Ways of Computing Diverse Collaborative Recommendations*

Derek Bridge and John Paul Kelly

University College Cork,
Cork, Ireland
d.bridge|jpk2@cs.ucc.ie

Abstract. Conversational recommender systems adapt the sets of products they recommend in light of user feedback. Our contribution here is to devise and compare *four* different mechanisms for enhancing the diversity of the recommendations made by collaborative recommenders. Significantly, we increase diversity using collaborative data only. We find that measuring the distance between products using Hamming Distance is more effective than using Inverse Pearson Correlation.

1 Introduction

Recommender systems suggest products, services or information sources to their users. They differ in the way they find the items they recommend:

Content-based systems: The system stores a description of each available item. A user describes the item that she wants as a query or she describes the kinds of items that she likes as entries in a user profile. The system compares the user's descriptions against the store of item descriptions and recommends items that match.

Collaborative systems: Item descriptions are not used. A user's profile stores user opinions against item identifiers. The system compares other users with the active user and recommends items that were liked by users whose profiles are similar to the active user's profile.

Recommender systems differ also by the extent to which they engage in dialogue with the user:

Single-shot systems: In response to a user request, the system delivers a set of recommendations. Each request is treated independently of previous ones.

Conversational systems: Users elaborate their requirements over the course of an extended dialogue. In particular, the user can supply feedback on the recommended items. Her feedback influences the next set of recommendations.

* This material is based on works supported by Science Foundation Ireland under Grant No. 03/IN.3/136. We are grateful to Professor Barry Smyth for his advice and to the GroupLens project team for making their data available.

	Ann	Bob	Col	Deb	Edd	Flo
Cape Fear	⊥	⊥	3	5	5	5
Naked Gun	3	2	⊥	2	4	⊥
Aliens	⊥	5	⊥	⊥	2	4
Taxi Driver	⊥	⊥	3	4	3	⊥

Table 1. A ratings matrix

Conversational systems can more easily adapt their recommendations to the user's short-term interests. By dint of mood changes or other special circumstances, short-term interests may not coincide with long-term interests.

There is a mature body of research on conversational *content-based* systems. But research into *collaborative* systems has focused on single-shot recommenders. The work of Rafter & Smyth [1] is a recent exception. Section 3 describes their work on conversational collaborative recommenders. Section 4 describes how we have enhanced the diversity of the recommendations made by a conversational collaborative recommender. Doing this requires a way of measuring the distance between two products; Section 5 proposes four ways of doing this using collaborative data only. Section 6 compares the diversity-enhanced conversational collaborative recommenders with a single-shot recommender and with a standard conversational collaborative recommender. But first Section 2 summarises the operation of the class of collaborative recommenders used in this work.

2 Single-Shot Collaborative Recommender Systems

In a collaborative recommender, given m items, $I = \{i : 1 \dots m\}$, and n users, $U = \{u : 1 \dots n\}$, preferences are represented using a $m \times n$ matrix of ratings $r_{i,u}$. Note that it is possible and common that $r_{i,u} = \perp$, signalling that the user has not yet rated that item. An example of a ratings matrix for movies is shown as Table 1. Each column in the matrix is a user's long-term profile. We will write u^{LT} for the item identifiers that have non-$\perp$ ratings in user u's long-term profile. For example, $\text{Bob}^{LT} = \{\text{Naked Gun}, \text{Aliens}\}$. We will write u^{LT^+} for the set of items in u's long-term profile for which the rating is greater than or equal to the mid-point of the rating scale. For example, assuming a 5-point rating scale, $\text{Edd}^{LT^+} = \{\text{Cape Fear}, \text{Naked Gun}, \text{Taxi Driver}\}$.

There are many ways of building collaborative recommenders, most of which are compatible with the research reported in this paper. Here we describe just the one we have implemented; for details, see [2]:

- The similarity $w_{u_a,u}$ between the active user u_a and each other user, $u \neq u_a$, is computed using Pearson Correlation, $\text{correl}(u_a, u)$, over their co-rated items.
- After computing the similarity between u_a and each other user, u, the N (in our case, 20) *nearest neighbours* are selected, i.e. the N for whom $w_{u_a,u}$ is highest.

– For each item i that has not been rated by u_a but has been rated by at least one of the neighbours, u_a's rating for i is predicted, p_{i,u_a}. This is essentially an average of the neighbours' ratings for item i weighted by the values for correl(u_a, u).
– These items are then sorted into descending order of p_{i,u_a}. This is the order in which items will be recommended. For example, if in a single-shot system we want to recommend three items, then the first three items in this sorted list are selected.

