Increasing Consumers' Understanding of Recommender Results: A Preference-based Hybrid Algorithm with Strong Explanatory Power

Paul Marx
Department of Marketing and Media
Bauhaus-University Weimar
Bauhausstr. 11,
99423 Weimar, Germany
paul.marx@uni-weimar.de

Thorsten Hennig-Thurau
Department of Marketing and Media
University of Muenster
Am Stadtgraben 13-15,
48143 Muenster, Germany
thorsten@hennig-thurau.de

André Marchand
Department of Marketing and Media
Bauhaus-University Weimar
Bauhausstr. 11,
99423 Weimar, Germany
andre.marchand@
uni-weimar.de

ABSTRACT
Recommender systems are intended to assist consumers by making choices from a large scope of items. While most recommender research focuses on improving the accuracy of recommender algorithms, this paper stresses the role of explanations for recommended items for gaining acceptance and trust. Specifically, we present a method which is capable of providing detailed explanations of recommendations while exhibiting reasonable prediction accuracy. The method models the users' ratings as a function of their utility part-worths for those item attributes which influence the users' evaluation behavior, with part-worth being estimated through a set of auxiliary regressions and constrained optimization of their results. We provide evidence that under certain conditions the proposed method is superior to established recommender approaches not only regarding its ability to provide detailed explanations but also in terms of prediction accuracy. We further show that a hybrid recommendation algorithm can rely on the content-based component for a majority of the users, switching to collaborative recommendation only for about one third of the user base.

Categories and Subject Descriptors
H.3.3 [**Information Storage and Retrieval**]**:** Information Search and Retrieval – *Information filtering, Search process, Selection process.* H.2.8 [**Database Management**]**:** Database applications – *Data mining.*

General Terms
Algorithms, Design, Human Factors, Measurement

Keywords
Recommender systems, explanation of recommendations, user preferences, constrained optimization, hybrid algorithms

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
RecSys'10, September 26–30, 2010, Barcelona, Spain.
Copyright 2010 ACM 978-1-60558-906-0/10/09...$10.00.

1. INTRODUCTION
Stimulated by the Netflix Prize Competition, recommender research has focused on recommender algorithms accuracy, whereas topics of recommender acceptance and trust received less attention [14]. Although movie research provides evidence that movie characteristics such as stars and budgets significantly influence the movie success as a result of consumers' preferences for them [11], such characteristics were not adequately handled by recommender researchers.[1]

We argue that incorporating such item characteristics in the recommendation process can be fruitful, as it allows recommender systems to provide users with reasons underlying recommendations [15], which will increase recommender transparency and credibility, two established performance criteria [6, 16, 19]. We further argue that explanations can lead to higher choice efficiency [22] and even satisfaction [2] with the recommendations. This argument is consistent with Aksoy et al. [1] who show that consumers make better choices when using recommendation agents which use attribute weights and decision strategies that are similar to their own.

We present a method that extracts user attribute-related preferences from movie rating data of a commercial movie recommender system and show that the derived preference information is suitable not only for providing users with meaningful explanations of recommendations, but also for generating reliable recommendations. Adding to recent developments on hybrid recommenders [4, 5], our method combines the extracted preference-related information with traditional collaborative techniques.

[1] When preferences towards movie attributes were used in extant work (eg. [23]), the choice of the attributes was based on information availability, not a thorough study of relevant attributes. Movie attributes were used for post processing of predictions, but were not directly involved in the process of recommendation generation (eg. [21]).

2. DATA

We develop and evaluate our method with data from the movie recommender platform Moviepilot.com. We preferred this data over Netflix because it does not suffer from artifacts based on scale and interface changes which are known for the Netflix data [13] and is newer, encompassing ratings provided between August 2006 and April 2008 which should adequately reflect contemporary consumer attitudes and behaviors. Also, our cooperation with Moviepilot gave us complete insight into the processes and algorithms underlying the data. The raw dataset contains 1,389,749 ratings of 15,593 movies by 9,788 users of the platform.

