Research Repository UCD

Title	What have the neighbours ever done for us? A collaborative filtering perspective
Authors(s)	Rafter, Rachael, O'Mahony, Michael P., Hurley, Neil J., Smyth, Barry
Publication date	2009-06
Publication information	Rafter, Rachael, Michael P. O'Mahony, Neil J. Hurley, and Barry Smyth. "What Have the Neighbours Ever Done for Us? A Collaborative Filtering Perspective." Springer, 2009.
Conference details	Paper presented at the First and Seventeenth International Conference on User Modeling, Adaptation and Personalization (UMAP 09), Trento, Italy, 22-26 June, 2009
Publisher	Springer
Item record/more information	http://hdl.handle.net/10197/1336
Publisher's version (DOI)	10.1007/978-3-642-02247-0_36

Downloaded 2023-04-07T10:52:29Z

The UCD community has made this article openly available. Please share how this access benefits you. Your story matters! (@ucd_oa)

© Some rights reserved. For more information

What Have The Neighbours Ever Done for Us? A Collaborative Filtering Perspective.*

Rachael Rafter, Michael P. O'Mahony, Neil J. Hurley, and Barry Smyth

CLARITY: Centre for Sensor Web Technologies,
School of Computer Science and Informatics,
University College Dublin, Belfield, Dublin 4, Ireland
`{rachael.rafter,michael.p.omahony,neil.hurley,barry.smyth}@ucd.ie`

Abstract. Collaborative filtering (CF) techniques have proved to be a powerful and popular component of modern recommender systems. Common approaches such as *user-based* and *item-based* methods generate predictions from the past ratings of users by combining two separate ratings components: a *base estimate*, generally based on the average rating of the target user or item, and a *neighbourhood estimate*, generally based on the ratings of similar users or items. The common assumption is that the neighbourhood estimate gives CF techniques a considerable edge over simpler *average-rating* techniques. In this paper we examine this assumption more carefully and demonstrate that the influence of neighbours can be surprisingly minor in CF algorithms, and we show how this has been disguised by traditional approaches to evaluation, which, we argue, have limited progress in the field.

Key words: Recommender Systems, Collaborative Filtering, Predictive Accuracy

1 Introduction

Collaborative filtering (CF) [?] has become a popular recommendation technique and has been applied successfully in many online applications. Different types of CF techniques all share an ability to harness the past ratings of users (over some catalog of items) in order to predict a user's likely rating for an unseen item. In a recommender system, CF techniques can be used to recommend items with high predicted ratings while suppressing items with low predicted ratings.

In this paper we will focus on two common flavours of collaborative filtering, so-called *user-based* and *item-based* based techniques. Given a target user t and a target item i , a rating $r_{t,i}$ is computed as a combination of a *base estimate* (B) and a *neighbourhood estimate*, where the former is generally taken to be the average user or item rating and the latter is some function of the ratings assigned by the target's nearest neighbours, $\mathcal{N}$, see Eq. 1. The neighbourhood estimate is essentially a way to refine the rather blunt, initial base estimate in a way that should improve the accuracy of the resulting prediction.

* This work is supported by Science Foundation Ireland under grant 07/CE/I1147

$$r_{t,i} = B + f(t, i, \mathcal{N}) \quad (1)$$

User-based and item-based approaches differ principally in the way that they compute and combine the neighbourhood estimates into the overall prediction process. For example, the classic approach to user-based collaborative filtering is presented by [?] and shown in Eq. 2. Here the base estimate is based on the target user’s average rating $\bar{r}_t$, and the neighbourhood estimate is based on a weighted-average of the extent to which similar neighbours $u \in \mathcal{N}$ appear to like or dislike the target item. Neighbour similarity, $S_{t,u}$, is usually calculated using Pearson’s correlation (comparing target user and neighbour ratings). The extent to which a neighbour likes or dislikes the target item is based on whether their rating for i is greater than or less than their average rating $\bar{r}_u$.

$$r_{t,i} = \bar{r}_t + \frac{\sum_{u \in \mathcal{N}} (r_{u,i} - \bar{r}_u) \times S_{t,u}}{\sum_{u \in \mathcal{N}} |S_{t,u}|} \quad (2)$$

