Available online at www.sciencedirect.com
ScienceDirect
Information Sciences 178 (2008) 37–51
INFORMATION SCIENCES
AN INTERNATIONAL JOURNAL
www.elsevier.com/locate/ins

A new similarity measure for collaborative filtering to alleviate the new user cold-starting problem

Hyung Jun Ahn *

Department of Management Systems, Waikato Management School, University of Waikato, Private Bag 3105, Hamilton 3240, New Zealand

Received 8 February 2007; received in revised form 25 July 2007; accepted 25 July 2007

Abstract

Collaborative filtering is one of the most successful and widely used methods of automated product recommendation in online stores. The most critical component of the method is the mechanism of finding similarities among users using product ratings data so that products can be recommended based on the similarities. The calculation of similarities has relied on traditional distance and vector similarity measures such as Pearson's correlation and cosine which, however, have been seldom questioned in terms of their effectiveness in the recommendation problem domain. This paper presents a new heuristic similarity measure that focuses on improving recommendation performance under cold-start conditions where only a small number of ratings are available for similarity calculation for each user. Experiments using three different datasets show the superiority of the measure in new user cold-start conditions.

Keywords: Similarity measure; Collaborative filtering; Cold-starting

1. Introduction

With the advancement of electronic commerce, automated product recommendation has been perceived as a critical tool for boosting sales in online stores. By providing personalized recommendation of products to users, online stores have been able to increase revenue through up-selling and cross-selling.

There have been numerous ways of product recommendation methods that utilize various types of data and analysis tools [5,6,8,9,12,14,15,17,20,32]. One of the most successful methods is collaborative filtering (CF) that recommends products based on the similarity of users in online stores that is calculated using users' ratings on items. CF has been proved to be successful by numerous studies and has been implemented by many real-world businesses [4,6,8,9,14,15,32–34].

The most critical component of the CF mechanism is finding similarities between users effectively. Studies and real-world implementations so far have relied on traditional vector similarity measures, mainly Pearson's

* Tel.: +64 7 858 5108; fax: +64 7 838 4270.
E-mail address: hjahn@waikato.ac.nz

0020-0255/$ - see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.ins.2007.07.024

correlation or cosine. Despite the overall success of CF recommendation systems, there have been few questions asked if the traditional measures are most suitable for the recommender systems domain. This paper shows that the traditional measures are limited 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 [7,19,21].

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 automated product recommendation is given. Next, the traditional measures are diagnosed showing their limited use or misuse of information contained in ratings data. Next, the new heuristic measure is presented, followed by experiments that evaluate the performance of the measure in various settings. Finally, discussion, conclusion, and further research issues are presented.

2. Literature review

2.1. Overview of research on recommender systems

Because vast amount of research has been produced in the recommendation systems field, it is not easy to classify all the systems and studies into a small number of clear-cut categories. Researchers, however, have often classified the studies broadly into content-based methods and collaborative ones. As the name suggests, content-based methods use content information of items to find the match between items and users. For example, keywords of purchased books of a user might be used to find other books that contain the same or similar keywords for recommendation [1,3,13,17,20,22]. In contrast, the collaborative methods, also called collaborative filtering, use other users' ratings on items to find similarities between users or between items, where recommendation is made using the similarities. The GroupLens system [14,26] and Amazon.com [18] show good examples of this type of recommendation. One can note that the clear difference between the two categories is the type of information used for recommendation. An extensive review based on these two categories can be found in [2].

A further classification of systems has been also attempted by Burke [5]. In this study, the author presents three additional types that constitute in total five categories of recommendation systems: collaborative, content-based, demographic, utility-based, and knowledge-based. Demographic recommendation systems are similar to content-based methods except that similarities are calculated using demographic data, rather than ratings on items. Utility-based methods provide utility functions so that trade-offs can be considered taking into account many factors of utility in making recommendations. However, since we can regard all recommender systems are based on some form of underlying utility assumption, the distinction between this category and others may not be always clear. Knowledge-based methods present knowledge models, usually those that enable inferences, especially for recommendation of complex products so that users' purchasing process could be guided.

Based on the basic categories introduced so far, many hybrid systems have been suggested attempting to overcome the limitations of the methods or to combine benefits of different approaches. The categorization of hybrid systems is again found in [5] where seven categories are presented as: weighted, switching, mixed, feature combination, cascade, feature augmentation, and meta-level. Since repeating the detailed explanation of the categories in this paper might be redundant, interested authors are referred to the original article.

2.2. Collaborative filtering methods

CF methods have been very popular for both researchers and practitioners alike evidenced by the abundance of publications and actual implementation cases [6,8,9,14,15,32]. Although there have been many variations, the basic common idea is to calculate similarity among users using some measure to recommend items

Table 1
Similarity measures often used for CF

Measures	Definition
Pearson's Correlation (COR)	$\mathrm{sim}(u_x,u_y)=\frac{\sum_{h=1}^{n'}(r_{u_x,i_h}-\overline{r_{u_x}})(r_{u_y,i_h}-\overline{r_{u_y}})}{\sqrt{\sum_{h=1}^{n'}(r_{u_x,i_h}-\overline{r_{u_x}})^2}\sqrt{\sum_{h=1}^{n'}(r_{u_y,i_h}-\overline{r_{u_y}})^2}}$, where $r_{u,i}$ is the rating of the item i by user u, $\overline{r_u}$ is the average rating of user u for all the co-rated items, and n' is the number of items co-rated by both users
Cosine (COS)	$\mathrm{sim}(u_x,u_y)=\frac{\sum_{h=1}^{n'}r_{u_x,i_h}r_{u_y,i_h}}{\sqrt{\sum_{h=1}^{n'}r_{u_x,i_h}^2}\sqrt{\sum_{h=1}^{n'}r_{u_y,i_h}^2}}$, where $r_{u,i}$ is the rating of the item i by user u and $n'$is the number of items co-rated by both users
Adjusted Cosine (ACOS) for similarity between items	$\mathrm{sim}(i_x,i_y)=\frac{\sum_{j=1}^{m'}(r_{u_j,i_x}-\overline{r_{u_j}})(r_{u_j,i_y}-\overline{r_{u_j}})}{\sqrt{\sum_{j=1}^{m'}(r_{u_j,i_x}-\overline{r_{u_j}})^2}\sqrt{\sum_{j=1}^{m'}(r_{u_j,i_y}-\overline{r_{u_j}})^2}}$, where $r_{u,i}$is the rating of the item i by user u, $\overline{r_u}$ is the average rating of user u for all the items rated by the user, and m' is the number of users who rated both of the items
Constrained Pearson's Correlation (CPC)	$\mathrm{sim}(u_x,u_y)=\frac{\sum_{h=1}^{n'}(r_{u_x,i_h}-r_{\mathrm{med}})(r_{u_y,i_h}-r_{\mathrm{med}})}{\sqrt{\sum_{h=1}^{n'}(r_{u_x,i_h}-r_{\mathrm{med}})^2}\sqrt{\sum_{h=1}^{n'}(r_{u_y,i_h}-r_{\mathrm{med}})^2}}$, where $r_{u,i}$ is the rating of the item i by user u, r_{med} is the median value in the rating scale (e.g. 3 in the rating scale of 5), and n' is the number of items co-rated by both users
Spearman's Rank Correlation (SRC)	$\mathrm{sim}(u_x,u_y)=1-\frac{6\sum_{h=0}^{n'}d_h^2}{n'(n'^2-1)}$, where d_h is the difference in the ranks for item h by the two users and n' is the number of items co-rated by both users

