The Fourteenth International Symposium on Artificial Life and Robotics 2009 (AROB 14th '09),
B-Con Plaza, Beppu, Oita, Japan, February 5 - 7, 2009

A Movie Rating Prediction System based on Personal Propensity Analysis

Soojin Lee, Taeryong Jeon, Gyeongdong Baek, Jaewoo Cho, and Sungshin Kim

School of Electronics and Electrical Engineering, Pusan National University, Busan, Korea
Pusan National University Changjeon 2 dong, Keumjeong-ku, Busan 609-735, Korea
(Tel : 82-51-510-2367; Fax : 82-51-513-0212)
(Email address: {pooh3887, jtr, gdbaek, jwjw777, and sskim}@ pusan.ac.kr)

Abstract: Most of all the movie recommendation methods need the data for genre of movies and private information of user. As the kinds of movies are getting various, recommendation system for analyzing user's propensity can be required because movie recommendation method by data for genre of movies is limited. The recommendation system for analyzing user's propensity is recommending movies through the movie records and evaluation of the past. This paper proposed active recommendation algorithm which analyzed propensity of user and predicted evaluation of the movies before users look up the data for preferred recent movies.

Keywords: prediction system, data mining, movie recommendation, user propensity analysis

I. INTRODUCTION

As internet service and computer technology are getting better, the personalized recommendation system is actively researched by analyzing user's personal information such as propensity. Web sites about e-commerce are investing development of excellent recommendation system because additional information can be added to user's personal information in the ideal personalized recommend system. Amazon, which is the company merchandising product by online system, and Netflix, which is the company renting movies by online system, are recommending products that considered user's expectation and satisfaction.

Data-mining techniques that are included pattern recognition and information filtering methods have been applied to develop recommendation system. Representative technology of recommendation system is Information filtering technology that divides content-based recommendation and collaborative recommendation [3-4]. Content-based recommendation extracts user's propensity and predicts user's preference. But this recommendation is difficult from missing information about item at that time of the first attempt [5-6]. Collaborative recommendation which is the most popular recommendation analyzes other user information based on similarity to target user.

In this study, the suggestion is predicting evaluation from target movie about target user that Netflix offers data about movies and users. Data consist of users (480,189), movies (17,770), dates of evaluation and points from 1 to 5. Also prove data and qualifying data are given for the performance evaluation of developed system.

II. MOVIE RATING PREDICTION SYSTEM

Each user's personal propensity is analyzed to predict system watched movie evaluation. The analysis method for personal propensity usually is normalizing personal information such as age, sex and occupation. This kind of method can be caused security problems by personal information. And user's satisfactions are lower than expected because user's data also wasn't perfectly matched with personal taste and individuality. This movie rating prediction system that analyzes personal propensity through the previously watched user's movie records is developed.

1. Reorganization of the data set

The raw data of the Netflix are used to build movie rating prediction system. The structure of raw data is shown in Fig. 1 and all data are consisted of text type. The training data is consisted of evaluation and data of 17,770 movies that all users of approximately 4.8 million peoples watched. But the training data isn't properly used by analyzed personal information because classification is about kind of movies. It is the reason why the reorganization is needed before the rating is predicted. The reorganization of the training data that are properly aimed at analysis of personal propensity offered from the Netflix is shown in Fig. 2.

The probe and qualifying files are offered by the Netflix that purposed evaluation of the performance.

The rating about evaluated movies will be obtained till the first decimal number by estimation algorithm. And each detailed item of evaluation is constituted with ID of the movie, date and ID of the user to predict a point of target movie. In other words, rating will be predicted after target movie is watched by target user. There are difference between probe and qualifying data. The probe data is offered as detailed result of each item of evaluation and qualifying data is offered as calculated total error through the on-line. The performance of system is evaluated to improve and evaluate confidence of movie rating prediction system.

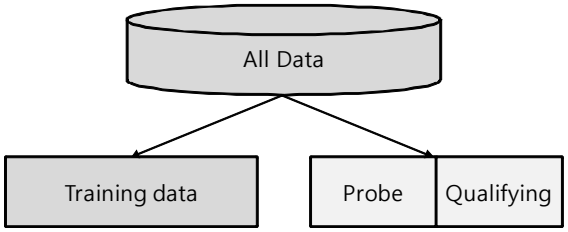

Fig.1. The data structure from the Netflix

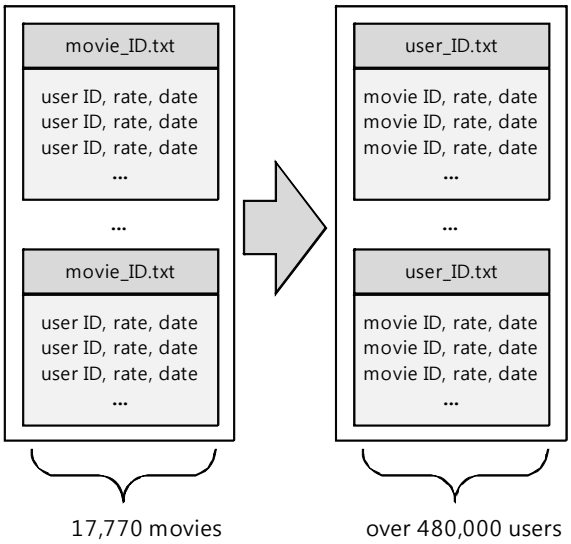

Fig.2. The reorganization of Data set

2. The process of the movie rating prediction system

The prediction process for rating of the movies after target user watched target movie is shown in Fig. 3. First of all, information of target user about target movie is extracted through the probe data. And then related users with target user are subjected for rating of the movie. Probe data is evaluated and predicted through the repetition of the process.