3 Conversational Collaborative Recommender Systems

In 2004, Rafter & Smyth described their conversational collaborative recommender: the system recommends items to the user; the user gives feedback on the recommendations; and the feedback influences the next set of recommendations [1]. We use CCR$^+$ to designate our implementation of their ideas [3].

In CCR$^+$, the active user has a long-term profile (based on a column in the ratings matrix), u_a^{LT}, as do all other users. But, for the duration of her interaction with the system, the active user also has a short-term profile, $u_a^{ST^+}$.

Initially, the short-term profile is empty and the first set of k (typically, three) recommendations is made in the fashion described in Section 2. At this point, the system solicits user feedback. The user can terminate the dialogue, with or without having chosen an item for purchase or consumption. Or, if she wishes to continue the dialogue, she can optionally indicate which recommended item best matches her short-term interests. If she does, the selected item's identifier is added to her short-term profile, $u_a^{ST^+}$. Nothing is done with the other items.

If the dialogue has not been terminated, the system now recommends another set of items. New recommendations never repeat ones made previously in the dialogue. But, additionally, through the way it computes user similarity, the system attempts to steer new recommendations towards the kind of items in $u_a^{ST^+}$; see below for details. This recommendation-and-feedback cycle continues until either the user finds an item she wishes to consume, she abandons the dialogue having found no such item, or the system can make no fresh recommendations.

It remains to say how $u_a^{ST^+}$ influences subsequent recommendations. When finding neighbours, the similarities between users will no longer be based just on the Pearson Correlation between their long-term profiles. The idea in conversational collaborative recommending is that the selection of nearest neighbours is "... directed towards users that have liked the items in the target user's [short-term profile]" [1, p.152]. Specifically, correl(u_a, u), the correlation between the long-term profiles of the active user u_a and each other user u, will be boosted by an amount based on the size of the intersection between u's long-term positive profile u^{LT^+} (the item's in u's long-term profile for which the rating equals or exceeds the mid-point of the rating scale) and u_a's short-term profile u_a^{ST} [3].

We have found that enhancing the diversity of the recommendations improves results (Section 6), so this is the topic of the next section.

4 Diversity-Enhanced Conversational Collaborative Recommender Systems

This section introduces the CCR$^+$Div(b, k) system. In its name, Div indicates a concern for the diversity of recommendations; b and k are parameters, which are explained below.

For *content-based* recommender systems, the argument has been convincingly made that items should be selected for *diversity* (relative to each other) as well as *similarity* (to the query or the user's profile) [4]. Too much similarity between the recommended items (e.g. three Woody Allen movies) can be undesirable. But, when recommendations are diverse, if the user is not satisfied with the most highly recommended item, for example, the chances of her being satisfied with one of the alternative recommendations is increased.

There is a body of research that addresses diversity for *content-based* recommenders, e.g. [4–6]. It is only now that we are seeing the first work that attempts to improve the diversity of the items recommended by *collaborative* recommenders [3, 7]. We hypothesise that a direct concern for diversity may be important, especially in *conversational* collaborative systems: diverse recommendations increase the chances of positive feedback (where an item is preferred over the others), and this helps the system target the recommendations towards the user's short-term interests.

To investigate this, we implemented the Bounded Greedy selection algorithm (henceforth BG) from [4]. To recommend k items, BG finds bk items. In [4], these are the bk items that are most similar to the query (content-based recommending). Here, they are the bk items with the highest prediction values p_{i,u_a} (where neighbours are computed by the CCR$^+$ system). From these bk items, BG selects k to recommend to the user. It selects the k in a greedy fashion, based on ones selected so far; see Algorithm 1.

In the algorithm, the quality of item i relative to the result set so far R is defined as follows:

$$\mathit{Quality}(i, R) =_{\text{def}} (1 - \alpha) \times p_{i,u_a} + \alpha \times \mathit{RelDiv}(i, R) \tag{1}$$

i.e. it is a weighted combination of the predicted rating for item i and the diversity we will attain if we include i in R. α is a factor that allows the importance

Algorithm 1 The Bounded Greedy selection algorithm. Adapted from [4].

Candidates ← bk items recommended by CCR$^+$
$R \leftarrow \{\}$
for $j \leftarrow 1$ to k **do**
  best ← the $i \in$ *Candidates* for which *Quality*(i, R) is highest
  insert *best* into R
  remove *best* from *Candidates*
end for
return R

of the predicted rating and relative diversity to be changed. In this paper, we only investigate the case where the two factors are given equal weight. Hence, we normalise both factors so that they fall in $[0, 1]$ and we then use $\alpha = 0.5$.

