A Psychology-Based Mixture Data Clustering Approach for Shared Account Problem in Movie Recommendation Systems

In a collaborative filtering scenario, it is assumed that each row of the user-item rating matrix represents a single user preference. However, one account is usually shared among household members, and thus, the ratings data of users in such accounts will be mixed. Consequently, recommendations would severely fail to follow each user's preferences. To solve this problem, we leverage the correspondence of movie and user features from media psychology research and coin a user character concept, a common latent factor in movie and account features. After establishing the movie feature matrix in the character representation, we can identify the presence of different characters in each account by factoring out the account feature binary matrix from the rating matrix. Minimizing the estimation error of a given mixed data matrix leads to a binary quadratic optimization model. Considering scalability, we relax the binary constraint, approximate the solutions to a convex problem, and solve the model via a modified gradient descent algorithm. Finally, based on the identified characters in each account, the rating matrix will be reconstructed so that each row represents a single user preference. Furthermore, a shared account dataset was generated from MovieLens ratings based on CAMRa2011 statistical features to evaluate the method. Experiments on this dataset demonstrate the efficiency of our proposed method. In summary, to tackle the shared account problem, we offer a fast yet efficient mixture data clustering approach of ratings that hinges on the psychology of entertainment.

Keywords:

Movie Recommendation System, Shared Account, Mixture Data Clustering, Convex Optimization

1 INTRODUCTION

A Recommender System is vital in e-commerce, media streaming, and social platforms, creating substantial value for customers and businesses [1], [2]. In such systems, a single user interacts with available items via an account, and the most desirable items for the user are predicted based on the user's preferences. The preference data of users are generally collected explicitly expressed, e.g., as a point picked from a finite scale by users, or are inferred by interpreting the actions of the users [3]. Either rating data are arranged in a user-item matrix, in which each row contains a user's preferences for interacted items or missing values. The main problem in RS is finding these missing values [4].

Various approaches have been developed to solve this problem. These approaches can be categorized as Content-Based and Collaborative Filtering. The CB approach generates recommendations based on similarities between active user preferences and item attributes. The CF methods extract the user's preference pattern from past ratings to estimate future preferences; the two major approaches are Neighborhood-Based and Model-Based [5]. Matrix Factorization and Matrix Completion are among the most acclaimed methods in Model-Based approaches due to their capacity to incorporate auxiliary information, be scalable, and avoid the cold start problem [6].

Besides, there is a common belief that the user's and the item's features correlate, and only a few latent features influence how much a user likes an item [7]. Mathematically, the user-item matrix can be seen as a high-dimensional vector space where users and items can be linearly represented in fairly lower-dimensional subspaces [8]. This amounts to the fundamental modeling assumption of CF scenarios. In the MF method, we assume ratings follow a linear model over items' and users' features [9], and equivalently, in the MC, we presume the rating matrix is low-rank [9].

Nevertheless, in reality, the user who was supposed to be single often shares her account. This is particularly true for subscription media streaming platforms like Netflix and Spotify [10]. Recently, account sharing has become a prominent social phenomenon, to the extent that Netflix announced a password crackdown policy to hinder sharing beyond a household [11]. Sharing practice immediately affects the rating matrix so that its rows may contain a mixture of different individual preferences. In this situation, simply applying a recommendation algorithm leads to inferior results.

Consequently, the abovementioned fundamental assumption will no longer be valid, i.e., the ratings would not follow a linear model over users' and items' features. As a result, MF methods would fail [12]. Also, the rating matrix would be a high rank that cannot be completed via conventional methods [13]. This problematic situation is called the shared account problem. It incurs a loss of value for users and businesses by polluting recommendations [10], [14], [15].

This problem has been tackled mainly through the User Identification task in the RS literature. The UI goal is to attribute the items a particular user consumes to the user while the mapping is unknown. Proper recommendations can be generated knowing just a few favorite items of each user and desirably diminish the negative effect of account sharing. As we elaborate in the next section, most studies addressed this task in a content-based or session feed setting by excessively exploiting geo-temporal users' data. These data are not always available, and incorporating them in the algorithms leads to over complexity.

This work focuses on movie recommender systems since sharing movie accounts is more prevalent among other fields [10]. We regard the problem as a mixture of data clustering while the data is not fully observed and look for a simple, fast, and effective remedy. Before making recommendations, the exact identification of users within a shared account is not necessary. It is enough to differentiate the ratings of users sharing an account. To generalize users in a given account, we need some shared characteristics. Inspired by the demixing approach in [13] and incorporating cognitive science in mathematical modeling in [16]–[18], we seek a predictive variable that explains movie-consuming behavior in the psychology of entertainment studies. In this sense, we will choose a proxy variable to identify distinct users' presence in an account—section 2 reviews shared account problem literature and the psychological relationship between users and movies.

In Section 3, the proposed method is explained. After elaborating on the main problem, we describe the link between users and their proxies. We will develop a model for characterizing users based on their interest in movie genres. To do so, we bring psychology theories in line with computational methods. These proxies are chosen in an account’s content articulation, providing a common feature space for movies and users, allowing the linear relationship of ratings over movies and accounts that reassure the fundamental assumption of the matrix factorization approach again. Factoring out the unknown part of this linear equation, the account feature desirably illuminates the presence of distinct users in an account. Then, we cast the problem in a binary quadratic formulation and propose a convex computational procedure to solve it. Next, we use a simple partitioning approach to decompose ratings. Finally, based on the identified characters in each account, the rating matrix will be reconstructed so that each row represents a single user preference. In Section 4, we evaluate the proposed method. Due to the lack of a standard shared account dataset, we generated one using the MovieLens 100K dataset [19]. Finally, we discuss our approach and future research directions in Section 5.

2 LITERATURE REVIEW

Before starting the review, an analogy and explanation need to be made. A Group Recommender System aims to aggregate the preferences of individual users within a group and generate recommendations for a group of users collectively [20]. In contrast, herein, we separate the preferences of individual users within a group and create recommendations for a group of users individually. Besides, the practice of account sharing has also been investigated in [10], [21]–[23], answering questions like why people share their accounts. Notably, these studies reported that a couple, a family, or roommate friends share most accounts. These small social groups could be referred to as a household. However, a gap within these studies is investigating the demographic distribution of account users.

This review is twofold. First, we look at the developed solutions for the shared account problem in RS. Then, motivated by leveraging the psychological aspect of humans to identify users in a shared account, the second part aims to shed light on the relationship between movies and their audiences.

