Improving Rating Estimation in Recommender Systems Using Aggregation- and Variance-based Hierarchical Models

Akhmed Umyarov
New York University, NY, USA
aumyarov@stern.nyu.edu

Alexander Tuzhilin
New York University, NY, USA
atuzhili@stern.nyu.edu

ABSTRACT

Previous work on using external aggregate rating information showed that this information can be incorporated in several different types of recommender systems and improves their performance. In this paper, we propose a more general class of methods that combine external aggregate information with individual ratings in a novel way. Unlike the previously proposed methods, one of the defining features of this approach is that it takes into the consideration not only the aggregate average ratings but also the variance of the aggregate distribution of ratings. The methods proposed in this paper estimate unknown ratings by finding an optimal linear combination of individual-level and aggregate-level rating estimators in a form of a hierarchical regression (HR) model that is grounded in the theory of statistics and machine learning. The proposed HR model is general enough so that the standard individual-level recommender systems and naive aggregate methods constitute special cases of this model. We show that for the general HR model, the presence of the aggregate variance, surprisingly, does not significantly improve estimation of unknown ratings vis-a-vis the case when only aggregate average ratings are considered. In the paper, we experimentally show that the optimal linear combination approach significantly dominates all other special cases, including the classical non-aggregated case and our previously studied aggregate methods, and therefore is the method of choice.

Categories and Subject Descriptors: H.1.2 [Models and Principles]: User/Machine Systems - Human information processing.

General Terms: Algorithms, Design, Theory

Keywords: predictive models, aggregate ratings, aggregate variance, hierarchical models

1. INTRODUCTION

Consider a Netflix recommender system [7] and assume that it is augmented with the aggregate ratings from the IMDB database [10], such as the one specifying that females in the age category of 18 to 29 gave an average rating of 6.9 (out of 10) to the movie "Madagascar." Can such additional aggregate rating information, provided from the external sources, improve the quality of individual ratings? More generally, a traditional recommender system providing individual ratings to individual users can be supplemented with an *externally* provided OLAP-based [2] system of aggregate ratings, such as the aggregate ratings for "Madagascar" provided by females vs. provided by females in the age category of 18 to 29 years.

In our prior work [16] and [17], we studied this question by assuming certain probabilistic models for recommender systems and incorporating the aggregate information as constraints into these models. Although useful, one of the major limitations of these papers is the assumption about the specific model form.

In this paper, we present an alternative approach that is independent of the specific model assumptions and estimates the unknown individual-level ratings with the optimal linear combination of two estimators: individual- and aggregate-level estimators. We present two theorems about optimal linear combination of two rating estimators and, based on these theorems, we incorporate into the model not only the aggregate average rating but also the variance of the aggregate rating distribution. The resulting model constitutes a certain type of a hierarchical linear regression model [14] that we call the *HR* model. We show that our new approach has certain nice theoretical grounds and prove experimentally that it outperforms rating estimations produced by both the individual rating and the aggregate rating models under a wide range of assumptions about these models. We also consider several special cases of this HR model, including the non-hierarchical simple regression model (SR) from which the aggregate variance information is removed, empirically compare them with each other and also with our previously proposed method from [18] on several datasets and under a wide range of experimental conditions. As a result of this comparison, we experimentally demonstrate that

1. the optimal combination of two estimators outperforms each of the estimators in practice;
2. for the optimal linear combination of the estimators, the aggregate variance does not improve predictive performance of the model, even though improvement possibility is suggested from theoretical considerations;
3. the optimal linear combination performs at least as

RecSys'09, October 23–25, 2009, New York, New York, USA.

well as all other known methods for incorporating external aggregate ratings.

From all this, we conclude that for the optimal linear combination of the rating estimators, the aggregate variance does not improve performance of the model, as compared to a simpler regression-based model without variance (SR). Furthermore, this simple SR model performs as well as the hierarchical regression model (HR) and dominates all other models, including our previously studied model from [18]. Therefore, it is clearly the best choice for incorporating aggregate ratings into the individual rating models.

2. RELATED WORK

The usage of aggregate ratings has been previously studied in the recommender systems literature. An idea of using an OLAP-based multidimensional approach to recommender systems was proposed by [3]. This approach was subsequently extended by [2] by incorporating additional contextual information on ratings, such as when, how and with whom a movie was watched. Also, [13] presents a method for providing recommendations to a group of users. [11] discusses new issues that arise when one considers web-based personalization involving groups for a certain subclass of group recommender systems. Both these methods deal with the bottom-up approach to recommendations that use aggregate ratings as a basis for recommendations to groups of users. In contrast to this, [9] presents a top-down method for using aggregate information about traversal of hypertext pages by a group of users in order to provide better recommendations of hypertext pages to individual members of the group. [6] presents a two-level rating estimation method where at the lower level ratings are estimated using collaborative filtering deploying local scale neighborhood information. At the upper level, [6] uses SVD-style factorization based on global scale information to improve predictions. However, this work does not use any information on prespecified taxonomy of users or items, nor does it use externally specified aggregate ratings. [4] uses pre-existing taxonomy of webpages and advertisements in order to better estimate the click-through rate and combat the sparsity of the data. However, this work is only tangentially related to recommender systems, and also does not use any externally specified aggregate information and does not deal with aggregate ratings.