Although Moviepilot.com presents ratings in its user interface on the scale varying from 0 to 10 points in .5 steps, ratings are stored in the database as integer numbers from the interval between 0 and 100 (i.e., a rating of 7.5 point is stored as 75). We left out the six latest ratings for validation purposes and six randomly drawn ratings for each user as a holdout for out-of-sample predictions; users for whom there was not enough data to generate the both holdouts were discarded. Both holdouts comprise of 47,610 ratings each. The data about movie attributes (genres, year of production, country of origin, budget, admissions, box office, acting stars, directors, writers and production companies) was taken from the Internet Movie Database (IMDb, see www.imdb.com).

3. MODELLING CONSUMER PREFERENCES

In the context of movie recommendations ratings determine the value of a particular movie for the user and allow comparisons of users' liking of different movies. The ratings can be interpreted as a normalized utility, which allows comparing item utilities between users and makes normalization unnecessary.

We model the rating r from user u for a movie i as an inner product of the binary vector of movie features $\boldsymbol{m}$ and the vector of users' part-worths $\boldsymbol{p}$, as shown in Equation 1:

$$r_{u,i} = \mu + \boldsymbol{m}_i^T \boldsymbol{p}_u \qquad (1)$$

Here the movie ratings and movie features vector are known from the users' rating records and the IMDb. The mean movie rating (μ) serves as baseline on which the part-worths are centered. The vector $\boldsymbol{p}$ is to be estimated. Once estimated, the part-worths can be used both for predictions of the user's ratings to new items and for providing the explanations to recommendations. Moreover, the explanations can be presented in a "pros-and-cons" style, such as "Titanic is recommended to you because it matches your preferences highly. Pros: High budget Hollywood movie. Cons: you don't like the movie's drama genre and its star Leonardo Di Caprio. Taking these factors into account, we expect that you will rate this movie 8 out of 10."

This simple model of user preferences does not account for the effects which occur independently of user-item interactions. Specifically, some users give higher/lower average ratings than the average user, something we refer to as user bias. Also, some movies generally receive higher/lower ratings than others (item bias [3, 13]). Users also differ in their reaction to average movie ratings; while some users adapt to mainstream judgments, others react overly positive, and a third group reacts skeptical. Incorporating these effects leads to Equation (2):

$$r_{u,i} = \mu + b_u + b_i s_u + \boldsymbol{m}_i^T \boldsymbol{p}_u \qquad (2)$$

where b_u and b_i indicate the user bias and the item bias, respectively. User and item bias are defined as deviations of a user's and a movie's mean rating value from the overall mean μ, respectively. The users' reactions to the movie bias are captured by scale factor s_u.

We also consider the changes of movie popularity as well as the individual users' changing preferences and rating behavior over time [13]. Specifically, we incorporated temporal dynamics for each of the biases. We replaced b_u by $[b_u + \alpha_u t]$, where b_u constitutes only the static part of the user's rating, t is time, and α_u is the slope of the user's rating trend, The movie bias b_i is replaced analogously by $[b_i + \beta_i t]$ which leads to Equation (3):[2]

$$r_{u,i} = \mu + b_u + \alpha_u t + (b_i + \beta_i t)\, s_u + \boldsymbol{m}_i^T \boldsymbol{p}_u \qquad (3)$$

4. ESTIMATING PREFERENCES

The scarcity of data is the biggest challenge for recommender research. We suggest a combination of statistical and optimization techniques for parameter identification. The procedure encompasses two steps, estimation and optimization. The optimized parameter values are then used for predicting movie ratings which are new to the users, and for explaining these predictions both in "keyword" and "influence" as well as in "pro-and-con" style. We also report the results of a post-hoc integration of our model with traditional algorithms into a hybrid recommender to further increase prediction accuracy.

4.1 Step 1: Estimation

For each parameter the initial value and its confidence interval are estimated through univariate OLS regression analysis. We utilize OLS regression, as it provides inferences about parameter significance. The latter information is used for dropping parameters that are statistically meaningless for describing users' movie preferences and for generating and explaining rating predictions. For example, if the parameter for star "George Clooney" does not reach significance, this actor is considered neutral for the user's movie preference formation and can be excluded from the estimation process ($p < .10$ was used as cut-off criterion).