2.1 Shared Account problem in RS review

The problem of identifying users within a shared account was first addressed in a supervised manner at the RecSys 2011 ACM conference [24]. However, the first formal attempt that realistically modeled and solved the problem was [12]. Their method merely takes the user-movie rating matrix into account. The item's feature vector is available via rating matrix factorization. But to estimate the user profile (the number of users within an account and each one rated which movies), they have proposed a mean square error minimization program over some observed user-related points. Each point was made by extending the rating given to a film with one and its feature vector. Geometrically, these points lie on a union of some hyperplanes, called subspace arrangement. The number of hyperplanes in each arrangement and the closest hyperplane to each point and its normal vector would be equivalent to the number of users in an account, the users who rated this movie, and the user's feature vector, respectively. This insight enabled them to solve the problem in a subspace clustering manner via Spectral Clustering and Expectation Maximization methods. They evaluated their approach on the CAMRa2011 dataset, focusing on 272 rows of its rating matrix, which entails two-user household accounts. To measure their proposal performance, they introduced a similarity criterion, which is a weighted sum. If each movie's identified user matched the ground truth map, one would be added to the sum. Similarity near 1 shows excellent performance, while similarity near 0 reflects poor performance.

In another pioneering work, Verstrepen et al. [15] have considered a Top-N Item-Based collaborative recommendation system with binary rating input and no available contextual information. They devised a solution by tackling three problems arising from account sharing. First, the K Nearest Neighbor items for the preferred items of an account are determined. Then, the recommendation score for those K items is computed. The generality problem arises when the system cannot distinguish between a score that is the sum of a few significant or many small contributions. This problem will be solved by a length adjustment procedure, in which the recommendation score for all possible subsets of preferred items in each account will be approximated. Then, it will be divided by the size of each subset. This procedure will balance all scores for users within an account. The second problem happens because the system does not consider that a specific user may have more available ratings than others. To solve this dominance problem, a disambiguating procedure is proposed, which is vague itself. The performance of the approach is measured by a fraction of users who do not get any individualized recommendations. Their approach is limiting and only accountable in an Item-Base Top-N RS with binary input.

Another idea to solve the SA problem is to look for different consumption patterns, such as geographical and temporal. Wang et al. in [25] proposed a method that decomposes users based on mining other preferences over different periods. Similarly, Lesaege et al. [27] tried to differentiate the active user by learning the periodic temporal consumption pattern, such as news on weekday afternoons. To do so, they assumed that movie consumption followed a Dirichlet distribution and maximized the likelihood of observed data. They paired users of MovieTweetings to synthesize an SA dataset. They also used the similarity criterion as in [12].

Jiang et al. [26] proposed an interesting, rather too complex method[27]. Initially, a graph of items is constructed from their metadata, such as genre, artist, and publication year. Then, a feature representation of nodes will be learned by maximizing the likelihood of observations. Unlike the previously mentioned study [28], Jiang et al. assumed that the user's behavior follows the Poisson distribution. They also defined a function to capture each session's feature representation based on the occurrence of each item in a given session. Finally, user identification is promised via clustering and grouping these session features.

Lian et al. [29] look for interest incoherence in an account by computing the average of the item's similarity preferred therein. Yang et al. [30] took a broader approach, assuming users sharing an account have different tastes in movie features like genre, which vary depending on user-related contextual information such as device and location. Accordingly, present personas in each account will be learned as distinct user preference patterns. To do so, the user's consumption vectors are projected into a lower dimensional subspace by PCA. If these lower dimensional vectors were not close to each other, it reflects the presence of different personas. Like Jiang et al. [27], clustering these vectors will determine personas. Eventually, mapping the active user to the predicted personas will be done via learning a regression model.

There is also a line of solutions that leverage the perceptional aspect of the user. Mao et al. [31] developed a probabilistic method that models users' patience in an account. Furthermore, they devised a recommendation system that makes a tradeoff between recommending discriminating items for user identification and preferable ones to keep the user interested. Wen et al. [32] proposed using the human attention mechanism.

Pimentel [13] was the first to consider the rating matrix of a shared account as a mixture of data structures. Although the main application of his study is image processing, he has proposed an alternating cluster-complete method to solve such a problem. It considers a mixture matrix with missing entries that need to be completed. This matrix is a high rank. First, it needs to be decomposed into low-rank partitions. Then, each part can be completed by conventional methods. This method is rather too complex to be applied in real-world RS. It is worth noting that most studies compared their propositions with a simple RS algorithm, where the presence of multiple users in an account has been neglected. All in all, there is no thorough evaluation between developed methods.

2.2 Media-Audience Literature Review

As mentioned in Section 1, instead of identifying users via complex mathematical tasks, we try to do it via the link between people within a household (users within an account) and media preference. We try to extract a quantitative association between users and movie features that determine their preferences. To do so, we review the studies incorporating individual differences in media preference.

Generally, these studies can be categorized into two lines: psychology-informed RS and psychology of entertainment. Psychology-informed RS provides a deeper understanding of user behavior and uses psychological relations between items and users to improve the recommendations [33]. This relation can be based on personality [34], emotions [35], values [36], arousal [37], or demographic traits. In this context, personality refers to enduring patterns of mental and behavioral nuances [38]. Demographic characteristics are rule-based individual differences such as role, age, and gender.

Regarding movies, the genre is the most inclusive movie feature that the user's preference depends on, according to the literature. Furthermore, regarding media selection, it is believed that gender can be a powerful predictor [39]. Hence, we leverage demographic traits, particularly gender and age features, as the proxy for each household member because a general, role-based variable is needed to associate household members with the user's preference within an account so that it could intensify the presence of each member. Thus, we scope out personality- and emotion-related traits and narrow our review report to studies incorporating demographics. In the following, we concisely review studies that aim to predict movie genre preference from user traits.

Modeling users via demographic characteristics was approached in a cognitive science study by Rich in 1979 [40], outwardly. Cantador et al. [41] extracted the association rules between movie genre and personality traits of male and female users. Motamedi planned to devise an RS incorporating User Centric Item Characteristics approach [42]. This approach models the items in a user perceptional way. Interestingly, Hass et al. [16] model the operation of human memory by adopting theories of human cognition to predict the preferable time to listen to a music track to improve recommendations. However, there were also shortcomings. For example, although children and families comprise a large proportion of movie consumers, no comprehensive study incorporates their preferences. Nevertheless, it can be seen that the most preferred genres to watch together in a family are comedy, sci-fi, and animation by a quick search [43].

