An Improved Recommendation Algorithm for Micro-blog Network Advertisement

Yanxia Yang
Faculty of Information Engineering,
City College Wuhan University of Science and Technology,
Wuhan, China
yxy_job@163.com

***Abstract*—In this paper, a layered hybrid network advertisement recommendation algorithm which is based on the micro-blog platform is proposed and implemented. In this algorithm, two kinds of recommendation algorithms are combined in a stacked way, and on the basis of results of classification recommendation algorithm, recommendation is conducted by employing the algorithm based on user clustering. The experiment proves that the algorithm can solve the data sparsity problem and optimize the recommendation results.**

Keywords-Micro-blog Marketing; Recommended Algorithm; Collaborative Filtering; Naive Bayesian Classification

Ⅰ. INTRODUCTION

The key to micro-blog marketing is accuracy and efficiency. The precison of micro-blog's recommendation in the early stage of micro-blog marketing has important significance, which requires a personalized recommendation to accurately predict the target audience and to recommend users for products which meet their interest preferences. And a truly good personalized recommendation is not only to recommend products to users, in addition, but also virtually, establishing some kind of close connection with users, making users dependent on this tailored personalized service in the premise of satisfying them, so as to enhance the using experience[2]. In the personalized recommendation, choosing appropriate and effective recommendation algorithm is of great concern for the recommendation effect.In this paper,on the results of the rough classification matching by using naive Bayesian classification algorithm, optimizing the recommendation algorithm by further employing the collaborative filtering algorithm based on user clustering, a hybrid recommendation algorithm of higher accuracy is obtained.

Ⅱ. HYBRID RECOMMENDATION ALGORITHM

A. Recommendation Algorithm Based on Classification

In this paper, the naive Bayesian classification algorithm is used to classify the micro-blog advertising, the basic method of which is looking for classified feature words of micro-blog, on the basis that the feature item appears, calculating the probability of each category according to the feature item, so as to realize the classification [4].

Bayesian classification can be divided into three stages:

The first stage—preparation stage, which mainly including data acquisition, Chinese segmentation, denoising, feature extraction, and then labeling and classifying class attribute in the training set, in this stage, inputing all the samples to be classified, outputing classified classification attributes and training set [4].Bayes theorem is the basis of Bayesian classification, and it is a theorem about conditional probability and marginal probability of random event *A* and *B*[5]. Bayes theorem is shown in (1).

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)} \tag{1}$$

Conditional independence is the basic assumption of naive Bayes, whose formalized expression is shown in (2):

$$P(B \mid A) = P(b1 \mid A) * P(b2 \mid A) * \ldots * P(bn \mid A) \tag{2}$$

The second stage—classifier training stage. This is the process of classifier's generation, which mainly calculates probability of each class in training samples, that is the number of occurrences of each category and the total number of samples, and also calculates conditional probability of features arisen in each class.

The third stage—application stage. The naive Bayesian algorithm is applied in practice, the data set being classified and its accuracy being estimated.

In this paper, the naive Bayesian classifier based on document frequency is adopted to classify micro-blog network advertising and users, whose specific steps are as follows:

Step 1:Text data feature extraction.Segmenting collected micro-blog text and tagging the part of speech, and denoising the data which has been segmented.

Step 2:Calculating log P(a|y), the probability value of each feature word in every category after word segmentation.

Step 3: Calculating the probability of each category.

Step 4:After the trainer is completed, finding out the probability of the largest category of max{ $P(y_1|x)$,$P(y_2|x)$,$P(y_n|x)$} under the characteristic X.

Recommendation algorithm based on classification can be regarded as a semi personalized recommendation. Although it can't deviate from the user interest and recommend product advertising that users may be of interest in to them, recommendation precision is not enough, which means it can only recommend a certain class or a few categories of product advertising to users, and it will lead to a phenomenon that network advertising appears too frequently in front of users, causing antipathy, the loss overweighing the gain. Thus, to get a more precise recommendation result, further improvement in calculating algorithm is essential.

