DLRRS: A New Recommendation System Based on Double Linear Regression Models

Chenglong Li[1], Zhaoguo Wang[2], Shoufeng Cao[1](✉), and Longtao He[1]

[1] National Computer Network Emergency Response Technical Team/Coordination Center of China (CNCERT/CC), Beijing 100029, China
{lichenglong, csf, hlt}@cert.org.cn

[2] School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150006, Heilongjiang, China
wangzhaoguo@tsinghua.edu.cn

Abstract. Recently, it is difficulty for ordinary users to find their own points of interest when facing of massive information accompanied by the popularity and development of social networks. Recommendation system is considered to be the most potential way to solve the problem by profiling personalized interest model and initiatively pushing potential interesting contents to each user. However, collaborative filtering, one of the most mature and extensively applied recommender methods currently, is facing problems of data sparsity and diversity and so on, causing its effect unsatisfactory. In the article, we put forward DLRRS, a new recommendation system depending on double linear regression models. Compared with the traditional methods, such as item average scores, collaborative filtering, and rating frequency, DLRRS has the best predictive RMSE accuracy and less fluctuation. DLRRS also has high real-time performance, which makes the system complete all the calculations in the time of Ω(n).

Keywords: Recommendation system · Linear regression · RMSE

1 Introduction

The popularity and growing of social networks have changed the way people passively access information in last several years. And the content generated by users has exploded. For ordinary users, it is difficult to find their own points of interest when facing of massive information. The web portals, such as Yahoo, USA.gov, etc., help users quickly index by sorting information with their attributes. And the search engines, such as Google, Baidu, etc., return the most relevant content by analyzing the queries entered by the user. Although they greatly improve the efficiency of the information accessing, they need for users' close participation, and cannot automatically perceive the users' interests. Moreover, the users are often confused of their real demands, or cannot use keywords to describe his/her own interests. In addition, the results returned from classification and searching technology lack personality which causing poor user experience. By analyzing the user's historical behavior, recommendation system [1]

G. Sun and S. Liu (Eds.): ADHIP 2017, LNICST 219, pp. 242–249, 2018.
https://doi.org/10.1007/978-3-319-73317-3_29

profiles personalized interest model for each user, and initiatively pushes potential interesting content to the user. Therefore recommendation system is as being the most potential method for solving the information overload issue.

In the article, DLRRS, a new recommender system depending on double linear regression models is presented. With the prepared inputs, DLRRS establishes double linear regression models of the certain score and the highest frequency score of user or item by using the frequency information of the user or the item, and then uses the models to predict the unknown score directly according to the historical score frequency as the system outputs. Compared with the traditional methods, DLRRS has the best predictive accuracy in the term of RMSE and less fluctuation. DLRRS greatly reduces the computational complexity, which makes the system complete all the calculations in the time of $\Omega(n)$. So it is easy to be applied to the actual industrial production. Using the group wisdom and the statistical parameters to estimate the model parameters, DLRRS has a good anti-noise ability. DLRRS also has a good capability of the incremental update, which makes the system complete update to the new user behavior in constant time, leading to high real-time performance.

2 Related Work

A recommender system is defined as: "attempt to recommend the most suitable items (products or services) to particular users (individuals or businesses) by predicting a user's interest in an item based on related information about the items, the users and the interactions between items and users" [2]. Currently, the most widely used personalized recommendation systems mainly depend on the collaborative filtering-based methods. Collaborative filtering systems mainly use two kinds of methods [3]: heuristic-based approaches [4–7] and model-based approaches [8–11].

The heuristic-based methods obtain the user score matrix by making use of the hidden or explicit behavior of the user firstly. And then it calculates the similarity between items or users. Finally, according to the score and similarity of neighbor users or items, the forecast score and the recommended results are achieved. The heuristic-based method could be further divided into user-based approach [12] and object-based approach [13]. Because of its ease of deployment and efficient features, heuristic-based methods are now widely used in commercial systems such as Amazon. However, the sparseness, diversity of data, and other issues make the recommendation performance of heuristic-based methods difficult to improve.

To elevate recommending preciseness, the model-based approaches exploit item scoring matrix for training more accurate scoring models, such as clustering [14, 15], Bayesian belief network [16], Markov decision process [17] and the potential semantic model [18], etc. Although the model-based approach improves the prediction accuracy, they also face problems such as complex model, various parameters and strong dependency on large statistical properties of the data set. The above reasons also cause model-based methods difficult to apply to practical recommendation systems.