Leaning on demographics for analysis should be vigilant due to its stereotypical effect. Although one might claim it causes deductive conclusions negatively, it has positive aspects. To illuminate, the negative aspect happens when generalizing is established upon this assumption that if a few people in a group have a characteristic, all of them do [44]. On the contrary, a recent systematic study reviewed papers on information retrieval and recommender systems incorporating gender [45]. They reported how gender is used and tried to orientate the community towards thoughtful, conscientious use and non-use of gender. Furthermore, AlRossais et al. used the positive aspect of stereotyping to solve the cold start problem in RS. They stereotypically modeled the users via the item's features [46]. Similarly, we seek a quantitative correspondence between movie genre preference and gender. The following table presents a quick review of studies that measure how much each gender prefers a different genre.

Table 1- Gender-Genre Correspondence empirical studies

Article	**Correspondence (**media vs. audience**)**	**participants Age group**
Redfern-2012 [47]	Movie genre preference vs. gender age	15-55
Romans-2015 [48]	Movie genre preference vs. gender and personality	18-22
Veenstra-2017 [49]	Movie genre preference vs. gender and age	16-18
Wühr-2017 [50]	Movie genre preference vs. gender	18-45

Infortunata-2021[51]	Movie genre preference vs. gender, age, and mood	18-65

The Gender-Genre correspondence of the abovementioned studies is all in accordance. As we accessed the survey's data of Wühr et al. [48], we utilize its experiment’s data to derive and generalize the actual Gender-Genre correspondence.

3 MODELING

3.1 Problem Definition

Consider a collaborative filtering movie recommender system with a 5-star user feedback setting, where, realistically, each account could be used by multiple users. Let n_A and n_M denote the number of accounts and the number of movies in the system. Although conventionally, we had a user-movie rating, herein, the available data $R^{\Omega} \in \mathbb{R}^{n_A \times n_M}$ is an account-movie rating matrix. Ω shows the set of observed entries, and the missing entries are filled with zeros. The ratings in each account are indeed a mixture of ratings of users within the account. For example, if the users' ratings in the i^{th} account are R_{i^1} and R_{i^2}, shown in Figure 1. What the system can see is R_i, which is the mixture of R_{i^1} and R_{i^2}.

$$\boxed{\begin{matrix} R_{i^1} = [1 & 0 & 0 & 5 & 3 & 0 & 0 & 0 & \dots & 0] \\ R_{i^2} = [0 & 4 & 0 & 0 & 0 & 1 & 0 & 0 & \dots & 0] \end{matrix}} \Longrightarrow \boxed{R_i = [1 \quad 4 \quad 0 \quad 5 \quad 3 \quad 1 \quad 0 \quad 0 \quad \dots \quad 0]}$$

Figure 1: An example to illuminate the mixture. Left: Ratings of two distinct users sharing their account and right: What system saves as ratings of account *i*

Our first assumption is the ratings in each account have no collision. By that, we assert the system solely collects and saves one specific feedback for each movie from the users in an account. For example, if each one of the two users within an account liked a movie and gave a rating of 5, the system would not add these two ratings. It saves a preference of 5 for that specific movie in that account. Our second assumption is common in shared account studies [12], [15], [22]. Each account is being used by people within a household.

Upon these assumptions, the rating matrix is a mixture of different household members' preferences. By a mixture, we mean that every nonzero entry is equal to the entry of one of the low-rank matrices associated to each household member. Mathematically, the rating matrix lies in a union of (unknown) low-dimensional subspaces, while the data is not fully observed. Now, demixing the shared rating matrix will identify users.

This demixing problem is rather complex. Although there are solutions like [52], [53], and the inspiring work of [13], they are too complicated to be embedded in an RS. We look for a lower dimensional user-related common subspace between accounts and movies to avoid over-complexity and let the user identification be scalable, such that the high-rank account-movie rating matrix R^{Ω} could be clustered in some low-rank user-movie submatrices, each of which follows a linear model over users and movies. Eventually, completing each submatrix produces recommendations.

Based on the fundamental assumption of CF, the account-movie rating matrix cannot be described over users. But, it can be factorized in an account feature $A \in \mathbb{R}^{n_A \times k}$ and a movie feature matrix $M \in \mathbb{R}^{n_M \times k}$ given by the following equation:

$$R = \langle A.M \rangle + Z + E \qquad (1)$$

Where the symbol $\langle . \rangle$ denotes the inner matrix product, and Z and E are the bias and error terms. Simplifyingly, we ignore the terms of bias and error. Since we need to identify users, Equation (1) will not take us anywhere. As a remedy, let’s force the account feature matrix entries to be user-related. These features cannot be seen directly but are observed in.

We need to learn them by inferring users' recurring actions. Herein, we only have ratings. Suppose we can fix the latent features to be some characters and describe movies by those characters in M. Then, factoring out the unknown account feature matrix A reveals the presence of each character within an account. Desirably, we would have identified users in a general account characters manner. A quantitative approach for describing the movies in that user-related articulation will be proposed in the following.

3.2 Movie Modeling

As mentioned in Section 2, media psychology suggests a correspondence between the most critical movie feature, which is the genre, and a user's demographics, which is gender. Also, remember that it is assumed that each account is shared by people within a household. Generally, each household member could be related to gender and age. Hence, each account could comprise a Female, a Male, and a Child, which we call characters. Since we focus on their preferences, another character should be added to this list. When a preference is neutrally related to all of them simultaneously, it could be characterized as the family taste or the so-called watch together. Although there is a lack of studies on children and family genre preference in the literature, we use an intuition explained at the end of this section. All in all, the set of characters within accounts is defined as set below and depicted in Figure 1.

$$S_C = \{Female, Male, Family, Children\}$$

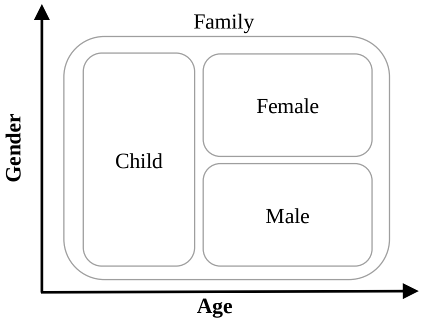

Figure 1- Household and account model

Asserting gender pluralism, we will determine a spectrum of each character's presence within accounts. To model the movies in a character articulation, we measure the weight of each character preference to each movie genre w_{cg} by the following equation:

$$w_{cg} = \frac{P_{cg}}{\sum_{c \in S_C} P_{cg}} \qquad (2)$$

Where c and g are, respectively, character and genre indices, and P_{cg} is the average of the actual preference of character c to genre g. Applying Equation 2 to the user's actual preference for each genre from the survey data of Wühr et al. [50] would lead to Table 2. Notably, although these characters' preferences are statistical aggregates, naturally, exceptions may exist.