The idea of incorporating aggregate rating information into a model of a recommender system was described in [16], where it is done for the statistical model of the recommender system described in [5]. [16] theoretically demonstrates that these incorporated aggregate ratings indeed provide for better estimation of unknown ratings than the standalone model of [5]. However, [16] shows this only for the recommender system from [5]. Further, [17] studies how to incorporate aggregate rating information into collaborative filtering models. The paper presents experimental results showing that the aggregate information indeed improves estimations of unknown ratings, and suggests the ways to make the proposed method more scalable. Additionally, [18] unites [16] and [17] into a single framework of utilizing the aggregate information by imposing constraints on the underlying model, provides more scalable method for statistical estimation of model parameters and demonstrates the experimental results on larger datasets.

In this paper, we improve the initial approaches to incorporating aggregate ratings information in recommender systems presented in [16] and [17] by proposing and studying more advanced and better performing models that consider not only the aggregate average ratings, but also the aggregate variances. Although, [1] proposed to use variance of neighbors in neighborhood-based collaborative filtering techniques and demonstrated the increased precision and diversity of individual predictions based on precision-in-top-N and diversity-in-top-N metric, our results are based on externally provided variance of aggregate rating distribution rather than internally computed variance of distribution of ratings participating in the computation, therefore, our results take place in a different domain of applying variance than [1].

In the following sections, we provide a detailed description of our approach.

3. OUR APPROACH

Assume that we have a set of N users and M items and let r_{ij} be a rating of user i for item j (either observed or unobserved). Also assume that a recommender system RS estimates unknown individual ratings r_{ij} using some estimator $\hat{r}_{ij}$, such as classical collaborative-filtering. We call it an *Individual Rating Estimator (IRE)* in this paper.

Also, let r_j^a be the aggregate average rating for the item j that is determined from the externally provided data sources[1], such as the ones described in Section 1. Note that this externally specified aggregate rating r_j^a also constitutes an estimator of an unknown rating r_{ij} of user i for item j, which we call *Aggregate Rating Estimator (ARE)*. Note that this is a rather "simple" estimator, as compared to various possible individual-level estimators $\hat{r}_{ij}$; however it *is* an estimator.

These two estimators can be combined into one new estimator using various methods, including the linear combination, such as

$$r_{ij}^* = \alpha + \beta \hat{r}_{ij} + \gamma r_j^a \quad (1)$$

We, next, demonstrate that, as long as two estimators are not perfectly correlated, this linear combination (1) produces a "better" estimator in the following sense.

THEOREM 1. *Assume that $\hat{x}_1$ and $\hat{x}_2$ are two **unbiased uncorrelated** estimators of unknown quantity x with the following properties:*

$$\begin{cases} E\hat{x}_1 = x, & Var(\hat{x}_1) = v_1 \\ E\hat{x}_2 = x, & Var(\hat{x}_2) = v_2 \\ cov(\hat{x}_1, \hat{x}_2) = 0 \end{cases}$$

Assume that we create a new estimator $\hat{x}$ as a linear combination of $\hat{x}_1$ and $\hat{x}_2$ in the form

$$\hat{x} = \alpha + \beta \hat{x}_1 + \gamma \hat{x}_2$$

Then, the best unbiased estimator $\hat{x}$ that achieves the lowest variance is

$$\hat{x} = \beta \hat{x}_1 + (1 - \beta)\hat{x}_2$$

where $\beta = \frac{v_2}{v_1 + v_2}$

Furthermore, $Var(\hat{x}) \leq \min(Var(\hat{x}_1), Var(\hat{x}_2))$

[1]The fact that the aggregate ratings r_j^a are given at the item-level and depend only on index j does not limit our analysis and is used only for the purpose of compatible notation with the aggregate dataset that we used.

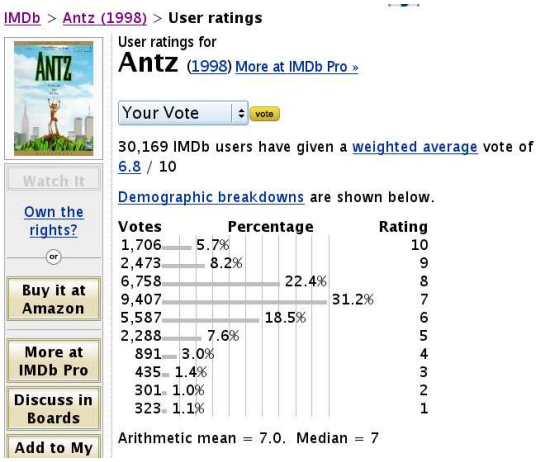

Figure 1: Break down of aggregate rating distribution for a particular movie in IMDB database

The next theorem states a similar result but for the biased and correlated estimators.

THEOREM 2. *Assume $\hat{x}_1$ and $\hat{x}_2$ are two **linearly biased correlated** estimators of unknown quantity x with the following properties:*

$$\begin{cases} E\hat{x}_1 = a_1 x + b_1, & Var(\hat{x}_1) = v_1 \\ E\hat{x}_2 = a_2 x + b_2, & Var(\hat{x}_2) = v_2 \\ cov(\hat{x}_1, \hat{x}_2) = c_{12} \end{cases} \quad (2)$$

where a_1, a_2, b_1, b_2, v_1, v_2, c_{12} are known values.

