Handling Incomplete Data Problem in Collaborative Filtering System

Hyunju Noh[a], Minjung Kwak[b] and Ingoo Han[c]

[a] Graduate School of Management, Korea Advanced Institute of Science and Technology
[b] Department of Information Statistics, Pyongtaek University
[c] Graduate School of Management, Korea Advanced Institute of Science and Technology
([a] *hjnoh@kgsm.kaist.ac.kr*, [b] *mjkwak@ptuniv.ac.kr*, [c] *ighan@kgsm.kaist.ac.kr*)

Collaborative filtering is one of the methodologies that are most widely used for recommendation system. It is based on a data matrix of each customer's preferences of products. There could be a lot of missing values in such preference data matrix. This incomplete data is one of the reasons to deteriorate the accuracy of recommendation system. There are several treatments to deal with the incomplete data problem such as case deletion and single imputation. Those approaches are simple and easy to implement but they may provide biased results. Multiple imputation method imputes m values for each missing value. It overcomes flaws of single imputation approaches through considering the uncertainty of missing values. The objective of this paper is to suggest multiple imputation-based collaborative filtering approach for recommendation system to improve the accuracy in prediction performance. The experimental works show that the proposed approach provides better performance than the traditional Collaborative filtering approach, especially in case that there are a lot of missing values in dataset used for recommendation system.

Key words: Multiple Imputation, Collaborative Filtering, Incomplete data

Received: October 2003 Accepted: November 2003 Corresponding Author: Hyun Ju Noh

1. Introduction

The movement toward E-commerce has been rapidly grown up as Internet becomes increasingly popular. Companies have been collecting and providing a lot of product information to meet the various needs of different customers as a means of surviving in the new business environments. However such situation has also brought out the information overload problem. Recommendation system has a function to offer a personalized service to a customer by recommending products that are likely to meet his/her needs. Collaborative filtering (CF) algorithm is one of the most widely used methodologies for making personalized recommendation. It is based on the assumption that a good way to find interesting products for a target customer is to find other customers who have similar preference patterns, and then recommend products that those similar customers like. An important task in CF-based recommendation system is to calculate the similarity between customers so that it is used to select neighborhoods similar to a target customer in terms of product preference. However, there is a fundamental data problem in using CF algorithm, incomplete data. CF is based on a data matrix of each customer's preferences of products. Most of customers may not rate all products because they do not have any experience about some products or they do not want to provide their preferences. This incomplete data is one of the reasons of deteriorating the accuracy of the recommendation system ([1]). There are several treatments to deal with the incomplete data problem such as case deletion and single imputation. Case deletion uses cases with available information in an analysis. Single imputation replaces a single value with each missing data. These approaches are simple and easy to implement but they may provide biased results. Single imputation does not reflect the uncertainty about the predictions of unknown missing values, and variances of parameter estimates is underestimated ([7]). Multiple imputation (MI) method imputes m values for each missing value so m complete dataset can be made. A following analysis task is based on those m complete dataset. MI considers the uncertainty about the prediction to impute using m imputed data and helps to attain exact confidence interval to estimate a parameter by adjusting the underestimation of variability in single imputation approaches ([9]).

The objective of this paper is to suggest Multiple Imputation-based Collaborative Filtering (MICF) approach for recommendation system. It solves the incomplete data problem for computing similarities of CF algorithm and helps to improve the prediction performance of recommendation system. The rest of this paper is organized as follows. Section 2 provides a brief overview of collaborative filtering algorithm and the incomplete data problem in CF algorithm. Section 3 explains the proposed MICF approach in detail. Section 4 shows the experimental works that are implemented on a rating dataset of movie preferences. The final section provides some concluding remarks and directions for future research.

