A Hybrid Collaborative Filtering Recommender System Using a New Similarity Measure

HYUNG JUN AHN
Management Systems, Waikato Management School
University of Waikato
Private Bag 3105, Hamilton 3240
New Zealand
http://mngt.waikato.ac.nz

Abstract: - This paper presents a hybrid recommender system using a new heuristic similarity measure for collaborative filtering that focuses on improving performance under cold-start conditions where only a small number of ratings are available for similarity calculation for each user. The new measure is based on the domain-specific interpretation of rating differences in user data. Experiments using three datasets show the superiority of the measure in new user cold-start conditions.

Key-Words: Hybrid recommender systems; collaborative filtering; similarity measure

1 Introduction

The most critical component of the collaborative filtering (CF) mechanism is finding similarities between users effectively. However, studies and real-world implementations so far have relied on traditional vector similarity measures, mainly Pearson's correlation or cosine, without questioning the effectiveness of them in the recommender systems domain. This paper is based on the observation that the two measures are significantly flawed for use in CF methods in that they are not properly utilizing the domain specific meanings of the ratings data, especially when the ratings data are not sufficient, often leading to the cold-starting problem which refers to the serious degradation of recommendation quality when only a small number of purchasing records or ratings are available [1,6,7].

In order to address the above problem, this paper designs a heuristic similarity measure based on the minute meanings of co-ratings. In comparison with the generic traditional similarity measures, the suggested measure looks at the ratings data in the context of product recommendation, and hence, better utilizes the ratings in cold-start conditions. The measure is tested with experiments on multiple datasets for completeness.

The remainder of this paper is organized as follows: first, a brief overview of related literature on cold starting problem is given. Next, the new heuristic measure is presented, followed by an experiment that evaluates the performance of a hybrid recommendation method using the measure under an artificial cold-starting condition. Finally, discussion, conclusion, and further research issues are presented.

2 Literature Review

Among the problems of CF briefly, this paper is focusing on the cold-start problem for new users with small number ratings or purchase records. Considering that the average number of purchases per user in a single Internet shopping mall, even over a long period, is usually very limited and that there are always significant portion of new users or less-active users in every shopping mall, the "new user cold-start problem" is a very serious issue for most real-world e-retailers.

The focus of the studies so far addressing the problem has been on targeting cold-start situations where no rating record is available at all for new users or new items. All the studies, to the best of the author's knowledge, present hybrid recommender systems that combine both content information and ratings data [2,5,9-11] to circumvent the problem, where usually content-based similarity is used for new users or new items for whom ratings-based similarity cannot be calculated.

The aim of this research is slightly different from the above stream of studies in that it attempts to improve recommendation performance when the number of rating records is small, but not zero, by developing a new similarity measure for CF systems. One advantage of this approach is that no additional

information is required other than the rating data that CF systems basically utilize and that existing CF recommender systems can be easily updated by only replacing the similarity calculation part. Note, however, that the two types of studies are aiming at different goals that it is difficult to compare one against the other, and hence, this study can be regarded as complementing existing studies on cold-starting problems. Other benefits and limitations of the approach will be discussed further later in this article.

3 New Similarity Measure

In order to address the problem introduced previously, this article presents a heuristic measure based on the following specific goals:

A. The measure should utilize domain specific meanings of data, rather than just employing traditional similarity or distance measures, in order to be more effective in cold-start recommendation conditions.

B. In order to be more practical, the measure should allow easy plug-in to existing collaborative filtering systems by replacing only the similarity measures of the systems, not requiring huge re-implementation or additional data collection.

C. The measure should not only show better results in new user cold start conditions but also comparable results to other popular measures in non-cold-start conditions

The measure is composed of three factors of similarity, *Proximity*, *Impact*, and *Popularity*, and hence, was named PIP. With the PIP measure, the similarity between two users ui and uj is calculated as:

$SIM(u_i, u_j) = \sum_{k \in Ci,j} PIP(r_{ik}, r_{jk})$, where r_{ik} and r_{jk} are the ratings of item k by user i and j respectively, $C_{i,j}$ is the set of co-rated items by user u_i and u_j, and PIP (r_{ik}, r_{jk}) is the PIP score for the two ratings r_{ik} and r_{jk}. For any two ratings r_1 and r_2,

PIP (r_1, r_2) = Proximity (r_1,r_2) × Impact (r_1, r_2) × Popularity (r_1,r_2).

