The Effect of the Personalized Settings for CF-Based Recommender Systems

Il Im
School of Business, Yonsei University
(il.im@yonsei.ac.kr)

Byung Ho Kim
School of Business, Yonsei University
(goodbear@yonsei.ac.kr)

In this paper, we propose a new method for collaborative filtering (CF)-based recommender systems. Traditional CF-based recommendation algorithms have applied constant settings such as a reference group (neighborhood) size and a significance level to all users. In this paper we develop a new method that identifies optimal personalized settings for each user and applies them to generating recommendations for individual users. Personalized parameters are identified through iterative simulations with 'training' and 'verification' datasets. The method is compared with traditional 'constant settings' methods using Netflix data. The results show that the new method outperforms traditional, ordinary CF. Implications and future research directions are also discussed.

Received : June 09, 2012 Revision : June 12, 2012 Accepted : June 19, 2012
Type of Submission : Excellent Paper of Conference Corresponding author : Im Il

1. Introduction

Collaborative filtering (CF) has become a popular method for recommender systems (Im and Hars, 2007; Xiao and Benbasat, 2007; Park et al., 2011). CF, in a nutshell, is a method that finds a group of users or neighbors who have similar preferences with the active user and generates recommendations based on those users' ratings.

There are several parameters of CF-based recommender systems that need to be set in order to generate recommendations. These parameters include the number of people in the neighbor group of a user (reference group size or neighborhood size), the required minimum number of common items to be included in the reference group (significance level), etc. Most current CF-based recommender systems identify a

* This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government (MEST) (SSK Grant B00080).

set of optimal settings for recommendation and apply the settings to the task of generating recommendations for all users (Breese et al., 1998; Herlocker et al., 1999; Cho and Bang, 2011).

This paper explores the impact of personalized settings for each user on the performance of recommender systems. As settings become more fine-tuned for each user, the recommendations are more likely to be accurate. In this paper, the details of the new method (i.e., personalized setting collaborative filtering, or PS-CF) are described, and the new method is compared with traditional methods.

2. CF-Based Recommender Systems

One stream of past research on CF algorithms has focused on improving recommendations. Examples of those efforts are improving similarity measures (Burke, 2000), replacing null ratings (blanks) with proper estimations, and adding implicit ratings from users (Lee et al., 2007). Another stream of research has focused on combining the CF algorithm with other recommendation methods such as content-based recommendation (Shahabi and Chen, 2003), agent-based recommendation (Riedl, 1999), or meta-recommender systems (Schafer et al., 2002).

Most of these studies, however, applied the same system settings to all users. Settings include reference group size (or neighborhood size), significance level, and other things that can be changed. The underlying assumption of applying the same settings is that only one optimal setting exists for all users.

In previous studies, it was shown that the optimal settings for CF vary depending on the domain in which the recommendations are generated, the characteristics of users, and occasions of system use (Im and Hars, 2007; Konstan et al., 1997). It is likely that users have different optimal settings because their rating and preference patterns differ from one another.

2.1 Settings for CF-based Recommender Systems

There are numerous parameters that can be adjusted for CF-based recommendations. Theoretically there can be as many possible parameters as the number of available information categories in the system. Since a Netflix dataset was used for this study, the settings discussed and investigated in this study are limited to the type of data available from the Netflix database.

The most significant parameters in the Netflix dataset are reference group size, confidence level, and rating timing. Although only these three parameters are discussed in this paper, other parameters can be personalized following similar logic to that proposed in this paper, if available in the dataset.

2.1.1 Reference Group Size

Reference group size or neighborhood size refers to the number of users whose ratings are used to generate recommendations for the active user (Im and Hars, 2007). There are two methods to set reference group size – ‘best-N neighbors’ and ‘thresholding’ (Im and Hars, 2007; Breese et al., 1998; Herlocker et al.,1999). The best-N neighbors method takes the top N (a pre-determined number) neighbors as a reference group, those who have the highest correlations with the active user. The thresholding method selects as a reference group users who have correlations higher than a certain pre-set value. Generally, the best-N neighbors method has higher coverage, while thresholding is more accurate (Im and Hars, 2007; Herlocker et al., 1999).

Previous studies examined the effect of reference group size. For example, Herlocker et al. (1999) found that there exists an optimal reference group size in a given circumstance. Another study found that optimal reference group size varies across domains (Im and Hars, 2007), which implies that the differences in the characteristics of these data may cause differences in optimal reference group size.