2. Collaborative Filtering and Incomplete Data Problem

Collaborative Filtering Algorithm

Collaborative filtering approach in recommender system recommends products to a target customer through considering of other customers' experience and opinions. It is used in most successful recommender systems such as Tapestry ([4]), Ringo ([10]) and so on. CF-based recommendations have been also begun to use in several web sites including book, music, movie and information. In the early stage of CF-based recommender systems, it usually required the explicit customers' votes to express their preference about the products. More recently a number of systems have begun to infer user preferences from actions rather than requiring the user to explicitly rate a product.

The CF algorithm is generally composed of 3 steps. Step 1 is to weigh all users with respect to similarity with the target customer. There are several similarity weighting measures. The most common weighting measure used is the Pearson correlation coefficient. Other similarity measures are the vector similarity cosine measure, the entropy-based uncertainty measure, the mean-squared difference algorithm and so on. However, such measures have been known that they do not perform well compared to Pearson correlation coefficient in CF-based recommendation system ([2], [5]). Step 2 is to find a neighborhood of customers with similar preferences to use their preference in the computation of a prediction for the target customer. The systems must select the best neighbors because most collaborative filtering systems should handle a lot of customers, so considering of every neighbor is infeasible. It is useful, both for accuracy and performance, to select the neighborhood to use in computing a prediction and discarding the remaining customers' opinions. Step 3 is to combine the rating from neighbors for computing a prediction of the target customer' preference about a product. For that purpose, it is useful to use the similarity measure obtaining from step 1. That is, a weighted average of the ratings is calculated using the Pearson correlation coefficients as the weights.

The similarity between the target customer and neighbors could be measured as follows ([5]). The similarity weight, $W(a,i)$ between the target customer a and neighbor i as defined by the Pearson correlation coefficient:

$$W(a,i) = \frac{\sum (v_{a,j} - \bar{v}_a)(v_{i,j} - \bar{v}_i)}{\sqrt{\sum_j (v_{a,j} - \bar{v}_a)^2 \sum_j (v_{i,j} - \bar{v}_i)^2}}$$

where $v_{a,j}$ means the preference of the target customer a to item j and $\bar{v}_a$ is the preference mean of items rated by customer a. A final prediction, $P_{a,j}$, by performing a weighted average of deviations from the neighbor's mean;

$$p_{a,j} = \bar{v}_a + \frac{\sum_{i=1}^{n} w(a,i)(v_{i,j} - \bar{v}_i)}{\sum_{i=1}^{n} |w(a,i)|}$$

$P_{a,j}$ means the prediction for the target customer 'a' for item 'j'. n is the number of neighbors.

Incomplete data problem in Collaborative Filtering System

Collaborative filtering algorithm calculates the similar preferences through correlations among customers' profiles. Correlation technique is based on the pairwise deletion approach in handling of missing values. That is, correlation between two customers' profiles can only be computed based on items that both users have rated. If correlation coefficients are computed using just a few overlapping observations among customers, it may be inaccurate similarity estimates ([1]). Such a loss of data can reduce the power of a test and the statistically inferred results may incur bias issues. Table 1 shows a simplified example of an incomplete data matrix, where each cell represents a customer's preference about a specific product, movie in this example.

Table 1. Example of customer's preferences about movies

	Titanic	Love Letter	Ghost	Load of the Rings	Minority Report	Chicago	Diehard	Rambo	007
Customer X	6	5	4	4	5	6			
Customer Y	1	2	3				3	2	1
Customer Z	???			4	5	6	3	2	1

Note: each cell represents a customer' preference value about a specific movie on a six-point scale from the worst (1) to the best (6). '???' is the target preference to be predicted by CF algorithm