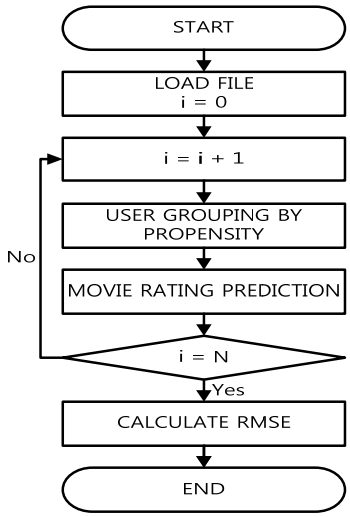

Fig.3. The flow chart of movie rating prediction

3. User grouping classified by user propensity

Users having similar propensity with target user are classified to related user group in order to analyze target user's propensity. The process for each user's propensity is analyzed by target user and related user group about rating records of movie.

The process of classifying related user group is as follows. At first, target user and movie are extracted in the probe data in order to evaluate performance of movie rating prediction system. Secondly, all users that watched target movie except target user are classified related user group.

4. The prediction of movie rating

After user grouping classified by user propensity, similarities are derived from related propensity between target user and related users. The calculating method for similarity is shown in Eq. (1).

$$\mu = \frac{N\left(r_{t,i}, r_{u,i}\right)}{n_t} \tag{1}$$

The rate of evaluated movie *i* by target user is $r_{t,i}$, the rate of evaluated movie *i* by related user is $r_{u,i}$ and $N\left(r_{t,i}, r_{u,i}\right)$ is the same number between $r_{t,i}$ and $r_{u,i}$. And n_t is the number of movies that target user watched. If the value of similarity is high, related users have more similar propensity. Each similarity is estimated as follows.

$$Estimated\ Rate = \frac{\sum_{i=1}^{n} \mu_i \times R}{\sum_{i=1}^{n} \mu_i} \tag{2}$$

The similarity of related user is μ_i. R is evaluated rate of related user about target movie and n is the number of related users, Eq. (2). In other words, the number of related users with target user is i. The estimated rate is reflected with evaluated rate through the related user's movie information watched and evaluated.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2} \qquad (3)$$

5. Error calculation of prediction system for movie evaluation

Probe data are applied to evaluate performance of the movie rating prediction system and error calculation is taking advantage of the RMSE (root mean square error). The equation of system error, RMSE, is showing in Eq. (3). The actual rate of target user about target movie is y_i and also $\hat{y}_i$ is estimated rate through the analysis of personal propensity. And n is the number of items to evaluate in the probe data.

III. AN EXPERIMENT AND RESULTS

1. Experimental Method

Training data is utilized in order to develop the movie rating prediction system that offered movie data (17,770) and evaluation data (480,189) from the Netflix. It takes a lot of times to reorganize data that fits the purpose of analysis of personal propensity. The process of reorganization is accomplished only once per each user transforming database that isn't repeatedly accomplished.

As the amount of training data is getting much, the samples reflected population's attributes are experimented through the random sample method of probabilistic sampling method. And also, the samples are classified as 3 parts, e.g. (100, 300 and 500 users) in order to compare system performances with different sizes of the samples repeatedly 10 times.

2. Experimental results

The visual C++ is utilized for evaluation and development of proposed the movie rating prediction system. The performance of prediction system is evaluated through the calculated estimation error for movie rating of the system. The results of proposed movie rating prediction system, RMSE, are presented to Table 1.

Table 1.RMSE of movie rating prediction system

	100	300	500
1	1.094094	1.057966	0.961277
2	1.084670	0.987677	0.970496
3	1.156681	1.080526	0.940614
4	1.089129	1.062151	0.983205
5	1.025894	1.020257	0.961183
6	1.118859	1.006011	0.974777
7	1.175142	0.991430	0.951108
8	1.076674	1.009890	0.949743
9	1.179335	0.979645	0.981149
10	1.061592	1.008651	0.987640
Avg.	**1.106207**	**1.020420**	**0.966119**

As the number of users is increasing, RMSE is conversely decreasing. In other words, the more users are included, the more detailed prediction is achieved, because information of propensity about related user is also extended.