Table 2- Adult Character-Genre Correspondence

W	Male	Female
Action	0.59	0.41
Adventure	0.56	0.44
Animation	0.53	0.47
Comedy	0.51	0.49
Crime	0.48	0.52
Drama	0.45	0.55
Erotic	0.65	0.35
Fantasy	0.57	0.43
Series	0.46	0.54
History	0.5	0.5
Horror	0.59	0.41
Mystery	0.57	0.43
Romance	0.37	0.63
Sci-Fi	0.67	0.33
Thriller	0.54	0.46
War	0.63	0.37
Western	0.64	0.36

The intuition for relating family and children's characters to genres is as follows. From the table above, we can see that male and female characters prefer some genres almost equally. Neural genres intuitively are considered those of male-female weight in a bin of range [0.49, 0.51]. Movies with many of these neural genres, such as comedy, can be categorized as "watch together" or family movies. For children, there is a specific genre in our movie data set: children. If a movie has this genre, we assign a weight of 1 in favor of the children's character in the Character-Genre Correspondence Matrix. The final weight matrix is delineated in Table 3. Now, accumulating all of these weights for each movie genre composition, we can compute the amount of each character feature available in each movie by the following equation:

$$M = \langle G, W \rangle \qquad (3)$$

Where $G \in [0,1]^{\boldsymbol{n}_M \times \boldsymbol{n}_g}$ is the Movie-Genre composition matrix, which is easily available in the system and $W \in [0,1]^{\boldsymbol{n}_c \times \boldsymbol{n}_g}$ is the Character-Genre correspondence matrix produced by Equation (2). Each row of M is a vector representation of each movie in a character manner, as the movie's latent factors.

Table 3- Character-Genre Correspondence

Genre	Male	Female	Children	Family
Action	0.59	0.41	0	0
Adventure	0.4	0.27	0	0.33
Animation	0.15	0.1	0.63	0.13

Children	0	0	1	0
Comedy	0.26	0.24	0.25	0.25
Crime	0.35	0.38	0	0.27
Documentary	0.33	0.33	0	0.33
Drama	0.45	0.55	0	0
Fantasy	0.57	0.43	0	0
Series	0.3	0.37	0	0.33
Horror	0.59	0.41	0	0
Musical	0.05	0.35	0.45	0.15
Mystery	0.57	0.43	0	0
Romance	0.37	0.63	0	0
Sci-Fi	0.5	0.16	0	0.33
Thriller	0.54	0.46	0	0
War	0.63	0.37	0	0
Western	0.64	0.36	0	0

3.3 Problem Formulation

Consider the simplified version of Equation (1), given R and having M via Equation (3). If we specify A as a binary matrix, each array shows the presence or absence of each character in a given account. We will use this binary format at the end of this section to cluster ratings. Specifying A in can be done by minimizing the Frobenius norm error over a binary domain, as follows:

$$\min_{\{0,1\}^{n_A \times n_c}} \|R - O \odot X.M\|_F^2 \qquad (4)$$

Where $\odot$ is the elementwise multiplication, and O is the observations matrix, which is defined as follows:

$$O_{ij} = \begin{cases} 1 & (i,j) \in \Omega \\ 0 & (i,j) \notin \Omega \end{cases} \qquad (5)$$

Equation (4) denotes a Binary Quadratic Program, which is NP-hard. Although there are suitable methods to solve it, due to the complexity and high dimensionality of the problem and the necessity of an on-demand response for the solution, any approach should be fast and scalable. So first, we relax the binary constraint by adding a term to the objective function that makes the solutions near {0, 1} based on the penalizing algorithm [54]. Thus, equation (4) will become the following convex program:

$$min\left(\|R - O \odot X.M\|_F^2 + \frac{p}{2}\|X^2 - X\|_F^2\right) \qquad (6)$$

Where regularization parameter p controls the effect of the penalty term. To solve (6), we approximate the solutions via a modified version of the gradient descent algorithm [54], which is described in Algorithm (1). Where the gradient of the objective function is calculated as follows:

$$\nabla = \left(R - O\odot(X.M)\right).M^T + p(X^2 - X)\odot(2X-1) \qquad (7)$$

So far, we have extracted the amount of each character's presence in each account from the user's ratings generated in that account. But still, there is a problem. The number of different user characters in the system is not equal. This affects and impairs the result of our algorithm. For instance, in the survey data of Wühr et al. [50] and MovieLens100K, the female character goes under the domination of male characters. As a result, the total number of identified female characters would be deficient. It happens because there are just a few numbers of one user character and their favorite movies. We perceive it as the dominance problem introduced by [15]. Thereby, we propose another simple remedy for it, a balancing procedure in which we try to balance all columns of account-feature matrix A.

$$b_i = \mu - \mu_i \qquad (8)$$

$$B_i = A_i + b_i \qquad (9)$$

Consider account-feature matrix A. Herein μ is the overall mean of its arrays, μ_i denotes the mean of the i^{th} column. Thus, adding b_i to the A_i will balance it. Equation (9) will balance A_i and put it in B_i. Wrapping all balanced columns in a matrix format to have a balanced account-user type feature matrix without dominance defect. Embedding this procedure in the algorithm will solve the dominance problem.

Now, we look for the mapping of identified characters and the movies they liked, done via clustering. Since all parameters and variables of the problem are intrinsically nonnegative, it is wise to see that the program in Equation (6) is indeed a non-negative matrix Factorization method. Strohmeier and Needell [55] proposed a well-suited clustering approach. There is an insightful geometric interpretation of the factorization process. It is the orthogonalization effect of NMF on the original subspaces. It is said that the rows of $\hat{A}$ obtained via NMF applied to the mixture matrix R have minimal values at entries corresponding to basis vectors for the lower-dimensional subspaces (character) in which it does not belong.

