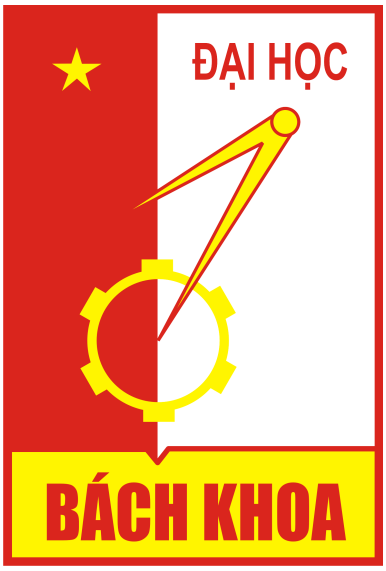
HA NOI UNIVERSITY OF SCIENCE AND TECHNOLOGY
SCHOOL OF INFORMATION AND COMMUNICATION TECHNOLOGY
FACULTY OF COMPUTER SCIENCE

o0o



PROJECT REPORT:
MOVIE RECOMMENDATION SYSTEM

Course: Machine Learning - IT3190E
Class Code: 141178
Supervisor: Associate Prof Than Quang Khoat
Member of our group:

Trinh Hoang Giang	20214893	giang.th214893@sis.hust.edu.vn
Ho Ngoc Anh	20214877	anh.hn214877@sis.hust.edu.vn
Hoang Thanh Dat	20214899	dat.ht214889@sis.hust.edu.vn
Lang Van Quy	20214928	quy.lv214928@sis.hust.edu.vn
Vu Lam Anh	20214876	anh.vl214876@sis.hust.edu.vn

Ha Noi 2023

Abstract

Industry 4.0, recommender systems play an important role in numerous sectors. The use of recommender systems to enhance user experience is widespread in social media, e-commerce, and online platforms. Users are to receive personalized recommendations from the recommender system based on their preferences, behaviors, or previous interactions. The goal of this project is to explore two common methods for movie recommendation systems: collaborative filtering and content-based filtering. Additionally, we investigate several forms of collaborative filtering, such as matrix factorization-based Singular Value Decomposition and Gradient Descent. We also go through how these strategies' performance was judged using a variety of evaluation criteria that were selected and used with recommender systems in mind.

Contents

1 Introduction to Movie Recommendation System

1.1 Short description

A movie recommendation system, or a movie recommender system, is an ML-based approach to filtering or predicting the users' film preferences based on their past choices and behavior. It's an advanced filtration mechanism that predicts the possible movie choices of the concerned user and their preferences toward a domain-specific item (movie).

1.2 Input and output

There are two main elements in every recommendation system: **users** and **items**. In this case, the system generates movie predictions for its users, while items are the movies themselves. The movie recommendation system takes **input** is the data about the user's past choices and also the features of the movie, then it generates the **output** is the filtration or prediction for the user's film preferences.

When building this system, we have to deal with many data about the features of the movie (actor, rating, genre,....) and also the past choice of users. Therefore, the data type of input may be a string (actor name, genre,...) or real value(rating, year,...), or also a boolean(like/dislike,...).

2 Dataset

2.1 About the data

The primary dataset used in this project comes from MovieLens generated by GroupLens. It has two versions, the full version has a collection of 26 million reviews over 45,000 different movies by 270,000 users. Due to limited computing power, a smaller version, which contains 1,000,209 anonymous ratings of approximately 3,900 movies made by 6,040 MovieLens users who joined MovieLens in 2000, is mainly used in this project instead.

The dataset, put in dataset, consists of the following files:

- ratings.dat: contains all ratings following format UserID::MovieID::Rating::Timestamp.
- users.dat: contains user information regarding User ID, Gender, Age, Occupation, and Zip-code.
- movies.dat: contains movie information regarding Movie ID, Title, and Genres.

2.2 Data pre-processing

We divided the dataset into test and training on the basis of UserID, using 90% of each user's reviews for training the model and the remaining 10% for validating the model's accuracy. Using the train-test-split technique from the Scikit-Learn package, we accomplish the aforementioned. The feature on which the dataset is split into training and testing is specified by the function's stratify property.