based on the similarity. The approaches that use similarities among items instead are called item-based collaborative filtering (IBCF). Note that, in the remainder of this paper, the term CF is used exclusively for user-based CF.

The definitions of similarity measures that have been used for CF recommender systems are summarized in Table 1. The first one, Pearson's correlation, measures the linear correlation between two vectors of ratings. The cosine measure looks at the angle between two vectors of ratings where a smaller angle is regarded as implying greater similarity. The third one, adjusted cosine, is used in some IBCF methods for similarity among items where the difference in each user's use of the rating scale is taken into account [28].[1] The fourth one, constrained Pearson's correlation, is a slightly modified version of Pearson's correlation that allows only the pairs of ratings on the same side, e.g. both being positive or both being negative, to contribute to the increase in the correlation [31]. The fifth one is called Spearman's Rank Correlation where similarity between two vectors is calculated based on the similarity of ranks of values in the vectors [9]. There are some variations of CF methods that use more or less modified formulae, but this paper uses the formulae in Table 1 for all experiments.

Once similarities are ready, prediction of a rating of an item i_a by user u_a can be calculated as follows for user-based CF methods [14]:

$$p(u_a,i_a)=\overline{r_{u_a}}+\frac{\sum_{h=1}^{m'}\mathrm{sim}(u_a,u_h)(r_{u_h,i_a}-\overline{r_{u_h}})}{\sum_{h=1}^{m'}|\mathrm{sim}(u_a,u_h)|}, \tag{1}$$

where m' is the number of other users who have also rated item i_a.

Similar prediction formulae are also available in literature for IBCF and other variations of CF.

Literature provides rich evidence on the successful performance of CF methods. However, there are some shortcomings of the methods as well. First, CF methods are known to be vulnerable to data sparsity and to have cold-start problems. Data sparsity refers to the problem of insufficient data, or sparseness, in the ⟨user × item⟩matrix [4,29,35]. Cold-start problems refer to the difficulty of recommending new items or recommending to new users where there are not sufficient ratings available for them [4,7,19,21].

[1] Readers should note that the Adjusted Cosine (ACOS) is not used in the experiments of this paper since this paper is focused on improving user-based collaborative filtering methods.

2.3. Approaches to cold-start problems

Among the problems of CF briefly introduced in Section 2.2, 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 Internet store, 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 [10,16,25,27,30] 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 use 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 on.

3. A new similarity measure

3.1. Diagnosis of the limitations of existing similarity measures

Although the two most-widely used similarity measures for CF, COR and COS, have been proved to be successful in many studies, they are limited to be used in new user cold-start situations where only a small number of ratings are available for similarity calculation. The problem is even more amplified by the sparsity of available data. For example, if the sparsity level is 0.90, which means only 10% of all possible ratings are present in the dataset, and if a user has only five purchase records available, then the average number of ratings available for similarity calculation per each reference user is 0.5, which makes it infeasible to derive similarity values with many of the reference users using the traditional measures. Along with this, major limitations can be briefly summarized as follows:

(1) Very limited number of co-ratings under data sparsity.
(2) If the number of co-rated items is 1, COR cannot be calculated and COS results in 1 regardless of differences in individual ratings.
(3) If all the available ratings of a given user are flat, e.g. $\langle 1,1,1\rangle$, $\langle 3,3,3\rangle$ or $\langle 4,4,4\rangle$, COR cannot be calculated between the user (and, hence, often regarded as 0) and another since the denominator part of the correlation formula becomes zero.
(4) If two vectors are on the same line, e.g. vectors $\langle 2,2\rangle$ and $\langle 3,3\rangle$, COS results in 1 regardless of the difference between the two.
(5) Both COR and COS can be sometimes misleading, where very different users may appear to be very similar to each other by the similarity measures, and vice versa (see Fig. 1 for examples).

All the above problems become more serious when the number of ratings is limited and the dataset is sparse, because the probability of them occurring by chance is bigger when sufficient ratings are not available for similarity calculation.