In the article, we put forward DLRRS, a new recommender system depending on double linear regression models. DLRRS establishes double linear regression models, and then uses the models to predict the unknown score directly according to the historical score frequency. Compared with the traditional methods, DLRRS has the best predictive accuracy in the term of RMSE and less fluctuation. DLRRS also improves time performance which is easy to be applied to the actual industrial production.

3 The System Implementation of DLRRS

According to the system workflow, DLRRS consists mainly of three parts: data preparation, establishing double linear regression models and prediction result output, described as followings.

3.1 Data Preparation

Firstly, DLRRS requires a certain size of known data to prepare for subsequent modeling and output of the results. Specifically, the data which should be prepared in advance mainly need to include three elements: users, items and user ratings of items. Through the known data, after the subsequent modeling, and calculation the predicted unknown user ratings could be obtained consequently.

3.2 Double Linear Regression Models

In the second step, we use linear regression method to establish double models: a model between the user's highest frequency score and items' scores, and a model between scores of all items and scoring frequency of the corresponding items.

The standard linear regression model is described as following formulas [19]. Given a dataset $\{y_i, x_{i1}, \ldots, x_{ip}\}_{i=1}^{n}$ of n statistical blocks, the model assumes that the relationship between the dependent variable y_i and the p-vector of regressors x_i is linear. With error variable ε_i, the model takes the form:

$$y_i = \beta_0 1 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \varepsilon_i = x_i^T \beta + \varepsilon_i, i = 1, \ldots, n. \qquad (1)$$

where T represents the transpose, so that $x_i^T \beta$ is the inner product between vectors x_i and β.

Using the above linear regression method, the first model between the user's score for the items and the user's highest frequency score could be established. And then we could use the model to predict and score non-rated items of the target users. First of all, traverse all users, and each user u makes an n-dimensional vector of the historical ratings of all the evaluated items, where n is the items' count evaluated by the customer u, i.e. $\mathrm{Y_u} = \left[\mathrm{r_{u,i_1}}, \mathrm{r_{u,i_2}}, \ldots, \mathrm{r_{u,i_k}}, \ldots, \mathrm{r_{u,i_n}}\right]$, where $\mathrm{r_{u,i_k}}$ represents the customer u's score on item $\mathrm{i_k}$. Then it calculates the highest score in the historical score of the item involved in $\mathrm{Y_u}$ and divides the result into the vector $\mathrm{X_u}$ in the order of the items in $\mathrm{Y_u}$, i.e. $\mathrm{X_u} = [\mathrm{x_{i_1}}, \mathrm{x_{i_2}}, \ldots, \mathrm{x_{i_k}}, \ldots, \mathrm{x_{i_n}}]$, where $\mathrm{x_{i_k}}$ indicates the highest score for the historical score of the item $\mathrm{i_k}$. Assume $\mathrm{Y_u}$ and $\mathrm{X_u}$ satisfy the relation $\mathrm{Y_u} = \beta_\mathrm{u} \mathrm{X_u} + \varepsilon_\mathrm{u}$, where β_u and ε_u are real numbers. Applying the least squares method [20], the empirical fitting equation of the above relation is described as followings:

$$Y_u = \beta_u X_u + \varepsilon_u \\ \begin{cases} \beta_u = \frac{L_{uy}}{L_{ux}} \\ \varepsilon_u = \bar{y}_u - \beta_u \bar{x}_u \end{cases} . \tag{2}$$

where $\bar{x}_u = \frac{1}{n}\sum_{j=1}^{n} x_{i_j}$, $\bar{y}_u = \frac{1}{n}\sum_{j=1}^{n} r_{u,i_j}$, $L_{ux} = \sum_{j=1}^{n}\left(x_{i_j} - \bar{x}_u\right)^2 = \sum_{j=1}^{n} x_{i_j}^2 - n\bar{x}_u^2$ and $L_{xy} = \sum_{j=1}^{n}\left(x_{i_j} - \bar{x}_u\right)\left(r_{u,i_j} - \bar{y}_u\right) = \sum_{j=1}^{n} x_{i_j} r_{u,i_j} - n\bar{x}_u\bar{y}_u$.