Due to unintentional duplicate entries and test entries, we discovered that several MovieIDs in this collection do not relate to movies. We decided to treat these erroneous movies as movies with no ratings in order to preserve consistency.

2.3 Data exploration

- **Distribution of movies genres**
  In the dataset, there are 18 genres total, but they have unequal representations. The top three genres with the most movies are Drama, Comedy, and Action.

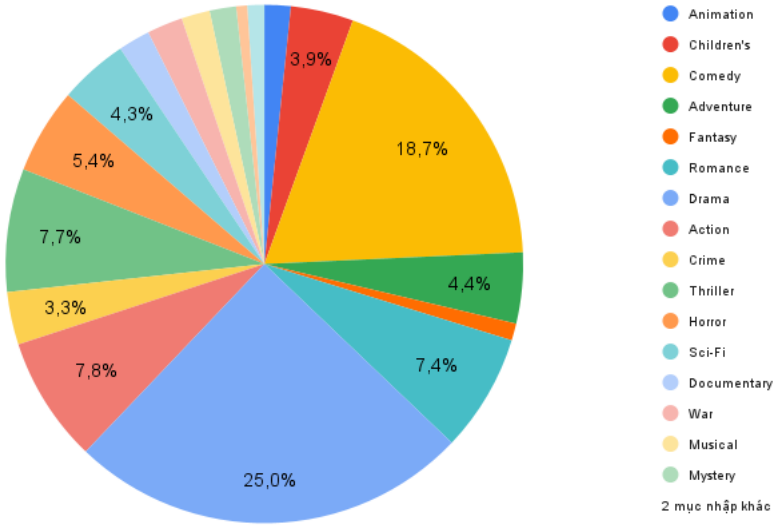


Figure 1: Distribution of movies genres

- **Distribution of ratings**
  Reviews are mostly towards the higher end, 3, 4 or 5. The average rating is 3.58.

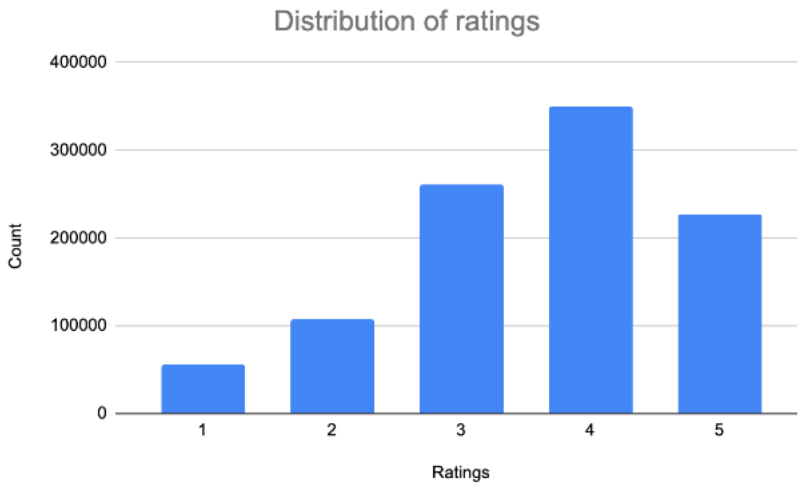


Figure 2: Distribution of ratings

- **Distribution of ratings per users**
  Most users review under 100 movies. On average, a user rates 165 movies, and a movie receives 253 reviews. This is caused by extreme outliers not displayed in the figure below (a user rates well over 3000 movies).