Assume each customer rates 6 movies among 9 as in Table 1. Customer X does not rate action movies, 'Die hard', 'Rambo' and '007.' Customer Y does not answer his/her preferences on 'Load of the rings', 'Minority Report' and 'Chicago' because he/she does not watch those recent movies. Customer Z (a target) does not answer his/her preferences on 'Titanic', 'Lover letter' and 'Ghost'. In order to determine whether an online movie site should recommend 'Titanic' to customer Z or not, the collaborative filtering algorithm considers the correlations of preferences between target customer and other customers, X and Y. Under this incomplete dataset, both of customers X and Z commonly rate 3 movies, 'Load of the rings', 'Minority Report' and 'Chicago'. The movie preferences of two customers seem to be entirely consistent based on those 3 movies, that is, the correlation between X and Z, *Corr (X, Z) = 1*. The movie preferences of two customers Y and Z also show same result, *Corr(Y, Z)= 1*, based on the rating values of 3 movies answered by both of them. Therefore, the recommendation system based on collaborative filtering algorithm may determine both customers X and Y are the nearest neighbors who have same preferences with customer Z and use their preferences to calculate the target's rating. However, the two neighbors show the opposite pattern to the target movie, 'Titanic'. Although *Corr (X,Z) = Corr (Y,Z) =1, Corr(X,Y) = -1* not *1*. As a result, this correlation-based analysis with incomplete data generates significantly biased results.

Methodologies for handling incomplete data: Single imputation

There are some treatments to deal with the incomplete data problem. The most simplest way is case deletion, or list deletion, that deletes the case with missing values and this approach is widely used in most of analytical procedures. However it can lead to huge amount of data loss. It may provide biased results if there are numerous missing values because the complete data matrix after deleting the case with missing values is often not representative of the real population. Another simple method is single imputation approach that fills each missing value with a predicted or simulated value. Mean imputation, hot-deck, and regression approaches are representative single imputation methodologies that have been popularly used up to now. Mean substitution is an easy way to replace missing values with mean of observations. However it attenuates the relationship of variables because of altering marginal distribution. Hot-deck approach is to replace a missing value with one of another observations matched on all other characteristics with that missing case. It may preserve marginal distribution but distort correlations. Regression approach is to replace the missing values with the predicted one using regression. It may inflate correlation and is difficult to apply to multivariate data when more than one variable have missing values. Another single imputation method is maximum likelihood approach using EM (expectation maximization) algorithm. It has been used for many missing data problems and well known that the estimated parameters are mostly consistent and efficient ([7]).

Those single imputation methods produce one plausible value for each missing cell to make complete data analysis instead of deleting cases. However, those approaches do not consider the uncertainty about the predictions of the missing values and result in underestimation of variance ([9]).

Methodologies for handling incomplete data: Multiple imputation

Multiple imputation (MI) is a simulation-based approach. The idea of multiple imputation is that each missing value is replaced with *m* >1 plausible values from their predictive distribution and converts an incomplete dataset into *m* complete dataset like Figure 1. MI preserves the merits of single imputation and corrects its major flaws through considering the uncertainty of missing values. MI produces *m* complete datasets and each dataset is used for analysis. Overall estimate is obtained by combining these *m* estimates. MI approach is known that it can be highly efficient even for small values of *m*.

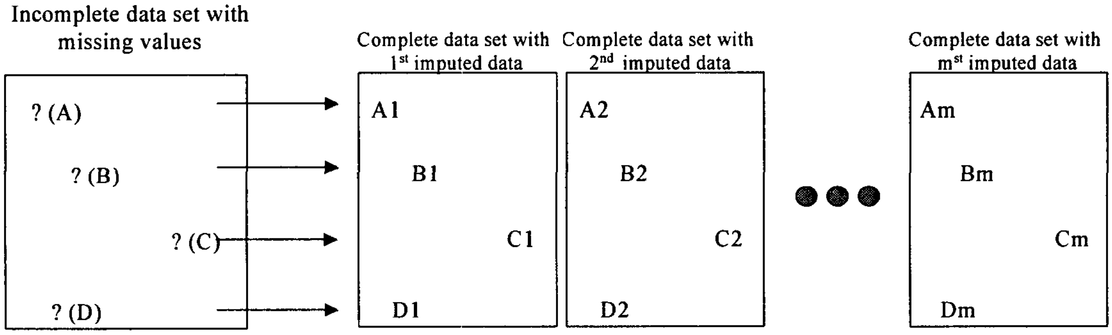
Incomplete data set with missing values
? (A)
? (B)
? (C)
? (D)
Complete data set with 1st imputed data
A1
B1
C1
D1
Complete data set with 2nd imputed data
A2
B2
C2
D2
Complete data set with mst imputed data
Am
Bm
Cm
Dm

