CRLRM: Category based Recommendation using Linear Regression Model

Conference Paper · August 2013

DOI: 10.1109/ICACC.2013.11

3 authors, including:

Gourav Jain
Indian Institute of Technology Roorkee
2 PUBLICATIONS **0** CITATIONS

SEE PROFILE

2013 Third International Conference on Advances in Computing and Communications

CRLRM: Category based Recommendation using Linear Regression Model

Gourav Jain
School of Information Technology,
Rajiv Gandhi Proudyogiki Vishwavidyalaya,
Bhopal, M.P., India, 462036.
jaingourav3010@gmail.com

Nishchol Mishra
School of Information Technology
Rajiv Gandhi Proudyogiki Vishwavidyalaya,
Bhopal, M.P., India, 462036.
nishchol@rgtu.net

Sanjeev Sharma
School of Information Technology
Rajiv Gandhi Proudyogiki Vishwavidyalaya,
Bhopal, M.P., India, 462036.
sanjeev@rgtu.net

Abstract

A system that suggests list of most popular items to a set of users on the basis of their interest is named as recommendation system. Recommendation system filters the unnecessary information by applying knowledge discovery techniques for online users and has become the most powerful and admired tools in E-Business. ERPM is one of the easiest movie recommendation method, which overcomes the limitations of scalability and sparsity of recommendation system, but it generates predictions on the basis of probability model, which are less accurate and requires more time for calculations. This article presents a novel method named CRLRM (Category based Recommendation using Linear Regression Model) which is based on linear regression model that improves the prediction accuracy and speed up the calculations. Performance of proposed method is evaluated on the basis of MAE (Mean Absolute Error) comparison, and result obtained is far much better than ERPM and shows improvement in 30-40% of user ratings.

Keywords: Collaborative Filtering, ERPM, MAE, Recommendation system, Regression model.

I. INTRODUCTION

With the increase in amount of information across the world, it is necessary to process data more quickly in the exigent environment. For processing the data, two terms i.e. data mining and recommendation system plays a key role. Data mining is defined as the process of mining unnecessary information and summarizes them into useful form, while recommendation system [1] is suggesting that information to user and incorporates the data mining techniques.

Jian. et. al [2] proposed an ERPM technique provide the easiest way to compute the prediction for a movie from entire databases but it has a problem of poor prediction. This article proposes CRLRM method, which is based on regression model and overcomes the above mentioned problem by computing the predictions from category of movie.

Remainder of the sections is organised as follows: section 2 described related work, section 3 describes the regression based recommendation approaches, section 4 describes the experimental result, performance study and lastly section 5 concludes paper.

II. RELATED WORK

Collaborative filtering [3] is the fundamental method used for recommendation system and applied in various domains. In collaborative filtering similarity between user and item is determined, and recommending the item to similar user.

Another method used for recommendation based on collaborative filtering method proposed by Yi. et. al [5] named as Typicality based collaborative filtering recommendation (TyCo), which overcome the problem of data sparsity, recommendation inaccuracy and big error in prediction by taking the ideas of object typicality from cognitive psychology.

Clustering is one of the techniques used for recommendation system which create the cluster of user based on the similar user preferences. Xiang. et. al [4] proposed a method of collaborative filtering algorithm based on uncertain user interest, in which cluster is made based on uncertain features.

Toine. et.al [6] proposed a context walk algorithm based on random walk which is used for recommendation and solve the problems related to contextual information such as collection of context information and generating a computable formalization of contextual information. Author creates the contextual graph in which each node (context) is connected to other node (context) and applies random walk on them.

Jian. et.al [2] overcomes the limitation of scalability and sparsity of collaborative filtering recommendation system and proposed a method which is based on probability model named as ERPM (Easy Recommendation based on probability model). ERPM directly compute the predicted rating by find out the probability of user gave rating to the number of movie and probability of movie gets rating by number of users and manipulating them. For checking the effectiveness of method MAE (Mean Absolute Error) is calculated for both the method and lower value of ERPM MAE show that it is better than the traditional collaborative filtering method.

