

## Abstract

**What?':** An assumption that demographic information holds implicit information about users' taste and interests. This information can be used to improve recommendation results.

**How?':** A comparison of recommendation results when using different demographic features (age, location and gender) which are commonly available in online recommendation communities. The comparison is performed through a simple method that extends standard collaborative filtering algorithms to include the demographic features.

**Results:** Results imply that simple demographic data (age, sex, location) can be beneficial for recommendation.

## Dataset

The data comes from German movie recommendation community Moviepilot<sup>a</sup> and contains users, movies, the ratings by users on movies, user related data such as age, location and gender.



**Demographic relations from Moviepilot** In our experiments, we chose to evaluate three basic demographic properties:

- ▶ *Location* - users from the same cities
- ▶ *Age* - users born in the same decade
- ▶ *Gender* - users of the same gender

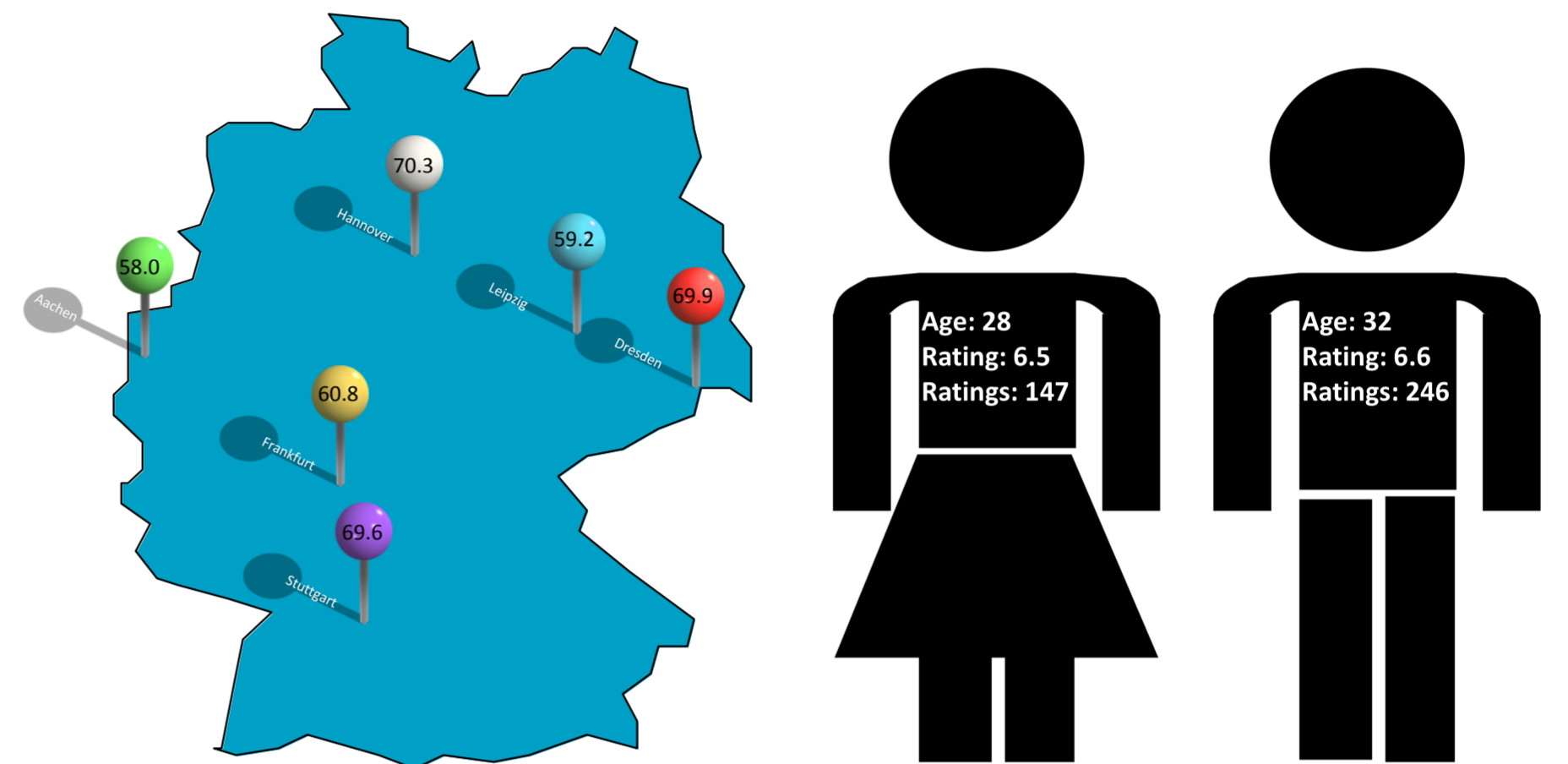
Type	# of ratings	%
City	991,845	64%
Age	1,144,761	74%
Gender	1,398,732	91%
Total	1,539,393	100%

Table: The number and percentage of total ratings available for every demographic property.

<sup>a</sup><http://www.moviepilot.com>

## Data characteristics

There are very clear differences between the different demographic groups, e.g. the difference between the most "positive" city and the most "negative" city is 12.3 on a scale of 0-100, women tend to have 100 ratings less than men. Similarly, number of ratings increases by age up to young adulthood and decreases by age afterwards.



(a) The top three and bottom three German cities according to average movie ratings.

(b) Average age, rating and number of friends for women and men.

Feature	Number of users
City	4,400
Age	1,292
Sex	6,583
Total	10,000

Table: Number of users with each specific feature.

## Approach

- ▶ A  $k$ -NN algorithm generates neighborhoods of most *similar* users.
- ▶ Similarity is based on usage-similarity (ratings) as well as group belonging. The similarity of users within the same group is multiplied by a factor proportional to the number of users.
- ▶ Now, users from the same demographic groups are more important when finding the most popular items in the neighborhood (i.e. the items to recommend).

## Results

Type	P@10 <sup>10K</sup>	P@10	%	MAP <sup>10K</sup>	MAP	%
City	2.79E-4	2.66E-4	4.7%	3.89E-3	3.81E-3	2.2%
Age	2.45E-4	2.45E-4	0.0%	3.93E-3	3.93E-3	0.0%
Sex	2.86E-4	2.33E-4	<b>22.9%</b>	4.22E-3	3.82E-3	<b>10.4%</b>

Table: Results for different demographic features.

Our initial tests show that our assumption, "demographic data has an impact on CF", is true. Gender, especially, seems to have a large impact with a resulting increase of 10% for MAP (and 22% Precision at 10) are very promising. Age seems, however, not to have a very high impact.