Figure 1. Multiple Imputation Process

Multiple imputation method is implemented under the assumption that missing data are missing at random (MAR) in the sense that the missing probability of an observation may depend on observed values but not missing ones. Under MAR assumption, MI is generated from certain data distribution with prior distribution of model parameters, which is a parametric Bayesian approach. As a prior distribution for parameter, both non-informative and informative priors can be selected.

Suppose that $Y=(Y_{abs}, Y_{mis})$ follows a parametric model $P(Y|\theta)$ where θ has a prior distribution and Y_{mis} is MAR. Then the distribution of missing is defined as follows:

$$P(Y_{mis}|Y_{obs}) = \int P(Y_{mis}|Y_{obs},\theta)P(\theta|Y_{obs})d\theta \tag{1}$$

Multivariate normal assumption for data is used to generate the imputations for missing values. It is known that MI is not sensitive to departures from the distribution assumption ([8]).

3. Imputation-based Collaborative Filtering

As it is described in the above section, traditional collaborative filtering algorithm based on correlation technique may result in inaccurate prediction of product preference under incomplete dataset. This study proposes multiple imputation-based CF approach to solve this incomplete data problem of collaborative filtering algorithm. Single imputation-based CF approach is proposed for the comparative study, too.

Single Imputation-based Collaborative Filtering

There are several single imputation approaches. We adopt the EM algorithm for single imputation-based collaborative filtering (SICF). EM is a general imputation technique through finding maximum likelihood estimates and known relatively consistent and efficient than other single imputation approaches ([7]). There are two stages in SICF approach. First, incomplete preference dataset is converted into a complete dataset through imputation module using EM algorithm. Then CF algorithm is applied to the complete data for prediction of target customers' preferences.

Multiple Imputation-based Collaborative Filtering

Multiple imputation-based collaborative filtering (MICF) consists of 3 modules as in Figure 2: multiple imputation, multiple collaborative filtering, and combining module.

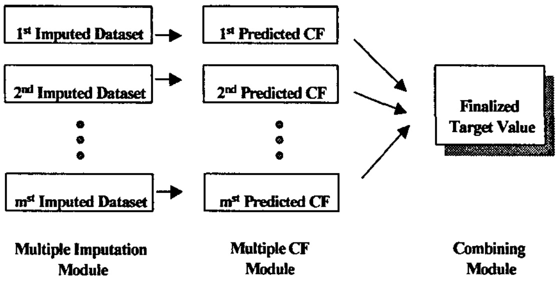

Figure 2. MICF process

Multiple imputation stage includes the following steps.

[Step 1: Initialization] Use single imputation approach through Expectation Maximization to generate initial values for $Y_{mis}^{(0)}$ and $\theta^{(0)}$. In most cases, non-informative prior works well but this study uses the ridge prior for parameter by adding small amount to variance matrix for avoiding singularity in case of high missing rate.

[Step 2: Generation of Distribution] Generate missing values from $Y_{mis}^{(t+1)} \sim P(Y_{mis} \mid Y_{obs}, \theta^{(t)})$.

[Step 3: Generation of Parameters] Draw unknown parameters from $\theta^{(t+1)} \sim P(\theta \mid Y_{obs}, Y_{mis}^{(t+1)})$.