Part of the above problems can be clearly observed through simple experiments using a sample dataset as seen in Fig. 2. Regarding Pearson's correlation, the sample statistics in Fig. 2 illustrate how much of limited but valuable information is discarded when using the measure under an artificial cold-start condition. The statistics were generated assuming that there are only five ratings per each buyer for 7650 pairs of randomly

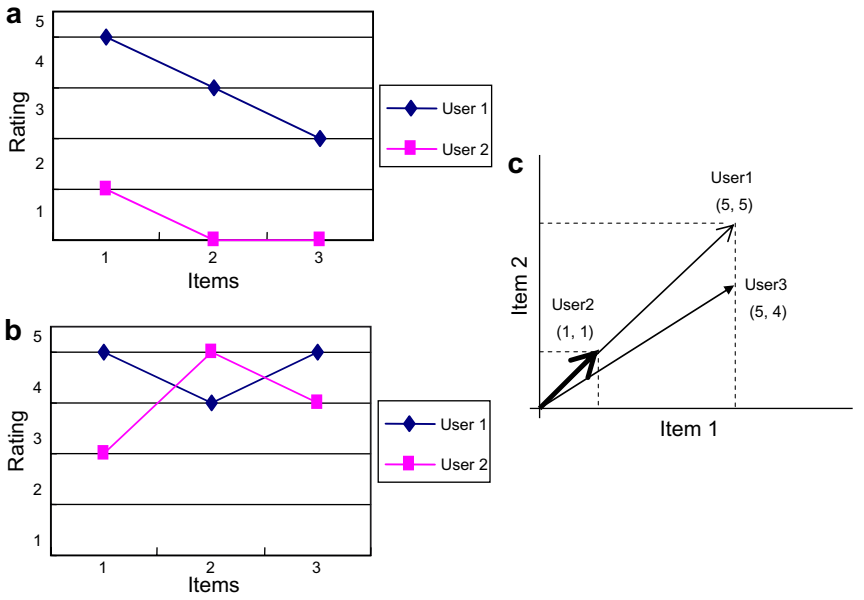


Fig. 1. (a)–(c) show examples of misleading values of Pearson's correlation and cosine. (a) Shows high correlation regardless of the difference in the ratings of the two users and (b) shows low correlation regardless of the similar ratings by two users. Users 1 and 3 are showing very similar ratings for the two items in (c), but the cosine value between them are smaller than that of users 1 and 2 whose rating vectors are on the same line.

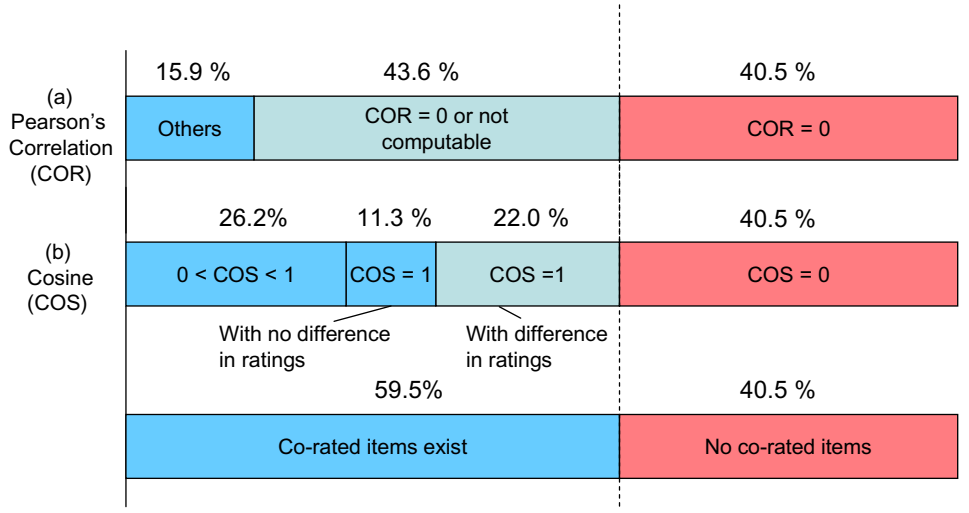


Fig. 2. Sample statistics that show the low utilization of meaningful information in an experimental cold-start condition where it is assumed that there are only five ratings available for buyers.

chosen users in the MovieLens dataset (see Table 3 for the summary of the datasets used in this article). According to (a) of Fig. 2, among all the pairs of users, 40.5% appear to have no co-rated items leading to zero or incomputable similarity values. However, additional 43.6% are also showing zero or incomputable similarities even when there are one or more co-rated items. Thus, only 15.9% of the pairs are generating non-zero similarities, showing very low utilization of available information under the harsh condition. The same statistics were also collected for the cosine measure as with the Pearson's correlation, using the same MovieLens dataset. In (b) of the figure, because the measure only considers the angle between two given vec-

tors, 22.0% of the similarities have value 1 even when there is difference between the ratings of two users, again not fully utilizing available information.

3.2. A new similarity measure for CF systems: PIP (Proximity–Impact–Popularity) measure

The diagnosis in the previous section has shown that the significant underutilization of available information by the two similarity measures is contributing to the cold-start problem of CF systems. In order to address this problem, 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 CF 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 u_i and u_j is calculated as:

$$\mathrm{SIM}(u_i, u_j) = \sum_{k \in Ci,j} \mathrm{PIP}(r_{ik}, r_{jk}),$$

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 $\mathrm{PIP}(r_{ik}, r_{jk})$ is the PIP score for the two ratings r_{ik} and r_{jk}. For any two ratings r_1 and r_2, $\mathrm{PIP}(r_1, r_2) = \mathit{Proximity}(r_1, r_2) \times \mathit{Impact}(r_1, r_2) \times \mathit{Popularity}(r_1, r_2)$.

Fig. 3 illustrates the basic ideas behind the three factors. First, the *Proximity* factor is based on the 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. That is, if two ratings are on the same side of a given rating scale which is divided by its median, they are regarded to be in agreement. For example, in (a) of Fig. 3, the pair (a_1, a_2) is given further penalty since the two ratings a_1 and a_2 are in disagreement, one preferring the item, the other disliking, compared with the (b_1, b_2) pair located on the same side. The penalty is given by doubling the distance between the ratings, which is then squared (see details in Table 2).

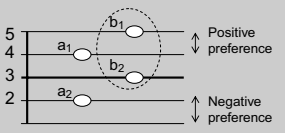
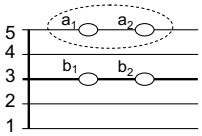
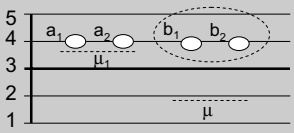
Proximity	**I**mpact	**P**opularity
		
Both pairs (a_1, a_2) and (b_1, b_2) have the same distance of 2. However, in the first pair, a_1 is positive while a_2 is negative, and hence, the distance between them is given further penalty. On the other hand, in the second pair, b_1 is positive and b_2 is neutral, and hence, no further penalty is given.	Both pairs (a_1, a_2) and (b_1, b_2) have zero distance. However, the first pair shows an agreement at a stronger preference level, and hence, is regarded as having more *impact* than the second pair.	Both pairs (a_1, a_2) and (b_1, b_2) have zero distance and the same *impact* factor. However, when μ_2 is the average rating for the co-rated item, the similarity between the two can be regarded as showing stronger evidence of similarity compared with when the average is μ_1, because the two users are showing more differentiated preference compared with average users.