Assume that we create a new estimator $\hat{x}$ as a linear combination of $\hat{x}_1$ and $\hat{x}_2$, that is

$$\hat{x} = \alpha + \beta\hat{x}_1 + \gamma\hat{x}_2$$

Then, the estimator $\hat{x}$ is unbiased and achieves the lowest variance if

$$\begin{cases} \beta = \dfrac{a_1 v_2 - c_{12} a_2}{a_1^2 v_2 - 2c_{12} a_1 a_2 + a_2^2 v_1} \\ \gamma = \dfrac{a_2 v_1 - c_{12} a_1}{a_1^2 v_2 - 2c_{12} a_1 a_2 + a_2^2 v_1} \\ \alpha = -\beta b_1 - \gamma b_2 \end{cases}$$

Furthermore, $Var(\hat{x}) \leq \min(Var(\hat{x}_1), Var(\hat{x}_2))$

PROOF. The detailed proof of both theorems is available in [18], because of the space limitations. Note that both theorems make *no assumptions about the shape of the distributions* for estimators $\hat{x}_1$ and $\hat{x}_1$, except for the semiparametric assumptions (2) about the expected values of these estimators and their variances. □

When applied to our settings, Theorems 1 and 2 show that

$$r^*_{ij} = \alpha + \beta\hat{r}_{ij} + \gamma r^a_j$$

is an estimator with a smaller variance than that of $\hat{r}_{ij}$ and r^a_j *if* we choose correct weights α, β and γ. Therefore, *these theorems provide a theoretical way to combine the two estimators in order to create a more precise estimator, if* we know the properties of these two estimators, such as their bias, variance and correlation between the two estimators.

For the aggregate-level estimator r^a_j, the external sources can provide not only aggregate average rating r^a_j itself, but also the complete aggregate distribution of the ratings. For example, Figure 1 displays the aggregate distribution information for movie "Antz" from the IMDB database by providing the amount of users who gave a particular rating for every possible rating level. More specifically, let C_{jl} be the number of users who gave a particular rating l to the item j, $l \in L$, where L is the discrete and finite set of ratings. Let S be the set of all the users who gave a rating to item j. Then, the variance of the aggregate-level estimator can be computed as an aggregate variance of the distribution as follows

$$\mathrm{Var}(r^a_j) \approx \frac{1}{|S|}\sum_{i \in S}(r^a_j - r_{ij})^2 = \frac{1}{|S|}\sum_{l \in L} C_{jl} \cdot (r^a_j - l)^2$$

Therefore, the variance of aggregate-level estimator r^a_j can be directly computed from the aggregate distribution of ratings.

In contrast, for the individual-level estimator $\hat{r}_{ij}$, the classical techniques, such as the traditional collaborative filtering [15], typically, do not determine the "uncertainty" levels associated with the predicted rating value. Most of the recommender systems cannot determine how far the true rating value r_{ij} is likely to be from the predicted value $\hat{r}_{ij}$. This "uncertainty" can be measured by the variance of the individual-level estimator $\mathrm{Var}(\hat{r}_{ij})$. Unfortunately, the established classical recommender systems methods typically do not provide ways to compute such variances.

Without having a good model for determining the uncertainty (in the form of variance $\mathrm{Var}(\hat{r}_{ij})$) of individual-level rating estimator $\hat{r}_{ij}$, we cannot apply any of the theorems in practice directly, since the optimal coefficients depend on this unknown parameter.

Because of the difficulty and ambiguity with defining and estimating the variances and covariances of individual-level estimators, we propose a different and more practical approach based on the ideas suggested by Theorems 1 and 2. Intuitively, what Theorems 1 and 2 state is that when the aggregate variance gets larger with other things being equal, the aggregate-level estimator r^a_j becomes less helpful, and therefore the theorems put less weight on the aggregate-level estimator in (1). The converse is also possible: when the aggregate variance is very small (that is, almost all the users tend to agree on the rating for that item), Theorems 1 and 2 put larger weight on the aggregate rating.

More specifically, consider what these theorems imply for weights α, β, γ. Assume that the aggregate-level rating estimator variance $v_2 = \mathrm{Var}(r^a_j)$ is a known number; then, after some algebraic manipulations, the optimal weighting α, β, γ can be represented as a function of v_2 as follows:

$$\begin{cases} \gamma = \dfrac{E}{C + Dv_2} \\ \beta = A + \dfrac{B}{C + Dv_2} \\ \alpha = F + \dfrac{G}{C + Dv_2} \end{cases}$$

where A, B, C, D, E, F, G are independent of v_2.

That is, according to Theorems 1 and 2, the optimal weighting α, β, γ that achieves the best MSE prediction, depends on v_2 in a monotonous fashion. Therefore, given that individual-level estimator variance v_1 is not observed by us, but aggregate-level estimator variance v_2 is directly available, the effect of interest for us is how the predictive

power of a recommender system is affected by the redistribution of weights caused by aggregate variance v_2.