Item-based CF can be presented similarly (see [?]) such that predictions are computed according to Eq. 3. This time the base estimate is the target item’s average rating $\bar{r}_i$ across all users and the neighbourhood estimate is based on a weighted-average of the extent to which the user’s existing item ratings differ from the average rating received by those items across all users. The similarity $S_{i,j}$ between items i and j is computed using the adjusted cosine metric [?] and the neighbourhood $\mathcal{N}$ consists of each item j previously rated by the user.

$$r_{t,i} = \bar{r}_i + \frac{\sum_{j \in \mathcal{N}} (r_{t,j} - \bar{r}_j) \times S_{i,j}}{\sum_{j \in \mathcal{N}} S_{i,j}} \quad (3)$$

The assumed power of collaborative filtering is derived largely from its neighbourhood estimate which must perturb the base estimate by the *correct magnitude* and in the *correct direction*. It is surprising, to us at least, that there has been no detailed examination of these common collaborative filtering techniques, that focuses on the individual base and neighbourhood estimates.

In this paper we argue that a more principled approach to CF design and evaluation is merited and that it is important to consider more carefully the influence of base and neighbourhood estimates if we are to significantly advance the current state-of-the-art. The main contribution of this paper is an initial analysis of these estimates across three standard CF datasets using both user-based and item-based techniques, with the surprising result that the neighbourhood estimates plays a relatively minor, and often unreliable, prediction role. Moreover, we argue that traditional evaluation methodologies have served only to disguise this effect, and we propose a return to an *analysis of the extremes* as originally proposed by [?], which seems to have been largely forgotten by the community.

2 The Importance of Good Neighbours

In this section we focus on the assumption that underlies user-based and item-based CF—namely, that the neighbourhood estimate, in general, improves the

Datset	# Users	# Items	# Ratings	Sparsity	Rating Scale
MovieLens	943	1,682	100,000	93.695%	1–5
Netflix	24,010	17,741	5,581,775	98.690%	1–5
Book-Crossing	77,805	185,973	433,671	99.997%	1–10

Table 1. Dataset Statistics

	User-based			Item-based		
	Mag.	Cor. Dir.	MAE	Mag.	Cor. Dir.	MAE
MovieLens	0.43	66%	0.73	0.34	64%	0.73
Netflix	0.41	66%	0.7	0.35	67%	0.69
Book-Crossing	0.99	53%	1.53	0.94	63%	1.34

Table 2. Magnitude, correct direction (%) of neighbourhood estimate and overall MAE

prediction accuracy of the base estimate. We test this assumption on three commonly used, large-scale, real-world datasets: MovieLens (100K)¹, Netflix² and Book-Crossing [?]³. Since the trends observed for MovieLens and Netflix are similar, we at times report on just one. Dataset statistics are given in Table 1.

We first examine the *direction* of the neighbourhood estimate, i.e. how often the base estimate is pushed closer towards the true rating, see Table 2. For MovieLens and Netflix, we find that the neighbourhood estimate produces an adjustment in the correct direction (on average across both user- and item-based CF) in only 63% of cases. This means that 37% of the time, the neighbourhood estimate is actually pushing the prediction from the base estimate in the wrong direction, thus making it less accurate. For the Book-Crossing dataset, user-based CF performs particularly poorly, with only 53% of neighbourhood estimates contributing in the correct direction, and actually performing only slightly better than chance. This implies that CF is contributing to poorer quality predictions in just under half the cases; we will return to this in Section 3 where we will discover more positive results in different regions of the ratings space. The differences in results for user-based CF across the datasets correlate well with dataset sparsity but a deeper analysis is left for future work.

Table 2 also shows the average (absolute) magnitude of shift produced by the neighbourhood estimate. For both MovieLens and Netflix, neighbourhood estimates contribute less than 1/2 point on the 5-point scale. For Book-Crossing, the average magnitude is approximately 1 point on a 10-point scale. Clearly, the ability of the neighbourhood estimate to significantly influence the final prediction is limited. The cumulative distribution functions (CDF) of neighbourhood

¹ <http://www.grouplens.org/>

² <http://www.netflixprize.com/>

³ In Netflix, we performed our analysis on a randomly selected 5% of users and associated ratings; for Book-Crossing, we ignored implicit ratings. All results are obtained using 10-fold cross validation to make predictions for randomly selected test ratings.

Fig. 1. (a) Cumulative distribution functions showing the magnitude of the neighbourhood estimate and (b) frequency of ratings in MovieLens and Netflix datasets

estimate magnitudes in Figure 1 (a) show that in fact this aspect of the CF algorithm has ultimately little influence on the predicted rating. This is true across the three different datasets and both the user- and item-based algorithms.