Similarly, we propose the second model between the score of all items and scoring frequency of the corresponding items depending on score conditions to predict scores. First of all, traverse all the items, each object i constitutes an m-dimensional vector Y_i of all the historical score of i, i.e. $Y_i = [r_{u_1,i}, r_{u_2,i}, \ldots, r_{u_k,i}, \ldots, r_{u_m,i}]$, where $r_{u_k,i}$ represents the customer u_k's score on item i. Then it calculates the highest score of the user's historical scores involved in Y_i. And the results make up the vector X_i in the order of the users in Y_i, i.e. $X_i = [x_{u_1}, x_{u_2}, \ldots, x_{u_k}, \ldots, x_{u_m}]$, where x_{u_k} is the highest rated score in the user u_k's historical scores. Assume Y_i and X_i satisfy the relation $Y_i = \beta_i X_i + \varepsilon_i$, where β_i and ε_i are real numbers. Applying the least squares method, the empirical fitting equation of the above relation is described as followings:

$$Y_i = \beta_i X_i + \varepsilon_i \\ \begin{cases} \beta_i = \frac{L_{iy}}{L_{ix}} \\ \varepsilon_i = \bar{y}_i - \beta_i \bar{x}_i \end{cases} . \tag{3}$$

where $\bar{x}_i = \frac{1}{m}\sum_{j=1}^{m} x_{u_j}$, $\bar{y}_i = \frac{1}{m}\sum_{j=1}^{m} r_{u_j,i}$, $L_{ix} = \sum_{j=1}^{m}\left(x_{u_j} - \bar{x}_i\right)^2 = \sum_{j=1}^{m} x_{u_j}^2 - m\bar{x}_i^2$, and $L_{iy} = \sum_{j=1}^{m}\left(x_{u_j} - \bar{x}_i\right)\left(r_{u_j,i} - \bar{y}_i\right) = \sum_{j=1}^{m} x_{u_j} r_{u_j,i} - m\bar{x}_i\bar{y}_i$.

3.3 Prediction Result Output

In the final step, using the recommendation method of linear regression, the previously obtained predicted outcome is merged as the result of the user's evaluation of the items. Firstly, we use the most frequently occurred score X_i of the historical scores of the predicted items as the input of formula (2). Thus the forecast score Y_u is calculated as output. Secondly, we use the most frequently occurred score X_u of the historical scores of the predicted users as the input of formula (3). Thus the forecast score Y_i is calculated as output. In the end, the user u's rating vector on the all undisclosed items is as followings:

$$P_u = \frac{Y_u + Y_i}{2}. \tag{4}$$

Based on the requirements of DLRRS, the system is able to sort and select N items with the highest predicted values in P_u as the final output.

3.4 The Performance Analysis of DLRRS

The actual production environment, especially the large real system with over 100 million users and commodities, which is more time-sensitive, always has a certain demand for the response time of the recommended results. By the analysis of DLRRS system, the modeling process based on the average score of the item only needs to calculate the average score of each item. And the method based on the user and the item rating frequency also requires only a simple calculation of the highest frequency score of each user and the highest frequency of the item. So the modeling time of the system is short. Overall, the method greatly reduces the computational complexity, so that the algorithm is able to complete all the calculations in $\Omega(n)$ time. In summary, the modeling time and prediction time of DLRRS are highly competitive and can meet the requirements of the real system for forecasting time performance.

4 Experiments and Evaluations

4.1 The Dataset of Experiments

To test the performance of DLRRS, the published real dataset MovieLens [21] is used for experimental evaluation is. The MovieLens dataset is a set of film scores graded by a group of users, which is collected by the GroupLens research team at the University of Minnesota from the MovieLens website. The group published three different sizes of datasets. We select the 1 M dataset of MovieLens for experiments, which includes 1 million scores on 3952 films from 6040 users. Each score is an integer between 1 to 5. The value size indicates the users' preference for the certain film. Each user graded at least 20 movies. And the users and the movies are numbered with consecutive integers.

