A Movie Genre Prediction Based on Multi-Variate Bernoulli Model and Genre Correlations

Eric Arnaud Makita Makita, Artem Lenskiy
School of Electronics, Electrical and Communication Engineering,
Korea University of Technology and Education.
lensky@koreatech.ac.kr

***Abstract*—In this paper, a movie category based on Bayesian model and categories correlations is proposed. Although several methods have been reported on improving the user satisfaction based on unexpectedness metric, to the best of our knowledge, predicting items' categories rather than predicting items' rating is a new attempt. This in turn completes the items' categories given by experts and improves user satisfaction by providing a surprise effect in the recommendations given to users. We employ Bernoulli multi-variate model to estimate a likelihood of a movie given category and the Bayes rule to evaluate the posterior probability of a genre given a movie. Experiments with the MovieLens dataset validate our approach.**

***Index Terms*—Recommender Systems; Genre Prediction; Movie Recommender; Multivariate Bernoulli Model; Naïve Bayes Classifier.**

I. INTRODUCTION

Nowadays, web users are no longer considered as consumers of information but also as active sources that generate large volume of data online. Consequently, the amount and diversity of information on Internet increase exponentially. The entire information cause losses of time to users browsing through information without any guarantee of finding what they are looking for. Aiming to solve these problems, researchers in the academia and/or industries have suggested the use of recommender systems [1] that overcome the information overload by facilitating search and access to information by providing to users relevant items at possibly shortest time. In this context, items can be of any kind, namely a movie to watch, a soundtrack to listen, a webpage to click on, or else. Among the widely proposed recommendation techniques, content-based filtering [3, 4] and collaborative filtering algorithms [5, 6] have been the most famous one in the literature [2]. The content-based filtering is made under the assumption that users' future preferences are similar to those they liked in the past, while the collaborative filtering recommendation is made under the assumption that if two users have similar tastes in the past, they will have similar taste or preferences in the future. Between them, collaborative filtering is the most widely used, therefore attracts more interests from researchers [7].Collaborative filtering techniques are ratings-oriented and count a huge number of users who provide fewer ratings than the items they consume. Taking that into consideration, questions such as how to alleviate the data sparsity while increasing the recommendation accuracy are the main concerns in the related works.

Recently, some approaches considering factors outside users' ratings have been proposed in the literature [7]. Since recommender systems can naturally be applied in various fields where items are categorized, their associated datasets provide not only users' ratings but also items categories. Based on this information, many recommender systems extension have been made.

In this paper, we propose a movie genre prediction based on users' ratings. We apply multi-variate Bernoulli model to estimate likelihoods that are used in naïve Bayes rule to predict movies' genres. We also calculated the genres correlations, to check if incorrectly predicted genre is correlated the correct one. In general, prediction item category recommender is important in sense that, it can enhance the items' categories given by expert therefore increase the user satisfaction by providing surprising recommendations.

The proposed approach has the following combination of contributions in order to expand traditional recommender systems.

1. We proposed a new approach that expands the traditional recommender systems by predicting a category of an item under evaluation.
2. Bernoulli multi-variate models are used to learn movies' likelihoods of belonging to a particular genre.
3. Bayesian probabilistic reasoning is applied to predict genres. To the best of our knowledge, item's category prediction recommendation is a new attempt.
4. We provide an experimental study of our technique using the MovieLens dataset. Experimental results show the correctness of our proposed approach.

In the remainder of this paper, we present the details of our model with the following organization. Section 2 presents various item genre-oriented recommendations. In Section 3, we describe the data model and the mechanism of our proposed technique. Section 4 contains the performance study. Finally in Section 5, we summarize our work.

II. RELATED WORK

A large body of research on incorporating information about genres or category for recommending items has been performed.