The basic idea behind the three factors is as follows. First, the *Proximity* factor is based on simple arithmetic difference between two ratings, but it further considers whether the two ratings are in agreement or not, giving penalty to ratings in disagreement. The penalty is given by squaring the distance between the ratings (see details in Table 1).

Second, the *Impact* factor considers how strongly an item is preferred or disliked by buyers. When it is strongly preferred or disliked, we can regard that a clearer preference has been expressed for the item, and hence, bigger credibility can be given to the similarity.

Third, the *Popularity* factor gives bigger value to a similarity for ratings that are further from the average rating of a co-rated item. In other words, two users showing the same positive preference for a well-made blockbuster movie may not provide much information regarding their similarity, while two users showing the same positive preference for a cult move may provide much stronger hint about their similarity.

Table 1 Formal description of formulas

Agreement	For any two ratings r_1 and r_2, let R_{max} be the maximum rating and R_{min} the mini mum in the rating scale, and let $R_{med} = \frac{R_{max} + R_{min}}{2}$. A Boolean function *Agreement* (r_1, r_2) is defined as follows: *Agreement* (r_1, r_2) = **false** if ($r_1 > R_{med}$ AND $r_2 < R_{med}$) or ($r_1 < R_{med}$ AND $r_2 > R_{med}$), and *Agreement* (r_1, r_2) = **true** otherwise.
Proximity	A simple absolute distance between the two ratings is defined as: $D(r_1,r_2) = \lvert r_1 - r_2 \rvert$ if *Agreement* (r_1, r_2) is **true**, and $D(r_1,r_2) = \lvert r_1 - r_2 \rvert^2$ if *Agreement* (r_1, r_2) is **false**. Then the *Proximity* (r_1, r_2) is defined as: $Proximity(r_1,r_2) = \{\{2 \cdot (R_{\max} - R_{\min}) + 1\} - D(r_1,r_2)\}^2$
Impact	Impact *Impact* (r_1, r_2) is defined as: $Impact(r_1,r_2) = (\lvert r_1 - R_{med} \rvert + 1)(\lvert r_2 - R_{med} \rvert + 1)$ if *Agreement* (r_1, r_2) is **true**, and

	$Impact(r_1,r_2) = \frac{1}{(\lvert r_1 - R_{med}\rvert + 1)(\lvert r_2 - R_{med}\rvert + 1)}$ if *Agreement* (r_1, r_2) is **false**.
Popularity	Let μ_k be the average rating of item *k* by all users. Then *Popularity* (r_1, r_2) is defined as: $Popularity\ (r_1,r_2) = 1 + (\frac{r_1 + r_2}{2} - \mu_k)^2$ if ($r_1 > \mu_k$ AND $r_2 > \mu_k$) or ($r_1 < \mu_k$ AND $r_2 < \mu_k$), and $Popularity(r_1,r_2) = 1$ otherwise.
Prediction of ratings (Note that this function is used for all three similarity measures.)	A rating for item *k* by user *j* is predicted as [4]: $r'_{ik} = \bar{r}_i + \frac{\sum_j Sim(i,j)(r_{jk} - \bar{r}_j)}{\sum_j \lvert Sim(i,j)\rvert}$ where $\bar{r}_i$ and $\bar{r}_j$ are the average ratings of all items by user *i* and *j* respectively, r_{jk} is the rating of item *k* by user *j*, and *Sim(i, j)* refers to the similarity value between user *i* and *j* calculated using either the Pearson's correlation, cosine, or PIP.

4 Experiment Results

4.1. Overview

In order to prove the effectiveness of the PIP measure, two experiments were performed.

The first experiment compares the performance of three measures, COR (Pearson's Correlation), COS (Cosine), and PIP using the full ratings available for each dataset. The second one experiments a hybrid approach combining COR and PIP. In each of the experiments, sixty percent of the users were used for training (or similarity calculation) and forty percent for testing. Eighty percent of the items were used for training and twenty percent for testing.