[Step 4: Stabilization of Distribution] Repeat step 2 and step 3 to create the following Markov Chain. $Y_{mis}^{(1)}, \theta^{(1)}, Y_{mis}^{(2)}, \theta^{(2)}, \Lambda\ , Y_{mis}^{(t)}, \theta^{(t)}, \Lambda\ \Lambda$

After a sufficient number of steps to stabilize the distribution, $P(Y_{mis}, \theta \mid Y_{obs})$ imputations can be drawn from a Markov chain at every *k* cycle.

Next step is the multiple collaborative filtering stage. After making *m* complete dataset through multiple imputation stage, *m* complete sets are analyzed by CF.. After performing the same CF analysis on *m* imputed datasets, we have *m* equally plausible correlation estimates $P_{a,j}^{(1)}, P_{a,j}^{(2)}, \Lambda\ , P_{a,j}^{(m)}$ for customer *a* on item *j*.

Finally, the MICF approach implements combining module to make a prediction from *m* prediction values. In this study, the MICF estimate is given by $P_{a,j} = \sum_{i=1}^{m} P_{a,j}^{(i)} / m$ in combining module.

4. Experiments

In this section we present experimental results of the proposed algorithms, SICF and MICF, and compare the performance of prediction to traditional collaborative filtering methodology.

Dataset

The initial dataset used for this experiment is EachMovie data available on Internet ([6]). The EachMovie data is originated from a research project at the System Research Center of Digital Equipment Corporation. EachMovie data is in the form of a sparse matrix whose rows mean users, columns mean items (movies), and elements of the matrix are users" preferences to the corresponding movies.

The prediction performance of traditional collaborative filtering methodology and proposed SICF and MICF approach could show different effects according to the different missing rates. This study is implemented on three types of datasets for low to high missing rates of 25%, 50% and 75%. We randomly select 30 persons and extract three sets of 100 movies regarding to the level of missing rate for making training sets proper to three levels of missing rates. Those training sets are used for searching neighbors to predict target customer's preferences. Then, we randomly select other 5 persons for making test set and select 50 observations from target customers to be predicted. There are 6 points rating values from 0 (worst) to 5 (best) in those datasets.

Measurement Criteria

Mean absolute error (MAE) and variance of mean absolute error are used to compare the prediction performance. The accuracy of the system is high as MAE of a prediction system is low. The variance of mean absolute error should be also minimized for well-performed prediction system. In addition to the above measurements, we consider availability of prediction" that means how many prediction values can be provided by the proposed approach. If there are many missing values in the dataset, the prediction algorithm may not provide estimates to the targets and it could not be useful for recommendation system.

Experimental Results

Most researches dealing with incomplete data problem focus merely on comparing imputation methods themselves on the belief that if the method to create imputations is proper, then the resulting inferences will be statistically valid ([7]). However, imputation method is based on the model assumption so it is not guaranteed that imputed data analysis performs always better than standard complete data analysis. Hence, this study compares the improvement on prediction performance among the suggested MICF (CF after multiple imputations), SICF (CF after single imputation), and traditional CF according to the missing rates.

We vary the number of multiple imputations for comparing the efficiency of MI as 3, 5, 7 and 10. The experimental results for traditional CF, SICF and MICF approaches are described in table 2 ~ 4. MICF_3, MICF_5, MICF_7 and MICF_10 imply the MICF approaches based on imputed values of 3, 5, 7, and 10 respectively. Lastly, we make another test set with higher missing rate than that of training set. We consider that a shopping mall wants to recommend some products to its new customers using the preferences of existing customers. However, new customers do not have enough experience of products so they may have higher missing rates than the existing customers. Table 5 shows the experimental results in case of test set with missing rate 80% and missing rate of training set 75%.