Fig. 3. Description of the three factors of PIP using example ratings.

Table 2
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 minimum in the rating scale, and let $R_{\text{med}} = \frac{R_{\max} + R_{\min}}{2}$ A Boolean function $Agreement(r_1, r_2)$ is defined as follows: $Agreement(r_1, r_2) =$ **false** if ($r_1 > R_{\text{med}}$ and $r_2 < R_{\text{med}}$) or ($r_1 < R_{\text{med}}$ and $r_2 > R_{\text{med}}$), 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 $D(r_1, r_2) = 2 \cdot \lvert r_1 - r_2 \rvert$ 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: $Impact(r_1, r_2) = (\lvert r_1 - R_{\text{med}} \rvert + 1)(\lvert r_2 - R_{\text{med}} \rvert + 1)$ if $Agreement(r_1, r_2)$ is **true**, and $Impact(r_1, r_2) = \frac{1}{(\lvert r_1 - R_{\text{med}} \rvert + 1)(\lvert r_2 - R_{\text{med}} \rvert + 1)}$ 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: $Popularity(r_1, r_2) = 1 + \left(\frac{r_1 + r_2}{2} - \mu_k\right)^2$ 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	Formula (1) of Section 2.2 is used for prediction for all measures

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. For example, in (b) of Fig. 3, both pairs (a_1, a_2) and (b_1, b_2) are showing zero distance in agreement, hence having the same *Proximity* factor. However, the ratings a_1 and a_2 are showing clearer and stronger preference than the others, and hence, a bigger impact factor is given to the similarity between the two. Third, the *Popularity* factor gives bigger value to a similarity for ratings that are further from the average rating of a co-rated item. As in (c) of Fig. 3, when two ratings a_1 and a_2 are close to the average rating of the item, this agreement between the two might not provide much information about the similarity between the two users, because the similar ratings can be just a result of being close to the average alike. In contrast, if two ratings are close to each other but are very far from the average rating of the item as with the ratings b_1 and b_2 in the example when the average is μ_2, this can be signaling stronger similarity between the two users. 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 movie may provide much stronger hint about their similarity.

Table 2 provides detailed description on how the three factors are calculated exactly, along with the prediction function that predicts ratings of users after similarities are calculated.

4. Experiments

4.1. Overview

In order to prove the effectiveness of the PIP measure, several experiments were performed focusing on testing recommendation performance in new user cold-starting conditions.

The first experiment compares the performance of the five measures, COR, COS, CPC, SRC, and PIP using the full ratings available for each dataset. The second experiment simulates artificial cold-starting situations and compares the performance of each measure. The third one creates virtual cold-start situations with

different percentage of new users and performs recommendations. The fourth one experiments a hybrid approach combining COR and PIP.

All the experiments are repeated for the three datasets shown in Table 3 for completeness and better generalization of results. As shown in the table, the datasets are publicly open for research purpose and downloadable at the locations shown at the fifth column. Subsets of the datasets were used and the sizes are given at the third column. The MovieLens and Netflix datasets provide ratings on movies in the scale of 1–5 and the Jester dataset provides ratings on jokes in the scale of −10 to 10. However, all the experiment results using the Jester dataset were scaled down to be equivalent with the other two datasets for easy comparison. For all the experiments, 80% of the users were used as reference users for similarity calculation and actual recommendation was conducted to 20%. Similarly, 80% of the movies or jokes were used for similarity calculation, while 20% were actually recommended to users. Table 4 briefly describes each experiment.

4.2. Experiments with full ratings

The first experiment simply compared the recommendation performance of the CF method using the three datasets for five similarity measures: COR, COS, PIP, CPC and SRC. This experiment also included two baseline recommendations, Base1 and Base2, which perform recommendations based on just the average item ratings and the average user ratings respectively. Note that ACOS introduced in Section 2 is not defined for

Table 3
Summary of the datasets

Dataset	Description	Profile	Ratings per user	Availability
MovieLens [23]	Ratings of movies in scale of 1–5	943 users 1682 movies 100,000 ratings	On more than 100 movies per user	Available at http://www.grouplens.org/
Jester joke recommender dataset [11]	Ratings of jokes in scale of −10 to 10	1001 users 100 jokes 24,761 ratings	On more than 24 jokes per user	Available at http://www.ieor.berkeley.edu/~goldberg/jester-data/
Netflix [24]	Ratings of movies in scale of 1–5	885 users 1000 movies 113,885 ratings	On more than 100 movies per user	Available at http://www.netflixprize.com/

Table 4
Description of the experiments

No	Experiment	Description	Tested variables
1	With full ratings	Testing the measures with all available ratings with each dataset	–
2	For cold-starting users (new users)	Testing the measures for artificial cold-start conditions where only 1–20 ratings are used for similarity calculation	1. Number of ratings used for similarity calculation
3	For different percentage of cold-starting users	Testing the measures for different proportions of new users and non-new users	1. Percentage of new users (new users are defined to have less than N_n ratings, where N_n is a small number)
4	Hybrid approach combining PIP and COR	Testing a hybrid recommendation method switching from PIP to COR. The test is repeated for different proportions of new users and non-new users	1. Threshold of switching from PIP to COR 2. Percentage of new users (new users are defined to have less than N_n ratings, where N_n is a small number)

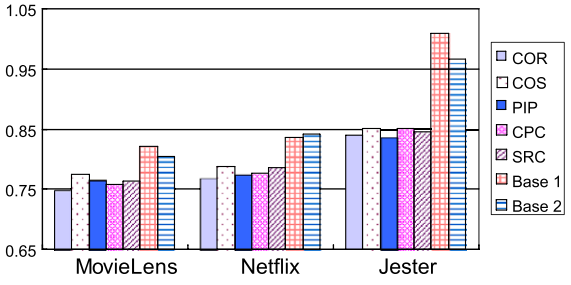


Fig. 4. Comparison of the measures using full ratings.

user-based CF and, hence, was not included in this and following experiments. The result in Fig. 4 shows that there is not much difference among the five measures when applied to full ratings. COR is showing the best performance for the MovieLens and Netflix datasets while PIP is the best for the Jester dataset. All the measures are showing better performance compared with the two baseline recommendation methods.