3.4 Computation

As the binary format of X was desirable, inspired by the Optimization with Bounded ℓ_∞-Norm of [56], we propose using an indicator function after approximation of X so that its arrays become exactly {0, 1}. Since at least one user is in each account, at least one array in each row of A must be 1. As a result, we define the indicator function Π as follows, and the step of applying it to the approximated solutions of (5) will be appended at the end of Algorithm 1.

$$\Pi(x) = \begin{cases} 1 & x \geq l * argmax(x) \\ 0 & x < l * argmax(x) \end{cases} \qquad (10)$$

ALGORITHM 1: User-Character Identification Algorithm

```
Convex function and its Gradient, Initial Point $X^{(0)} = 0, p = 1.1,$ Max Iteration $=10$ or $\varepsilon>0$
Normalize columns of $M$     # so that each movie has the same total weight.
For $t = 0, \ldots,$ Max iteration do
        Set $\alpha^{(t)} \coloneqq 1/1 + t$
        Update $X^{(t+1)} \coloneqq X^{(t)} - \alpha^{(t)} \nabla f(X^{(t)})$
        Update $p \coloneqq p * p$
        Normalize rows of $X^{(t+1)}$
        If the stopping criterion becomes true
End for
Return $\hat{A}$      # as $argmin\ f(X)$
Calculate and return $\hat{A}\_B$     # as $\Pi(\hat{A})$
```

Algorithm (1) output is the User-Character Membership matrix in order of [Male, Female, Children, Family]. For example, suppose the first row of $\hat{A}\_B$ is determined as [1, 0, 0, 1]. It means that two distinct characters are sharing the first account. Thus, a male and a family preference could be extracted from the first row of the rating matrix. Generally, to do that extraction, we need to cluster account-movie ratings. This amounts to identifying the preferences of users behind each account. To do so, we propose a partitioning algorithm for the rating matrix based on the result of clustering in Algorithm (1). The following Algorithm (2) partitions R into $\{R^k\}_{1\leq k\leq K}$, each a low-rank matrix. Herein, we have K $= n_C = 4$ partitions. Its theoretical guarantee comes from the orthogonalization effect inherent in NMF [55].

ALGORITHM 2: Partitioning Mixture Matrix

```
Specify $\{\Omega_M^k\}_{1\leq k\leq K} \leftarrow j$      $if\ \widehat{M}\_B_{kj} > 0$   # indices set of nonzero entries of the $k^{th}$ column of $\widehat{M}\_B = \Pi(M)$.
Specify $\{\Omega_A^k\}_{1\leq k\leq K} \leftarrow i$      $if\ \hat{A}\_\mathrm{B}_{ik} > 0$   # indices set of nonzero entries of $k^{th}$ column of $\hat{A}\_\mathrm{B}$
Specify $\{R_{ij}^k\}_{1\leq k\leq K} \leftarrow R_{ij}\ for\ i \in \Omega_{M_A}^k\ and j \in \Omega_A^k$   # partition k of $R$ that is a low-rank matrix.
```

Now we have *k* submatrices of ratings. Each corresponds to one character. Finally, we need to demix the account-movie ratings based on the characters identified in User-Character Membership $\hat{A}\_B$ and the ratings in each corresponding submatrices to acquire the character-movie rating matrix. To reconstruct the rating matrix, for all found characters in each row of $\hat{A}\_B$, we append the row of the corresponding submatrix. Desirably, the resulting matrix would possess the low-rank property as the fundamental assumption behind the collaborative filtering method. To sum up, we will go over the notations in Table 4.

Table 4-List of notations

notation	**description**
$\boldsymbol{n_A}$	Number of accounts
$\boldsymbol{n_M}$	Number of movies
$\boldsymbol{R}^{\mathbf{sa}}$	Shared ratings matrix ∈ $\mathbb{R}^{n_A \times n_M}$
$\boldsymbol{R}^{\mathbf{re}}$	Reconstructed rating matrix as final output ∈ $\mathbb{R}^{n_{identifed\ characters} \times n_M}$
$\boldsymbol{n_U}$	Number of users
$\boldsymbol{S_C}$	Set of characters =$\{Female, Male, Child, Family\}$
$\boldsymbol{n_C}$	Number of characters
$\boldsymbol{k}$	Submatrices or Character's index
$\boldsymbol{S_G}$	Set of movie genres =$\{Action, \dots, war\}$
$\boldsymbol{n_G}$	Number of genres
$\boldsymbol{g}$	Genre's index
$\boldsymbol{w_{cg}}$	Preference of character *c* for genre *g*

$\boldsymbol{G}$	Movie-Genre composition matrix $\in [0,1]^{\boldsymbol{n_M \times n_g}}$
$\boldsymbol{W}$	Character-Movie genre preference association matrix $\in [0,1]^{\boldsymbol{n_M \times n_g}}$
$\boldsymbol{M}$	Movie-Character feature matrix
$\boldsymbol{\widehat{M}\_B}$	Movie-Character Membership $\in \{0,1\}^{n_A \times n_C}$
$\boldsymbol{j}$	Movie's index
$\boldsymbol{A}$	Account-Character feature matrix $\in [0,1]^{n_A \times n_C}$
$\boldsymbol{i}$	Account's index
$\boldsymbol{O}$	The observations matrix $\in \{0,1\}^{n_A \times n_M}$
$\boldsymbol{\widehat{A}}$	Estimated Account-Character feature matrix $\in [0,1]^{n_A \times n_C}$
$\boldsymbol{\widehat{A}\_B}$	User-Character Membership $\in \{0,1\}^{n_A \times n_C}$

4 NUMERICAL RESULTS AND EVALUATION

4.1 Shared Account Dataset Generation

A standard shared account dataset was needed to evaluate the performance of our proposed method. The CAMRa2011 was released at the Context-Aware Movie Recommendation (CAMRa) challenge. Afterward, the publisher removed it from public access. However, based on [12], we extracted and summarized its statistical characteristics in Table 2.

Table 5: Statistical Characteristics of CAMRa2011

Ratio	Amount
Total number of shared account members per total users	0.33
Shared account of size 2 per total number of shared accounts	0.94
Shared account of size 3 per total number of shared accounts	0.05
Shared account of size 4 per total number of shared accounts	0.01
Number of users per number of accounts	1.2

MovieLens is the most renowned dataset in the recommendation research community. Consider MovieLens 100K rating in a user movie matrix format, named *R*. Then we try to generate a shared account dataset by randomly adding some specific rows of *R* based on ratios from Table 2 above. Recall that our shared account assumption was "no collision". It means two users who rated a common movie could not share an account. The specific rows have no two nonzero elements in the same column. To do so, we benefit from the Algebra of sets. Note that Python code for all mentioned algorithms and procedures is available. Figure 3 compares the density of MovieLens 100K and the generated shared account dataset. For example, the ratings of user 339 and user 573 got mixed, making the shared account 336.