Table 2. Experimental results in case of dataset with missing rate = 25%

	CF	SICF	MICF_3	MICF_5	MICF_7	MICF_10
MAE	1.76	1.74	1.74	1.72	1.72	1.74
% change		1.1% ·	1.1%	2.3%	2.3%	1.1%
VAR	1.53	1.38	1.46	1.43	1.51	1.46

CF: traditional CF using incomplete data, SICF: CF after single imputation, MICF_*m*: CF after multiple imputation, where *m* =3, 5, 7, and 10

Table 3. Experimental results in case of dataset with missing rate = 50%

	CF	SICF	MICF_3	MICF_5	MICF_7	MICF_10
MAE	1.52	1.5	1.46	1.46	1.44	1.46
% change		1.3%	3.9%	3.9%	5.3%	3.95
VAR	1.32	1.56	1.36	1.4	1.35	1.31

CF: traditional CF using incomplete data, SICF: CF after single imputation, MICF_*m*: CF after multiple imputation, where *m* =3, 5, 7, and 10

Table 4. Experimental results in case of dataset with missing rate = 75%

	CF	SICF	MICF_3	MICF_5	MICF_7	MICF_10
MAE	1.64	1.51	1.46	1.46	1.43	1.43
% change		7.7%	11.1%	11.1%	12.9%	12.9%
VAR	1.24	1.79	1.14	1.26	1.13	1.13

CF: traditional CF using incomplete data, SICF: CF after single imputation, MICF_*m*: CF after multiple imputation, where *m* =3, 5, 7, and 10

Table 5. Experimental results in case of test set with missing rate = 80%

	CF	SICF	MICF_3	MICF_5	MICF_7	MICF_10
MAE	1.85	1.37	1.37	1.36	1.31	1.31
% change		25.9%	25.9%	26.6%	29.0%	29.0%
VAR	1.5	1.08	1.31	1.42	1.29	1.35

CF: traditional CF using incomplete data, SICF: CF after single imputation, MICF_*m*: CF after multiple imputation, where *m* =3, 5, 7, and 10

The experimental results are summarized as follows. First of all, imputation-based CF approaches show better performance than traditional CF approach with incomplete dataset. Overall, MICF approaches outperform SICF approach in prediction performance but the difference is not significantly large. Secondly, when the missing rate is high, the gain is high. As table 4 describes, the improvement in prediction performance of MICF is up to 12.9% compared to the traditional CF approach when missing rate is 75 %. As table 5 shows, the improvement of MICF is up to 29% in comparison with traditional CF approach in case of missing rate 80% of test set. The MAE of these approaches by missing rate is depicted in Figure 3. Thirdly, as the missing rate increases, the number of imputations increases to attain similar efficiency. However, the results reveal that 5 or 7 imputations are enough, even for high missing rate like 75%. Fourth, in terms of the variances of MAE, those methods are not significantly different.

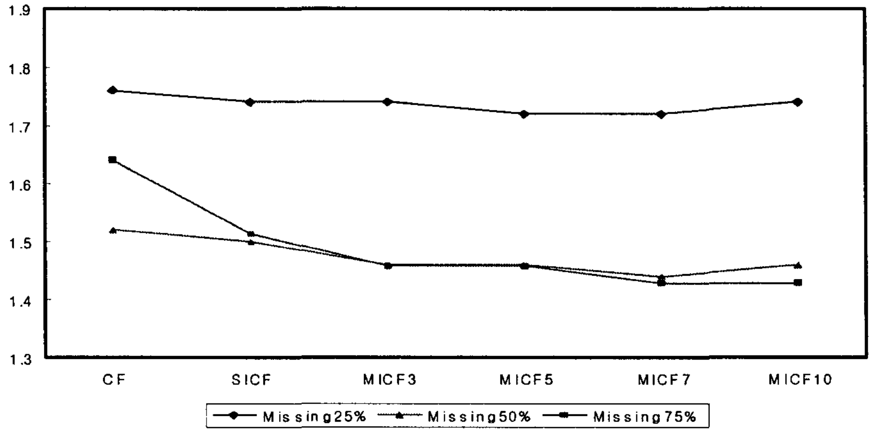

Figure 3. MAE for CF, SICF and MICFs

Regarding the availability of prediction, in case of the dataset with relatively low missing rates of 25% and 50%, all approaches provide whole prediction estimates for the targets. However, in case of dataset with high missing rate of 75%, traditional CF approach fails to provide prediction values for 50% of target values to be estimated but SICF and MICF approaches provide the values of 70% of targets. Such low availability in prediction of traditional CF approach is remarkable in the case of test sets with high missing rates of 80%. Traditional CF approach provides prediction values for only 20% of targets while SICF and MICF approaches answer their prediction values for 70% of targets.