We can attribute the lack of contribution of the neighbourhood estimate to one of two potential factors. Firstly, since most of the ratings in the datasets fall around the average rating, see MovieLens for example in Figure 1 (b), the required neighbourhood estimate is in fact small. However, we cannot rule out the possibility that CF only appears to be working well because there are many ratings that are close to the average rating. More importantly, the key challenge for the algorithm surely is how it performs when a rating lies closer to the extremes, i.e. when it is rated very high or very low. In the next section, we will examine our results in more detail and look at these cases individually.

3 MAE Evaluation Metric

Prediction accuracy in CF is usually measured using mean absolute error (MAE) across a set of predicted ratings.⁴ The MAE data in Table 2, shows that the user- and item-based techniques perform reasonably, with predicted ratings within 20% of true ratings. However these results are misleading and, as first proposed in [?], it is vitally important to consider the distribution of prediction errors at the rating extremes, a fact seemingly often ignored in conventional CF evaluations.

Figure 2 presents a more fine-grained analysis by calculating both the mean and standard deviation of the prediction error at each point on the ratings scale. This time the results tell a very different story. We can immediately see that, while the algorithms perform reasonable well in the mid-range of the ratings scale, they perform very poorly at the extremes, particularly at the low end of the ratings scale. This means that these algorithms are not capable of reliably

⁴ We focus on MAE here as it is the most common metric used for evaluating prediction accuracy. We leave an analysis of other metrics used, e.g. RMSE, for future work.

Fig. 2. Mean and standard deviation of prediction error across the ratings scale

predicting items that will be loved or hated, with the risk that mediocre items will be recommended in practice. Predictions using the Book-Crossing dataset are especially poor, probably due to the high sparsity of this dataset; for example, user-based CF (UB) is only able to predict the rating of a disliked item to within 5 or 6 ratings points on average (and with high variation) on a 10-point scale.

These results suggest that CF is performing poorly exactly when it is most needed: at the extremes. This is reinforced by Figure 3(a), which shows the difference between the required shift needed to make a correct rating prediction, and the actual shift that the neighbourhood estimate delivers in practice (here we focus on user-based CF for MovieLens). At the extremes, the actual shift is far from what is required. Interestingly however, the direction of shift is at least more accurate at the extremes. For example, the shift is in the correct direction about 80% of the time for MovieLens items that are rated as 1 or 5.

In Figure 3(b) we get a sense of this for predictions computed across different ratings, where the ratio of the neighbourhood estimate to the base estimate is plotted. As expected, the relative contribution of the neighbourhood estimate is minor across the ratings scale, but there is an interesting effect at the ratings extremes where neighbours exert a stronger influence. For example, with the MovieLens dataset and user-based CF, there is a neighbourhood-to-base estimate ratio of 0.2 at the first rating point, meaning that the neighbourhood estimate is contributing 20% to the predicted rating. This is twice the contribution that is noted for higher points of the ratings scale, but it is still low.

4 Conclusions and Future Work

CF techniques generate predictions by relying, in part, on the ratings of a neighbourhood of similar users or items. In this paper, we have explored just how important a role neighbours play in prediction; something that has not been examined in detail before. What we have found is surprising. Notwithstanding the significant research that has been invested in neighbourhood selection techniques, the influence of neighbours remains relatively minor (neighbours not usually exerting enough of a shift on the base estimate) and often unreliable (neighbours

Fig. 3. (a) Difference between the actual and the required neighbour estimate for user-based CF for MovieLens, and (b) ratio of the neighbourhood estimate to base estimate for user- and item-based CF for MovieLens and Netflix

often shifting the base estimate in the wrong direction). This has a number of important implications. Firstly, as a community, we need to better understand the factors that influence the ability of neighbours to improve a baseline prediction. Secondly, from an evaluation perspective we need to recognise that simple MAE-style evaluations serve only to disguise important prediction errors, especially for extreme ratings. At the very least a more fine-grained error analysis is required in order to highlight the significant variations in error across a given ratings scale. As a final point, we need to emphasise the importance of developing new CF algorithms that offer prediction improvements on extreme ratings because, ultimately, users need to receive reliable recommendations containing items they strongly like and avoiding items they strongly dislike.

References