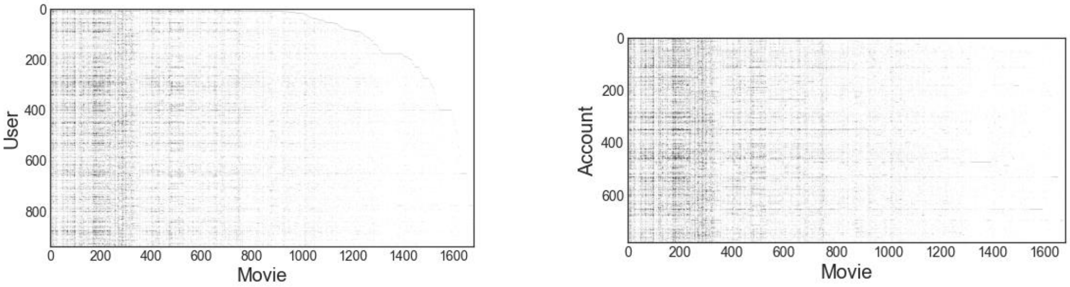

Figure 3: shared account generation from MovieLens100K ratings. Left: before, Right: After, visualization codes are adopted from [57].

4.2 Numerical results

Experimentally, we tune the parameter *l* of the indicator function depicted by equation (7) to be 0.98. After solving the abovementioned problem, results obtained as follows:

Table 6 Part of $\widehat{\mathrm{A}}$

Shared account number	Male	Female	Children	Family
0	0.264	0.251	0.227	0.256
1	0.274	0.223	0.246	0.256
...	...	...	...	...
782	0.249	0.249	0.258	0.244
mean	0.25	0.25	0.25	0.25

Table 7: Part of $\widehat{\mathrm{A}}\_B$

Shared account number	Male	Female	Children	Family
0	1	0	0	0
1	1	0	0	0
...	...	...	...	...
782	0	0	1	0
mean	0.25	0.25	0.25	0.25

The final result is the demixed rating matrix, as described before.

4.3 Evaluation

To measure the performance of our approach, we produce recommendations via the Surprise library [58]; first, we compare the RMSE of the shared rating matrix and the demixed rating matrix, then compute the metric as in [15] based on Top-N recommendation.

$$RMSE\ R^{ML100k}\ = 0.9360$$

$$RMSE\ R^{sa}\ = 0.9452$$

$$RMSE\ R^{re} = 0.9324$$

The decline in $RMSE$ is promising. Furthermore, even the reconstructed ratings of the mixed version produce even better results than the MovieLens100k ratings. The other two metrics are based on the recall of a true user. Recall is the percentage of true users' top-5 recommendations that are also present in the top-5 recommendations for its shared account. The first one indicates the total recall for all users, i.e. the latter measures the portion of users who do not give a proper recommendation, i.e., that do not find a single one of its top-5 individual recommendations in the top-5 recommendations of the shared account to which it belongs.

$$Recall = \frac{|Top5 - true\ users\ in\ account\ a\ in\ true\ ratings\ \cap\ Top5 - account\ a\ in\ shared\ account\ ratings|}{5}$$

$$Rate\ of\ Recall\ R^{sa}\ = 0.2595$$
$$Rate\ of\ Recall\ R^{re} = 0.2207$$

$$Rate\ of\ Recall\ at\ Zero\ R^{sa}\ = 0.1219$$
$$Sum\ of\ Recall\ \ at\ Zero\ R^{re} = 0.1823$$

5 CONCLUSION

We aim to learn a vector representation of each account that reveals different users in each by projecting movies and accounts on some user-typical subspaces. Then, the collaborative filtering model is solved via a non-negative matrix factorization. The user-typical was shaped by gender norms in media psychology studies, a way of individualizing a group. It is not the actual user gender prediction but merely a movie taste-associated property that has been exhibited in an account. First, we propose a novel procedure that computes the movie's user-typical feature vectors. A movie's content is complicated. Digging into the literature of media psychology, we have come to an aggregation of the genre-gender correspondence, which aids in modeling movies and accounts in a joint articulation. We acknowledge gender pluralism. Thus, a spectrum of gender is being considered. After casting the problem in a mixture format, we could automatically group users within an account via a simple clustering task.

Incorporating other information like tags, cast age groups, cast gender frequency, or the movie protagonist available could be added to feature vectors and amplify the exposing effect of different characters in each account. It is a gap in the literature and could benefit from future studies. Our approach labels users in an account. It has no limits for the number of users in each account. The merits of user identification in this manner are beyond solving the SA problem. For instance, in an apparel marketing scenario, this demographic approach aids in targeting advertisements. Furthermore, this approach could be used in other RSs regarding different items such as music. If the genre preference for children and families were reliable enough, more characters in a household could be identified. Furthermore, incorporating the movie genre composition instead of a predetermined binary genre format is advantageous. Some methods quantify the amount of each genre in a given movie based on viewer opinion and perception, such as [59], [60].

Future research directions by incorporating:

- The movie genre composition that is based on users' opinion
- Children and family genre correspondence
- Movie information like tags casts age groups, cast gender frequency, or the protagonist

REFERENCES

[1] N. Ghanem, S. Leitner, and D. Jannach, “Balancing consumer and business value of recommender systems: A simulation-based analysis,” *arXiv Prepr. arXiv2203.05952*, 2022.

[2] D. Jannach and M. Jugovac, “Measuring the business value of recommender systems,” *ACM Trans. Manag. Inf. Syst.*, vol. 10, no. 4, pp. 1–23, 2019.

[3] F. Ricci, “Recommender Systems: Models and Techniques,” *Encycl. Soc. Netw. Anal. Min.*, pp. 2147–2159, 2018, doi: 10.1007/978-1-4939-7131-2_88.

[4] D. Roy and M. Dutta, “A systematic review and research perspective on recommender systems,” *J. Big Data*, vol. 9, no. 1, p. 59, 2022, doi: 10.1186/s40537-022-00592-5.

[5] X. Su and T. M. Khoshgoftaar, “A Survey of Collaborative Filtering Techniques,” *Adv. Artif. Intell.*, vol. 2009, p. 421425, 2009, doi: 10.1155/2009/421425.

[6] R. Mehta and K. Rana, “A review on matrix factorization techniques in recommender systems,” in *2017 2nd International Conference on Communication Systems, Computing and IT Applications (CSCITA)*, 2017, pp. 269–274, doi: 10.1109/CSCITA.2017.8066567.

[7] M. Udell, C. Horn, R. Zadeh, and S. Boyd, “Generalized Low Rank Models,” 2015.