III. REGRESSION BASED RECOMMENDATION APPROACH

A. Data Representation

Typical recommendation system which provides the E-Business facility, contain the list of m users represented by set $[u_1,u_2,u_3.....u_n]$. User select a item from the list of *m*

978-0-7695-5033-6/13 $26.00 © 2013 IEEE
DOI 10.1109/ICACC.2013.11

items represented by a set $[i_1,i_2,i_3....i_m]$ and relationship between user and item is represented by a $n \times m$ matrix as shown in figure 1. Entry of matrix is the rating given by the user $u_i \epsilon U$ for items $i_j \epsilon I$ and represented as $r_{i,j}$ means rating given by user *i* on the item *j*.

	i_1	i_2	…	i_m
u_1	$r_{1,1}$	……	……	$r_{1,m}$
u_2		$r_{2,2}$	……	
…	……	……	……	……
u_n	$r_{n,1}$	$r_{2,n}$	……	$r_{n,m}$

Fig. 1: Representation of user item matrix

B. Regression model

A model that has both deterministic as well as probabilistic components is called a regression model [7]. In deterministic model, with the help of one variable, value of other variable can be predicted and represented by y=f(x) which means value of y depends upon x, this is the reason why this model is known as deterministic model. In real scenario, the chances of y being fully dependent upon x are very slim hence we use probabilistic model.

Probabilistic model [7] or probability model are used to predict the value of a variable on the basis of previous information and represented by $Y \sim p(y)$ where *Y* is randomly generated from probability distribution *p(y)*. Prediction generated by the both the model is need not compulsorily occur in the past, future or even in the present. Probability model does not exactly tell what the value of *Y* will be, hence for increasing the prediction accuracy, we combine the feature of both the models (deterministic and probability) that builds up regression model.

Like deterministic model [7], Regression model [8] also predicts the value of one variable based on other variable, represented by $Y \sim p(y|x)$, where *Y* is generated at random from the probability distribution for known *x*. The regression model has proven to be a powerful tool that makes prediction about past, present or future events with the help of information about past or present events.

C. Problem Definition

The main goal of recommendation system is to suggest the available item(s) to a listed user according to user interest and generate the high quality prediction along with high speed for an active user-movie pair. Instead of calculating the predicate rating from entire databases in ERPM, we evaluate the rating from the category of movie using linear regression model. Predicted rating is evaluated by finding the probability of movie gets rating by number of users and probability of user gave rating to number of movie of particular category on the basis of their interest for each user movie pair. After calculating these probabilities, manipulate them for finding the predicted rating with the help of regression value which is computed by the regression model.

A movie is watched by a number of users who rate them between 1(for bad movie) and 5(for good movie). Quality of movie is depends upon the number of user and cannot be evaluated by a single user. Movie m is rated by *n* different users and the number of user who rates the movie *m* as 1 is represented as Q_1. Q_2 is the number of users who rated movie m as rating 2 and so on.
Prior probability $P_{m,r}$ which represent the probability of movie gets rating r by number of user is computed as follows:

$$P_{m,r} = \frac{Q_r}{N_m} \quad (1)$$

Where Q_r the number of user who gave rating r to movie m, N_m is the total number of user who rated movie m, $N_m = Q_1+Q_2+Q_3\ldots\ldots+Q_{r_{max}}$, *r* is the rating and *m* is the movie number.

Similarly to find out the probability of movie, now we calculate the probability of user gave rating to a number of movies of particular category. This can be evaluated by the counting the no of movie of category *c* gets rating *r* given by user. User U_d rated N_d different movie and $S_{c,1}$= number of movie of category c rated 1 by user. $S_{c,2}$= number of movie of category c rated 2 by user and so on.
Prior probability $P_{d,c,r}$ which represent the probability of user gave rating r to a movie of category c as follow:

$$P_{d,c,r} = \frac{S_{c,r}}{N_d} \quad (2)$$

Where $S_{c,r}$ represent the number of movies in category c rated as r by user d, N_d represent the total number of movie rated by user d in category c, $N_d = S_{c,1}+S_{c,2}+S_{c.3+\ldots\ldots\ldots\ldots}+S_{c,r_{max}}$ and r_{max} is the highest rating given by user *d* for movie *m* and *c* denotes the category of movie.

D. Proposed methodology

CRLRM method computes the prediction from the movie category using linear regression model while ERPM method computes the prediction from entire databases by probability model.
Predicted rating $Pr_{d,m}$ is calculated by:

$$Pr_{d,m} = \frac{\sum_{r=1}^{r_{max}} [r \times P_{d,m,c,r}]}{\sum_{r=1}^{r_{max}} [P_{d,m,c,r}]} \quad (3)$$

Where $P_{d,m,c,r}$ is defined as the probability of user *d* gives rating *r* on movie *m* of category *c* and is calculated as

$$P_{d,m,c,r} = [P_{m,r} \times P_{d,c,r}]^{[\text{Regression}]} \quad (4)$$

Regression values are used for prediction and value of regression is calculated by the equation:

$$\text{Regression (y')} = a+bx \quad (5)$$

Regression model finds the value from equation 5 Where *x* is the independent variable, *y* is a dependent variable (not shown in equation but used for computing the value of *y'*) and *y'* is the regression value computed with the help of *x* and *y*. In database the value of *x* is taken as the number of users for each category and value of *y* is the ratio of summation of ratings given by users for one category to the rating giving for movie of all categories.