All the experiments are repeated for three datasets shown for completeness and better generalization of results. As shown in Table 2, the datasets are publicly open for research purpose and downloadable at the locations shown at the fourth column. Subsets of the datasets were used of which the size is given at the third column.

Table 2 Summary of Datasets

Dataset	Description	Profile
MovieLens [8]	Ratings of movies in scale of 1 to 5.	943 users 1,682 movies 100,000 ratings
Jester joke recommender dataset [3]	Ratings of jokes in scale of -10 to 10.	1,001 users 100 jokes 24,761 ratings
Netflix (Netflix, 2006)	Ratings of movies in scale of 1 to 5.	885 users 1,000 movies 113,885 ratings

4.2. Experiments with Full Ratings

The first experiment simply compared the recommendation performance of the user-based CF method using the three datasets for each similarity measure COR, COS, and PIP. Note that ACOS introduced in Section 2 is not defined for user-based CF and, hence, was not included in this and following experiments. The result in Fig. 1 shows that there is not much difference among the three measures when applied to full ratings. COR is showing the best performance for the MovieLens and Jester datasets while PIP is the best for the Netflix dataset.

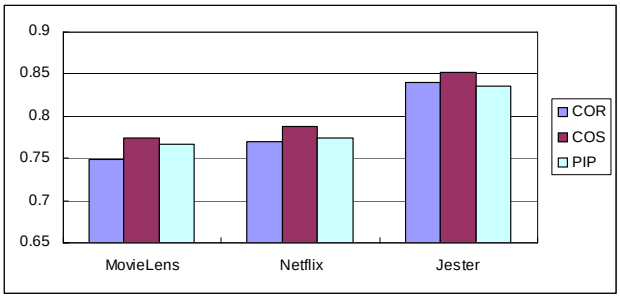
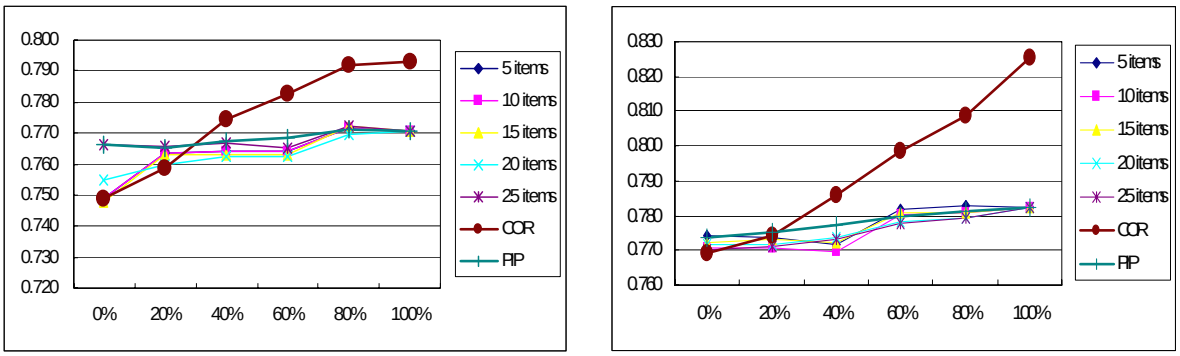


Fig.1 Results with Full Ratings

4.3. Hybrid Recommender

Based on the observation that the COR measure provides better results for the MovieLens and Netflix datasets when all ratings are used, and that the COR measure begins to outperform PIP when the number of

ratings increases, a hybrid approach combining the two measures was tested. Simply put, the hybrid approach uses PIP when the number of ratings for similarity calculation is smaller than or equal to a given threshold, and applies COR when it is greater than the threshold. Threshold values of 10, 15, 20, and 25 were used and the test was repeated for different percentage of cold-start users as shown in the X axis of the graphs of Fig. 2, where cold-start users were defined to have only *k* ratings. *k* was assumed to be uniformly distributed between 1 to 5.