One way to capture this effect is to allow the data "speak for itself" by finding the best weights that fit the data and verifying whether the dependence of weights on aggregate variance changes the predictive power. In order to accomplish that, we employed the following *hierarchical regression model (HR)*

$$r^*_{ij} = \alpha + \beta \hat{r}_{ij} + \gamma r^a_j \tag{3}$$

$$\text{where } \begin{cases} \alpha = \alpha_0 + \alpha_1 \text{Var}(r^a_j) \\ \beta = \beta_0 + \beta_1 \text{Var}(r^a_j) \\ \gamma = \gamma_0 + \gamma_1 \text{Var}(r^a_j) \end{cases}$$

The variables $\hat{r}_{ij}$, r^a_j, $\text{Var}(r^a_j)$ constitute known parameters of the model, while α_0, α_1, β_0, β_1, γ_0, γ_1 are unknown. The variable r^*_{ij} constitutes a dependent variable that we are trying to estimate; so it is known for the training sample, but unknown for the testing purposes.

This model not only incorporates the fact that the change in the aggregate variance affects the weighting scheme as in Theorems 1 and 2, but also that the aggregate variance affects the weight in monotonous fashion.

Furthermore, the classical recommender system algorithms constitute special cases of model (3). In this paper, we consider the following cases:

1. *Individual Rating Estimator (IRE):* $\beta_0 = 1$ and all other coefficients are zero. This case results in the following expression: $r^*_{ij} = \hat{r}_{ij}$. This is the case of usual individual-level recommender system that does not take into account any aggregate information.

2. *Aggregate Rating Estimator (ARE):* $\gamma_0 = 1$, all other coefficients are zero, resulting in the following expression: $r^*_{ij} = r^a_j$. This is the case of the predictor that predicts a rating for each item as the item's aggregate rating.

3. *Blending Rating Estimator (BRE):* $\gamma_0 = 1 - \beta_0$ and all other coefficients are zero. This case results in the expression compatible with Theorem 1, that is $r^*_{ij} = \beta \hat{r}_{ij} + (1 - \beta) r^a_j$, that is based on the belief that the estimators $\hat{r}_{ij}$ and r^a_j are unbiased and uncorrelated.

4. *Simple regression (SR):* $\alpha_1 = \beta_1 = \gamma_1 = 0$. This case results in the simple regression model of the form $r^*_{ij} = \alpha + \beta \hat{r}_{ij} + \gamma r^a_j$, such that coefficients α, β and γ are constant across all items and do not depend on aggregate variance.

Note that the coefficients α_0, α_1, β_0, β_1, γ_0 and γ_1 are not known in advance and need to be estimated from the training sample. There are several different ways to do it using the training sample [8] and the specifics of this process are described in Section 4.4.

In this section, we presented two theorems that provide a theoretical way to find an *optimal* linear combination of the two rating estimators assuming the variances of both estimators are known. Since for the individual-level rating estimator the variance is unknown, we proposed an alternative way to utilize the intuition of the theorems: a hierarchical regression model *HR* that captures the effect of weight redistribution in the linear combination of rating estimators.

In the following sections, we apply this model and its special cases (1) - (4) to real life datasets in order to see the effect of weight redistribution caused by aggregate variance on the predictive performance of recommender systems.

4. EXPERIMENTAL SETTING

In this Section, we describe the data used in our experiments, partitioning of the data into the training and the testing sets, and the performance measures used in our experiments.

4.1 Individual rating datasets

We used the following "real-life" datasets for learning individual ratings and empirically validating our methods.

4.1.1 MovieLens Dataset

We used the full MovieLens dataset [12] consisting of more than 1 million ratings of 3900 movies provided by 6040 users.

4.1.2 Subsample #1 of the Netflix Prize Dataset

We also used a random subsample[2] of the Netflix Prize dataset [7] consisting of 10,000 users, 10,000 movies, and 1,000,000 ratings. This subsample was produced using the following procedure:

1. Select 10,000 random users from the set of all the Netflix users ranked between #10,000 and #300,000 based on the total number of ratings they gave.

2. Select 10,000 random movies out of the movies that these 10,000 users watched.

3. Select 1,000,000 random ratings out of the ratings that these 10,000 users provided for these 10,000 movies.

This dataset contains the release year for the movies and no attributes for the users since the Netflix Prize dataset [7] contains *no data at all* about their customers beyond the customer ID number. For movies, the Netflix Prize dataset [7] provides the movie title and the release year. For users' ratings, the dataset contains the timestamp when the rating appeared on the website.

4.1.3 Subsample #2 of the Netflix Prize Dataset

This is another random subsample of the Netflix Prize dataset [7] consisting of 1,000 users and 1,000 movies with 5,000 ratings. The subsample was produced using the same procedure as described in Section 4.1.2.

4.1.4 Subsample #3 of the Netflix Prize Dataset

This is also a subset of the Netflix Prize dataset [7] consisting of 1,000 users and 1,000 movies with 5,000 ratings, that is produced using the same procedure as described in Section 4.1.2. The reason for including both Subsample #2 and Subsample #3 is that we used different kinds of training/test split for these datasets as described in Section 4.4.