With regard to the user bias parameters b_u and α_u, we run a simple regression $r_{u,i} = b'_u + a_u t$ for each user. Whereas the user's rating trend parameter a_u is derived directly from this regression, the baseline b_u is taken from b'_u by subtracting the overall rating mean, i.e. $b_u = b'_u - \mu$. After experimenting with different time frames we found that setting t = one day produced good estimates when letting the standard deviation of the user's rating time be at least 60. In other words, we require the user to have been rating the movies for at least 120 days in order to be able to capture his or her drifting rating behavior. For the users who do not meet this condition, b'_u is the mean of the correspondent user's ratings.

The item biases are estimated in the same way, using auxiliary regressions of the form $r_{u,i} = b'_i + \beta_i t$. Again, the time resolution is here set to one day. In contrast to user bias, we expect movie popularity to change slower and thus require the time frame

[2] Please note that we tested also for short-term changes [12], but found none. We also tested for but found no temporal dynamics in the user rating scaling factor as well as in the user part-worths. This might be the result of the relatively short time frame of ratings covered by the data set. As a result, we did not consider any of those effects in the empirical estimation process.

between a movie's first and last ratings to be at least 240 days. Auxiliary regressions were also used to estimate the user part-worths. However, we have to deal with concurrent parameters, where two shortcomings of OLS regressions surface: its sensitivity toward model misspecifications and its tendency toward overfitting under certain conditions. To avoid overfitting we discard information which has non-zero values in less than five percent of rated items. Overestimation of the parameter values due to model misspecification was prevented by multi-level correction of the auxiliary regression results.

The estimate for the scale factor s_u, which reflects the user's reaction to the movie bias can then be calculated by fixing all remaining model parameter at their estimated values in Equation (4):

$$s_u = \sum_{(i \in I)} \left(\left(r_{u,i} - \mu - b_u - \alpha_u t - \boldsymbol{m}_i^T \boldsymbol{p}_u \right) / (b_i + \beta_i t) \right) \quad (4)$$

The confidence limits for s_u are set to $s_u \pm \sigma t$, with σ being the standard deviation and t drawn from the Student's t-distribution for p = .10 and degrees of freedom equal to the number of user's ratings minus one.

4.2 Step 2: Optimization

We then performed an optimization of the model parameters, allowing them to vary inside their respective confidence intervals we have obtained in the estimation step. Initialized with the point estimates, Equation (3) fits to the test data with an RMSE of 24.67. Whereas the point estimates represent the most probable values of the model parameters, they are not necessarily the "true" values. Finding these "true" values constitutes an optimization problem, which we solve through the conjugate gradient method for multiple dimensions [17]. We modify this method so that the parameter values are not allowed to exceed their confidence limits. In order to prevent overfitting parameter learning is stopped when the error on the holdout data increases. Using the parameter values gained through this procedure results in a RMSE of 24.17, which represents an accuracy improvement of about 2 per cent. This improvement is significant at $p < .05$.

4.3 A Hybrid Approach to Further Increase Prediction Accuracy

The non-zero RMSE of our method indicates that Equation (3) does not capture all the user ratings variance. An inspection of the absolute deviations of our predictions from the test ratings revealed that a considerable part of the overall error stems from a small number of data points. Table 1 summarizes the distribution parameters of the absolute error.

Table 1. Distribution Parameters of the Absolute Prediction Error

Mean	Mode	Curtosis	SE of Curtosis	25 Percentil	50 Percentil	75 Percentil	SD
18.19	0	2.434	.022	6.03	13.60	25.48	16.36

The high curtosis (over 2) and relatively low standard deviation indicate that the distribution is peaked and positive skewed. Further, the absolute prediction error for particular ratings is lower than the value of the RMSE and exceeds it in only about 30% of the cases. This means that the RMSE value mainly constitutes from a low number of points with large deviations, rather than from large number of points with nearly equal deviations. Although most of the data points with large deviations belong to the same group of users, we were unable to find patterns which would allow us to identify users with high prediction error a priori. Thus, those users form their movie preferences using information not captured by the preference function shown in Equation (3).

