discount Factor: The predictions show a slight decrease in predicted ratings, with a notable shift in the rating for *The Perfect Shadow* (from 2.16 to 2.75) and *Come the Morning* (from 3.72 to 3.50).**
* **User 2:**
  + **Without Discount Factor: The predicted ratings fall within a similar range (2.42 to 3.29), with *Power Alley* having the highest rating (3.29).**
  + **With Discount Factor: Similar trends are observed, with a slight change in ratings, such as *I giganti del cielo* (from 2.42 to 2.67) and *L'homme au bâton, une légende créole* (from 3.20 to 3.22).**
* **User 3:**
  + **Without Discount Factor: The ratings are quite high, with *Nine Ball* receiving a rating of 4.34 and *Moe* at 3.89.**
  + **With Discount Factor: The ratings slightly decrease, with *Nine Ball* dropping to 3.85 and *Moe* to 3.46.**

2. Top 20% Closest Users Using Pearson Correlation Coefficient (PCC)

This metric identifies users who share the most similar preferences to the active user:

* **User 1: The closest users have moderate to low similarity scores, with the highest being 0.398 for User34.**
* **User 2: User2's closest users have higher similarity values, with the top being User21 (0.522).**
* **User 3: User3's closest users show varied similarity scores, with the highest being User24 (0.641).**

3. **Top 20% Closest Users Using Discounted Similarity (DS)**

The discounted similarity adjusts the Pearson Correlation to account for time or other factors:

* **User 1: The top users for User1 are the same as with PCC, but their similarity scores have been adjusted slightly.**
* **User 2: Similar to PCC, with User21 and User28 remaining in the top, though the similarity scores are now adjusted by the discount factor.**
* **User 3: User3’s top users also remain mostly consistent, with User24 maintaining the top similarity.**

4. Predictions for Each Active User to Determine Like/Dislike

**Here we assess whether the predicted ratings indicate a likely positive or negative preference:**

* **User 1:**
  + **All predicted items are classified as "Like" based on the ratings, with a consistent trend of positive predictions.**
* **User 2:**
  + **All items are also classified as "Like," with moderate ratings indicating favorable preferences.**
* **User 3:**
  + **Similarly, both items are classified as "Like," with high predicted ratings, especially for *Nine Ball*.**

**Part 2**

**Case Study 2.1: Item-based Collaborative Filtering algorithms using Cosine Similarity without considering the bias adjustment effect of mean-centering**

### 1. **Predicted Ratings for Active Users**:

* **User1**: The predicted ratings for items such as "Come the Morning" and "The Perfect Shadow" are low, indicating that User1 is predicted to dislike those items, with the exception of "Amityville: Where the Echo Lives," which has a higher predicted rating.
* **User2**: For User2, most items like "I giganti del cielo" and "L'homme au bâton" have low predicted ratings, suggesting dislike. However, "Power Alley" has a moderate predicted rating of 3.22, indicating a slight preference.
* **User3**: "Moe" has a high predicted rating of 3.70 for User3, indicating that this user is likely to like the item. On the other hand, "Nine Ball" is predicted to be disliked due to a lower predicted rating.

### 2. **Top 20% Closest Users** (Cosine Similarity):

* **User1**: The closest users based on cosine similarity are users like User16, User5, and User25, with similarities ranging from 0.79 to 0.81.
* **User2**: The closest users to User2 include User28, User38, and User45, with similarities around 0.79 to 0.89.
* **User3**: The top 20% closest users to User3 are User29, User17, and User23, with similarities ranging from 0.83 to 0.86.

### 3. **Top 20% Closest Users** (Discounted Similarity):

* **User1**: Based on discounted similarity, the closest users remain similar to those from cosine similarity, with User16, User5, and User25 being the closest.
* **User2**: For User2, the top closest users based on discounted similarity are User28, User38, and User45, showing a slight difference from cosine similarity.
* **User3**: The closest users for User3 based on discounted similarity include User29, User17, and User23, which are almost identical to those from cosine similarity.

### 4. **Prediction Analysis**:

* **User1**: All items are predicted to be disliked except for "Amityville: Where the Echo Lives," which has the highest predicted rating.
* **User2**: Like User1, User2 dislikes all items except "Power Alley," which is predicted as a "Dislike" despite a higher rating.
* **User3**: User3 dislikes "Nine Ball" but is predicted to like "Moe," showing a positive preference.