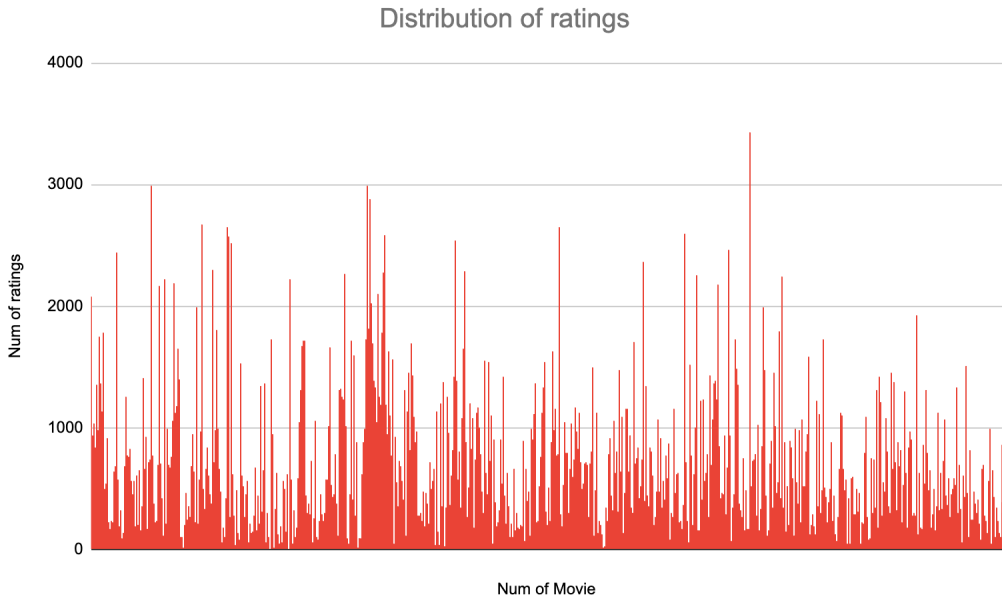


Figure 3: Distribution of ratings per users

3 Content-Based Filtering

3.1 Introduction to Content-Based Filtering

Content-based filtering is a type of recommender system that attempts to guess what a user may like based on that user's activity. Content-based filtering makes recommendations by using keywords and attributes assigned to objects in a database and matching them to a user profile. The user profile is created based on data derived from a user's actions, such as purchases, ratings (likes and dislikes), downloads, items searched for on a website and/or placed in a cart, and clicks on product links. Advantages of content-based recommender system are following:

- Because the recommendations are tailored to a person, the model does not require any information about other users. This makes scaling of a big number of people more simple.
- The model can recognize a user's individual preferences and make recommendations for niche things that only a few other users are interested in.
- New items may be suggested before being rated by a large number of users, as opposed to collective filtering.

3.2 Content-Based Filtering Implementation

3.2.1 Utility Matrix

In Recommender Systems, building the Utility Matrix is paramount. There are two main entities in Recommendation Systems, users and items. Each user will have a degree of preference for each item. This interest level, if known in advance, is assigned a value for each user-item pair. Assuming that interest is measured by the user rate value for the item, let's call this value rating. The collection of all ratings, including the unknown values that need to be predicted, forms a matrix called the Utility Matrix.

	u_0	u_1	u_2	u_3	u_4	u_5	u_6
i_0	5	5	2	0	1	?	?
i_1	4	?	?	0	?	2	?
i_2	?	4	1	?	?	1	1
i_3	2	2	3	4	4	?	4
i_4	2	0	4	?	?	?	5
	↓	↓	↓	↓	↓	↓	↓
$\bar{u}_j$	3.25	2.75	2.5	1.33	2.5	1.5	3.33

Figure 4: Utility Matrix

3.2.2 Items Profiles

In content-based systems, an item profile is a record including important information of that item. In simple cases, this profile illustrates some characteristics, those are easily discovered. In movie recommender system, item profile may contains:

- The set of actors in the movie. Some viewers prefer movies with their favorite actors.
- The director. Some viewers have a preference for the work of certain directors.
- The year in which the movie was made. Some viewers prefer old movies; others watch only the latest releases.
- The genre or general type of movie. Some viewers like only comedies, others dramas or romances.