It has been shown that users' rating patterns are quite diverse (Bell et al., 2007). Because of this diversity, it is very likely that users have different optimal reference group sizes. For example, users who have unique tastes about movies, such as cult movie lovers, will have only a small number of people who share this common interest. For those users, a small reference group will lead to better results. On the other hand, for those users who have more common tastes, a larger reference group will be preferable because a larger reference group will be less biased than smaller groups. If the optimal reference group size for each user can be identified, more accurate recommendations can be generated.

2.1.2 Weights for Significance Level

One issue in determining who should be included in a reference group is the level of significance. Level of significance refers to the number of co-rated items on which the similarity measure of two users is based (Herlocker et al., 1999). For example, suppose the similarity of two users was calculated using ten movies that they both rated. Then, this similarity measure would be more reliable (would have a higher significance level) than the similarity of two users who rated only three movies in common.

CF-based recommender systems usually set a weight for significance level, and apply it to all users (Herlocker et al., 1999). It is probable, however, that the effect of significance level may vary across users. For example, a higher significance level would be better for users with volatile rating patterns because making predictions with a small number of co-rated items would be inaccurate for these users. On the other hand, for users having stable or plain rating patterns, a low significance level would be fine because their ratings are more predictable with a small number of co-rated items. Therefore, if the optimal confidence level (or optimal weight for it, if weights are applied instead of a cut-off point) can be obtained and applied for each user, it would improve the accuracy of recommendations.

2.1.3 Weights for Rating Timing

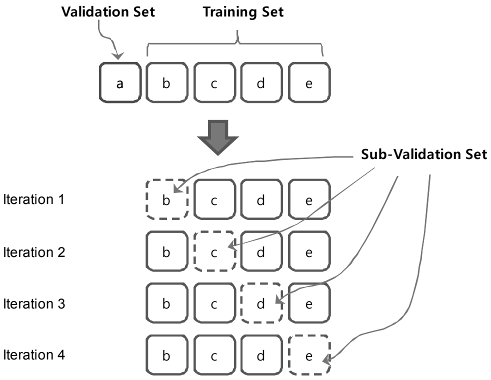
Users' ratings may be affected by timing. Users who watch and rate a movie right after its release rate it differently from those users who rate the movie long after its release. This is because movies just released have the 'new movie' effect, which includes the influence of the ads and the buzz around them. This effect will affect users' ratings. Therefore, it is desirable to weight ratings according to timing.

There is another issue related to rating timing. Users may have different levels of sensitivity about rating timing. For example, some users may be affected greatly by the 'new movie' effect, while other users may not. Thus, it is also expected that each user has different optimal weight levels for rating timing. The weight for rating timing is also personalized–i.e., the optimal weight for rating timing is identified for each user.

2.2 Identifying Personalized Settings

The most common method to calculate the accuracy of recommender systems is simulation (Breese et al., 1998; Herlocker et al., 1999; Christakou and Stafylopatis, 2005). The dataset is divided into two subsets-'training' and 'validation.' The training set is used to generate recommendations (calculate estimated ratings) for the validation set. The accuracy of the recommendations is measured by calculating how close the recommendations are to the real ratings in the validation set.

The size of the training and validation sets can vary. Most often there is only one item in the validation set, which is called 'all but one' (Herlocker et al., 1999). In our simulations, the 'all but one' method was used.



<Figure 1> Training and Validation Datasets

For ordinary CF, a movie was selected from each user as the validation set. Different reference group sizes (1 to 100) were applied to all users and the one that resulted in the best accuracy was set as the official 'reference group size.'

There is one more step to identify personal settings. The training dataset is divided again into two－the sub-training and sub-validation sets. Suppose a user has rated five movies－a, b, c, d, and e－as shown in <Figure 1>. Movie a is selected as the validation set and the other four movies are selected as the training set. The training set is divided again into sub-training and sub-validation sets. For example, in iteration 1 in <Figure 1>, movie b is selected as the sub-validation set and the other three movies become the training set. The sub-training set (movies c, d, and e) is used to calculate estimated ratings for the sub-validation set (movie b). When reference group size is being personalized, the rating for movie b is estimated using the data in the sub-training set (movies c, d, and e) with different reference group sizes. The one producing the highest accuracy is recorded as the optimal reference group size for iteration 1. Then, the same procedure is repeated with different sub-validation sets. For example, in iteration 2, movie c is selected as the sub-validation set and the other three movies are set as the sub-training set. The average of all optimal reference group sizes from the different iterations is set as the final personalized reference group size for the user. The final personalized reference group size is then used to calculate predictions for the validation set (movie a in <Figure 1>) for the user.