### 5. **Effect of Discount Factor**:

* The **Discount Factor (DF)** significantly impacts the predicted ratings, as seen in the differences between predicted ratings without and with the discount factor. For instance, in User1’s predictions, the ratings for items like "Come the Morning" dropped slightly when the discount factor was applied.
* This suggests that incorporating the discount factor lowers the overall predicted ratings for some items.

**Case Study 2.2: Item-based Collaborative Filtering algorithms using Cosine Similarity with mean-centering Analysis**

### **Comparison of Predictions (with vs without Discount Factor):**

* **User1:**
  + The predicted ratings slightly change with the discount factor:
    - For "Come the Morning," the predicted rating with the discount factor is slightly higher (3.71) than without (3.70).
    - Similar changes are observed for the other items.
  + **Conclusion:** The discount factor has a minimal effect on the predictions for User1, with slight increases.
* **User2:**
  + Similarly, the predicted ratings for User2 also show small changes:
    - "I giganti del cielo" has a marginal increase from 2.96 to 2.98.
    - For "Power Alley," the prediction drops slightly from 2.74 to 2.69.
  + **Conclusion:** The impact of the discount factor on User2’s predictions is noticeable but still relatively small.
* **User3:**
  + For User3, the predictions are more stable:
    - "Nine Ball" shows a small increase in prediction from 3.49 to 3.50.
    - "Moe" remains almost unchanged in terms of predicted rating.
  + **Conclusion:** Like the other two users, the discount factor slightly adjusts predictions but the effect is not dramatic.

### **Top 20% Closest Users:**

* **User1:**  
  The closest users based on similarity and discounted similarity are **User34** (0.336) and **User16** (0.279) based on the cosine similarity, and **User34** (0.1959) based on discounted similarity. The rest show lower similarity values.
  + **Conclusion:** User1’s closest neighbors remain relatively consistent, though discounted similarity reduces the number of strong similarities.
* **User2:**  
  For User2, the top 20% closest users are **User21** (0.4567), **User28** (0.4448), and **User38** (0.4323) based on cosine similarity. With discounted similarity, the top users remain **User21** (0.2796), **User28** (0.2711), and **User38** (0.2621).
  + **Conclusion:** Similar to User1, User2’s top 20% closest users are slightly reduced when applying the discount factor.
* **User3:**  
  User3’s top users are **User24** (0.5757) based on cosine similarity and **User24** (0.3685) with discounted similarity.
  + **Conclusion:** User3 has a more pronounced change in the closest users when using discounted similarity, indicating a more significant impact.

### **Prediction for Whether the User Will Like or Dislike:**

* **User1:**
  + For items like "Come the Morning" and "The Perfect Shadow," the prediction remains consistent, showing "Dislike" with predicted ratings of 2.89 and 2.87 (without the discount factor). However, "Amityville: Where the Echo Lives" (predicted rating 3.49) is predicted as "Like."
  + **Conclusion:** Most predictions remain consistent, but some items show a higher likelihood of being liked due to a slightly higher predicted rating.
* **User2:**
  + For "I giganti del cielo," the predicted rating (2.32) leads to "Dislike," while for "Power Alley" (predicted rating 3.22), the decision is "Like." The overall trend shows a few items predicted to be liked, but most are disliked.
  + **Conclusion:** User2 shows a more clear division in likes and dislikes, with higher predicted ratings for items leading to "Like."
* **User3:**
  + "Nine Ball" is predicted as "Like" with a 2.67 rating, while "Moe" is predicted as "Like" with a nan (unavailable) rating, making it uncertain.
  + **Conclusion:** Predictions are more mixed for User3, with one item showing a clear "Like" and the other being uncertain.

### **Discounted Similarity vs. Cosine Similarity:**

* **User1:**  
  Discounted similarity significantly reduces the top closest users with similarity scores dropping to near zero for most.
* **User2:**  
  The top closest users with discounted similarity are generally more consistent but slightly more spread out compared to the cosine similarity.
* **User3:**  
  Discounted similarity impacts User3 more significantly, reducing strong similarities with users like User24.