As computers can only process numeric data, the generation of these features is a very crucial step to make sure that high-quality recommendations can be produced. In our project, we utilize the binary value of genres for each movie, where 1 represents the presence of a genre in a movie and 0 represents the absence. Finally, we construct the feature vector with TF-IDF (Term Frequency-Inverse Document Frequency).

3.2.3 Loss Function

We construct a linear model following the Ridge Regression model. Assume that, we have N users, M items, and Utility matrix Y. Furthermore, R is rated-or-not matrix, namely r_{mn} equals 1 if item m is rated by user n, and 0 otherwise.

Linear model:
Assume that, we find a model for each user, illustrated by weight w_n and bias b_n so that the interest of nth user for the m^{th} item is calculated by a linear function:

$$y_{mn} = x_m w_n + b_n \tag{1}$$

where x_m is a row vector representing a feature vector, w_n is a column vector.
The formula of the loss function for n^{th} user is:

$$L_n = \frac{1}{2s_n} \sum_{m:r_{mn}=1} (x_m w_n + b_n - y_{mn})^2 + \frac{\lambda}{2s_n} \|W_n\|_2^2 \tag{2}$$

where s_n is the number of items rated by n^{th} user.
As loss function only depends on rated items. Therefore, we extract all variables used according to rated items. The formula of the loss function is rewritten as:

$$L_n = \frac{1}{2s_n} ||\hat{X}_n w_n + b_n e_n - \hat{y}_n||_2^2 + \frac{\lambda}{2s_n} ||w_n||_2^2 \tag{3}$$

4 Collaborative Filtering

A collaborative filtering (CF) system provides recommendations for items based on user and movie similarity metrics. The algorithm suggests products that comparable types of people have a preference for. The benefits of collaborative filtering are numerous:

- It is content-independent, relying solely on connections.
- Since users provide explicit ratings in CF, true quality assessment of things is performed.
- It offers serendipitous suggestions since it bases them on user's or movie's similarity rather than movie content's similarity.

4.1 Neighborhood-based Collaborative Filtering

In neighborhood-based approaches, a selection of users are chosen based on how closely they resemble the active user, and predictions for this person are generated using a weighted combination of their ratings. The algorithm outlined in the following stages may be used to generalize the majority of these techniques:

1. Give each user a weight based on how similar they are to the active user.
2. Choose the k users—often referred to as the neighborhood—who are most comparable to the current user.
3. Create a forecast using a weighted average of the ratings of the chosen neighbors.

4.1.1 NB Collaborative Filtering Implementation

- **Similarity metrics**

1. In the first stage, determining the similarity between two users is the task that has to be completed first and foremost. The Utility matrix (which contains all user ratings) is the only piece of information we have, thus it is necessary to compare the two users' columns in this matrix to see how comparable they are.

   The weight $sim(u_i, u_j)$ is utilized as a measure of how comparable the user i and the active user j. The Pearson correlation coefficient between the evaluations of the two users is the most often used metric of similarity and is defined as follows:

$$sim(u_i, u_j) = \frac{\sum_{m \in M}(r_{i,m} - \overline{r}_i)(r_{j,m} - \overline{r}_j)}{\sqrt{\sum_{m \in M}(r_{i,m} - \overline{r}_i)^2 (r_{j,m} - \overline{r}_j)^2}} \tag{4}$$

   Where M represents the group of movies that both users reviewed $r_{i,m}$ represents the rating that user i gave to movie m, and $\overline{r}_i$ represents the mean rating that user i provided.

2. In the next step, predictions are generally calculated as the weighted mean of variances from the neighbor's mean, as in:

$$p_{i,m} = \overline{r}_i + \frac{\sum_{u \in K}(r_{i,m} - \overline{r}_i) \times sim(u_i, u_j)}{\sum_{u \in K} |sum(u_i, u_j)|} \tag{5}$$

   where $sim(u_i, u_j)$ is the degree of similarity between users i and j, $p_{i,m}$ is the prediction for the active user i for item m, K is the group or neighborhood of users who are most similar, and so on.

