Recommender System – Collaborative Filtering

Exercise sheet

Ralph Peeters & Alexander Brinkmann

For this exercise, we will be using real-world data from Last.fm[1] to see how Collaborative Filtering can be used to recommend artists to users.

1 Dataset

We start with the data from Table 1. The table shows the play counts for each band in the data set of the 10 users – assume empty cells to correspond to no (0) plays for the user-artist pair. The number of times a user has played a song by an artist is used as an implicit rating (rather than asking them to explicitly rate the artists).

	The Beatles	Radiohead	Coldplay	Pink Floyd	Muse
User 1	39655				
User 2		903	962		44076
User 3		489	6051		47468
User 4	14975			31957	
User 5	31526			5882	
User 6					42970
User 7	33685	2304		2351	
User 8		18652	31121		690
User 9	4		118857		
User 10			168		44036

We are next given the ratings for two users for whom we want to make recommendations:

	The Beatles	Radiohead	Coldplay	Pink Floyd	Muse
User 21	3344	?	?	22458	?
User 101	?	6293	2286	?	5156

Tools

To generate recommendations to the two users, we will use the excel spreadsheet cf-extended.xlsx. The spreadsheet contains the data set as well as an implementation of user-based and item-based collaborative filtering.

Task 1: User-based collaborative filtering

Examine the excel by understanding the difference of simple and advanced prediction, which considers the average rating behavior of the users. Can you find a band, which receives a much worse result, when the rating behavior of the users is considered? What is the reason for this observation?

[1] `http://ocelma.net/MusicRecommendationDataset/lastfm-360K.html`

Next we are interested in examining different ks for the nearest neighbor selection. What is the impact of setting k=10? What is a proper value for k?

Task 2: Item-based collaborative filtering

Examine the excel by understanding the difference of cosine and adjusted cosine similarity, which considers the average rating behavior of the users. Can you find a band, which receives a much worse result, when the rating behavior of the users is considered? What is the reason for this observation?

Next we are again interested in examining different ks for the nearest neighbor selection. What is the impact of setting k=10? What is a proper value for k? What is the impact of setting a threshold instead of choosing k?