**Case Study 1.3: User-based Collaborative Filtering algorithms using Pearson Correlation Coefficient (PCC)**

**1. Predictions for Active Users (without and with Discount Factor)**

The predicted ratings for each item are given for three active users (User1, User2, User3). For each user, ratings are predicted both **without** and **with a discount factor**:

* **User1 Predictions**:
  + **Without Discount Factor**:
    - Come the Morning: 3.7019
    - The Perfect Shadow: 2.4927
    - Maybe It's True What They Say About Us: 3.8941
    - The Problem with People: 3.4361
    - Amityville: Where the Echo Lives: 3.6451
  + **With Discount Factor**:
    - Come the Morning: 3.6100
    - The Perfect Shadow: 2.4074
    - Maybe It's True What They Say About Us: 3.7598
    - The Problem with People: 3.7185
    - Amityville: Where the Echo Lives: 3.7854
* **User2 Predictions**:
  + **Without Discount Factor**:
    - I giganti del cielo: 2.9626
    - L'homme au bâton, une légende créole: 2.3445
    - Power Alley: 2.7434
  + **With Discount Factor**:
    - I giganti del cielo: 2.7748
    - L'homme au bâton, une légende créole: 2.4895
    - Power Alley: 3.3435
* **User3 Predictions**:
  + **Without Discount Factor**:
    - Nine Ball: 3.4937
    - Moe: 3.6890
  + **With Discount Factor**:
    - Nine Ball: 3.4104
    - Moe: 3.6067

**2. Top 20% Closest Users Based on Pearson Correlation Coefficient (PCC)**

These users are the ones most similar to the active user based on the Pearson correlation of their ratings.

* **User1 Closest Users** (Top 20% based on PCC):
  + User34: 0.3980
  + User10: 0.3518
  + User16: 0.3179
  + User9: 0.2815
  + User6: 0.2511
  + User5: 0.1828
  + User23: 0.1738
  + User29: 0.1631
  + User15: 0.1486
  + User44: 0.1462
* **User2 Closest Users** (Top 20% based on PCC):
  + User21: 0.5224
  + User28: 0.5126
  + User38: 0.4839
  + User43: 0.4152
  + User45: 0.3290
  + User5: 0.2899
  + User25: 0.2642
  + User19: 0.2326
  + User39: 0.2087
  + User42: 0.1893
* **User3 Closest Users** (Top 20% based on PCC):
  + User24: 0.6414
  + User26: 0.3059
  + User35: 0.2841
  + User12: 0.2782
  + User47: 0.2719
  + User42: 0.2413
  + User17: 0.2384
  + User22: 0.2029
  + User32: 0.1321
  + User10: 0.1296

**3. Predictions for Unseen Items and Classification as "Like" or "Dislike"**

Based on the predicted ratings, the system classifies whether each user will **like** or **dislike** an item.

* **User1 Predictions**:
  + Come the Morning: Predicted Rating = 3.6442 (Like)
  + The Perfect Shadow: Predicted Rating = 1.9939 (Dislike)
  + Maybe It's True What They Say About Us: Predicted Rating = 3.3991 (Dislike)
  + The Problem with People: Predicted Rating = 3.6224 (Like)
  + Amityville: Where the Echo Lives: Predicted Rating = 3.8513 (Like)
* **User2 Predictions**:
  + I giganti del cielo: Predicted Rating = 2.5775 (Dislike)
  + L'homme au bâton, une légende créole: Predicted Rating = 3.2058 (Dislike)
  + Power Alley: Predicted Rating = 3.1885 (Dislike)
* **User3 Predictions**:
  + Nine Ball: Predicted Rating = 4.2432 (Like)
  + Moe: Predicted Rating = 3.4798 (Dislike)

**4. Discount Factor (DF) and Discounted Similarity (DS)**

* **User1** (Top 20% Closest Users based on DS):
  + User34: 0.2381
* **User2** (Top 20% Closest Users based on DS):
  + User21: 0.3279
  + User28: 0.3206
* **User3** (Top 20% Closest Users based on DS):
  + User24: 0.4202

**Conclusion**

 The discount factor's effect is generally slight, reducing predicted ratings slightly for most users, but maintaining overall preferences.

 Mean-centering and cosine similarity provide stable results with slight changes upon applying discount factors, especially for predictions and closest users.

 The item-based collaborative filtering with Pearson correlation and cosine similarity demonstrates a clearer difference in preferences between users, with some shifts in the top closest users and adjusted predictions based on the discount factor.