Manzato [8] proposed a movie recommender system that enhances the accuracy and overcome the traditional recommendation by factorizing user-genre matrix rather than the user-item matrix. The factorized user-genre matrix model is used to discover latent factors from genres in order to enrich user's profiles. In [9] content-based filtering using

user category-based filtering was proposed in order to overcome one of the major issues of recommender systems named item cold star. Item cold star refers to new items that have not received enough users' feedback, thus could decrease the accuracy of the recommendation. Another example of category-based recommendation is proposed in [10], where authors presented a framework called SEP for overcoming recommender systems problems such as cold-start and sparsity. The authors in [11] used the available movie genre information to compute the recommendations by matching the users' preferred genre with the genre correlation matrix. In [12], authors proposed a recommender systems approach that uses genre information in order to address not only the coverage into the recommender systems algorithms but the redundancy. Most of the related works focus on designing new approaches to find similarities between users, while the prediction of movie's genres remain under studied, to the best of our knowledge. However, it could play an important role in recommending novel items to the user.

III. Proposed Method

The proposed method applies the well-known Bernoulli model for calculating the conditional probability of movie being of a particular genre. To describe our idea clearly, we initially give some definitions used in this paper: *user set*: $U = \{u_1, u_2, \dots\}$; *movie set*: $M = \{m_1, m_2, \dots\}$; *genre set*: $G = \{g_1, g_2 \dots\}$; *rating set*: $R = \{r_1, r_2, \dots\}$.

A movie $m_i \in M$ is characterized by binary feature vector, which components set to one if the corresponding user$u_t$rated the movie $m_i$as *r* otherwise zero. That is to say:

$$v_{t,i}(r) = \begin{cases} 1 & if\ \ u_t\ rated\ movie\ m_i as\ r. \\ 0 & otherwise \end{cases} \quad (1)$$

Assuming that the ratings of one user do not depend on ratings of other users, the conditional probability of a movie, m_i, given genre g_j, is computed according to multi-variate Bernoulli model as follows:

$$P(m_i|g_j, r) = \prod_{t}^{|U|} \left[v_{t,i}(r) P(u_t|g_j, r) + \left(1 - v_{t,i}(r)\right)\left(1 - P(u_t|g_j, r)\right) \right] \quad (2)$$

where $v_{t,i}(r)$ is either 0 or 1 indicating whether the useru_t rated the movie of genre$g_j$as $r$or not. A movie can be seen as a collection of multiple independent Bernoulli experiments, one for each user in the user set$U$with the probabilities for each of these rating events defined by each component$P(u_t|g_j, r)$. The probability $P(u_t|g_j, r)$defines the probability of user u_t given a rating r to a movie labeled as g_j.

$$P(u_t|g_j, r) = \frac{1 + \sum_{i=1}^{|M|} v_{t,i} P(g_j|m_i, r)}{2 + \sum_{i=1}^{|M|} P(g_j|m_i, r)} \quad (3)$$

The probability $P(u_t|g_j, r)$can be thought as a user's preference model towards movie genres. In other words, knowing the genre and the rating, Equation (3) describes the probability that user $u_t$match a hypothetical user that wouldrate a movie of genre g_j as r. The probability $P(g_j|m_i, r)$is *1* if $m_i$is marked only as genre g_j otherwise *0*. If m_i belong to *N* genres at the same time, the probability is 1/*N*. To avoid the zero probability that can occur for situations where a user $u_t$did not rate a movie i we added one to the numerator and two to the denominator according to the Laplace's law of succession.

The posterior probability of a genre, given a movie and the rating is calculated as follows.

$$P(g|m_i, r) = \frac{P(m_i|g, r)P(g, r)}{P(m_t)} \quad (4)$$

Given the posterior probability (4), for each rating we predict a movie $m_i$as a genre$g$according to the highest posterior probability as follows:

$$k_r = \underset{g_k \in G}{arg\,max}\, P(g_k|m_i, r) \quad (5)$$

IV. Experimental Evaluation

A. Dataset

We begin with the description of the MovieLens dataset and parameters used in our experiments.