- **User-user vs item-item neighborhood**
  Due to the computational difficulty of the search for comparable users, traditional neighborhood-based CF algorithms do not scale effectively when applied to millions of people and goods. Item-to-item Collaborative was suggested as an alternative. Filtering where they match a user's rated things to comparable items rather than matching similar persons. In fact, this strategy speeds up online systems and frequently yields better suggestions.

4.2 Matrix Factorization Collaborative Filtering

4.2.1 Introduction to MF Collaborative Filtering

Matrix factorization algorithms work by decomposing the user-item interaction matrix into the product of two lower-dimensionality rectangular matrices which are the item feature

matrix (MxK) and the user feature matrix (KxN). With M is the number of items, N is the number of users and K is the number of features.

- **Latent Feature**

  In the real life, the preferences of each users has the connections to the preferences of other users through features. Similarly, some movies have identical feature with other movies. But the problem is we just indentify the limited feature and alot of other connections we don't actually know. The matrix factorization help us to deal with it very well.

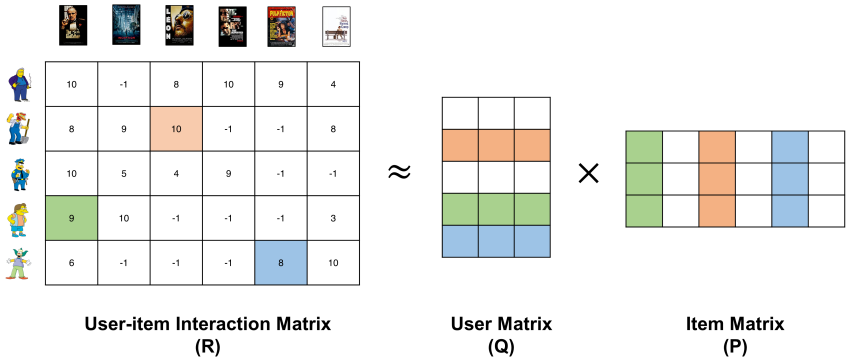


Figure 5: Matrix Factorization

The main idea behinh the matrix factorization is representing users and items in lower dimensional latent space. That means we expressed the orginal matrix into 2 lower dimensional matrix is Users matrix which will tell us the connections between users and Items matrix which tell us the connections between movies through K latent features. We call it the latent features since we don't actually know what exactly K features is, we only know that those will help us to express the relationship between Users and between Movies. Therefore, matrix factorization will indentify the connections between users and between movies based on known ratings and use it to figure out the unknown ratings, then make a recommendation to user. The high coefficient corresponds to one latent feature in both item and user matrix will also be high in the user-item matrix. That means the item has latent feature which the user likes, it will be suggested to the user.

- **Memory Reduction**

Matrix factorization decomposes the user-item matrix into 2 lower-dimensionality matrices will help the inference simpler since we only need to take the product of two vectors with length K which is much smaller than M, N. Also, storing 2 matrices: item and user