5. Conclusion

Collaborative filtering approach recommends products to a target customer through considering of other customers' experiences and opinions. There could be a practical limit in collecting the complete data from customers and this incomplete data set will cause an analyst to conflict the usage of the CF for recommendation system.

This study introduces imputation-based CF approaches. Single imputation method converts incomplete dataset into complete dataset using EM algorithm. Multiple imputation makes *m* complete datasets using MCMC method. The proposed CF approaches, SICF and MICF, adopt CF algorithm on the complete datasets after imputation to improve the prediction power. The experiments show that the imputation approaches outperform the non-imputation CF approach, especially in the dataset with high missing rate. And the MICF approach has slight improvement over SICF approach. The results also show that MI is highly efficient even for small values of *m* like 5 or 7. Therefore, MICF is useful and efficient for sparse data situation in CF.

The proposed approach in this study has a multivariate normality assumption. Multiple imputation is robust on the model assumption so that it is recommendable on mild violation of normality. For future study, we could employ Bootstrap-based MI as a non-parametric method to extend the usability of the proposed approach. The bootstrap-based MI does not depend on the missing-data mechanism ([3]). Hence Bootstrap-based MI is expected to improve the prediction performance of CF algorithm, especially in using the dataset which does not meet the normality assumption.

In addition to that, this experimental works employ only customers' preferences to make inferences on missing values. If the information of movie genre or demographics of customers is combined with preference data or used for clustering the customers with similar characteristics, imputed values for missing observations will be more accurate.

This incomplete data issue is not limited on CF application. This approach is helpful to solve the problem of missing values in the various fields such as public health research, bioinformatics, and so on.

References

[1] Billsus, D. and Pazzani, M. J. (1998). Learning collaborative information filters, *The 15th International Conference on Machine Learning (ICML).*

[2] Breese, J. S., Heckerman, D., and Kadie, C. (1998). Empirical analysis of predictive algorithms for collaborative filtering, In *Proceedings of Fourteenth Conference on Uncertainty in Artificial Intelligence (UAI-98),* pp. 43-52.

[3] Efron, B (1994). Missing data, Imputation, and the Bootstrap, *Journal of the American Statistical Association*, Vol. 89, pp. 463-479.

[4] Goldberg, D., Nichols, D., Oki, B. M., and Terry, D. (1992). Using collaborative filtering to weave an information tapestry, *Communications of the ACM*, 35, 12, 61-.

[5] Herlocker, J., Konstan, J., Borchers, A., and Riedl, J. (1999). An algorithmic framework for performing collaborative filtering, In *Proceedings of the 1999 Conference on Research and Development in Information Retrieval,* August.

[6] McJones, P. (1997). EachMovie collaborative filtering data set. DEC Systems Research Center.

[7] Rubin, D. B. (1987). Multiple imputation for nonresponse in surveys. New York: Wiley.

[8] Schafer, J. (1997). *Analysis of Incomplete Multivariate Data*, Chapman & Hall.

[9] Schafer, J. (1999). Multiple imputation: A primer. *Statistical Methods in Medical Research,* Vol. 8, pp. 3-15.

[10] Shardanad, U. and Maes, P. (1995). Social information filtering: Algorithms for automating "word of mouth," In *Proceedings of ACM CHI"95 Conference on Hunan Factors in Computing Systems*, 210-217.