(A) MovieLens **(B) Netflix**

Fig. 2 Hybrid recommendation experiments switching from PIP to COR measure with different percentage of cold-start users. X axis represents the percentage of cold-start users. Y axis represents the prediction accuracy in MAE. Note that COR and PIP results are presented with thicker lines for distinction from hybrid results.

The result shows superior performance of the hybrid approaches where they are showing better results than PIP and COR in most cases. However, when the percentage of cold-start users is 0, COR is showing as good performance as hybrid approaches of 10 or 15. Conversely, when the percentage is 100, PIP is showing equivalent results as hybrid approaches. Hence, it can be concluded that the hybrid approaches in the example dominated the non-hybrid ones except for only non-realistic extreme cases.

5 Conclusion

This paper presented a new heuristic similarity measure called PIP for collaborative filtering that is widely used for automated product recommendation in Internet stores. The PIP measure was developed utilizing domain specific interpretation of user ratings on products in order to overcome the weakness of traditional similarity and distance measures in new user cold-start conditions. PIP was tested using three publicly available datasets for completeness, where it showed superior performance for new user cold-start conditions. A hybrid CF approach was also tested that can combine the strengths of PIP and other similarity measure showing very successful result.

There are limitations of this work as well. First, the PIP measure is a heuristic one that lacks strict mathematical foundation, and hence, is not an optimal solution. Second, the significance of PIP is limited to the similarity calculation of the traditional user-based collaborative filtering, and hence, its effectiveness for other domains is unclear and needs to be tested. Third, although the paper used multiple datasets to allow more generalization of the results, the results may still vary depending on different characteristics of Internet stores or product types.

Further research issues include adapting and applying the PIP measure to other types of product recommendation. For example, item-based collaborative filtering or clustering approaches might modify and adopt the PIP measure and test the performance of it in comparison with other measures. Studying various performance characteristics of PIP can also be interesting and meaningful to see what characteristics make PIP perform better or worse.

Acknowledgement

The author acknowledges the kind generosity of the providers of MovieLens, Netflix, and Jester datasets for allowing the use of their valuable datasets for research.

References:

[1] Cylogy. Personalization Overview. 2005.

[2] Zan Huang, Chen Hsinchun, Zeng Daniel, Applying Associative Retrieval Techniques to Alleviate the

Sparsity Problem in Collaborative Filtering, *ACM Transactions on Information Systems*, Vol.22, No. 1, 2004, pp. 116-142.

[3] Jester. Jester Online Joke Recommender Dataset. 2006.

[4] Joseph A. Konstan, Miller Bradley N., Maltz David, Herlocker Jonathan L., Gordon Lee R., Riedl John, GroupLens: Applying Collaborative Filtering to Usenet News, *Communications of the ACM*, Vol.40, No. 3, 1997, pp. 77-87.

[5] Qing Li, Kim Byeong Man. Culstering Approach to Hybrid Recommendation. IEEE/WIC International Conference on Web Intelligence (WI'03), 2003.

[6] D. Maltz, Ehrlich E. Pointing the way: Active collaborative filtering. CHI 95 Human Factors in Computing Systems. Denver, USA, 1995. pp. 202-209.

[7] Stuart E. Middleton, Alani Harith, Shadbolt Nigel R. , De Roure David C. Exploiting Synergy Between Ontologies and Recommender Systems. Eleventh International World Wide Web Conference (WWW2002). Hawaii, USA, 2002.

[8] MovieLens. MovieLens dataset. 2005.

[9] Seung-Taek Park, Pennock David M., Madani Omid, Good Nathan, DeCoste Dennis. Naive Filterbots for Robust Cold-Start Recommendations. KDD'06. Philadelphia, Pennsylvania, USA: ACM, 2006.

[10] James Salter, Antonopoulos Nick, CinemaScreen Recommender Agent: Combining Collaborative and Content-Based Filtering, *IEEE Intelligent Systems*, Vol.21, No. 1, 2006, pp. 35-41.

[11] Andrew I. Schein, Popescul Alexandrin, Ungar Lyle H., Pennock David M. Methods and Metrics for Cold-Start Recommenadtions. SIGIR'02. Tampere, Finland: ACM, 2002. pp. 253-260.