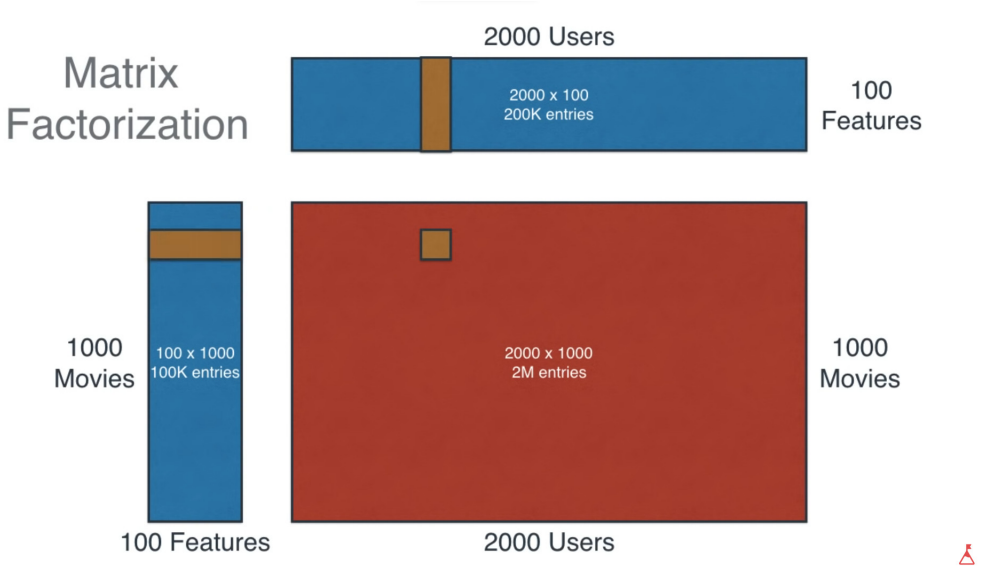


Figure 6: In this example, if we use original matrix to make recommendations, we need to store 2 milions entities. But by decompsing, we just need to store 2 matricies: One has 100K entities and one has 200K entities. So, in total, we just need to store 300K entities.

requires a smaller memory than the user-item matrix. Therefore matrix factorization is found to be the most accurate approach to large and high levels of sparsity datasets.

4.2.2 Matrix Factorization Implementation

There are many variations of the matrix factorization implementation. Because of the time limit of the project, our team decided to only chooses 2 ways to implement it:

1. Matrix factorization with Gradient Descent.
2. Matrix factorization with Truncated Singular Value Decomposition.

1. **Matrix factorization with Gradient Descent**

   In general, Gradient descent is an optimization algorithm which is commonly-used to train machine learning models. Training data helps these models learn over time, and the cost function within gradient descent specifically acts as a barometer, gauging its accuracy with each iteration of parameter updates. Until the function is close to or equal to zero, the model will continue to adjust its parameters to yield the smallest possible error.

   In particular in matrix factorization, we will initialize two random users and movies matricies. Then by each iteration, we will reduce the value of loss function by assigning new set of coefficients for user and movies matricies.

   - **Loss Function**

The loss function, we choose:

$$L(W, X) = \frac{1}{2s} \sum_{n=1}^{N} \sum_{m:r_{mn}=1} (y_{mn} - x_m w_n)^2 + \frac{\lambda}{2}(\|X\|_F^2 + \|W\|_F^2) \tag{6}$$

With X is the items matrix and W is user matrix. $r_{mn} = 1$ if movie m is rated by user n. Notation $\|\cdot\|_F^2$ is Frobineous norm is the square roots of the sum of the squares of all the entities in the matrix and s is the numbers of known ratings.

The simultaneous optimization of 2 matrices X and W is quite complicated.Therefore, the alternative method is in succession, we optimize one matrix while fixing another until converge.

- **Loss Function Optimization**
  When we fix matrix X:

$$L(W) = \frac{1}{2s} \sum_{n=1}^{N} \sum_{m:r_{mn}=1} (y_{mn} - x_m w_n)^2 + \frac{\lambda}{2}\|W\|_F^2 \tag{7}$$

When we fix matrix W:

$$L(X) = \frac{1}{2s} \sum_{n=1}^{N} \sum_{m:r_{mn}=1} (y_{mn} - x_m w_n)^2 + \frac{\lambda}{2}\|X\|_F^2 \tag{8}$$

We will optimize 2 above loss functions by using gradient descent

The updated formula for each column of W:

$$w_n = w_n - \alpha(-\frac{1}{s}\hat{X}^T(\hat{y}_n - \hat{X}_n w_n) + \lambda w_n) \tag{9}$$

Similarly, the updated formula for each column of X:

$$x_n = x_n - \alpha(-\frac{1}{s}(\hat{y}_n - x_n\hat{W}_n)\hat{W}_n + \lambda x_n) \tag{10}$$

2. **Matrix factorization with Truncated SVD**

- **Singular Value Decomposition (SVD)**
  The singular value decomposition of a matrix A is the factorization of A into the product of three matrices:

$$A = U\Sigma V^T$$

where the columns of U and V are orthonormal and the matrix Σ is diagonal with positive real entries whose i^{th} diagonal entry equals the i^{th} singular value σ_i for i =

1, . . . , r and all other entries of Σ are zero.