Diversity relative to the result set so far is defined as the average distance between i and the items already inserted into R:

$$RelDiv(i, R) =_{\text{def}} \begin{cases} 1 & \text{if } R = \{\} \\ \frac{\sum_{j \in R} \text{dist}(i,j)}{|R|} & \text{otherwise} \end{cases} \quad (2)$$

This will lie in $[0, 1]$ provided each $\text{dist}(i, j)$ lies in $[0, 1]$.

This leaves the issue of how to measure distance *between items* in Equation 2. In [4], the distance between items is the inverse of the *content-based* similarity. If item descriptions are available, the same approach can be used to enhance the diversity of collaborative recommendations. Ziegler, for example, uses taxonomic knowledge in his system [7]. But we choose to proceed on the assumption that item descriptions are not available. We enhance diversity using a measure of distance that is calculated using *collaborative data only*, i.e. we use only the ratings matrix.

Our approach to distance using collaborative data only is based on the following heuristic:

Two items are different if the users who rated them are different.

The intuition is that the community of users who have rated item i have a certain set of tastes. The more the membership of the community who rated item i differs from the membership of the community who rated item j, the more likely i and j satisfy different tastes and are different kinds of items. For example, according to this heuristic, a movie that is liked exclusively by adolescent males is likely to be distant from one that it liked exclusively by middle-aged women. (We stress, however, that, just as we are not using content-based data, we are not using demographic data either: our ways of computing distance will make use only of the ratings matrix.)

There are numerous ways to make this informal heuristic into something that can be implemented. We described one such way in [3], showing that it outperformed Rafter's & Smyth's original system, our minor variant of their system (CCR^+), and a system that enhances diversity by choosing items at random from the *bk* items with the highest prediction values. The contribution here is to compare that approach with three other ways of implementing the heuristic.

5 Definitions of Item Distance

In this section, we look at four ways of defining the distance between pairs of items, for use in Equation 2. The four definitions differ on two dimensions, which we explain below.

Nearest Neighbours or All Users. The first dimension is the way we choose the set of users on which the communities are defined. One possibility is to confine attention to the active user's nearest neighbours (NN). We could instead define the communities over the set of *all* users known to the system (All). In other words, in the former, distance is computed between vectors of length N, where N is the number of nearest neighbours; in the latter, distance is computed over vectors of length n, where n is the size of U.

Hamming Distance or Inverse Pearson. The second dimension is the way in which we compute by how much two communities differ.

One possibility is to define the distance between two items to be inversely related to the size of the intersection of the sets of users who rated the two items. This definition of distance can be computed quite efficiently using bit-vectors. In detail, then, we compute $\text{dist}(i, j)$ as follows:

- For both i and j, we create bit vectors I and J of length N (in the case where only nearest neighbours are being used) or n (in the case where all users are being used). Digit d in vector I is set if user d has a non-$\perp$ rating for item i; similarly for bits in J.
- $\text{dist}(i, j)$ is computed as the Hamming Distance (HD) between I and J, i.e. a count of the number of positions in which their bits differ. This value is normalised, so that it falls within $[0, 1]$, by dividing it by N or n as appropriate.

Figure 1 illustrates this process; it shows Naked Gun to be more different from Cape Fear than Taxi Driver is. In the figure, we are using NN, the nearest neighbours; we take the number of these to be three; and we assume these are Ann, Col and Deb. We take their ratings from Table 1 and set bits to show who rated what.

	Ann	**Col**	**Deb**
Cape Fear	0	1	1
Naked Gun	1	0	1
Hamming Distance:		2	
Normalised:		$\frac{2}{3}$	

	Ann	**Col**	**Deb**
Cape Fear	0	1	1
Taxi Driver	0	1	1
Hamming Distance:		0	
Normalised:		$\frac{0}{3}$	

Fig. 1. Hamming Distances.

This definition takes no regard of the magnitudes of the ratings. It considers only whether a user has rated an item or not; it counts how many of the users have rated one of the two items but not the other.

We could instead compute the Inverse of Pearson Correlation (IP) between the users' ratings of the two items. Effectively, this means computing Pearson Correlation between *rows* rather than columns in the ratings matrix (Table 1),

	Nearest Neighbours	All Users
Hamming Distance	HD-NN	HD-All
Inverse Pearson	IP-NN	IP-All

Table 2. Four definitions of item distance, using collaborative data only.

takings its inverse, and normalising so that it falls within $[0, 1]$. This would have the seeming advantage of being sensitive to the magnitudes of the ratings.

Summary. Two binary dimensions gives four definitions, as shown in Table 2. We emphasise that in all other respects these four systems are identical: only their computation of item-item distance within the definition of relative diversity varies.

