User-based Collaborative Filtering Worked Example

Michael O'Mahony
michael.omahony@ucd.ie

1 Introduction

This document presents a worked example using the user-based collaborative filtering algorithm. Cosine similarity and Resnick's algorithm are used to calculate similarities and predictions, respectively.

2 Collaborative Filtering

Here is a brief overview of the user-based collaborative filtering algorithm. Assume the objective is to compute a prediction, $p_{a,j}$, for user a on an item j.

1. Similarity: calculate the similarity between user a and all other users in the system
2. Neighbourhood Formation: select the users that will serve as neighbours
3. Prediction Computation: aggregate the neighbour's ratings for item j to compute the prediction (note – not all neighbours may have rated item j; such neighbours are ignored).

2.1 Similarity

With *cosine similarity*, users are represented by vectors in the n–dimensional item space. Thus, the similarity between any two users a and i can be calculated as the cosine of the angle between the vectors as follows:

$$w_{a,i} = \frac{\sum_{j \in I_a \cap I_i} r_{a,j}\, r_{i,j}}{\sqrt{\sum_{k \in I_a} r_{a,k}^2} \sqrt{\sum_{k \in I_i} r_{i,k}^2}} \tag{1}$$

where $r_{a,j}$ is the rating assigned to item j by user a and I_a and I_i are the set of items that are rated by users a and i, respectively. This weight gives a value between 0 and 1 (assuming all ratings

$r_{i,j} \geq 0$) and is 0 for any given pair of users if they have not rated any common items, i.e. $w_{a,i} = 0$ if $I_a \cap I_i = \phi$. The normalisation terms in the denominator are important to ensure that users who have rated many items are not *a priori* more similar to other users.

2.2 Prediction Calculation

Using Resnick's algorithm, a prediction, $p_{a,j}$, for user a (the active user) for an item j is given by:

$$p_{a,j} = \bar{r}_a + \frac{\sum_{i=1}^{n} w_{a,i}(r_{i,j} - \bar{r}_i)}{\sum_{i=1}^{n} |w_{a,i}|} \tag{2}$$

where n is the neighbourhood size, $\bar{r}_a$ is the mean rating for user a and $w_{a,i}$ is the similarity weight between users a and i.

3 Worked Example

We consider a simplified recommender system database to show how predictions are computed. Suppose we have the following database of users and items:

	i_1	i_2	i_3	i_4	i_5	i_6	i_7	i_8	i_9
u_1	2	5	**?**		4		1	2	3
u_2	4	3	4				2	5	
u_3	1	2		2		4		1	
u_4	5		2		3		5		4

Empty boxes indicate that users have not rated the corresponding items. Let's assume the objective is to make a prediction for user u_1 on item i_3. Assume users u_2, u_3, and u_4 are in the neighbourhood of user u_1. Note that only users u_2 and u_4 have rated item i_3 while u_3 has not; therefore, user u_3 is not considered when calculating the prediction.

The co-rated items between u_1 and u_2 are: i_1, i_2, i_7 and i_8. Using Equation (1), we have:
$w_{u_1,u_2} = \frac{(2\times4)+(5\times3)+(1\times2)+(2\times5)}{\sqrt{2^2+5^2+4^2+1^2+2^2+3^2}\sqrt{4^2+3^2+4^2+2^2+5^2}} = \frac{35}{\sqrt{59}\sqrt{70}} = 0.54$

The co-rated items between u_1 and u_4 are: i_1, i_5, i_7 and i_9. Using Equation (1), we have:
$w_{u_1,u_4} = \frac{(2\times5)+(4\times3)+(1\times5)+(3\times4)}{\sqrt{2^2+5^2+4^2+1^2+2^2+3^2}\sqrt{5^2+2^2+3^2+5^2+4^2}} = \frac{39}{\sqrt{59}\sqrt{79}} = 0.57$

Applying Equation (2):
$p_{u_1,i_3} = 2.83 + \frac{0.54\times(4.0-3.6) + 0.57\times(2.0-3.8)}{|0.54|+|0.57|} = 2.83 + \frac{-0.81}{1.11} = 2.1$