4.2 Aggregate rating dataset

In order to introduce aggregate rating information from the external sources into the Individual Rating datasets described above, we extracted the average ratings of the movies

[2]We decided to use subsamples instead of complete datasets for the sake of the speed of computations

used in those datasets from the IMDB database [10], i.e. for each movie in the Individual Rating datasets, we attempted to find a corresponding average movie rating from IMDB. The results of this matching process are presented in the following table:

Name of dataset	# of Movies	Movies Matched
MovieLens	3,952	2,162
Netflix #1	10,000	9,949
Netflix #2	1,000	998
Netflix #3	1,000	998

We also extracted the information from IMDB on the full distribution of given ratings similar to the one shown in Figure 1 which we used to compute the aggregate variance for each movie.

4.3 Individual-level estimators

For the individual-level estimators, we employed the following two models: classical item-based collaborative filtering [15] and hierarchical linear model (HLM) from [5].

More specifically, as a first method, we employed classical item-based collaborative filtering method from [15] with Pearson correlation measure of similarity and with the size of neighborhood equal to the complete number of items in the system.

As a second method, we employed HLM approach from [5] that estimates the unknown ratings r_{ij} in terms of user characteristics, item characteristics and the interaction effects between them. Interaction effects arise from the hierarchical structure of the model and are intended to capture effects such as, for example, how the age of a user changes his or her preferences for certain genres of movies.

4.4 Training and testing strategies

In order to empirically validate our approach, we split each dataset into 10 subsets for the 10-fold cross validation. The Netflix #1, #3 Subsets and MovieLens database were split randomly into 10 subsets with 40% of the data going into training set and 60% of the data going into the test set. The reason for this kind of split is that we would also like to validate generalizability of our techniques. Recommender systems are characterized by the fact that typically only a small percentage of all possible ratings is known, for example, in the Netflix Prize Dataset [7] only $\approx 1.2\%$ of all possible ratings is given. However, ideally, in order to recommend to each user the best set of items out of all available, the ratings for all the rest $\approx 98.8\%$ have to be estimated. Therefore, recommender systems are characterized by having very small training set as compared to the potential test set.

We used more traditional split for another subsample dataset to show that our results are not dependent on the type of the split. The Netflix #2 dataset was split randomly into 10 subsets with 90% of the data going into training set and 10% of the data going into the test set.

Note that as mentioned in Section 3, the coefficients α_0, α_1, β_0, β_1 and γ_0, γ_1 in equation (3) are not known in advance and need to be estimated *from the training sample.*

There are several different ways to do it using the training sample [8]. In this paper, we consider the following two methods:

1. Splitting the training set into two:
   (a) Randomly split each training set into disjoint "training #1" and "training #2" sets.
   (b) Train the individual-level rating model on "training #1" set
   (c) Provide predictions $\hat{r}_{ij}$ of the individual-level rating model for "training #2" set.
   (d) Based on the predictions $\hat{r}_{ij}$ on the "training #2" and aggregate ratings r_j^a, estimate the best parameters α_0, α_1, β_0, β_1 and γ_0, γ_1 that fit observed ratings to the model using linear regression.
   (e) These parameters are the output of the training procedure and will be used on the test set.
2. Choosing a designated test set from the usual split:
   (a) From the usual n-fold cross-validation training/testing split, choose a designated test set that will be used for fitting parameters.
   (b) Train the individual-level rating model on the training set corresponding to that designated test set
   (c) Provide predictions $\hat{r}_{ij}$ of the individual-level rating model for this one designated test set
   (d) Based on the predictions $\hat{r}_{ij}$ on the designated test set and aggregate ratings r_j^a, estimate the best parameters α_0, α_1, β_0, β_1 and γ_0, γ_1 that fit best the observed designated test set ratings using linear regression.
   (e) These parameters are the output of the training procedure and will be used on *other* test sets.

These two splitting methods present a trade-off between using more data and computational performance. The method #1 sacrifices some of the training data in order to have the complete range of test sets. The method #2 sacrifices some of the test sets in order to determine the model parameters on them, while verifying the results on the smaller number of test sets. We employed both of these methods, but these methods produced so similar results that there is no point to present both. Clearly, this is not a surprising result since the sizes of the datasets that we employed are more than enough to determine precisely only 6 free parameters of the model (3). Therefore, we plotted only the graphs for the method #2, that is clearly much faster.

4.5 Performance measures

In order to present performance of our models in intuitive way, we show the evolution of mean squared error (MSE) performance of the model as more and more aggregate information is introduced into the model.

More specifically, for each $k = 0, 1, 2, 3, \ldots$, we select only first k aggregate observations and incorporate them into the model as described in Section 5. Then, for each k, we calculate the mean squared error (MSE) of predictions of the models that use exactly k aggregate ratings across all the aforementioned test sets. Finally, we plot the graph of these MSEs for each value of $k = 0, 1, 2, 3, \ldots$, as is shown, for example, in Figure 2 for the case of the MovieLens dataset. Note that $k = 0$ means that *no* aggregate rating information is used at all, and we are dealing with the basic individual rating prediction model in this case.

5. EXPERIMENTAL RESULTS

The graphs in Figures 2 - 6 represent the mean squared error (MSE) performance of the item-based CF and HLM model as a function of the number of additional aggregate ratings introduced for 3 subsets of the Netflix Prize [7] and MovieLens datasets.