We performed our experiments on the MovieLens 100K dataset [13], which contains 1,682 movies, 943 users, and 100,000 ratings that range from range of 1 to 5. In this dataset, each user has rated at least 20 movies.18 movies genres were selected here and each movie as at least one genre where it belongs too. We carried out the experiment by dividing the dataset into two, a training set and a test set. During the training phase rows of the genre matrix and columns in the rating matrix that correspond to the testing set were removed. Thus, the users' preference models were built only using a portion of the available rated items with known genres.

B. Evaluation

In this section, we demonstrate the correctness of our proposed method.

Movies were selected randomly in both of our training and testing approaches. To test the prediction power of the proposed model we varied the size of the training set from 5% to 95%, with 5% step. For example, if 5% of the data were used for training then the remaining 95% were in the testing set, which will be used to assess the performance of our recommendation algorithm.

The first step in our algorithm is to estimate $P(u_t|g_j, r)$ based on the user ratings, that is described by (3). Figure 1, depicts the probability of the user u_t rating a movie of category g_j as rating $r = 1$. Figure 2., illustrates the probability of the selected users that rated the movie m_i as rating $r = 1$ among all the users or $v_{t,i}(r)P(u_t|g_j, r)$ that corresponds to the probability of success in the Bernoulli model. The probability of a failure is given by$\left(1\text{-}v_{t,i}(r)\right)\left(1\text{-}P(u_t|g_j, r)\right)$. Then the likelihood is calculated according to (2), which is used to predict the genre via Bayes rule (4).

When the prediction is incorrect, the genre correlation matrix (Figure 3) is used to check whether the incorrectly predicted category is correlated with the true category. If it is correlated, we accept this prediction as a prediction of a similar genre.

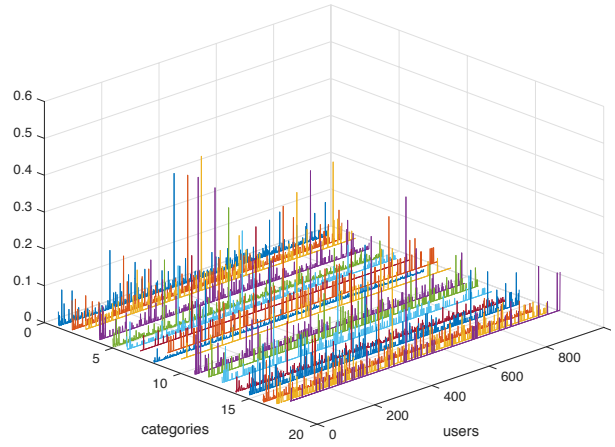

Figure 1: Preference models for rating 1

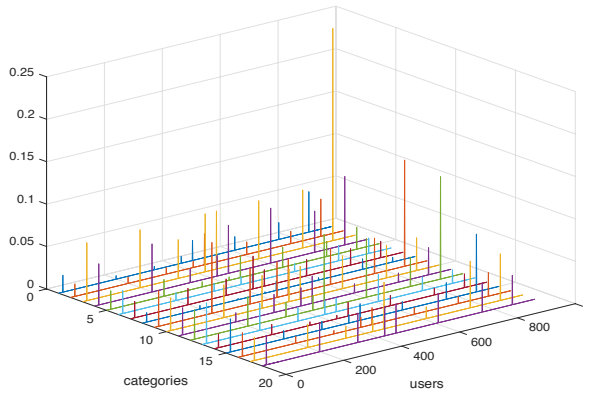

Figure 2: Probability for a movie m_i beingrated as r=1

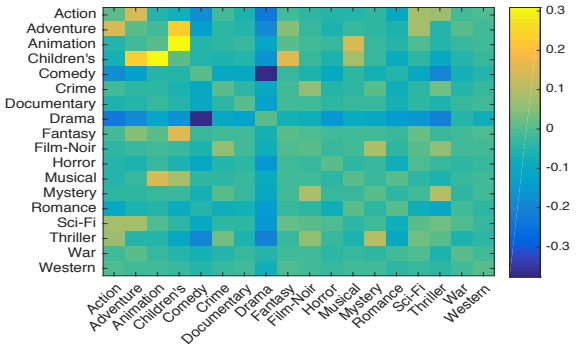

Figure 3: Movie correlation matrix

To measure the accuracy of the proposed approach on movie genre recommendations, we plot the predictions based on every rating from one to five. For every training size, we repeated the process of randomly selecting training samples 20 times.