4.3. Artificial cold-starting

Next, since PIP was designed to improve recommendation performance in cold-starting conditions, this experiment tested artificial new user cold-starting conditions by allowing the similarity computation to use only a small number of ratings per each user. The number of ratings, n_r, for similarity calculation was increased from 1 to 20. For each number of ratings, recommendation was performed for each dataset and each similarity measure. For the sake of clarity, the comparison with CPC and SRC is presented separately in Appendix 1.

The results are very positive for the PIP measure (see Fig. 5). First, for all the three datasets, PIP is showing the overall best performance with lowest mean absolute error (MAE) over the range of number of ratings used. However, for the two datasets MovieLens and Netflix, COS is showing slightly better performance at the very beginning. With the Jester dataset, PIP is clearly outperforming the other two measures. Second, it is observed that the performance of COR gradually improves and approaches that of PIP. This is consistent with the result generated from the full ratings experiment. From these results, it can be concluded that the PIP measure shows advantage over the other two measures under cold-starting conditions. The dominance of PIP is continuing until the number of ratings is 20 or slightly smaller in all three datasets. In the comparison with CPC and SRC, the dominance of PIP is clearer (Appendix 1).

A series of paired T test was performed to compare the results of PIP with those of COR and COS when $n_r = 5$, 10, and 15 (Appendix 3). The results of one-tailed hypotheses show that most of the differences appear to be statistically significant, although some of the small differences for the MovieLens dataset appear to be not.

4.4. Different percentage of cold-start users

The previous experiment in Section 4.3 showed the strength of the PIP measure in cold-starting conditions. However, in reality, there are always a mix of new users and extant ones, or users with few purchase records and users with sufficient records. In order to simulate this realistically, another experiment was performed changing the percentage of 'fairly new users'. Users with no more than five ratings were assumed to be new users and the percentage of these users was increased from 0 to 100 by 20. The ratings of these users for similarity calculation were randomly deleted so that the number of ratings is uniformly distributed between 1 and 5.

The results are shown in Fig. 6. As can be easily expected, the bigger the percentage of new users, the better the performance of the PIP measure compared with the other measures. In order for the COR measure to perform better than PIP for the MovieLens and Netflix datasets, roughly 70–80% of users need to have a sufficiently large number of records, which suggests that PIP can be very effective in most practical situations. In the case of Jester dataset, PIP is outperforming the other two measures in all levels of new user proportion. Again, for the sake of clarity, the comparison involving CPC and SRC is presented separately in Appendix 2, where we can observe the same results in favor of the PIP measure.

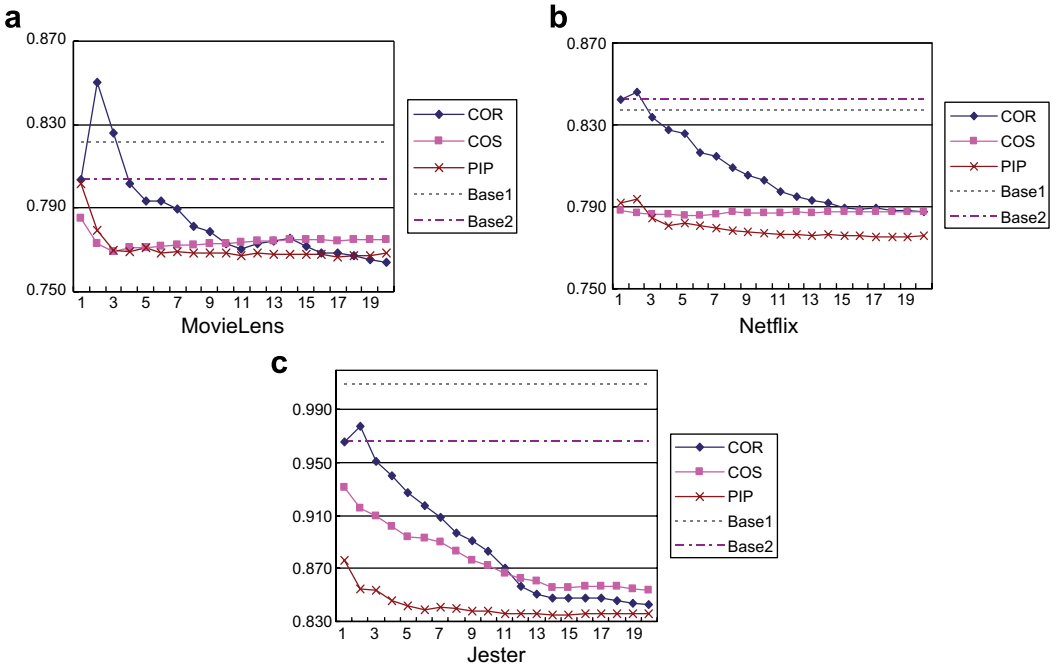


Fig. 5. Artificial new user cold-start experiments – *X*-axis represents the number of ratings used and *Y* the prediction accuracy measured in MAE.