Where *a* and *b* is the two parameters based on *x* and *y* and calculate as follows:

$$b = \frac{N\sum xy - \sum x \sum y}{N\sum x^2 - (\sum x^2)} \quad (6)$$

$$a = \bar{y} - b\bar{x} \quad (7)$$

Where $\bar{x}$ and $\bar{y}$ are the mean of x and y respectively and calculated by equation (8) and (9) respectively.

$$\bar{x} = \frac{\sum x}{n} \quad (8)$$

$$\bar{y} = \frac{\sum y}{n} \quad (9)$$

The value of b and a is computed using equation 6 and 7 respectively which further helps in calculating the value of regression from equation 5. With the help of number of user's for each category as shown in figure 2 and the value of *y*, regression value y' is computed, using which the rating is predicted for each user movie pair.

E. Comparison with ERPM model

The computed predictions on the basis of regression model in CRLRM method gives better prediction accuracy and also increase the speed of recommendation. The reason behind this is that regression model takes less time and/or resources for retrieving the information. Another important advantage of CRLRM techniques that it compute the prediction by mathematical equation and removes the dependency upon any parameter (α and β in ERPM) by which prediction is improved because a change in value of any parameter will change the result.

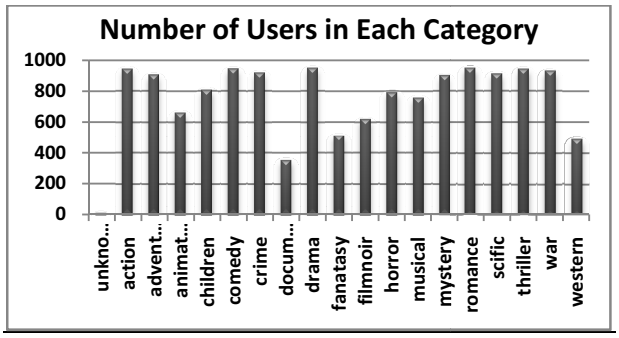

Fig.2. Representation of number of user in each category

IV. RESULT AND ANALYSIS

A. Dataset Description

We used the MovieLens data set [9] which was collected by the group lens research project at University of Minnesota. MovieLens data set has 943 User and 1682 movies and 100000 rating, by considering only those users who rated at least 20 movies out of 50000 users and more than 3000 different movie. Dataset was converted into user movie matrix which has 943 rows and 1682 column and entry in the matrix is rating given by 943 users to 1682 movie between 1(for bad movies) and 5(for good movies).

Dataset contain the information about the users, movies and rating in user, movie and data files of dataset respectively. Movie is classified into 19 genres which are used for computing the regression value. Experiment are performed on windows 7, 4GB RAM of main memory, Core i3 processor and jdk 1.7 of Java.

B. Performance Metrics

Quality of recommendation system can be evaluated by several types of measure like MAE (Mean Absolute Error), MSE (Mean Squared Error) etc., wherein Mean Absolute Error (*MAE*) is one of the popular methods among them and used to find out the error between actual rating and predicted rating for each user movie pair.

$$\text{MAE} = \frac{\sum_{i=1}^{N} |pr_m - r_m|}{N}$$

Where actual rating is denoted by the r_m, predicted rating is denoted by the pr_m, and N is the total number of items. The smaller *MAE* value reveals a better prediction system.

C. Experimental Result

We compare the CRLRM method with ERPM method on the data of 100 user and 1682 movies and find out the *MAE* value for both CRLRM and ERPM method. Table 1 presents the sum of rating for each 20 users which shows that CRLRM method predicts more efficiently than ERPM method and result show 30-40% improvement in rating prediction over 11000 rating for 100 users.

Table 1: Summation of ratings of 100 users for a group of 20 users

Number of Users	Sum of Rating		
	ACTUAL	*CRLRM*	*ERPM*
20	3582	4111.04	4669.5607
40	1452	1630.97	1795.074
60	2381	2769.52	3151.2033
80	1745	2072.56	2429.1058
100	3171	3698.88	4255.2204

The lower value of *MAE* means less error in prediction and Table.1 shows that CRLRM method gives lower MAE value in comparison to the ERPM and hence predicts more accurately. For better prediction, the difference between

actual and CRLRM should be low from the difference between actual and ERPM.