[8] M. D. Ekstrand, J. T. Riedl, and J. A. Konstan, “Collaborative filtering recommender systems,” *Found. Trends Human-Computer Interact.*, vol. 4, no. 2, pp. 81–173, 2010, doi: 10.1561/1100000009.

[9] A. Ramlatchan, M. Yang, Q. Liu, M. Li, J. Wang, and Y. Li, “A survey of matrix completion methods for recommendation systems,” *Big Data Min. Anal.*, vol. 1, no. 4, pp. 308–323, 2018, doi: 10.26599/bdma.2018.9020008.

[10] N. Sailaja and A. Fowler, “An Exploration of Account Sharing Practices on Media Platforms,” in *ACM International Conference on Interactive Media Experiences*, 2022, pp. 141–150, doi: 10.1145/3505284.3529974.

[11] “Netflix Terms of Use,” 2021. https://help.netflix.com/legal/termsofuse.

[12] A. Zhang, N. Fawaz, S. Ioannidis, and A. Montanari, “Guess Who Rated This Movie: Identifying Users through Subspace Clustering,” in *Proceedings of the Twenty-Eighth Conference on Uncertainty in Artificial Intelligence*, 2012, pp. 944–953.

[13] D. L. Pimentel-Alarcón, “Mixture matrix completion,” *arXiv Prepr. arXiv1808.00616*, 2018.

[14] W. Zhang and C. Challis, “Towards addressing unauthorized sharing of subscriptions,” *Appl. Intell.*, 2021, doi: 10.1007/s10489-021-02812-6.

[15] K. Verstrepen and B. Goethals, “Top-N Recommendation for Shared Accounts,” in *Proceedings of the 9th ACM Conference on Recommender Systems*, 2015, pp. 59–66, doi: 10.1145/2792838.2800170.

[16] M. Reiter-Haas, E. Parada-Cabaleiro, M. Schedl, E. Motamedi, M. Tkalcic, and E. Lex, “Predicting Music Relistening Behavior Using the ACT-R Framework,” in *Proceedings of the 15th ACM Conference on Recommender Systems*, 2021, pp. 702–707, doi: 10.1145/3460231.3478846.

[17] G. Hsieh, J. Chen, J. U. Mahmud, and J. Nichols, “You read what you value: understanding personal values and reading interests,” in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2014, pp. 983–986.

[18] L. V. Phan and J. F. Rauthmann, “Personality computing: New frontiers in personality assessment,” *Soc. Personal. Psychol. Compass*, vol. 15, no. 7, p. e12624, 2021, doi: https://doi.org/10.1111/spc3.12624.

[19] J. Konstan, “MovieLens 100K Dataset.” https://grouplens.org/datasets/movielens/100k/.

[20] S. Dara, C. R. Chowdary, and C. Kumar, “A survey on group recommender systems,” *J. Intell. Inf. Syst.*, vol. 54, no. 2, pp. 271–295, 2020, doi: 10.1007/s10844-018-0542-3.

[21] M. Jacobs, H. Cramer, and L. Barkhuus, “Caring About Sharing: Couples’ Practices in Single User Device Access,” in *Proceedings of the 2016 ACM International Conference on Supporting Group Work*, 2016, pp. 235–243, doi: 10.1145/2957276.2957296.

[22] B. Obada-Obieh, Y. Huang, and K. Beznosov, “The Burden of Ending Online Account Sharing,” in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 2020, pp. 1–13, doi: 10.1145/3313831.3376632.

[23] C. Y. Park, C. Faklaris, S. Zhao, A. Sciuto, L. Dabbish, and J. Hong, “Share and share alike? An exploration of secure behaviors in romantic relationships,” in *Fourteenth Symposium on Usable Privacy and Security ({SOUPS} 2018)*, 2018, pp. 83–102.

[24] A. Said, S. Berkovsky, E. W. De Luca, and J. Hermanns, “Challenge on Context-Aware Movie Recommendation: CAMRa2011,” in *Proceedings of the Fifth ACM Conference on Recommender Systems*, 2011, pp. 385–386, doi: 10.1145/2043932.2044015.

[25] Z. Wang, Y. Yang, L. He, and J. Gu, “User Identification within a Shared Account: Improving IP-TV Recommender Performance BT - Advances in Databases and Information Systems,” 2014, pp. 219–233.

[26] J.-Y. Jiang, C.-T. Li, Y. Chen, and W. Wang, “Identifying Users behind Shared Accounts in Online Streaming Services,” in *The 41st International ACM SIGIR Conference on Research &amp; Development in Information Retrieval*, 2018, pp. 65–74, doi: 10.1145/3209978.3210054.

[27] J.-Y. Jiang, C.-T. Li, Y. Chen, and W. Wang, “Identifying Users behind Shared Accounts in Online Streaming Services,” in *The 41st International ACM SIGIR Conference on Research &amp; Development in Information Retrieval*, 2018, pp. 65–74, doi: 10.1145/3209978.3210054.

[28] C. Lesaege, F. Schnitzler, A. Lambert, and J. Vigouroux, “Time-Aware User Identification with Topic Models,” in *2016 IEEE 16th International Conference on Data Mining (ICDM)*, 2016, pp. 997–1002, doi: 10.1109/ICDM.2016.0126.

[29] T. Lian, Z. Li, Z. Chen, and J. Ma, “The Impact of Profile Coherence on Recommendation Performance for Shared Accounts on Smart TVs BT - Information Retrieval,” 2017, pp. 30–41.

[30] S. Yang, S. Sarkhel, S. Mitra, and V. Swaminathan, “Personalized Video Recommendations for Shared Accounts,” in *2017 IEEE International Symposium on Multimedia (ISM)*, 2017, pp. 256–259, doi: 10.1109/ISM.2017.43.

[31] K. Mao, J. Niu, X. Liu, S. Tang, L. Liao, and T.-S. Chua, “A patience-aware recommendation scheme for shared accounts on mobile devices,” *IEEE Trans. Knowl. Data Eng.*, p. 1, 2021, doi: 10.1109/TKDE.2021.3069002.

[32] X. Wen, Z. Peng, S. Huang, S. Wang, and P. S. Yu, “MISS: A Multi-user Identification Network for Shared-Account Session-Aware Recommendation BT - Database Systems for Advanced Applications,” 2021, pp. 228–243.

[33] B. Ferwerda, L. Chen, and M. Tkalčič, “Editorial: Psychological Models for Personalized Human-Computer Interaction (HCI),” *Front. Psychol.*, vol. 12, 2021, doi: 10.3389/fpsyg.2021.673092.