For the validation set, all rated movies for each user were selected one at a time and simulations were run. For example, if a user has rated 20 movies, 20 iterations of the simulation were carried out and the accuracy measures were averaged.

When determining personalized settings, running iterations with all movies in the training set was too time-consuming. Therefore, 10 movies (or the total number of movies if there were less than 10 movies in the training set) were randomly selected one at a time from the training dataset as the sub-validation set. For each movie in the sub-validation set, different sizes were applied to determine the best personalized reference group size. In this study, group sizes of 1 to 100 were applied and the reference group with the highest accuracy was selected as optimal.

3. Evaluation

The data used to evaluate the new method in this study is from the Netflix contest (http://www.netflixprize.com/). The dataset contains ratings from a total of 480,189 users on 17,770 movies. The total number of ratings is 100,480,507 and the number of ratings per user is 209.3.

3.1 Dataset

Since the original Netflix dataset is too big to be handled on ordinary PCs, a sampled sub-dataset was used in this study. A total of ten sub-datasets (5,000 users each) were randomly sampled. In order to check if those sample datasets were biased, basic statistics were compared. As <Table 1> shows, the basic statistics of the sampled datasets are compatible with the original dataset in terms of number of ratings and other characteristics.

<Table 1> Comparison of the Original Netflix Dataset and Sample Datasets

	Full Data	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Sample 6	Sample 7	Sample 8	Sample 9	Sample 10
Total # of Movies	17,770	16,164	15,926	16,443	16,460	17,700	16,244	16,743	15,980	16,026	16,623
Total # of Ratings	100,480,507	1,082,624	1,050,965	1,032,734	1,074,120	1,056,430	1,034,707	1,073,998	1,063,262	1,026,716	1,073,348
Total # of Users	480,189	5,000	5,000	5,000	5,000	5,000	5,000	5,000	5,000	5,000	5,000
# of Ratings/User	209.3	216.5	210.2	206.5	214.8	211.3	206.9	214.8	212.7	205.3	214.7
AVG (Ratings)	3.60	3.62	3.61	3.61	3.61	3.59	3.61	3.59	3.63	3.62	3.58
AVG (Ratings)/User	3.67	3.67	3.67	3.67	3.68	3.68	3.68	3.68	3.68	3.67	3.67
AVG (Ratings)/Movie	3.23	3.32	3.31	3.27	3.29	2.90	3.33	3.20	3.37	3.30	3.11
STDEV(Ratings)	1.09	1.08	1.08	1.08	1.08	1.10	1.08	1.09	1.08	1.08	1.09

3.2 Algorithms

The ordinary CF in this study employed a common CF method. Pearson correlation was used as the similarity measure. For better accuracy, the 'bias-from-mean' adjustment was employed. In the 'bias-from-mean' adjustment, ratings are converted to 'deviations from mean' before being used for predictions, and the generated predictions are re-converted to original rating scales (Herlocker et al., 1999).

3.3 Simulation Results

Due to time constraints, we present only the results about personalized reference group sizes and level of confidence in this paper. The results are summarized in <Table 2>. For each sampled dataset, the accuracy of PS-CF was compared against that of ordinary CF. MAE (mean absolute error) and RMSE (root mean squared error) were used as the accuracy measures in this paper.

In the ordinary CF, reference group sizes of 1 to 100 were applied and the one showing the best result was selected as the final reference group size. For example, in Sample 1 in Table 3, the reference group size of 15 shows the best result (MAE = 0.741, RMSE = 0.897).

The MAE and RMSE of PS-CF for Sample 1 are 0.485 and 0.698, respectively. The average personalized reference group size for sample users is 21.2.