Table 2: Comparison of MAE value for 100 users and 1682 movies

	ERPM	***CRLRM***
MAE	3.364328638	2.035591765

A graph for 100 users and 1682 movies is drawn between users and sum of rating where users are taken on x axis and sum of rating on y axis as shown in Figure 3. The line in a graph represents the sum of rating for 100 users and in CRLRM method that line is almost near from the actual rating line in comparison to ERPM method. This concludes that CRLRM method predicted more accurate ratings in compared to ERPM method.

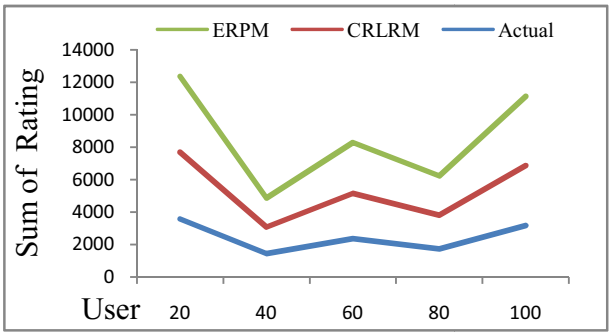

Fig.3. Rating comparison of ERPM and CRLRM method with Actual rating

Figure 4 and 5 represent the rating comparison graph between CLRLM and ERPM method. Both graph show that the better prediction of a movie is occurs for different user by CRLRM method.

V. CONCLUSION

Recommendation System is a computer based intelligence technology that helps users to finding the interesting product from the list of available item on the basis of user demand. The ERPM method used for movie recommendation have problem of poor prediction and taking long time for recommendation. To overcome the problem in ERPM, in this paper we proposed a novel method CRLRM (Category based Recommendation using Linear Regression Model) that predicts the rating from category of movie using regression model with dynamic data changes. Experiments prove that CRLRM method show better prediction accuracy and speed up the recommendation process.

REFERENCES

[1] Macro Gori,Augusto Pucci,"ItemRank: A Random-Walk Based Scoring Algorithm for recommender Engines",IJCAI,2007.

[2] Jian Chen ,Jin Huang, Huaqing Min, "Easy Recommendation Based on Probability Model",IEEE,2008,pp 441-444.

[3] Badrul Sarwar, George Karpis, Joseph Konstan and John Riedl,"Item based Collaborative filtering Recommendation Algorithm", ACM, 2001, pp-285-295.

[4] Xiang Cui,Guisheng Yin," Method of collaborative filtering based on uncertain user interests cluster",Journal of Computers ,Vol.8,No.1,January 2013, pp186-193.

[5] Yi Cai, Ho-fung Leung, Qing Li, Huaqing Min,Jie tang and Juanzi Li, "Typicality-based Collaborative Filtering Recommendation",IEEE Transaction on Knowledge And Data Engineering,Jan 2013.

[6] Toine Bogers, "Movie Recommendation uses Random Walk Over Contextual Graph", 2010.

[7] http://courses.ttu.edu/isqs5349-westfall/images/5349/deterministic_stochastic.htm

[8] http://www.psychstat.missouristate.edu/introbook/sbk16.htm.

[9] www.movielens.org.

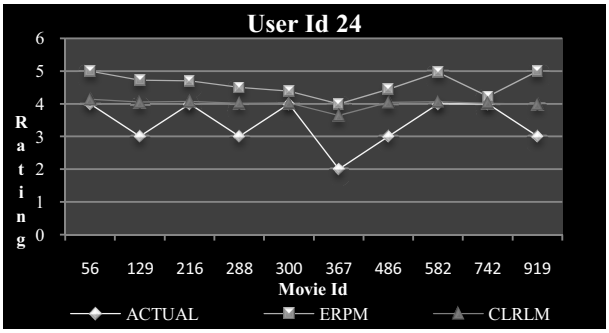

Fig 4: Rating Comparison for user 24

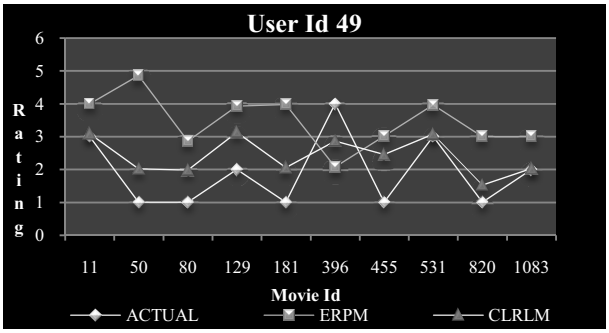

Fig 5: Rating Comparison for user 49