Exploiting Transparency in Recommendation Systems

Neil Getty and Mustafa Bilgic
Illinois Institute of Technology
ngetty@hawk.iit.edu, mbilgic@iit.edu

Abstract

There is a wealth of research showing the importance of explanations in recommendation system. Past research has primarily focused on comparisons of different explanation styles, how best to visualize the explanations, and the respective effects on the system and user. The use of explanations seem to be limited to the scope of promotion and satisfaction. What lacks is the use of user, model and prediction transparency to actively improve the model itself. This work surrounds a regression approach to a movie recommendation system utilizing GroupLens and IMDB datasets. Central to the implementation and results of the work is transparency. It will be shown that this transparency can be exploited to not only increase user satisfaction and trust, but as a means for improving future predictions and recommendations.

1. INTRODUCTION

Recommendation systems are ubiquitous, particularly as commerce and entertainment applications. Amazon and Netflix have probably the most well known examples of recommendation systems surrounding their products in these respective areas. Few recommendation systems exist without some level of explanation. Ppossible goals of explanations are described by Tintarev and Masthoff in their 2012 work on *Evaluating the Effectiveness Of Explanations For Recommender Systems*, see Table 1.

Table 1. Explanatory Aims

Aim	Definition
Transparency	Explain how the system works
Scrutability	Allow users to tell the system it is wrong
Trust	Increase users confidence in the system
Effectiveness	Help users make good decisions
Persuasiveness	Convince users to try or buy
Efficiency	Help users make decisions faster
Satisfaction	Increase the ease of use or enjoyment

The effect of explanations does not end with these aims, rather each aim may be viewed as an intermediary step in improving the performance of the system.

1.1. Motivation

While there are many commercial movie recommendation systems, most employ a black-box approach towards their recommendations. Even research based systems like MovieLens only offer a user profile (rating/year distribution), genre profile, and explanation of the methodology behind the 4 usable algorithms. No specific evidence is given for why a particular film is designated as a *top pick*. This trend may be indicative of a lack of improvement offered by further explanations, though the current research seems to disagree.

Another possibility is that the average user does not see the warrant for further explanation, would view it as adversely complex or any further interaction overly time consuming. It is imperative that these considerations be taken into account when developing recommendation systems that exploit explanations. These considerations make it clear that if a system is to improve through the use of explanations it must present the proper information and elicit the most meaningful information.

2. Method

The work was developed in python using primarily the following libraries: numpy, pandas, sklearn, matplotlib and pickle.

2.1. Data Sets

MovieLens Grouplens provides rating data sets of varying size. The smallest dataset of 100,000 ratings was used to maximize testing speeds. Each rating is an integer from 1 to 5. The set contains 943 users rating 1682 movies. Each user rated a minimum of 20 movies.

Along with ratings, the set also contains genre data for each movie. User demographic and occupation information is also accessible, though they were not used.

Internet Movie Database Movies in the Grouplens data set were matched to those in the IMDB set using regular expressions. Many title discrepancies exist between the two sets, with no unique identifiers in the later. 80% of the films were successfully matched.

The IMDB data set contains extensive information including genre, keyword, actor, director, plot and quotes. Only keyword and actor data was used. Only keywords which described atleast 50 films and actors which appeared in atleast 10 films were used.

It may be worth nothing that the IMDB data set is quite inconsistent. Entries do not follow a standard convention across sets or even internally. Delimiters proved to be unreliable and regular expressions were used to preprocess the sets.

2.2. Preprocessing

The following main steps were taken to extract the data sets used to train the classifiers.

1. Aggregate user rating vectors.
2. Match films between the two sources.
3. Generate keyword vectors for each film.
4. Generate actor vectors for each film.
5. Generate actress vectors for each film.
6. Combine features into a single matrix (genre, average rating, number of ratings, keywords, actors)
7. For each user intersect rated movie vector with feature matrix.
8. Scale non binary features.
9. Train regression model for each user using 5 fold cross validation.

2.3. Classification

Using 5 fold cross validation, linear, ridge and lasso regression classifiers were trained and compared with mean average error. Linear regression produced often poor and wildly inconsistent results. Ridge and lasso regression were fairly consistent and resulted in acceptable MAEs.

The models were trained and tested systematically with increasing alpha (regularization) values until optimal values were found. Using these optimal values, new models were then trained with a user's entire set of rated movies in place of cross validation. As the data was very sparse, this was important for obtaining nonzero coefficients and therefore more meaningful explanations.

2.4. Display of Results

Results were displayed using iPython notebook. Three main sections were organized, Dataset, Model, and User (or prediction) transparency.

3. Results

Dataset Transparency This section includes the rating and genre distributions and the most frequent keywords and actors. The rating distribution is left skewed, most movies were rated 3 or 4.

Model Transparency For the notebook 50 users are sampled using 5 fold cross validation with both lasso and ridge regression. The MAE over the entire sample is shown for each alpha tested in the tables below.

Table 2. Lasso Regression Results

Alpha	MAE
0.05	0.772237
0.10	0.750110
0.15	0.745936
0.20	0.751146
0.25	0.761458
0.30	0.772844
0.35	0.784824
0.40	0.796035
0.45	0.806092
0.50	0.814429
0.55	0.821334

The models were then trained with the optimal alpha value and the MAEs reported. For lasso regression the minimum MAE was 0.4814 and the maximum was 1.3591. For ridge regression the minimum MAE was 0.4695 and the maximum was

Table 3. Ridge Regression Results

Alpha	MAE
11	0.775779
13	0.772757
15	0.770548
17	0.768895
19	0.767714
21	0.766819
23	0.766149
25	0.765661
27	0.765309
29	0.765081
31	0.764963

1.3352. Clearly there is not a large difference between the two classifiers.

User Transparency A user is chosen and the average rating and number of movies rated is displayed along with charts/tables depicting the rating and keyword distributions. A lasso regression classifier is trained with 5 fold cross validation and the results are displayed as in Table 4.