Figures 4 to 8 show the prediction rate of our approach with and without including correlated genres.

These plots indicate that our movie category prediction based on the Bayesian model presented in Section 3.0 is effective for the lower rating. As the size of the training dataset increases, the prediction accuracy increases especially for r = {1,2}.

In contrast, the prediction accuracy for the rating r = 5 suddenly changes and presents the worse performance in our model. This situation lead us to focus on improving the accuracy of that rating r=5 in our future work.

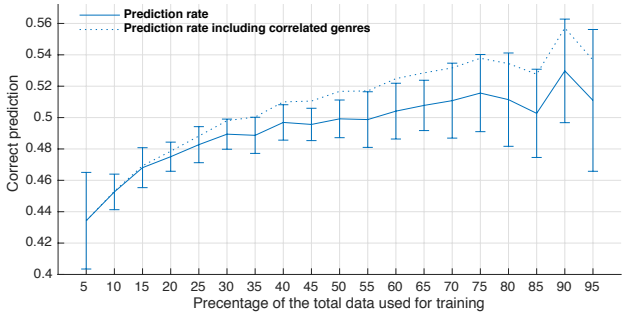

Figure 4: Prediction accuracy based on rating r=1.

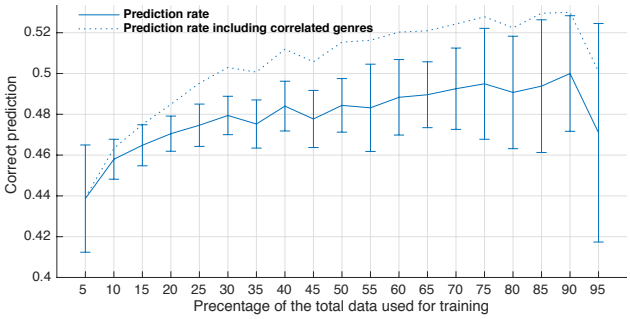

Figure 5: Prediction accuracy based on rating r=2.

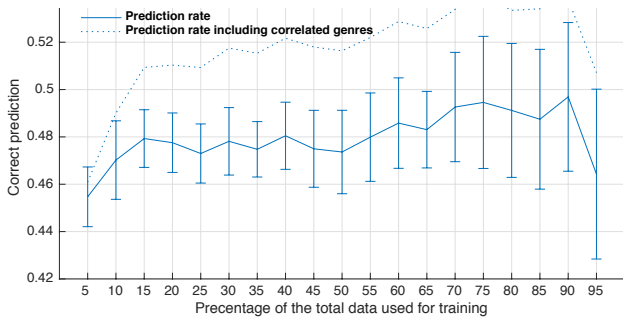

Figure 6: Prediction accuracy based on rating r=3.

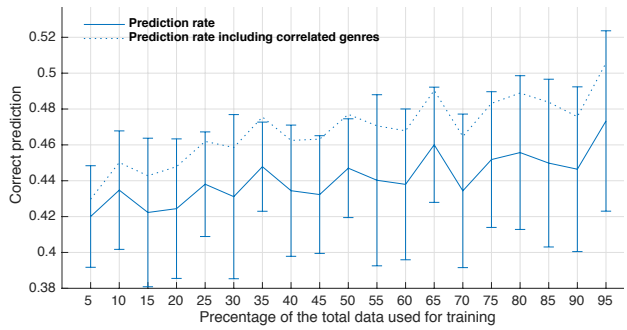

Figure 7: Prediction accuracy based on rating r=4.