More specifically, these figures plot on the x-axis the cumulative number of additional aggregate ratings introduced into the model. The 0-th tick corresponds to the plain basic recommendation model without any aggregate ratings. The 1st tick corresponds to adding just one aggregate rating of type (3) for the first item, all other ratings for all other items are predicted using individual-level estimator. The 2nd tick adds one more aggregate rating of type for the second item, and so on. On the y-axis we plot the MSE performance of the model based on 10-fold cross-validation described in Section 4.

Note that Figures 2 - 4 present the MSE performance improvements of the item-based CF, while Figures 5 - 6 present the MSE performance for the HLM model from [5] (presented in Section 4.3).

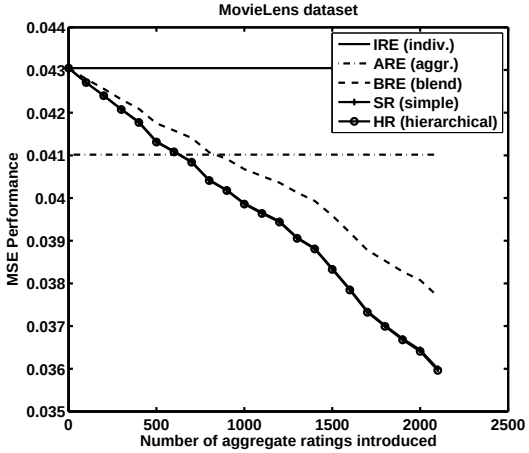

Figure 2: Item-based CF performance on MovieLens dataset

By examining Figures 2 - 6, we can make the following observations:

1. The MSE errors tend to go down on average with the number of additional aggregate ratings. Although the MSE performance is not a monotone function of the number of additional ratings [3], the drift is definitely downward across all the figures and the BRE, SR and HR methods.

2. The ARE method is better than the classical item-based CF, however worse than the HLM across all the datasets. This is a surprising result, stating that aggregate ratings, as estimators of unknown ratings, are *better* than the one produced by the item-based CF across *all* of our experimental settings. We want to put a disclaimer, however, that this result, obtained for the specific datasets pertaining to movies, cannot be generalized to other applications and datasets. It only shows that aggregate ratings themselves can often be good estimators of unknown ratings, and that they definitely should not be overlooked or discarded, but instead be incorporated into the models estimating unknown ratings, as we have done in this paper.

3. Combining individual and aggregate rating models, as we have done in this paper with the BRE, SR and HR models, results in superior performance since all the three models BRE, SR and HR eventually outperformed the pure individual (IRE) and the pure aggregate (ARE) models (when the number of aggregate ratings became sufficiently large) for *both* the item-based CF and the HLM methods across all the datasets.

4. The HR and SR methods outperformed BRE. This is definitely clear from Figures 2, 4, 6 that show a clear separation between the BRE line and the (merged) SR and HR lines. This result is not very surprising because the BRE model has an additional constraint

[3]Some of these graphs exhibit occasional upward jumps when new aggregate ratings are introduced, which cannot be truly observed in these figures because of large numbers of aggregate ratings on the x-axis of the graphs.