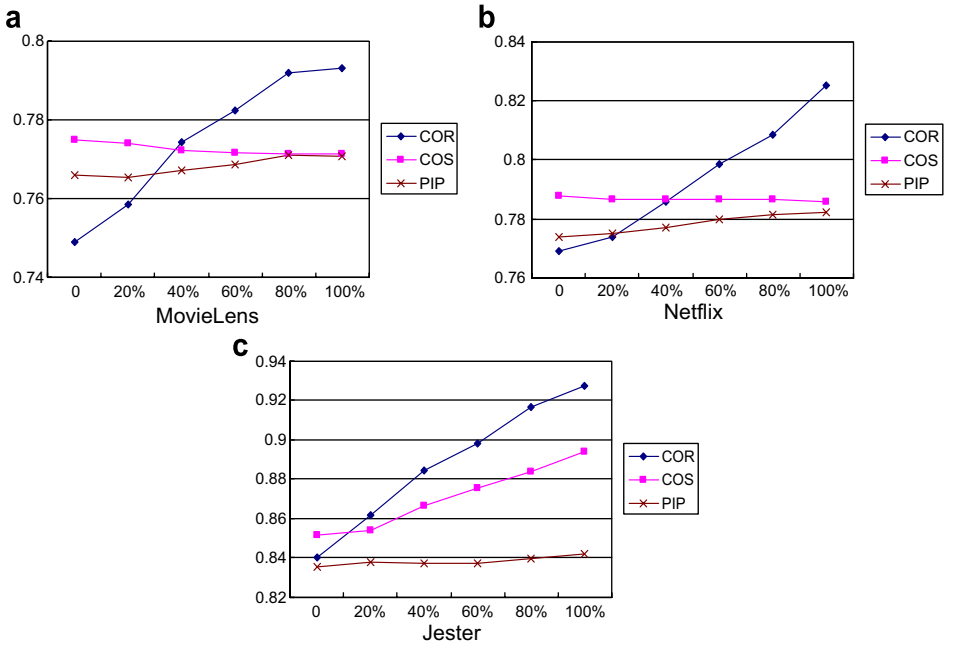


Fig. 6. Virtual cold-starting experiments using different percentage of cold-start users for each dataset. *X*-axis represents the percentage of cold-start users while *Y*-axis represents the recommendation accuracy in MAE.

One peculiarity observed is the increasing performance of the COS measure for the MovieLens and Netflix datasets. This might be due to either the inherent nature of the COS measure or the characteristics of the datasets. This issue, however, needs further investigation and is regarded as being out of scope of this research.

4.5. A hybrid approach

Based on the previous 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 as 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 5, 10, 15, 20, and 25 were used and the test was repeated for different percentage of cold-start users in the same way as in Section 4.4. The result is shown in Fig. 7.

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.

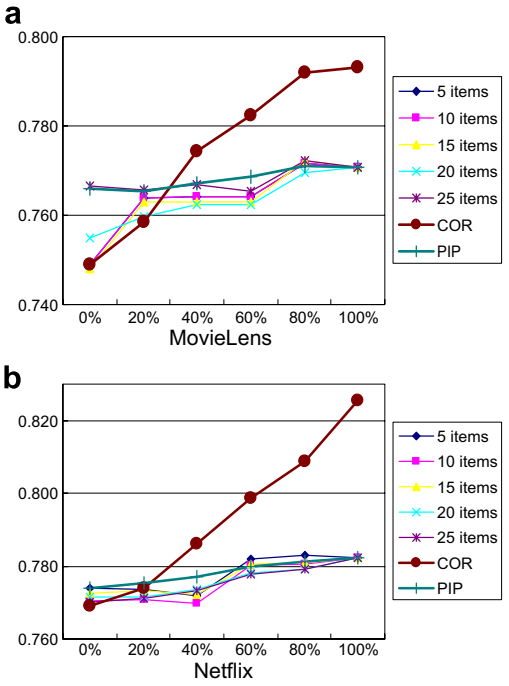


Fig. 7. Hybrid recommendation experiments switching from PIP to COR measure with different percentage of cold-start users. *X*-axis represents the percentage of cold-start users. *Y*-axis represents the prediction accuracy in MAE. The experiments were repeated for different threshold values of switching from PIP to COR. Note that COR and PIP results are presented with thicker lines for distinction from hybrid results.

Table 5
Summary of the experiment results

Experiment	Result
With full ratings	Different results for different datasets, but the difference is relatively small
For cold-starting users (new users)	PIP dominates other measures
For different percentage of cold-starting users	PIP dominates when the percentage of cold-starting users is not too small. (More than 25% for MovieLens and 20% for Netflix. Always for Jester.)
Hybrid approach combining PIP and COR	The hybrid approach dominates except for extreme cases

4.6. Summary of the findings and discussion

The summary of the experiment results is given in Table 5. It was observed that PIP has clear advantage over the other measures for new user cold-starting conditions. However, the results with full ratings varied depending on datasets but the difference was relatively small. In two of the datasets, COR showed the best performance, while, in one dataset, the PIP measure produced the best result. Considering that the average numbers of full ratings available in the two datasets where COR dominated are very large (see Table 3) compared with those of real Internet stores, the dominance of COR by a small margin might be insignificant in many real Internet shops. In more realistic settings when there were different portions of cold-start users, PIP still showed superior performance to the other measures except when the percentage of the new users was very small. The hybrid approach applied to the MovieLens and Netflix datasets showed very positive results where most hybrid approaches dominated the other approaches that used only PIP or COR alone. Since PIP has shown strong results for new user conditions and there might be other measures that perform better than PIP for non-cold-starting conditions, this result suggests that hybrid similarity combining PIP and others can often bring significant improvements.

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 CF, and hence, its effectiveness for other domains is unclear and needs to be tested. Third, although the paper used three datasets to allow more generalization of the results, the results may still vary depending on different characteristics of Internet stores, product types, etc. As we have seen from some of the results of the cold-starting experiments using the MovieLens dataset that appeared to be statistically insignificant, care should be taken in applying PIP to individual Internet stores by considering various factors and conducting pilot experiments.

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 suggested that can combine the strengths of PIP and other similarity measures, showing very successful results.

The academic contribution of this paper can be summarized as follows. First, to the best of the author's knowledge, the measure presented in this paper is the first work that designs a new similarity measure for CF methods that can replace traditional similarity and distance measures such as Pearson's correlation and cosine. Second, most studies that address the cold-starting problem have presented hybrid approaches that utilize both content-based information and rating data together focusing on cold-start situations where no rating is available at all. The focus of the present research is different from them in that it aims at improving existing CF methods by just replacing traditional similarity measures and also that it improves new user cold-start situations with one or more available ratings. Although direct comparison is not possible, the advantage of the present approach is that it requires no additional data and minimal additional implementation or modification of existing CF recommender systems. However, since the present research cannot be applied to completely new users without any rating record, this approach should be regarded as being complementary to existing studies.