6 Empirical Evaluation

To evaluate the systems that we have described, we adopt Rafter's & Smyth's methodology [1]. One hundred user profiles are selected at random and removed from the '100K MovieLens Dataset'[1]. Each of these acts in turn as an (artificial) active user. The item that the user is seeking is obtained through the leave-one-in methodology, i.e. given the active user's long-term profile, each item in turn is treated as the target item. Each of the systems that we evaluate repeatedly recommends sets of three items to the user until either the target item is one of the recommended items, there have been 100 recommendation cycles, or no further recommendations can be made to this user, whichever comes soonest. If the target item is recommended within 100 cycles, the number of items recommended is recorded. Results are subjected to three-fold validation, with a different 100 active users in each fold.

In each recommendation cycle, the (artificial) user's feedback needs to be simulated. For each movie, the MovieLens datasets record a set of genres, which allows a simple-minded content-based approach. If the target item's set of genres is G_t and a recommended item's set of genres is G_r, we compute $\frac{|G_t \cap G_r|}{|G_t \cup G_r|}$. If all recommended items score zero, then none is taken to match the user's short-term interests; otherwise, the highest-scoring item (with random tie-breaking) is taken to match the user's short term-interests and is inserted into $u_a^{ST^+}$.

In the diversity-enhanced systems, we have taken $k = 3$ and $b = 15$. In other words, a set of three items is chosen greedily from 45 candidates. In previous work, we have partially explored what happens when different values of b are chosen [3]. There we found better results for $b = 15$ than $b = 5$ and $b = 10$. Of course, it does not follow that results will continue to improve with ever larger values of b: at some point, the set of candidates will be so large that it will include items whose predicted ratings are so low that they will not be valuable

[1] `www.grouplens.org`

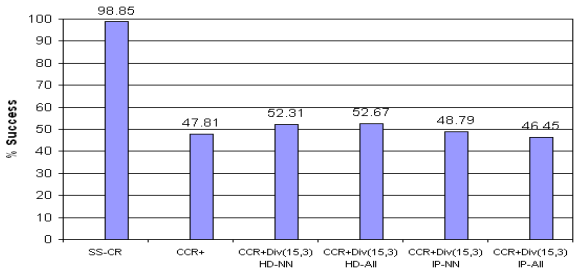

Fig. 2. Success rates

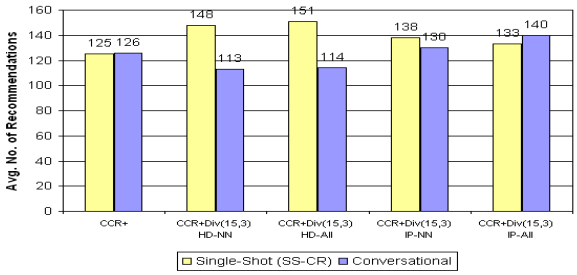

Fig. 3. Average number of recommendations

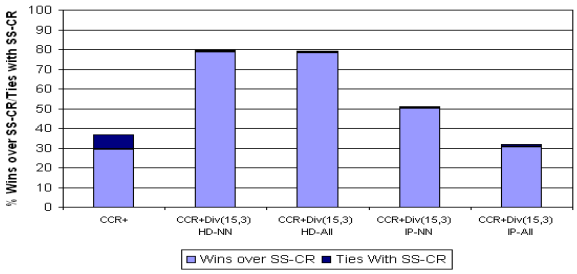

Fig. 4. Wins over SS-CR and ties with SS-CR

	Win	Lose
CCR^+	46	24
CCR^+Div(15, 3) HD-NN	51	29
CCR^+Div(15, 3) HD-All	56	33
CCR^+Div(15, 3) IP-NN	48	33
CCR^+Div(15, 3) IP-All	59	37

Table 3. Winning and losing margins

recommendations. In future work, we need to find the values of b for which this is the case.

Figure 2 shows, as a percentage of a total of 34759 dialogues, how often the target item was found. In addition to CCR$^+$ and the four versions of CCR$^+$Div(15, 3), we show the results for SS-CR, a single-shot recommender, which computes a ranked list of items in the way described in Section 2, and recommends them in decreasing order of predicted rating, k (= 3) items at a time. We regard SS-CR as successful if the target item is among all the possible recommendations it can make to the active user. The other systems are successful if the target item is recommended within 100 cycles of three recommendations each. Unsurprisingly, SS-CR has by far the highest success rate; encouragingly, most of the diversity-enhanced systems have higher success rates than CCR$^+$. Of the diversity-enhanced systems, the ones that use Hamming Distance are more successful than those that use Inverse Pearson Correlation.