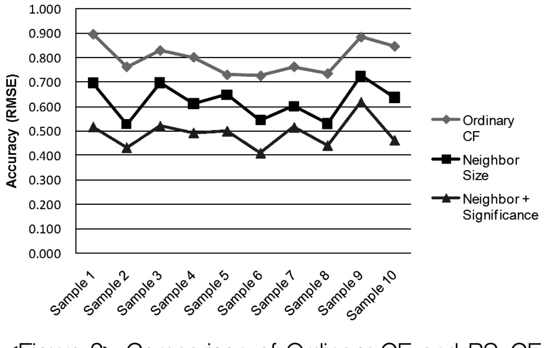
<Table 2> Simulation Results

	Accuracy					
	MAE			RMSE		
	Ordinary CF	Neighbor Size	Neighbor + Significance	Ordinary CF	Neighbor Size	Neighbor + Significance
Sample 1	0.741	0.485	0.310	0.897	0.698	0.518
Sample 2	0.591	0.338	0.228	0.763	0.528	0.433
Sample 3	0.667	0.460	0.282	0.830	0.698	0.521
Sample 4	0.607	0.399	0.262	0.802	0.612	0.493
Sample 5	0.606	0.388	0.256	0.730	0.649	0.501
Sample 6	0.563	0.357	0.228	0.726	0.544	0.410
Sample 7	0.574	0.376	0.248	0.763	0.602	0.516
Sample 8	0.584	0.347	0.239	0.735	0.530	0.442
Sample 9	0.705	0.485	0.323	0.885	0.726	0.618
Sample 10	0.665	0.407	0.238	0.847	0.637	0.463
Average	0.630	0.404	0.261	0.798	0.622	0.491
Improvements		35.9%	58.5%		22.0%	38.4%

<Table 3> Optimal Settings

	Optimal Neighbor Size		Avg. Cut-off Point (Significance Level)
	Ordinary CF	PS-CF (Average)	
Sample 1	15	21.2	23.6
Sample 2	15	23.9	26.6
Sample 3	65	19.5	25.0
Sample 4	50	26.1	28.2
Sample 5	35	23.1	20.6
Sample 6	25	26.5	21.8
Sample 7	20	20.6	27.8
Sample 8	5	22.9	23.6
Sample 9	35	18.1	32.9
Sample 10	20	23.0	26.3
Average	28.5	22.5	25.6

The accuracies of ordinary CF and PS-CF are compared in <Figure 2>. As shown in <Table 2> and <Figure 2>, the PS-CF method consistently outperforms the traditional method, as shown in the diagram. When both neighbor size and level of significance are personalized, the accuracy of the PS-CF method is 38.4% (RMSE) or 58.5% (MAE) better than ordinary CF.



<Figure 2> Comparison of Ordinary CF and PS-CF

The coverage of the PS-CF method is also compared with ordinary CF method in <Table 4>. The coverage of PS-CF is slightly lower than ordinary CF, but difference is not big.

<Table 4> Coverages

	Ordinary CF	Neighbor Size	Neighbor+ Significance
Sample 1	99%	98%	98%
Sample 2	98%	95%	95%
Sample 3	99%	97%	96%
Sample 4	100%	98%	98%
Sample 5	100%	98%	98%
Sample 6	98%	95%	95%
Sample 7	100%	99%	99%
Sample 8	99%	97%	97%
Sample 9	97%	96%	96%
Sample 10	99%	95%	95%
Average	98.9%	96.8%	96.7

For a more rigorous comparison, a nonparametric statistical analysis was conducted for the RMSE measures. Since the results from the ordinary CF and PS-CF are paired by samples, the Wilcoxon signed-rank test was employed. The z statistics are significant at 0.01 level for both (ordinary CF- PS-CF with personalized neighbor size) and (PS-CF with personalized neighbor size - PS-CF with personalized neighbor size and personalized significance level) pairs.

4. Conclusions

This paper presents PS-CF, a new method for CF-based recommender systems which utilizes personalized settings. Although only two parameters, neighbor size and level of confidence, were personalized, our test shows that the PS-CF method outperforms the traditional method that applies a global setting to all users. It is expected that the accuracy of CF will improve further if other parameters are personalized and PS-CF is combined with other methods.

This study has several limitations due to the dataset and methodology used. First, the PS-CF method was tested using only one dataset, Netflix. Although Netflix is a commonly used dataset, the method needs to be tested using other datasets. Second, the PS-CF method can only be applied when users have rated a substantial number of items because it requires additional data to determine personalized settings. Third, the PS-CF method requires intensive computing compared to ordinary CF because additional simulations need to be run to determine personalized settings. More research is needed to mitigate the computing requirements.

References

Bell, R., Koren, Y., and Volinsky, C., "Chasing $1,000,000 : How we won the Neflix progressive prize", *Statistical Computing and Graphics*, Vol.18, No.2(2007), 4~12.

Breese, J. S., Heckerman, D., and Kadie, C., Empirical analysis of predictive algorithms for collaborative filtering, in Fourteenth Conference on Uncertainty in Artificial Intelligence. Madison, WI, 1998.