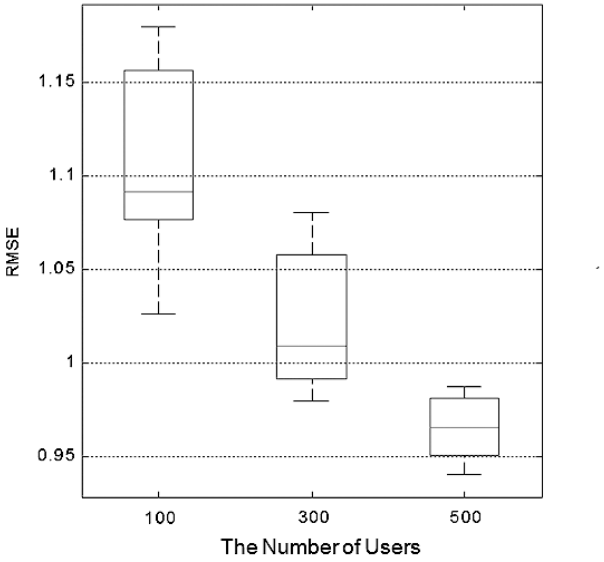

Fig.4.The results of RMSE

The evaluated results of system performance are presented by box-plot graph, Fig. 4. The number of users, e.g. (100, 300 and 500) is projected on the x-axis and the distribution of RMSE is projected on the y-axis. And the ranges of maximum and minimum value are presented through the graph. As the number of users is increasing, the range is decreasing that mean variation

of RMSE is also decreased. After the restriction is disappeared such as the number of users, RMSE will be converged to specific point. The performance of proposed system would be properly identified by reorganized structure and detected converging specific point.

The least predicted error among developed systems using opened data from the Netflix is 0.9841 which is estimated by team of Bellkor [11]. It is difficult to compare our system with team of Bellkor because there is difference between the two that our system experimented samples of population and Bellkor's system experimented whole population. But proposed system would contribute to improve performances of previous movie rating prediction system.

IV. CONCLUSION

After target user watched target movie through the analysis of personal propensity, the prediction system of evaluating target movie is proposed. The proposed system is based on opened data from the Netflix. The normalized analysis method through the personal information isn't considered in this paper, but the analysis method is developed with user's satisfaction through the personalized movie information.

The probe data from the Netflix is subjected to performance of movie rating prediction system through the extracted sample of random sample method. The extracted samples are examined by the number of users, e.g. (100, 300 and 500 users) and repeated 10 times per the number of users. The evaluated prediction error is calculated by means of RMSE.

As the similarity is calculated and analyzed by relation of target user and related users, the personal propensity is decided. The proposed movie rating prediction system can be using base technology to recommend movie that is properly capturing the public fancy through the analysis of user's propensity.

Further subject of the research will be considered about data reorganization of whole population for the comparison with previous movie rating prediction system and identified the performance of proposed system.

REFERENCES

[1] Herlocker J, Konstan J, Terveen L, and Riedl J , "Evaluating Collaborative Filtering Recommender Systems," ACM Transactions on Information Systems 22, ACM Press, 5-53, 2004.

[2] G. Adomavicius and A. Tuzhilin, "Towards the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions," IEEE Transactions on Knowledge and Data Engineering 17, 634-749, 2005.

[3] Ansari,A., Essegaier,S. and RKohli,R., "Internet Recommendation Systems," Journal of Marketing Reserch Vol.37, pp. 363-375, 2000.

[4] Il Im, "Augmenting Knowledge Reuse Using Collaborative Filtering Systems," A Dissertation Presented to the aculty of the graduate school USC (Information Systems), p.191, 2001.

[5] Basu,C., Hirsh,H. and Cohen,W, "Recommendation as Classification : Using Social and Content-based Information in Recommendation," Proc. of the Fifteenth National Conference on Artificial Intelligence (AAAI-98), pp.714-720, 1998.

[6] Pazzani,M., "A Framework for Collaborative , Content-Based and Demographic Filtering," Artificial Intelligent Review13(5-6), pp.393-408, 1999.

[7] D. Goldberg, D. Nichols, B. M. Oki and D. Terry, "Using Collaborative Filtering to Weave an Information Tapestry," Communications of the ACM 35, 61-70, 1992.

[8] Konstan,J., Miller,B., Maltz,D., Herlocker,J., Gordon, K. and Riedl,J. "GroupLens :Applying Collaborative Filtering to Usenet News," Communications of the ACM, Vol.40 No.3, pp.77-87, 1997.

[9] Rensnick,P., Iacovou,N., Suchak,M., Nergstorm,P. and Riedl.,J. "GroupLens : An Open Architecture for Collaborative Filtering of Netnews," Proc. of CSCW '94, pp. 175-186, 1994.

[10] Shardanand,U. and Maes,P., "Social information filtering : Algorithms for automating 'word of mouth'," Proc. of ACM CHI '95 Conference on Human Factors in Computing Systems, pp.210-217, 1995.

[11] Robert M. Bell and Yehuda Koren, "Improved Neighborhood-based Collaborative Filtering, " KDD 2007 Netflix Competition Workshop, 2007.