- **Truncated SVD**

  Truncated SVD is the lower-rank approximation. That means we use low-rank matrix to approximate the original matrix. In the diagonal matrix Σ the singular values σ_i in the diagonal is non-negative and decreasing: $\sigma_1 \geq \sigma_2 \geq \sigma_3 \geq ... \geq \sigma_r \geq 0$, show how important the various columns of U and rows of V are. Therefore, U_1 is more important than U_2 and U_2 is more important than $U_3, ...$ Similar to V_1, $V_2, ...$

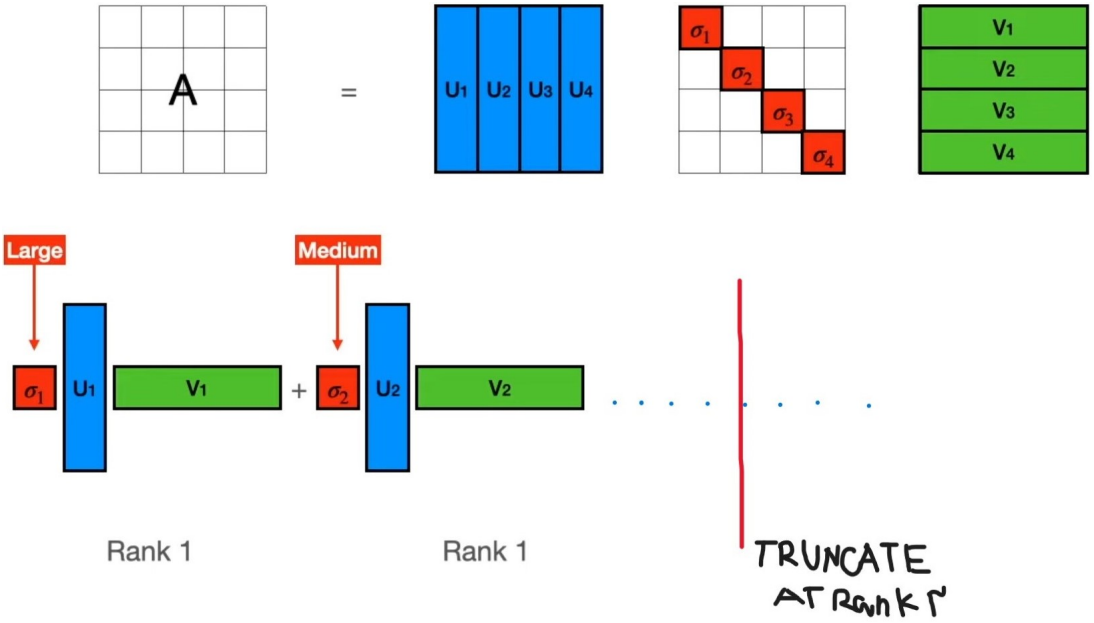


Figure 7: Truncated SVD

The word "important" here express which amount of values those columns or rows contribute to the original matrix. To be clearly, we can see the picture above. We can express the original matrix as the sum of many rank-1 matrices and the more to the right side, value of sigma is decreasing so it contribute value less to the original matrix.And maybe some last matrices, value of $\sigma \approx 0$. Now, truncated is applied, truncated is the action that we only keep the most important $r(r < rank(A))$ matrices or r matrices are left to right and all matrices after r we discard.So, from that, by using lower-rank matrix, we almost still have important information about the original matrix.

- **Matrix factorization with Truncated SVD**

  From the original ratings matrix, we use the truncated SVD to generate the low-rank r matrix which still have almost information, then use it make the recommendation.

5 Experiment and result

5.1 Evaluation Metrics

5.2 Content-based model

5.3 Collaborative Filtering model

5.4 Matrix Factorization model

6 Reference