Logistic recommendation algorithm based on collaborative filtering

Zhang Xiaoyu [a], Dai Chaofan[b] and Zhao yanpeng [c]

Science and Technology on Information System Engineering Laboratory, National University of Defense Technology, Changsha, Hunan 410073, P.R.China

[a]404338476@qq.com, [b]cfdai7318@qq.com, [c] 871029507@qq.com

Keywords: collaborative filtering, classification, Logistic

Abstract. Traditional collaborative filtering recommendation technology has lower accuracy and couldn't meet the actual needs. So we convert the recommendation problem into the classification problem. We introduce Logistic classification methods based on the collaborative filtering technology and determine whether the recommending units are recommendable according to personal feature of users and products. The result shows that the method could greatly improve the accuracy of recommendation under the premise to ensure recall.

Introduction

Collaborative filtering technology is currently the most widely used personalized recommendation technology[1], The core idea has two parts: The first, we calculate the similarity between users according to the user's personalized information; The second, we predict the target user preferences in relation to other products using his neighbors with higher similarity and recommend products for the target user according to his preferences. Collaborative filtering technology and content-based recommendation technology[2] is different. It has no special requirements for recommendation target and could handle music, movies and other unstructured objects. The systems based on collaborative filtering technology include Amazon books recommender system, Jester joke recommendation system, Ringo music recommender system , Phoaks WWW recommender system and so on.

Although collaborative filtering technology has been very widely used, it also faces many problems, such as cold start problem, sparsity scoring problem, algorithm scalability problemand recommendation accuracy problem[3]. To improve recommendation accuracy, we use collaborative filtering technologyget to get the recommending units, then we convert the recommendation problem into the classification problem. We introduce Logistic classification methods[4] to determine whether the recommending units are recommendable according to personal feature of users and products.

Algorithm

Collaborative filtering recommendation

Interest matrix about user an`d production is as follows.

Table 1. Interest matrix about user and production

	$item_1$	$item_2$	…	…	$item_m$
$user_1$	r_{11}	r_{12}	…	…	r_{1m}
$user_2$	r_{21}	r_{22}	…	…	r_{2m}
…	…	…	…	…	…
$user_n$	r_{n1}	r_{n2}	…	…	r_{nm}

r_{ij} indicates the rating of user i about product j. If r_{ij} =null, then we don't have the rating of user i about product j.

Similarity matrix about user is as follows.

Table 2.Similarity matrix about user

	$user_1$	$user_2$	⋯	$user_n$
$user_1$	s_{11}	s_{12}	⋯	s_{1n}
$user_2$	s_{21}	s_{22}	⋯	s_{2n}
⋯	⋯	⋯	⋯	⋯
$user_n$	s_{n1}	s_{n2}	⋯	s_{nn}

$$s_{ij} = \frac{\sum_{k=1}^{m} r_{ik} * r_{jk}}{\sqrt{\sum_{k=1}^{m} r_{ik}^{2} * \sum_{k=1}^{m} r_{jk}^{2}}} \quad , \text{ if } i \neq j \tag{1}$$

$$s_{ij} = 1, \text{ if } i = j \tag{2}$$

We could get t nearest neighbors based on similarity matrix. $D_i = \{u_1, u_2 ... u_t\}$, u_p is the highest similarity p-th user about user i.

we could complete the rating of r_{ij} =null using similarity matrix. Calculation method is as follows.

$$r_{ij} = \frac{\sum_{p}^{t} s_{ip} * r_{pj}}{\sum_{p}^{t} s_{ip} \partial_p} \tag{3}$$

∂_p=0， if r_{pj}=0 ,otherwise ∂_p=1 (4)

We choose the top-k products to recommend based on similarity matrix for user i. Recommendation set is B_i.

$$\text{UI} = \{(i, j) \mid i \in user, j \in item, r_{ij} \in B_i\} \tag{5}$$

Logistic regression(LR)

Logistic regression is a classic classification method that has been applied in many fields, and achieved good results[5]. We use logistic regression model to classify for UI collection obtained from the last step and determine whether unit(i,j) is worth to be reconmended.

Logistic regression model is as follows.

$$P(Y = 1 \mid X) = \frac{\exp(w * x)}{1 + \exp(w * x)} \tag{6}$$

$$P(Y = 0 \mid X) = \frac{1}{1 + \exp(w * x)} \tag{7}$$

$x \in R^q$ is the characteristic parameter, q is the characteristic parameter number, $w \in R^{q+1}$ is weight vector about characteristic, $y \in \{0,1\}$ is binary classification categories.

we use maximum likelihood estimation method to estimate the model parameters $\hat{w}$, Calculation method is as follows.

$$L = \prod_{i=1}^{N} [P(Y = 1 \mid x_i)]^{y_i} * [P(Y = 0 \mid x_i)]^{1-y_i} \text{ , } x_i \in R^n \text{ , } y_i \in \{0,1\} \tag{8}$$

$$\ln(L) = \sum_{i=1}^{N} [y_i * \log(P(Y = 1 \mid x_i)) + (1 - y_i) \log(P(Y = 0 \mid x_i))] \tag{9}$$

we could use gradient descent method to solve 8-th formula, then get $\hat{w}$.