4.2 Evaluation Methods and Indicators

The MovieLens 1 M dataset is randomly divided into training sets and test sets according to certain proportion. Based on the double linear regression models, DLRRS uses the training set of the film score of users, and trains model parameters. Then the system predicts the users' scores on the film in the test sets. The smaller the gap between the forecasted scores and real scores, the higher the prediction accuracy of the recommendation system provides. Therefore, we use Mean Absolute Error (MAE) [3, 22] and Root Mean Square Error (RMSE) [3, 23] to measure the performance of the recommended systems. If the test set contains the actual score r_{ui} of the user u for the movie i, the predicted score given by the DLRRS is p_{ui} ($p_{ui} \in P_u$), then the definitions of MAE and RMSE are as shown in following equations:

$$\mathrm{MAE} = \frac{\sum_{(u,i)\in T}\left|r_{u,i} - P_{u,i}\right|}{|T|}. \tag{5}$$

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{(u,i)\in T}\left(r_{u,i} - P_{u,i}\right)^2}{|T|}}. \tag{6}$$

where (u, i) indicates the customer movie pair. T indicates the set of all user movie pairs in the test set. And accordingly, |T| is the number of user movie pairs to be predicted in the test set.

Based on the indicators of MAE and RMSE, we choose some of the most commonly used recommendation methods including the recommendation method based on item average scores (IA), the collaborative filtering method based on items (ICF), the method directly using weighted user rating frequency and item rating frequency (RF) as the contrast references to the DLRRS in the experiments. And in order to compare the tolerances of data sparseness in different recommended methods, we divide the MovieLens 1 M dataset into different proportions of training sets and test sets. The training set ratio increases from 10% to 90% with 10% step size.

4.3 Evaluation Results

For the four recommendation methods, the evaluation of MAE and RMSE are displayed in Figs. 1 and 2. In above figures, each approach shows the maximum, minimum, average values (three horizontal lines) and distribution (the shadow area) of MAE/RMSE result under the conditions of different proportions of training sets.

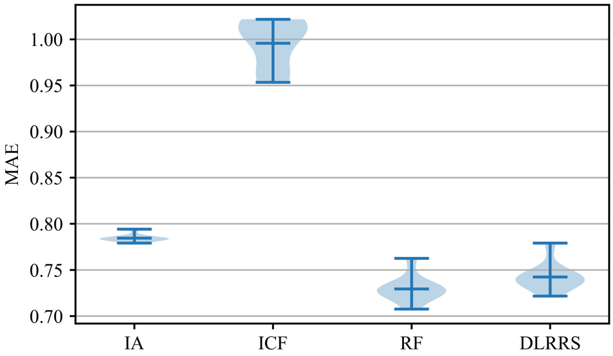

Fig. 1. The MAE comparison result.

From the experiment results, it shows that DLRRS provides the best performance in RMSE comparison, and has a small gap to RF method in MAE comparison. However, compared to the MAE, RMSE enlarges data fluctuation between the forecasted scores and actual scores by squaring. Therefore, DLRRS has the best predictive accuracy in

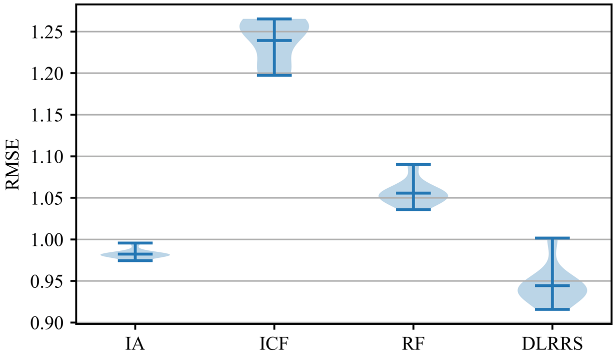

Fig. 2. The RMSE comparison result.

the term of RMSE and less fluctuation, which is superior to the existing collaborative filtering methods. At the same time, it can be seen that the RF accuracy is also high, indicating that the score frequency information has great value for the scoring forecast.

5 Conclusion

In the article, we put forward the DLRRS, a new recommendation system based on double linear regression models, and introduce its system implementation. Compared to other recommended methods such as collaborative filtering, DLRRS has the best RMSE predictive accuracy and less fluctuation. In the future, we will continue to optimize the DLRRS system to enhance its performance in multi-aspects.

Acknowledgments. This research is funded by National Key Research & Development Plan of China under Grant 2016YFB0801200 and 2016QY05X1000.