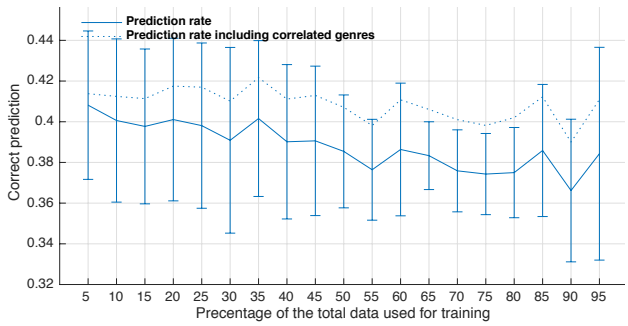

Figure 8: Prediction accuracy based on rating r=5.

V. CONCLUSION

Over the last decade, recommender systems have been successfully applied into various domains such as social networking, or online movie website, or e-commerce, etc. Until now, many recommender systems have been proposed and reported in the literature. However, all of them are item ratings' prediction-oriented. In this paper, we proposed an approach that expands the traditional recommender systems algorithms by predicting a category of an item under evaluation rather than predicting a rating of an item. Predicting a category of an item might help increasing the accuracy of the recommendation by complementing the categories of the items given by expert.

To show the correctness of our approach, we conduct an experiment study with MovieLens dataset. The experimental results show that predicting the category of an item under evaluation can achieve 50% accuracy rate based on 50% training set of users' rating 1. This work may find wide applications in practice. For instance, it can complement the genres given by the experts. It could significantly increase the accuracy and usefulness of recommendations.

We also show in our experimental analysis that predictions based the higher ratings do not follow the behavior the predictions based on the low ratings. This situation can be seen as an interesting open issue for our future where the focus will be on improving the prediction for higher ratings. Related study may lead to design new attempts in the field of recommender systems.

REFERENCES

[1] Lü, L. Medoe, M. Yeung, C. H. Zhang, Y-C. Zhang, Z-K. Zhou, T. 2012. Recommender systems, Physics Reports. 519: 1-49.

[2] Park, D. H. Kim, H. K. Choi, I. Y. and Kim, J. K. 2012. A literature review and classification of recommender systems research, Expert Syst. Appl. 39: 10059-10072.

[3] Bogers T. and Bosch, A. v. d. 2009. Collaborative and Content- based Filtering for Item Recommendation on Social Bookmark- ing Websites", In Proceedings of the ACM RecSys '09 workshop on Recommender Systems and the Social Web: 9–16.

[4] Basilico J. and Hofman, T. 2004. Unifying Collaborative and Content-Based Filtering, In Proceedings of the 21st International Conference on Machine Learning, Banff, Canada.

[5] Su X. and T. Khoshgoftaar, M. 2009. Review Article A Survey of Collaborative Filtering Techniques", Advances in Artificial Intelligence Volume 2009, 19 pages.

[6] Braida, F. Mello, C. E. Pasinato M. B. and Zimbrão, G. 2015. Transforming collaborative filtering into supervised learning", Expert Syst. Appl. 42: 4733-4742.

[7] Huang, Z. Chen H. and Zeng, D. 2004. Applying associative retrieval techniques to alleviate the sparsity in Collaborative filtering, ACM Transaction on Information Systems, 22(1):116-142.

[8] Manzato, M. G. 2012. Discovering Latent Factors from movies genres for enhanced recommendation, in Proceedings of the 6th ACM Conference on Recommender Systems, RecSys'12: 249-252.

[9] Sollenborn M. and Funk, P. 2002. Category-based filtering and user stereotype cases to reduce the latency problem in recommender systems, in 6th European Conference on Case Based Reasoning, ECCBR2002: 395-405, Springer.

[10] Tilwani R. and Tiwari, S. 2013. Implementation of category based recommendation module of SEP architecture using PBTA, International Journal of Advanced Research in Computer Science and Software Engineering, 3.

[11] Choi, S. M., Ko S. K. and Han,Y. S. 2012. A movie recommendation based on genres correlations", ExpertSyst.Appl. 39 7.

[12] Vargas, S. Baltrunas, L. Karatzoglou A. and Castells, P. 2014. Coverage, Redundancy and size-awareness in genre diversity for recommender systems, RecSys'14.

[13] http://grouplens.org/datasets/movielens/100k/