We assume that similarity among such "problematic" users is an appropriate information source for generating predictions when the explicit preference modeling not adequately captures the rating behavior. To test this assumption, we implemented a user-to-user collaborative filtering algorithm and performed a series of tests. Results show that the error distribution of the collaborative filter significantly differs from the one produced by Equation (3) ($p < .01$) which indicates that both algorithms capture different parts of user ratings' variance. Consistent with this, both approaches produce unequal errors for most users ($p > .1$) on the single user level.

As combinations of concurrent methods outperform the best individual predictions [7, 8, 20], we developed a hybrid approach which combined the predictions of both methods. We choose the individual predictor which performs best on the withheld data for each user and utilized an additional holdout set to compare the individual performance of the two prediction methods. The best performing method is determined through a t-test (two-sided) for paired samples for the significance level of $p < .10$. If the collaborative filter significantly outperforms Equation (3) for a user, it is used for generating his or her predictions; Equation (3) is used in all other cases. The overall RMSE of the hybrid method is 20.66, which constitutes a 16% improvement over Equation (3) used solely and a 10% improvement over the collaborative filter. It should be noted that the latter method was used for only 34% of the users, while the majority of the users (66%) received detailed explanation for the recommended items in "keyword", "influence" and "pros-and-cons" styles.

5. COMPARISON OF RESULTS WITH OTHER RECOMMENDERS

As the employed dataset has unique characteristics, we implemented some of the state-of-the-art recommenders and run them on our training data for comparability reasons. Specifically, we used the pure user-to-user k-means collaborative filter [12, 18] and the Singular Value Decomposition-like matrix factorization algorithm ("SVD") by Funk [9], the foundation for all matrix factorization recommenders. As matrix factorization is known to provide the best predictive accuracy for a single algorithm, we suggest that a comparison to their basis algorithm to be informative. The factor model of "SVD" is learnt for the dimensionality of 200. The predictive accuracy (RMSE) of these algorithms is measured using the same data. Results are presented in Table 2.

Table 2. Comparison of the Prediction Accuracy of Different Recommendation Algorithms

Algorithm	RMSE	Provided explanation modes
Pure	24.67	Influence + keyword + pros-and-cons
Optimized	24.17	Influence + keyword + pros-and-cons
Collaborative Filtering	22.86	nearest neighbor
SVD	21.49	n/a
Hybrid	20.66	Influence + keyword + pros-and-cons / nearest neighbor

Our models are denoted as “Pure” (estimation model) and “Optimized” (optimization model) as well as “Hybrid” (combination of ”Optimized” and collaborative filter). The “Hybrid” is found to be the most accurate of the considered methods, followed by “SVD”, “Collaborative Filtering”, “Optimized” and “Pure”. The difference in accuracy of 4% between the two best performing methods is substantial and significant ($p < .10$). We find this particularly notable, as the Hybrid algorithm outperforms state-of-the-art recommender algorithms in terms of prediction accuracy, while also providing the majority of users with explanations at the most detailed level.

6. DISCUSSION AND FUTURE WORK

Even the most accurate recommendation algorithm is subject to prediction errors. Hence an explanation facility should be made an integral part of recommender systems which help users to make better choices. Our proposed hybrid method outperforms both collaborative filtering and matrix factorization approach in terms of predictive accuracy, while providing all users with explanations of the reasoning behind recommendations. However, for the smaller fraction the users the explanations are given without such detail. This may be due to that our proposed model fails to adequately capture the preference structure and/or item evaluation behavior for a number of users which might point at missing item attributes.

Future research directions which we would like to explore affect primarily the modeling side of our method, such as extending of the list of item attributes and adding interaction terms. The algorithmic part might also benefit from robust techniques for mitigating overfitting and by improved handling of part-worths multicollinearity. Further, it seems reasonable to employ some similarity-based techniques for increasing of the users’ representation through imputation of the part-worths. We believe that these improvements are capable of achieving the overall best prediction accuracy while providing all users with the motivated explanations at the highest detail level.

7. ACKNOWLEDGMENTS

The authors thank Tobias Bauckhage for providing the data for our tests, Denis Rechkin for many fruitful discussions, and three anonymous reviewers for their helpful comments.

