MovieLens 10M is a movie database released by the GroupLens mainly used for research purpose. It is a collection of user ratings for movies and information about the movies in the form of tags. As the name suggests the dataset contains about 10 million ratings, given by 72 000 users

URM

The URM is a matrix with explicit ratings.

$$r_u^{\,i} \in \{0.5, 1., 1.5, 2., 2.5, 3., 3.5, 4., 4.5, 5.\} \in$$

Each user can express his dislike, $r_u^{\,i} = 0.5$ = 0.5 or love $r_u^{\,i} = 5.0$ towards the movies.

For this dataset we have over 70 thousands users that interact with over 65 thousand movies. The number of interactions is over 10 million however the sparsity of the matrix is very high.

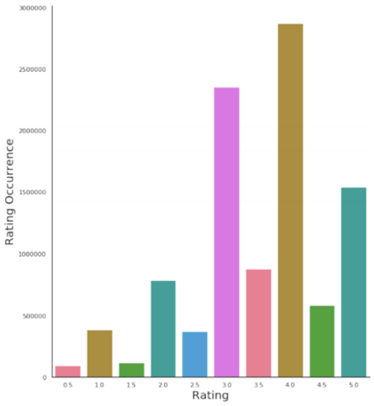

Figure 4.1: MovieLens 10M rating distribution

In Figure 4.1 we see the number of times each rating value was given by the users. The leading one is rating = 4 followed by rating = 3 with very few low ratings given to the users. We have an average rating of 3.5 which means that the users were rather positive towards the movies present in the database but not to overjoyed about them neither.

In Figure 4.2 we see the number of items rated by each user where few outliers are present, so very active users that have rated over 4000 movies. On average, a user has given about 140 ratings, which is a relatively high number of ratings to give. This however, is a good thing for a recommender system since it has a lot of information about the user preferences.

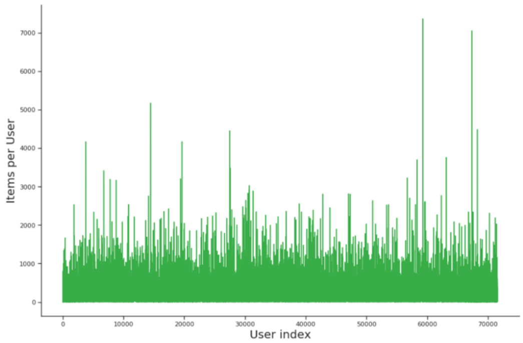

Figure 4.2: MovieLens 10M items rated by users

In Figure 4.3 instead we see the number of ratings given to an item. In this case we see that there are a large quantity of items that have little to no rating at all. However, on average an item has 153 ratings given to it. In this case the average

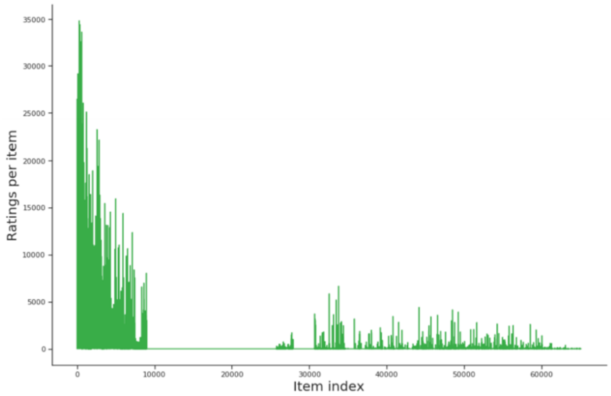

Figure 4.3: MovieLens 10M ratings per item

ICM

The information about the items instead is given in the form of social tags given by the users to the movies.

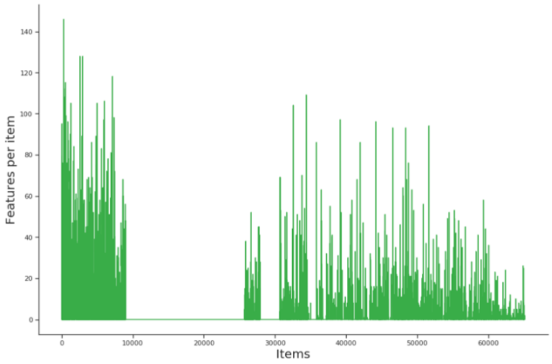

Figure 4.4: MovieLens 10M features per item

MovieLens 10M	ICM
Items	65134
Features	16529
Interaction	71155
Sparsity	0.9999

Table 4.2: Movielens 10M ICM statistics

In Figure 4.4 we see the distribution of features per items, where highest number of features an item has is 146. There are 57533 items with no features at all and 7601 items with features. On average we have that to each item it is associated a feature.

To get a clearer picture of the distribution of features per item, we see the extremes: the items with the highest number of features and the items with the lowest number of features.

F	Items with F features
146	1
128	2
122	1
118	1
115	1

Table 4.3: MovieLens 10M items most tagged items

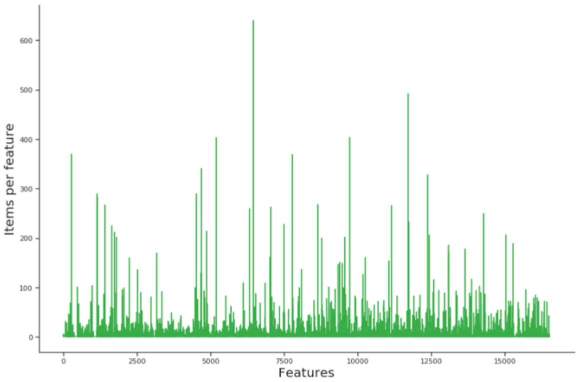

Figure 4.5: MovieLens 10M items per feature

In Figure 4.5 instead we see the distribution of the item per feature where the feature that describes most items goes up to over 600 items. On average we have that each feature is associated to 4 items.

F	Items with F features
0	57533
1	1519
2	1045
3	703
4	543

Table 4.4: MovieLens 10M least tagged items

Features examples
George Clooney
A nice romantic comedy
AFI #12
90s
3.5
9-2-2007
06 Oscar Nominated Best Movie
David O. Russell
2 endings
<3
===========

Table 4.5: MovieLens feature examples

In Table 4.5 we can see examples of the tags assigned by users to the movies. Given that social tags are assigned by users they vary a lot and can be from actor, directors to short reviews, ratings, or metadata about the movies as well as insignificant tags like the last example. It it fairly obvious that the quality of the features is low, very noise and most often redundant.