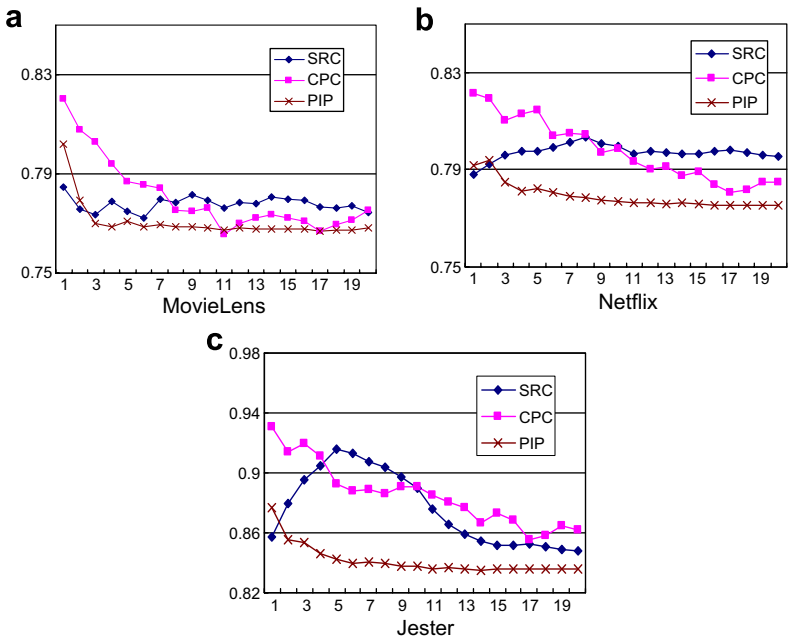
Further research issues include adapting and applying the PIP measure to other types of product recommendation. For example, item-based CF or clustering approaches might modify and adopt the PIP measure and test the performance of it in comparison with other measures. Studying various performance characteristics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

Acknowledgements

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. The author also acknowledges the very helpful feedback from three anonymous reviewers.

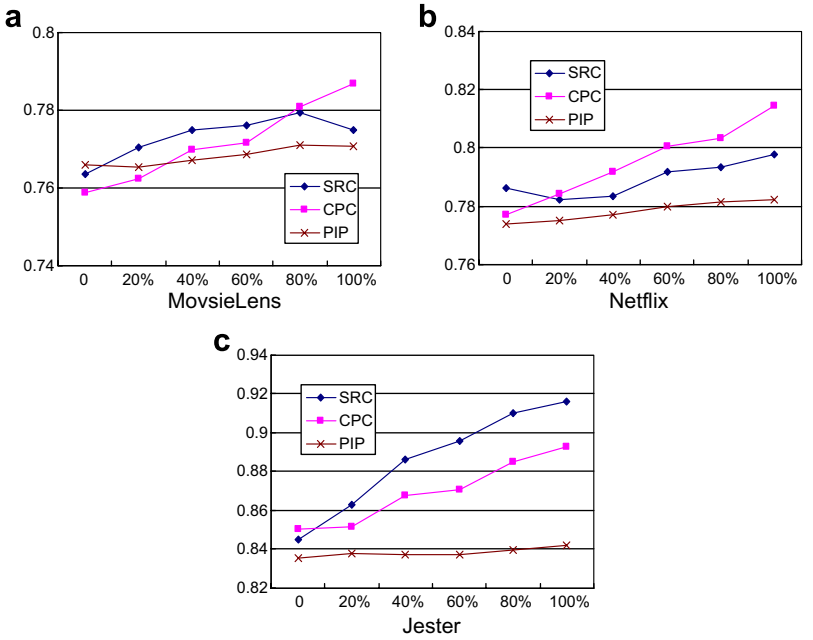
Appendix 1

Artificial cold-starting experiments in comparison with SRC and CPC. *X*-axis represents the number of ratings used and *Y* the prediction accuracy measured in MAE.



Appendix 2

Experiments for different percentage of cold-start users in comparison with SRC and CPC. *X*-axis represents the percentage of cold-start users while *Y*-axis represents the recommendation accuracy in MAE.



Appendix 3

Statistical significance of the differences in the artificial cold-start experiments

	$n_r = 5$		$n_r = 10$		$n_r = 15$	
	PIP vs. COS	PIP vs. COR	PIP vs. COS	PIP vs. COR	PIP vs. COS	PIP vs. COR
MovieLens	$t = -0.11$	$t = -1.60$	$t = -1.40$	$t = -0.33$	$t = -2.16$	$t = -0.32$
d.f. = 186	$p = 0.46$	$p = 0.06^{*}$	$p = 0.08^{*}$	$p = 0.37$	$p = 0.02^{**}$	$p = 0.37$
Netflix	$t = -1.00$	$t = -4.19$	$t = -2.99$	$t = -3.20$	$t = -3.37$	$t = -1.98$
d.f. = 175	$p = 0.15$	$p = 0.00^{***}$	$p = 0.00^{***}$	$p = 0.00^{***}$	$p = 0.00^{***}$	$p = 0.02^{**}$
Jester	$t = -3.57$	$t = -5.28$	$t = -2.88$	$t = -3.20$	$t = -1.87$	$t = -0.98$
d.f. = 198	$p = 0.00^{***}$	$p = 0.00^{***}$	$p = 0.00^{***}$	$p = 0.00^{***}$	$p = 0.03^{**}$	$p = 0.16$

PIP was compared with COR and COS when the number of ratings, n_r, is 5, 10, and 15 for each of the three datasets (d.f. = degree of freedom).
* For significance at 90%.
** For significance at 95%.
*** For significance at 99%.