Figure 3 shows how many items are recommended, on average, before the system recommends the target item. In this figure, each system is compared with the SS-CR in cases where both were successful in finding the target item. We see that SS-CR can rival CCR$^+$, which suggests that the user feedback has little value in CCR$^+$. The diversity-enhanced systems all outperform SS-CR and CCR$^+$, which confirms that diverse recommendations can elicit more useful user feedback. Of the diversity-enhanced systems, the ones that use Hamming Distance are by far the most successful; they require about 35 fewer recommendations to reach the target item than SS-CR.

Figure 4 shows how often each of the conversational systems makes the same or fewer recommendations than SS-CR (when both are successful). We see that the diversity-enhanced systems that use Hamming Distance make fewer recommendations than SS-CR nearly 80% of the time; the other systems are competitive with SS-CR between 30% and 50% of the time.

Finally, Table 3 shows winning and losing margins. The table shows, for example, that, when CCR$^+$ wins against SS-CR, it makes on average 46 fewer recommendations and, when CCR$^+$ loses against SS-CR, it makes on average 24 more recommendations. We see that IP-All wins by most when it wins but it also loses by most when it loses. By this measure, the systems that use Hamming Distance do well: when they win, they win by a respectable margin; when they lose, they lose by some of the smaller amounts.

7 Conclusions

Building on the seminal work reported in [1], we have developed a number of conversational collaborative recommender systems. In all these systems, the selection of neighbours is guided by overlap with the active user's short-term profile. In CCR$^+$Div(b, k), we introduce an explicit mechanism that uses collaborative data only to enhance the diversity of recommendations made by (conversational) collaborative recommender systems.

We have experimented with four definitions of distance, for use when computing relative diversity. We found, perhaps counter-intuitively, that approaches based on Hamming Distance work better than those based on Inverse Pearson Correlation. This is surprising: ignoring the magnitudes of the ratings is better than taking them into account! We suspect that this is because Hamming Distance, being more discrete, sharpens the definitions of the communities, which is important in our heuristic definition of item diversity, whereas Inverse Pearson, being more continuous, de-emphasises community differences.

In terms of success rates and average numbers of recommendations, there is little to choose between Hamming Distance over NN and Hamming Distance over All Users. More research, focusing on their relative efficiencies, is needed to choose between these two.

Conversational collaborative recommenders are a new line of research, and enhancing the diversity of their recommendations is a new departure too. Future work could include more systematic investigation of good values for α, b and k, and validation of our results on other datasets. It would also be interesting to compare content-based approaches to the approaches that we have reported in this paper, which use purely collaborative data. We would also like to investigate the role of diversity over the course of the dialogue. Diversity can be helpful in early cycles, when the user is exploring the space and making her short-term interests known; but in later cycles, when the user is homing in on a suitable item, diversity may be less appropriate [8].

References

1. R. Rafter and B. Smyth. Towards conversational collaborative filtering. In L. McGinty and B. Crean, editors, *Procs. of the 15th Artificial Intelligence and Cognitive Science Conference*, pages 147–156, 2004.
2. J. L. Herlocker. *Understanding and Improving Automated Collaborative Filtering Systems*. PhD thesis, University of Minnesota, 2000.
3. D. Bridge and J. P. Kelly. Diversity-enhanced conversational collaborative recommendations. In N. Creaney, editor, *Procs. of the 16th Irish Conference on Artificial Intelligence & Cognitive Science*, pages 29–38. University of Ulster, 2005.
4. B. Smyth and P. McClave. Similarity vs. diversity. In D. W. Aha and I. Watson, editors, *Procs. of the 4th International Conference on Case-Based Reasoning*, pages 347–361. Springer, 2001.
5. D. Bridge and A. Ferguson. Diverse product recommendations using an expressive language for case retrieval. In S. Craw and A. Preece, editors, *Procs. of the 6th European Conference on Case-Based Reasoning*, pages 43–57. Springer, 2002.
6. D. McSherry. Diversity-conscious retrieval. In S. Craw and A. Preece, editors, *Procs. of the 6th European Conference on Case-Based Reasoning*, pages 219–233. Springer, 2002.
7. C.-N. Ziegler, S. M. McNee, J. A. Konstan, and G. Lausen. Improving recommendation lists through topic diversification. In *Procs. of the 14th International World Wide Web Conference*, pages 22–32. ACM Press, 2005.
8. L. McGinty and B. Smyth. On the role of diversity in conversational recommender systems. In K. Ashley and D. Bridge, editors, *Procs. of the 5th International Conference on Case-Based Reasoning*, pages 276–290. Springer, 2003.