[34] S. Dhelim, N. Aung, M. A. Bouras, H. Ning, and E. Cambria, “A Survey on Personality-Aware Recommendation Systems,” *Artif. Intell. Rev.*, vol. 55, no. 3, pp. 2409–2454, Mar. 2022, doi: 10.1007/s10462-021-10063-7.

[35] M. Tkalčič and B. Ferwerda, “Eudaimonic modeling of moviegoers,” in *Proceedings of the 26th Conference on User Modeling, Adaptation and Personalization*, 2018, pp. 163–167.

[36] E. M. Khan, M. S. H. Mukta, M. E. Ali, and J. Mahmud, “Predicting Users’ Movie Preference and Rating Behavior from Personality and Values,” *ACM Trans. Interact. Intell. Syst.*, vol. 10, no. 3, Oct. 2020, doi: 10.1145/3338244.

[37] S. C. Banerjee, K. Greene, M. Krcmar, Z. Bagdasarov, and D. Ruginyte, “The role of gender and sensation seeking in film choice: Exploring mood and arousal.,” *Journal of Media Psychology: Theories, Methods, and Applications*, vol. 20. Hogrefe & Huber Publishers, Banerjee, Smita C.: 23 Carisbrooke Drive, Mapperley Park, Nottingham, United Kingdom, NG3 5DS, sbanerjee@lincoln.ac.uk, pp. 97–105, 2008, doi: 10.1027/1864-1105.20.3.97.

[38] C. Gerl, M. Stieger, and M. Allemand, “Developmental Changes in Personality Traits,” in *Encyclopedia of Personality and Individual Differences*, V. Zeigler-Hill and T. K. Shackelford, Eds. Cham: Springer International Publishing, 2020, pp. 1083–1092.

[39] B. P. Lange, P. Wühr, and S. Schwarz, “Of Time Gals and Mega Men: Empirical Findings on Gender Differences in Digital Game Genre Preferences and the Accuracy of Respective Gender Stereotypes,” *Front. Psychol.*, vol. 12, 2021, doi: 10.3389/fpsyg.2021.657430.

[40] E. Rich, “User modeling via stereotypes,” *Cogn. Sci.*, vol. 3, no. 4, pp. 329–354, 1979, doi: https://doi.org/10.1016/S0364-0213(79)80012-9.

[41] I. Cantador, I. Fernández-Tobías, and A. Bellogín, “Relating Personality Types with User Preferences in Multiple Entertainment Domains,” 2013.

[42] E. Motamedi, “User-Centric Item Characteristics for Modeling Users and Improving Recommendations,” in *Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization*, 2021, pp. 304–307, doi: 10.1145/3450613.3459659.

[43] “Top 50 Family Movies and TV Shows.” https://www.imdb.com/search/title/?title_type=feature&genres=family.

[44] B. Tucker *et al.*, *Exploring Public Speaking*, 4th Editio. University System of Georgia.

[45] C. Pinney, A. Raj, A. Hanna, and M. D. Ekstrand, “Much Ado About Gender: Current Practices and Future Recommendations for Appropriate Gender-Aware Information Access,” *arXiv Prepr. arXiv2301.04780*, 2023.

[46] N. AlRossais, D. Kudenko, and T. Yuan, “Improving cold-start recommendations using item-based stereotypes,” *User Model. User-adapt. Interact.*, vol. 31, no. 5, pp. 867–905, 2021, doi: 10.1007/s11257-021-09293-9.

[47] N. Redfern, “Correspondence analysis of genre preferences in UK film audiences,” *Particip. J. audience Recept. Stud.*, vol. 9, no. 2, pp. 45–55, 2012, [Online]. Available: http://www.participations.org/Volume 9/Issue 2/4 Redfern.pdf.

[48] A. Romans, “We Are What We Watch: Film Preferences and Personality Correlates,” 2015.

[49] A. Veenstra, *Watching film: An account of contemporary film consumption preferences and practices amongst youth in Flanders aged 16 to 18*. 2017.

[50] P. Wühr, B. P. Lange, and S. Schwarz, “Tears or Fears? Comparing Gender Stereotypes about Movie Preferences to Actual Preferences,” *Front. Psychol.*, vol. 8, p. 428, 2017, doi: 10.3389/fpsyg.2017.00428.

[51] C. Infortuna *et al.*, “The inner muses: How affective temperament traits, gender and age predict film genre preference,” *Pers. Individ. Dif.*, vol. 178, p. 110877, 2021, doi: https://doi.org/10.1016/j.paid.2021.110877.

[52] Y. Chen, C. Ma, H. V. Poor, and Y. Chen, “Learning Mixtures of Low-Rank Models,” *arXiv Prepr. arXiv2009.11282*, 2020.

[53] D. Li, C. Chen, W. Liu, T. Lu, N. Gu, and S. M. Chu, “Mixture-rank matrix approximation for collaborative filtering,” in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017, pp. 477–485.

[54] Z. Zhang, T. Li, C. Ding, and X. Zhang, “Binary Matrix Factorization with Applications,” in *Seventh IEEE International Conference on Data Mining (ICDM 2007)*, 2007, pp. 391–400, doi: 10.1109/ICDM.2007.99.

[55] C. Strohmeier and D. Needell, “Clustering of Nonnegative Data and an Application to Matrix Completion,” in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 8349–8353, doi: 10.1109/ICASSP40776.2020.9052980.

[56] M. Jaggi, “Convex Optimization without Projection Steps,” pp. 1–61, 2011, [Online]. Available: http://arxiv.org/abs/1108.1170.

[57] B. Lindsay, “Exploratory Data Analysis on MovieLens 100K Dataset.” https://github.com/benlindsay/movielens-analysis/blob/master/01_Exploratory-Analysis-on-100K-data.ipynb.

[58] N. Hug, “Surprise: A Python library for recommender systems,” *J. Open Source Softw.*, vol. 5, no. 52, p. 2174,

2020, doi: 10.21105/joss.02174.

[59] O. Gladfelter, “Measuring Film Genres,” 2019. https://cultureplot.com/film-genres/.

[60] A. Pal, A. Barigidad, and A. Mustafi, “Identifying movie genre compositions using neural networks and introducing GenRec-a recommender system based on audience genre perception,” in *2020 5th International Conference on Computing, Communication and Security (ICCCS)*, 2020, pp. 1–7, doi: 10.1109/ICCCS49678.2020.9276893.