A model is then trained with the full training set. Unfortunately as this user has relatively few ratings there are only three non-negative coefficients. Average Rating is positive, explosion and good-versus-evil are negative. It is intuitive that Average Rating is sufficient for the model if the user's tends to have ratings consistent with the average.

The significant explanations that follow are the movies with the highest and lowest ratings, the movies with the most positive and most negative evidence, the movie with the most conflicting evidence and the movie with the least evidence.

Evidence for the movie with the most positive evidence is displayed:

Movie Title: Shawshank Redemption, The (1994) User Rating: 5 Average Rating: 4.44210526316 Number of Ratings: 285

Prediction: 3.54895598899 Bias and evidences: 2.9845013378 0.0 0.56445465119 Positive Features Feature Weights 0 Average Rating 0.5645 Negative Features Empty DataFrame Columns: [Feature, Weights] Index: []

The importance of more data is clear. This explanation is meaningless under the circumstances, as there is barely any evidence one way or the other and the only significant feature here is average rating. It is also possible that a different alpha value should be calculated for each user, though this would be costly.

Finally the top and bottom recommendations for the user are displayed in a table for the user, again using 5 fold cross validation.

4. Future Steps

A large improvement in the results will likely come from the adaption of the larger GroupLens data set with 20 million ratings.

Besides gathering users with more ratings, more features may be added to the system. The most significant absent features include director, user demographic and occupation, and plot. Plot may be represented with a simple bag of words approach or a more complex method.

Additionally, better results as well as more descriptive and varied explanations may come from a hybrid system using collaborative filtering. Filtering may be done both user-user and item-item wise, with explanations of both.

The user must also be able to utilize any meaningful explanations through interaction with the system. If the user sees that the system believes them to like/dislike certain genre/keywords/actors yet they adamantly disagree the system should adapt.

Explanations may go beyond the rating scale with sufficient evidence. It may be the case that the user likes movies with strictly positive evidence rather than a movie with mixed evidence but a similar rating. These situations should be made transparent.

The user should also see what features are most significant to their predictions. By agreeing/disagreeing with or reordering the significant features the model may learn a great deal with minimal work from the user.

Finally the knowledge of the most positive/negative/conflicting/unknown films must be exploited to intelligently ask the user for more information. If a particular movie has no evidence it is because the user has not rated many movies with similar features. By recommending a film that the user may like but which also contains unknown features, the system may learn more definitively (for better or worse).

5. CONCLUSIONS

The results of the conducted experiments were not particularly meaningful, though the aim of the experiment warrants further exploration. Explanations and

Table 4. Lowest Errors

Movie ID	Average Rating	Error from Avg	User Rating	Model Prediction	Model Error
120.0	3.43619489559	0.436194895592	3.0	3.01855463838	0.0185546383841
393.0	3.30769230769	0.307692307692	3.0	2.96788745183	0.0321125481685

transparency still appear to be a viable means for a system to actively learn from the user without burdening or becoming overly complex. Once the previously discussed enhancements are enacted the results should allow for further experimentation.

References

[1] Al-Taie, Mohammed Z., and Seifedine Kadry. ”Visualization of Explanations in Recommender Systems.” JOAMS Journal of Advanced Management Science 2.1 (2014): 140-44.

[2] Bilgic, Mustafa, and Raymond Mooney. ”Explaining Recommendations: Satisfaction vs. Promotion.” In Proceedings of Beyond Personalization 2005, the Workshop on the Next Stage of Recommender Systems Research (2005): 13-18.

[3] Gedikli, Fatih, Dietmar Jannach, and Mouzhi Ge. ”How Should I Explain? A Comparison of Different Explanation Types for Recommender Systems.” International Journal of Human-Computer Studies 72.4 (2014): 367-82.

[4] Herlocker, Jonathan L., Joseph A. Konstan, and John Riedl. ”Explaining Collaborative Filtering Recommendations.” Proceedings of the 2000 ACM Conference on Computer Supported Cooperative Work - CSCW ’00 (2000):

[5] Rombouts, Jaldert, and Tessa Verhoef. ”A Simple Hybrid Movie Recommender System.” (2002)

[6] C. J. Kaufman, Rocky Mountain Research Lab., Boulder, CO, private communication, May 1995.

[7] Rubens, Neil, Dain Kaplan, and Masashi Sugiyama. ”Active Learning in Recommender Systems.” Recommender Systems Handbook (2010): 735-67.

[8] Tintarev, Nava, and Judith Masthoff. ”Evaluating the Effectiveness of Explanations for Recommender Systems.” User Model User-Adap Inter User Modeling and User-Adapted Interaction 22.4-5 (2012): 399-439. IEEE Trans. Electron Devices, vol. ED-11, pp. 3439, Jan. 1959.

[9] Tintarev, Nava, and Judith Masthoff. ”A Survey of Explanations in Recommender Systems.” 2007 IEEE 23rd International Conference on Data Engineering Workshop (2007)