References

[1] G. Adomavicius, R. Sankaranarayanan, S. Sen, A. Tuzhilin, Incorporating contextual information in recommender systems using a multidimensional approach, ACM Transactions on Information Systems 23 (2005) 103–145.
[2] G. Adomavicius, A. Tuzhilin, Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions, IEEE Transactions on Knowledge & Data Engineering 17 (2005) 734–749.
[3] A. Ahn, J.K. Kim, I.Y. Choi, Y.H. Cho, A personalised recommendation procedure based on dimensionality reduction and web mining, International Journal of Internet & Enterprise Management 2 (2004) 280–298.
[4] H.J. Ahn, Utilizing popularity characteristics for product recommendation, International Journal of Electronic Commerce 11 (2006) 57–78.
[5] R. Burke, Hybrid recommender systems: survey and experiments, User Modeling and User-Adapted Interaction 12 (2002) 331–370.
[6] W.W. Cohen, W. Fan, Web-collaborative filtering: recommending music by crawling the Web, Computer Networks 33 (2000) 685–698.
[7] Cylogy, Personalization Overview, http://www.cylogy.com/library/personalization_overview-kb.pdf (accessed April 2006).
[8] G. Greco, S. Greco, E. Zumpano, Collaborative filtering supporting web site navigation, AI Communications 17 (2004) 155–166.
[9] J.L. Herlocker, J.A. Konstan, L.G. Terveen, J.T. Riedl, Evaluating collaborative filtering recommender systems, ACM Transactions on Information Systems 22 (2004) 5–53.
[10] Z. Huang, H. Chen, D. Zeng, Applying associative retrieval techniques to alleviate the sparsity problem in collaborative filtering, ACM Transactions on Information Systems 22 (2004) 116–142.
[11] Jester, Jester Online Joke Recommender Dataset, http://www.ieor.berkeley.edu/~goldberg/jester-data/ (accessed September 29).
[12] B.-D. Kim, S.-O. Kim, A new recommender system to combine content-based and collaborative filtering systems, Journal of Database Marketing 8 (2001) 244–252.
[13] Y.S. Kim, B.-J. Yum, J. Song, S.M. Kim, Development of a recommender system based on navigational and behavioral patterns of customers in e-commerce sites, Expert Systems with Applications 28 (2005) 381.
[14] J.A. Konstan, B.N. Miller, D. Maltz, J.L. Herlocker, L.R. Gordon, J. Riedl, GroupLens: applying collaborative filtering to usenet news, Communications of the ACM 40 (1997) 77–87.
[15] D.-S. Lee, G.-Y. Kim, H.-I. Choi, A web-based collaborative filtering system, Pattern Recognition 36 (2003) 519–526.
[16] Q. Li, B.M. Kim, Clustering approach to hybrid recommendation, in: Proceedings of IEEE/WIC International Conference on Web Intelligence (WI'03), 2003.
[17] Y. Li, L. Lu, L. Xuefeng, A hybrid collaborative filtering method for multiple-interests and multiple-content recommendation in E-Commerce, Expert Systems with Applications 28 (2005) 67–77.
[18] G. Linden, B. Smith, J. York, Amazon.com recommendations: item-to-item collaborative filtering, IEEE Internet Computing 7 (2003) 76–80.
[19] D. Maltz, E. Ehrlich, Pointing the way: active collaborative filtering, in: Proceedings of CHI95 Human Factors in Computing Systems, Denver, USA, 1995, pp. 202–209.
[20] P. Melville, R.J. Mooney, R. Nagarajan, Content-boosted collaborative filtering for improved recommendations, in: Proceedings of the Eighteenth National Conference on Artificial Intelligence (AAAI-2002), Edmonton, Canada, 2002, pp. 187–192.

[21] S.E. Middleton, H. Alani, N.R. Shadbolt, D.C. De Roure, Exploiting synergy between ontologies and recommender systems, in: Proceedings of the Eleventh International World Wide Web Conference (WWW2002), Hawaii, USA, 2002.
[22] N. Mirzadeh, F. Ricci, M. Bansal, Feature selection methods for conversational recommender systems, in: Proceedings of 2005 IEEE International Conference on e-Technology, e-Commerce and e-Service, Hong Kong, China, 2005, pp. 772–777.
[23] MovieLens, MovieLens dataset, http://www.grouplens.org/ (accessed on April 2006).
[24] Netflix, Netflix movie dataset, http://www.netflixprize.com/ (accessed November 2006).
[25] S.-T. Park, D.M. Pennock, O. Madani, N. Good, D. DeCoste, Naive filterbots for robust cold-start recommendations, in: Proceedings of KDD'06, ACM, Philadelphia, PA, USA, 2006.
[26] P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom, J. Riedl, GroupLens: an open architecture for collaborative filtering of netnews, in: Proceedings of ACM 1994 Conference on Computer Supported Cooperative Work. ACM, Chapel Hill, NC, 1994, pp. 175–186.
[27] J. Salter, N. Antonopoulos, CinemaScreen recommender agent: combining collaborative and content-based filtering, IEEE Intelligent Systems 21 (2006) 35–41.
[28] B. Sarwar, G. Karypis, J.A. Konstan, J. Riedl, Item-based collaborative filtering recommendation algorithms, in: Proceedings of the 10th International World Wide Web Conference (WWW10), Hong Kong, 2001, pp. 285–295.
[29] A.I. Schein, A. Popescul, L.H. Ungar, CROC: a new evaluation criterion for recommender systems, Electronic Commerce Research 5 (2005) 51–74.
[30] A.I. Schein, A. Popescul, L.H. Ungar, D.M. Pennock, Methods and metrics for cold-start recommendations, in: Proceedings of SIGIR'02, ACM, Tampere, Finland, 2002, pp. 253–260.
[31] U. Shardanand, P. Maes, Social information filtering: algorithms for automating "Word of Mouth, in: Proceedings of ACM CHI'95 conference on human factors in computing systems, Denver, CO, 1995, pp. 210–217.
[32] R. Vezina, D. Militaru, Collaborative filtering: theoretical positions and a research agenda in marketing, International Journal of Technology Management 28 (2004) 31–45.
[33] M. Vozalis, K. Margaritis, Using SVD and demographic data for the enhancement of generalized collaborative filtering, Information Sciences 177 (2007) 3017–3037.
[34] B. Xie, P. Han, F. Yang, R. Shen, H.-J. Zeng, Z. Chen, DCFLA: a distributed collaborative-filtering neighbor-locating algorithm, Information Sciences 177 (2007) 1349–1363.
[35] C. Zeng, C.-X. Xing, L.-Z. Zhou, X.-H. Zheng, Similarity measure and instance selection for collaborative filtering, International Journal of Electronic Commerce 8 (2004) 115–129.