Data Selection and Preprocessing

Data set

Experimental use MovieLens[6] dataset, which is collected by GroupLens Group (www.grouplens.org). It contains 943 users and 1682 movies, It consists of three documents, The first is user ratings about product file in which user score 1-5 points for movies that had been seen, 1 point represents most dislike, 5 means a favorite, a total of 100,000 records. The second is user personalization profile in which there are 943 user's records and each record includes gender, age, occupation and other property. The third is product personalization profile in which there are 1682 product's records and record includes date,type and so on.

Data preprocessing

Experiments use 5-fold cross-validation[7] approach and the training set and test set is ratio of 4: 1, then get the mean of 5 times final results.

Missing values process, There is a large number of missing values in user and product characterized file. A portion of the user's occupation is empty, then add a "null" value category in the occupation attribute. There are some product categories missing, such as a total of 19 film categories, relevant parameters of a portion of films is only 18 categories, then the missing category is representation of "0".

User rating process, we assume that film score greater than or equal to 4 means that users like the movie.

Nominal property process, we convert the nominal property of the user and product into a binary property. For instance, occupation property is a kind of nominal property and includes 21 kinds of occupations, so we convert it to 21 binary, appropriate occupation of users is set to 1, the other is set to 0.

Experiment

We compare collaborative filtering (CF) and logistic recommendation algorithm based on collaborative filtering (CFLR) using MovieLens data set. We evaluate the effectiveness by three indicators include recall, accuracy and F1 value[8].

$$precision = \frac{|\ predictionSet \cap referenceSet\ |}{|\ predictionSet\ |} * 100 \tag{10}$$

$$recall = \frac{|\ predictionSet \cap referenceSet\ |}{|\ referenceSet\ |} * 100 \tag{11}$$

$$F1 = \frac{2 * precision * recall}{precision + recall} \tag{12}$$

The recall of CF and CFLR shown in Fig.1. The recall of CFLR is slightli lower than CF's, but little difference between them.

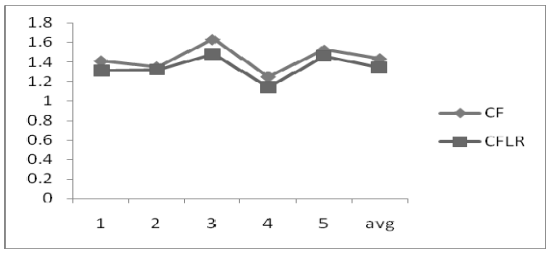

Fig.1. recall of CF and CFLR

The accuracy of CF and CFLR shown in Fig.2. The accuracy of CFLR is much higher than CF's.

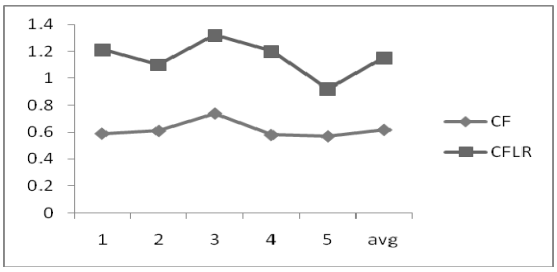

Fig.2.accuracy of CF and CFLR

The F1 value of CF and CFLR shown in Fig.3. The F1 value of CFLR is much higher than CF's.

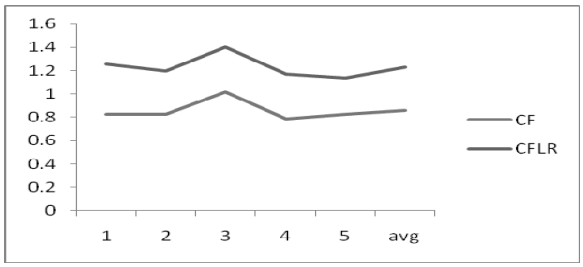

Fig.3. F1 value of CF and CFLR

Conclusion

Experimental results show that logistic recommendation algorithm based on collaborative filtering could get better recommendation effect than traditional collaborative filtering algorithm. We could significantly improve the accuracy of recommendation under the premise of general recall and have higher F1 value.

References

[1] Shi Y, Larson M, Hanjalic A. Collaborative filtering beyond the user-item matrix: A survey of the state of the art and future challenges[J]. ACM Computing Surveys (CSUR), 2014, 47(1): 3.

[2] Pazzani M J, Billsus D. Content-based recommendation systems[M]//The adaptive web. Springer Berlin Heidelberg, 2007: 325-341.

[3] Herlocker J L, Konstan J A, Terveen L G, et al. Evaluating collaborative filtering recommender systems[J]. ACM Transactions on Information Systems (TOIS), 2004, 22(1): 5-53.

[4] Hosmer Jr D W, Lemeshow S. Applied logistic regression[M]. John Wiley & Sons, 2004.

[5] Myller E, Hannola L. Logistic Overview of Finnish Transit Transportation and Foreign Trade between Finland and Russia[J]. 2010.

[6] Harper Y C F M, Konstan J, Li S X. Social comparisons and contributions to online communities: A field experiment on movielens[J]. The American economic review, 2010: 1358-1398.

[7] Golub G H, Heath M, Wahba G. Generalized cross-validation as a method for choosing a good ridge parameter[J]. Technometrics, 1979, 21(2): 215-223.

[8] Vargas-Govea B, González-Serna G, Ponce-Medellın R. Effects of relevant contextual features in the performance of a restaurant recommender system[J]. ACM RecSys, 2011, 11.