References

1. Resnick, P., Varian, H.R.: Recommender systems. Commun. ACM **40**(3), 56–58 (1997)
2. Bobadilla, J., Ortega, F., Hernando, A., et al.: Recommender systems survey. Knowl.-Based Syst. **46**, 109–132 (2013)
3. Adomavicius, G., Tuzhilin, A.: Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions. IEEE Trans. Knowl. Data Eng. **17**(6), 734–749 (2005)
4. Resnick, P., Iacovou, N., Suchak, M., et al.: Grouplens: an open architecture for collaborative filtering of netnews. In: Proceedings of the 1994 ACM Conference on Computer Supported Cooperative Work, pp. 175–186. ACM (1994)
5. Adomavicius, G., Kwon, Y.: Multi-criteria recommender systems. In: Ricci, F., Rokach, L., Shapira, B. (eds.) Recommender Systems Handbook, pp. 847–880. Springer, Boston (2015). https://doi.org/10.1007/978-1-4899-7637-6_25
6. Wang, F., Zhang, S., Henderson, L.M.: Adaptive decision-making of breast cancer mammography screening: a heuristic-based regression model. Omega (2017)

7. Sun, J., Wang, G., Cheng, X., et al.: Mining affective text to improve social media item recommendation. Inf. Process. Manag. **51**(4), 444–457 (2015)
8. Goldberg, K., Roeder, T., Gupta, D., et al.: Eigentaste: a constant time collaborative filtering algorithm. Inf. Retrieval **4**(2), 133–151 (2001)
9. Hofmann, T.: Collaborative filtering via gaussian probabilistic latent semantic analysis. In: Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 259–266. ACM (2003)
10. Jiang, S., Qian, X., Shen, J., et al.: Author topic model-based collaborative filtering for personalized POI recommendations. IEEE Trans. Multimedia **17**(6), 907–918 (2015)
11. Jiang, S., Qian, X., Shen, J., Mei, T.: Travel recommendation via author topic model based collaborative filtering. In: He, X., Luo, S., Tao, D., Xu, C., Yang, J., Hasan, M.A. (eds.) MMM 2015. LNCS, vol. 8936, pp. 392–402. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-14442-9_45
12. Veena, C., Babu, B.V.: A user-based recommendation with a scalable machine learning tool. Int. J. Electr. Comput. Eng. **5**(5) (2015)
13. Zhang, H.R., Min, F., Shi, B.: Regression-based three-way recommendation. Inf. Sci. **378**, 444–461 (2017)
14. West, J.D., Wesley-Smith, I., Bergstrom, C.T.: A recommendation system based on hierarchical clustering of an article-level citation network. IEEE Trans. Big Data **2**(2), 113–123 (2016)
15. Nilashi, M., Esfahani, M.D., Roudbaraki, M.Z., et al.: A multi-criteria collaborative filtering recommender system using clustering and regression techniques. J. Soft Comput. Decis. Support Syst. **3**(5), 24–30 (2016)
16. Ricci, F., Rokach, L., Shapira, B.: Introduction to recommender systems handbook. In: Ricci, F., Rokach, L., Shapira, B., Kantor, P. (eds.) recommender systems handbook. Springer, Boston (2011). https://doi.org/10.1007/978-0-387-85820-3_1
17. Shani, G., Brafman, R.I., Heckerman, D.: An MDP-based recommender system. In: Proceedings of the Eighteenth Conference on Uncertainty in Artificial Intelligence, pp. 453–460. Morgan Kaufmann Publishers Inc., Burlington (2002)
18. Hofmann, T.: Latent semantic models for collaborative filtering. ACM Trans. Inf. Syst. (TOIS) **22**(1), 89–115 (2004)
19. Linear Regression Method. https://en.wikipedia.org/wiki/Linear_regression
20. Yan, X.: Linear Regression Analysis: Theory and Computing, pp. 1–2. World Scientific, Singapore (2009)
21. The Datasets of MovieLens. https://grouplens.org/datasets/movielens/
22. Karatzoglou, A., Amatriain, X., Baltrunas, L., et al.: Multiverse recommendation: n-dimensional tensor factorization for context-aware collaborative filtering. In: Proceedings of the Fourth ACM Conference on Recommender Systems, pp. 79–86 (2010)
23. Jamali, M., Ester, M.: A matrix factorization technique with trust propagation for recommendation in social networks. In: Proceedings of the Fourth ACM Conference on Recommender Systems, pp. 135–142 (2010)