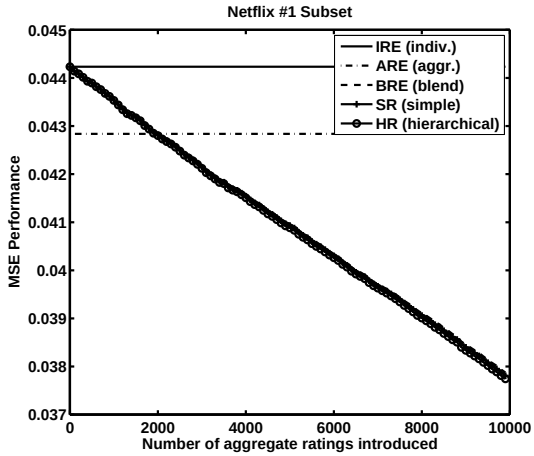

Figure 3: Item-based CF performance on Netflix #1 subset

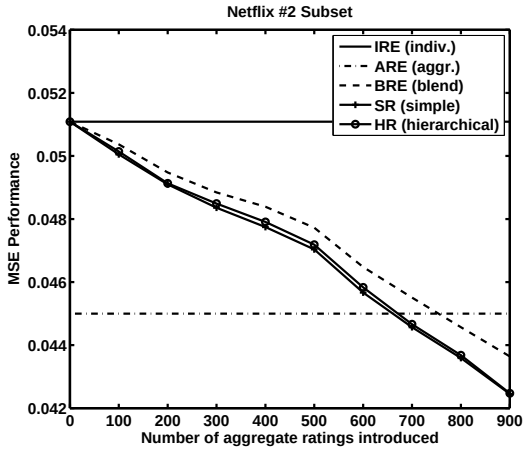

Figure 4: Item-based CF performance on Netflix #2 subset

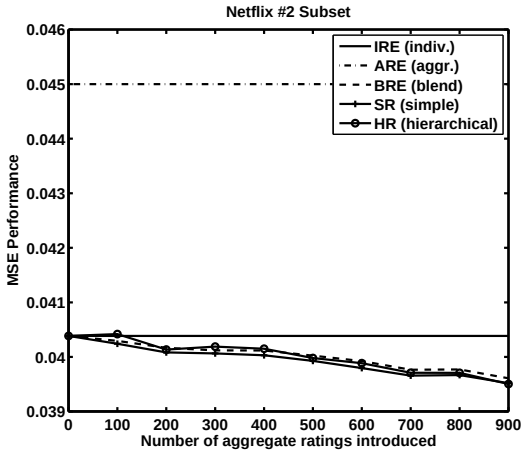

Figure 5: HLM performance on Netflix #2 subset

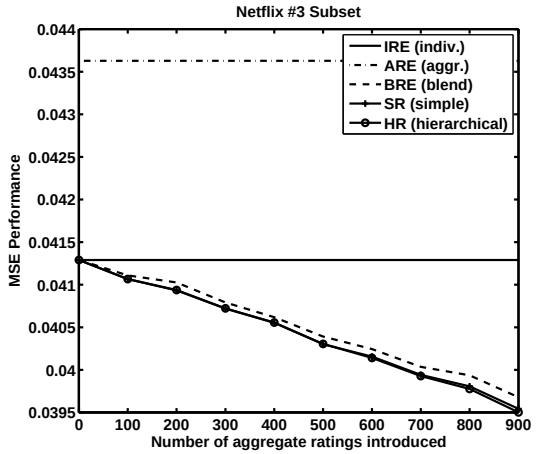

Figure 6: HLM performance on Netflix #3 subset

of $\alpha = 0$ and $\beta + \gamma = 1$ vis-a-vis the SR and the HR models. This extra "flexibility" of the HR and SR models results in better performance versus BRE.

5. The performance of the HR and the SR models is comparable. *All the HR and the SR lines in Figures 2 - 6 are almost indistinguishable and merge into one (thick) line.* This result implies that the aggregate variance $Var(r_j^a)$ used in equation (3) and being the main differentiator between the SR and HR models does not really contribute to the better performance of the HR model vs. SR, and using only the aggregate average rating r_j^a is enough to achieve good predictions of unknown ratings.

   Furthermore, we conjecture that this result is also true not only for the models of type (3) but for various types of models taking variance $Var(r_j^a)$ into consideration. We experimented with a significant number of such alternative models [4] and adding variance $Var(r_j^a)$ to these models did not result in any significant performance improvements. We believe that this is the case because the (unobserved) variance of individual-level estimator covariates in unison with the variance of aggregate-level estimator, therefore optimally there should be no redistribution of weights since both individual- and aggregate-level estimators get equally worse for the items that have large aggregate variance.

[4]Unfortunately, we cannot describe them in this paper because of space limitation.

From all these results reported above, we can draw the following conclusions:

- The simple linear regression (SR) model is the best choice among all the methods considered in this paper. It dominated all other methods, except HR. It is better to use it than HR because it is simpler and it is easier to estimate its parameters. It also provides a powerful linear combination of two fundamental models: individual- (IRE) and the aggregate-level (ARE) models. Further, we show in Section 6 that it also dominates in some cases other aggregate rating models previously considered by us [16], [18]. Therefore, we believe that the SR model is really the way to go.

- Aggregate rating variance $Var(r_j^a)$ did not improve performance of the rating estimation models considered in this paper. The particular form of the model that we employed used key insights from Theorems 1 and 2. These theorems suggest that the weighting that generates the best linear combination is monotonously dependent on aggregate variance, however our results showed that the monotonous dependence of weights on aggregate variance does not help to improve predictions in practice in this particular domain.

6. COMPARISON WITH PREVIOUS MODELS

In [16], we presented an alternative way to introduce the aggregate average rating information into the HLM model. We utilized the specific model properties of the HLM model and showed that this additional aggregate information is equivalent to imposing linear constraints on model parameters. We also demonstrated that in practice these additional model constraints improve predictive performance of the method presented in [18].

Since, SR method performed surprisingly well, we decided to compare it with our previous work [18]. Figure 7 demonstrates the performance comparison of the SR method presented in this paper and the "native" HLM model parameter constraint method presented in [18].

According to Figure 7, HR and SR model achieved approximately 2.5% MSE performance improvement, while the "native" HLM method achieved only 1.4% MSE performance improvement on the same Netflix #2 dataset. Note, however, that we cannot make any claims on superiority of either method, since "native" HLM method was not trained to the best MSE performance in the original paper [18], that method did not utilize its full potential in terms of MSE performance, since only the simplest weighting scheme for the newly aggregate information was used.