8. REFERENCES

[1] Aksoy, L., Bloom, P.N., Lurie, N.H., and Cooil, B. Should Recommendation Agents Think Like People? *Journal of Service Research 8*, 4 (2006), 297-315.

[2] Ariely, D. Controlling the Information Flow: Effects on Consumers' Decision Making and Preferences. *Journal of Consumer Research 27*, 2 (2000), 233-248.

[3] Austin, B.A. *Immediate Seating - A Look at Movie Audiences*. Belmont, California, Wadsworth Inc. 1989.

[4] Bao, X., Bergman, L., and Thompson, R. Stacking Recommendation Engines with Additional Meta-features. *RecSys '09*, (2009), 109-116.

[5] Burke, R. Hybrid Web Recommender Systems. In P. Brusilovsky, A. Kobsa and W. Nejdl, *The Adaptive Web*. Springer, Berlin, 2007, 377-408.

[6] Cramer, H., Evers, V., Ramlal, S., et al. The Effects of Transparency on Trust in and Acceptance of a Content-Based Art Recommender. *User Modeling and User-Adapted Interaction 18*, 5 (2008), 455-496.

[7] Elliott, G. and Timmermann, A. Optimal forecast combinations under general loss functions and forecast error distributions. *Journal of Econometrics 122*, 1 (2004), 47-79.

[8] Fildes, R. and Ord, K. Forecasting Competitions: Their Role in Improving Forecasting Practice and Research. In M.P. Clements and D.F. Hendry, *A Companion to Economic Forecasting*. Blackwell Publishers, Oxford, 2001, 322-353.

[9] Funk, S. Netflix Update: Try This at Home. 2006. http://sifter.org/~simon/journal/20061211.html.

[10] Gunawardana, A. and Meek, C. A Unified Approach to Building Hybrid Recommender Systems. *RecSys '09*, (2009), 117-124.

[11] Hennig-Thurau, T., Houston, M.B., and Walsh, G. The Differing Roles of Success Drivers Across Sequential Channels: An Application to the Motion Picture Industry. *Journal of the Academy of Marketing Science 34*, 4 (2006), 559-575.

[12] Herlocker, J., Konstan, J., Borchers, A., and Riedl, J. An Algorithmic Framework for Performing Collaborative Filtering. *SIGIR*, ACM (1999), 230-237.

[13] Koren, Y. Collaborative filtering with temporal dynamics. *KDD'09*, (2009), 447-456.

[14] McNee, S.M., Riedl, J., and Konstan, J.A. Being Accurate is Not Enough: How Accuracy Metrics have hurt Recommender Systems. *CHI'06*, (2006), 1097-1101.

[15] McSherry, D. Explanation in Recommender Systems. *Artificial Intelligence Review 24*, 2 (2005), 179-197.

[16] O'Donovan, J. and Smyth, B. Trust in Recommender Systems. *IUI '05*, (2005), 167-174.

[17] Press, W.H., Teukolsky, S.A., Vetterling, W.T., and Flannery, B.P. *Numerical Recipes: The Art of Scientific Computing*. Cambridge University Press, 2007.

[18] Sarwar, B., Karypis, G., Konstan, J., and Riedl, J. Analysis of Recommendation Algorithms for E-commerce. *Proceedings of the 2nd ACM conference on Electronic commerce*, (2000), 158-167.

[19] Sinha, R. and Swearingen, K. The Role of Transparency in Recommender Systems. *Conference on Human Factors in Computing Systems*, (2002), 830-831.

[20] Stock, J.H. and Watson, M.W. Forecasting Inflation. *Journal of Monetary Economics 44*, 2 (1999), 293-335.

[21] Symeonidis, P., Nanopoulos, A., and Manolopoulos, Y. MoviExplain: A Recommender System with Explanations. *RecSys'09*, (2009), 317-320.

[22] Tintarev, N. and Masthoff, J. A Survey of Explanations in Recommender Systems. *ICDE'07*, (2007), 1-10.

[23] Ying, Y., Feinberg, F., and Wedel, M. Leveraging Missing Ratings to Improve Online Recommendation Systems. *Journal of Marketing Research 43*, 3 (2006), 355-365.