Burke, R., Semantic ratings and heuristic similarity for collaborative filtering, in AAAI Technical Report WS-00-04, (2000), 14~20.

Cho, Y. and Bang, J., "Applying centrality analysis to solve the cold-start and sparsity problems in collaborative filtering", *Journal of Intelligence and Information Systems*, Vol.17, No.3(2011), 99~114.

Christakou, C. and Stafylopatis, A., A hybrid movie recommender system based on neural networks, in International Conference on Intelligent Systems Design and Applications (ISDA'05). IEEE : Wroclaw, Poland, 2005.

Herlocker, J. L., Konstan, J. A., Borchers, A., and Riedl, J., An algorithmic framework for performing collaborative filtering, in Conference on Research and Development in Information Retrieval. ACM Press : New York, NY, 1999.

Im, I. and Hars, A., "Does a one-size recommendation system fit all? : The effectiveness of collaborative filtering based recommendation systems across different domains and search modes", *ACM Transactions on Information Systems*, Vol.26, No.1(2007).

Konstan, J. A., Miller, B. N., Maltz, D., Herlocker, J. L., Good, N., and Riedl, J., "GroupLens : Applying collaborative filtering to Usenet news", *Communications of the ACM*, Vol.40, No.3(1997), 77~87.

Lee, T. Q., Park, Y., and Park, Y.-T., A similarity measure for collaborative filtering with implicit feedback, in ICIC, D.-S. Huang, L. Heutte, and Loog M. Springer-Verlag : Qingdao, China, (2007), 385~397.

Park, D. H., Kim, H. K., Choi, I. Y., and Kim, J. K., "A literature review and classification of recommender systems on academic journals", *Journal of Intelligence and Information Systems*, Vol.17, No.1(2011), 139~152.

Riedl, J., Combining collaborative filtering with personal agents for better recommendations, in Conference of the American Association of Artificial Intelligence. Orlando, FL, 1999.

Schafer, J. B., Konstan, J. A., and Riedl, J., Meta-recommendation systems : User-controlled integration of diverse recommendations, in CIKM'02. ACM : McLean, Virginia, USA, 2002.

Shahabi, C. and Chen, Y.-S., "An adaptive recommendation system without explicit acquisition of user relevance feedback", *Distributed and Parallel Databases*, Vol.14(2003), 173~192.

Xiao, B. and Benbasat, I., "E-commerce product recommendation agents : Use, characteristics, and impact", *MIS Quarterly*, Vol.31, No.1 (2007), 137~209.

Abstract

CF 기반 추천시스템에서 개인화된 세팅의 효과

임 일[*] · 김병호[**]

본 논문에서는 협업필터링(collaborative filtering : CF) 기반한 추천시스템의 정확도를 높일 수 있는 방법을 제안하고 그 효과를 분석한다. 일반적인 CF기반 추천시스템에서는 시스템 세팅(참조집단 크기, 유의도 수준 등)을 한 가지 정해서 모든 경우에 대해서 동일하게 적용한다. 본 논문에서는 개별 사용자의 특성에 따라 이러한 세팅을 최적화 해서 개별적으로 적용하는 방법을 개발하였다. 이런 개인화된 세팅의 효과를 측정하기 위해서 Netflix의 자료를 사용해서 일반적인 추천시스템과 추천 정확도를 비교하였다. 분석 결과, 동일한 세팅을 적용하는 일반적인 추천시스템에 비해서 개인화된 세팅을 적용한 경우 정확도가 월등히 향상됨을 확인하였다. 이 결과의 시사점과 함께 미래 연구의 방향에 대해서도 논의한다.

Keywords : Collaborative filtering, Personalization, Netflix

[*] 연세대학교 경영대학
[**] 연세대학교 경영대학

저자소개

Il Im

Il Im is an Associate Professor of Information Systems at School of Business, Yonsei University. He received his Ph.D. from Marshall School of Business, University of Southern California. Prior to joining Yonsei University, Dr. Im was an Assistant Professor in the Information Systems Department at New Jersey Institute of Technology. His current research focuses on personalization technologies and their impacts, technology acceptance, and electronic commerce.

Byung Ho Kim

Byung Ho Kim is a research fellow at the Yonsei Business Research Institute (YBRI). He received his BS in Computer Science and Industrial System Engineering from Yonsei School of Engineering and MS in Business from Yonsei School of Business. He has industry experiences for over five years in various areas such as system development, database, and website building.