7. CONCLUSIONS AND FUTURE WORK

In this paper, we presented a new method for incorporating aggregate rating information into the traditional indi-

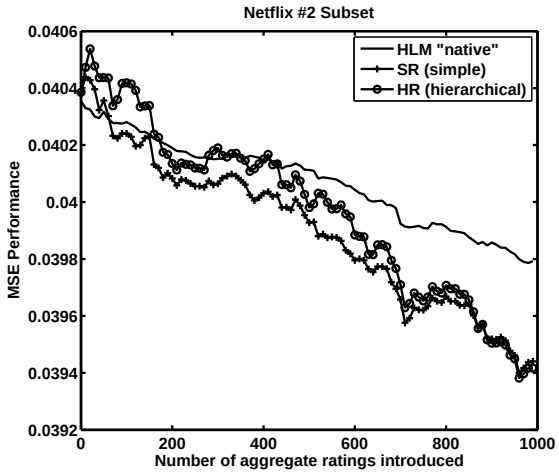

Figure 7: Comparison of simple regression approach with model-based approach

vidual rating estimation methods, such as the classical collaborative filtering or the hierarchical linear model (HLM) from [5]. Our approach is based on the optimal linear combination of the individual and the aggregate rating estimators that also incorporate aggregate variance information into the model, called *HR*, in a hierarchical manner. We also consider a simplified non-hierarchical version of the HR model, called SR, that drops the variance information and, therefore, constitutes an instance of a simple linear regression.

We show that, surprisingly, the simple regression model SR performs as well as the more complicated hierarchical linear model HR. This means that the additional aggregate variance information does not really help, even though it is suggested by statistical theorems. We also show that the SR model outperforms various other models considered in the paper, including some of our previously studied models [18], and therefore, provides the best approach of incorporating aggregate rating information into the individual-level models. One of the reasons that would create such a phenomenon is the fact that both the individual-level and the aggregate-level estimator variances increase in unison.

One of the reasons why the SR model outperforms some of the models from [18] is because these other models have certain types of restrictive assumptions imposed on them (in the form of the model form and fixed weights for the aggregate information). Therefore, as a part of the future work, we plan to extend the models in [18] so that the models use optimal weight assignments rather than a fixed constant. This would allow for additional more extensive comparisons between HR model, SR model and the models from [18].

Another direction for the future work is incorporating multiple levels of hierarchical aggregate information. In some cases several different aggregate ratings are available for a single individual rating on the different levels of aggregation hierarchy, such as, a known aggregate average rating for all comedies by all females in addition to known aggregate average rating for all comedies by all females in New York. The question of the optimal linear combination of those ratings with individual rating and the statistical theory behind it constitutes one of the important directions for the future work in this area.

8. REFERENCES

[1] G. Adomavicius and Y. Kwon. Overcoming Accuracy-Diversity Tradeoff in Recommender Systems: A Variance-Based Approach. *Proceedings of WITS'08, Paris, France*, 2008.

[2] G. Adomavicius, R. Sankaranarayanan, S. Sen, and A. Tuzhilin. Incorporating contextual information in recommender systems using a multidimensional approach. *ACM Transactions on Information Systems (TOIS)*, 23(1):103–145, 2005.

[3] G. Adomavicius and A. Tuzhilin. Multidimensional recommender systems: a data warehousing approach. In *2nd Intl. Workshop on Electronic Commerce. LNCS 2232*, 2001.

[4] D. Agarwal, A. Broder, D. Chakrabarti, D. Diklic, V. Josifovski, and M. Sayyadian. Estimating rates of rare events at multiple resolutions. *Proceedings of the 13th ACM SIGKDD*, pages 16–25, 2007.

[5] A. Ansari, S. Essegaier, and R. Kohli. Internet Recommendation Systems. *J. of Marketing Research*, 37(3), 2000.

[6] R. Bell, Y. Koren, and C. Volinsky. Modeling relationships at multiple scales to improve accuracy of large recommender systems. *Procs of the 13th ACM SIGKDD*, 2007.

[7] J. Bennett and S. Lanning. The Netflix Prize. *Proceedings of KDD Cup and Workshop 2007*, 2007.

[8] C. Bishop and N. Nasrabadi. Pattern Recognition and Machine Learning. *J. of Electronic Imaging*, 16:049901, 2007.

[9] J. Bollen. Group user models for personalized hyperlink recommendations. *Procs of the Int. Conference on Adaptive Hypermedia and Adaptive Web-Based Systems*, 2000.

[10] IMDB. http://www.imdb.com. 2006.

[11] A. Jameson and B. Smyth. *Recommendation to Groups*, chapter The Adaptive Web: Methods and strategies of web personalization. Springer, 2006.

[12] MovieLens. available at http://www.grouplens.org/node/73, 2006.

[13] M. O'Connor, D. Cosley, J. A. Konstan, and J. Riedl. Polylens: A recommender system for groups of users. *Procs of ECSCW*, 2001.

[14] S. W. Raudenbush and A. S. Bryk. *Hierarchical Linear Models: Applications and Data Analysis Methods*. Sage Publications, Inc, 2001.

[15] B. Sarwar, G. Karypis, J. Konstan, and J. Reidl. Item-based collaborative filtering recommendation algorithms. *Procs of the 10th int. conference on World Wide Web*, 2001.

[16] A. Umyarov and A. Tuzhilin. Leveraging aggregate ratings for better recommendations. *Procs of the 2007 ACM Recommender systems*, 2007.

[17] A. Umyarov and A. Tuzhilin. Improving Collaborative Filtering Recommendations Using External Data. *Proceedings of the IEEE ICDM 2008 Conference*, 2008.

[18] A. Umyarov and A. Tuzhilin. Leveraging aggregate ratings for improving predictive performance of recommender systems. Working paper. Stern School of Business. New York University. CeDER-08-03, 2008.