B. User - project Rating Matrix

After roughly classifying by using Bayesian

978-1-5386-7445-1/18/$31.00 ©2018 IEEE
DOI 10.1109/DCABES.2018.00080

classification algorithm, optimizing it by further adopting collaborative filtering algorithm, consequently, the recommended result being more accurate. Usually the first step is to build user-item rating matrix, counting the similarity degree among users, then, predicting the rating to target users' unknown term made by them, and producing the final recommendation matrix. From the point of view of the user's score, retweeting indicates that the user is interested in the product of the advertisement and the product of it, and the rating is three; Thumbing up means the user's interest is two; Evaluation can't determine whether users' interest is positive, so the rating is set as one; If the user doesn't do any treatment to the micro-blog they read, expressing as a missing value, and the rating is 0. Thus the user interest model based on the user－item rating can be obtained, user U={U_1,U_2,…,U_m}, project I={I_1,I_2,…,I_n}, there into, the value range of $R_{i,j}$ is [0, 3].

C．Recommendation Algorithm Based on User Clustering

In the previous micro-blog recommendation system, recommending micro-blog content to users' needs to traverse all the micro-blog in the system, the characteristics of real time and efficiency that recommendation system should have being challenged. Therefore, at first, cluster users, and then recommend content to users within the cluster range. This paper will look for users' nearest neighbors on the basis of the users' interest model, thus forming the cluster. The specific steps of the algorithm are illustrated as follows:

Step 1: Calculating the similarity of score vector.According to the user - item score matrix obtained from the last section, the similarity between the users is calculated by using the cosine similarity method which is shown in the (3).

$$sim(i,j)=\frac{\sum_{c\in I_{i,j}}(R_{i,c}-\bar{R}_i)(R_{j,c}-\bar{R}_j)}{\sqrt{\sum_{c\in I_i}(R_{i,c}-\bar{R}_i)^2}\sqrt{\sum_{c\in I_j}(R_{j,c}-\bar{R}_j)^2}} \tag{3}$$

Among them, *i*, *j*, respectively indicates different micro-blog network advertising, $R_{i,c}$ signifies the user's score to the micro-blog content *i* which is derived from the user's behavior .

Step 2: Constructing similar user group.Scaning the entire user set, conducting the comparison of similarity between target users and each user. Sorting other users out according to the similarity value, and then taking out some users of larger similarity value to form a similar user group.

Step 3: Generating recommended items.According to the neighbor users in user groups, generating recommended items for target users. Rating of the user's unknown item can be predicted through the project weighted rating of the user, whose calculation formula is as follows in the formula (4),*N* is the similar neighbor set of target users.

$$P(U_y,I_i)=\frac{\sum_{i\in N(U_y)}(r_{x,i}\times sim(U_x,U_y))}{\sum_{i\in N(U_y)} sim(U_x,U_y)} \tag{4}$$

D．Hybrid recommendation algorithm

Firstly,according to existing advertising categories, analysing and processing text and micro-blog posted by users by using naive Bayesian classification algorithm, then according if it is the same category, determining whether to recommend to users, obtaining the initial recommendation results through classification matching and feedback the results to users, users express their preferences for the recommendation through forms like retweeting, commenting and thumbing up, the system builds user-item matrix according to the users' feedback on the recommendation results. Recommendation steps are as follows.

Step1:According to the user-item score matrix, conducting the similarity calculation between users by using cosine similarity coefficient formula, then producing a similar set of neighbors of target users.

Step2:Predicting the rating of the unknown item from neighboring users in a similar set of neighbors.

Step3:Sorting the predicted items out, then recommending the former N items to users.

Ⅲ. IMPLEMENTATION OF HYBRID RECOMMENDATION ALGORITHM IN MICRO-BLOG NETWORK ADVERTISING

For implementing the hybrid recommendation algorithm of micro-blog advertising that is proposed in this paper, first of all, the collected micro-blog data will be preprocessed, mainly Chinese word segmentation, data denoising and feature extraction, then getting the training samples. And then using Bayesian classification to classify the training set, on the basis of classification matching results, the user item rating matrix is established according to the user's feedback, so that the final recommendation results are acquired by employing the collaborative filtering algorithm based on user clustering according to scoring matrix of the data.

A. Micro-blog Data Crawl

B. Chinese Word Segmentation Technology

ICTCLAS is used to conduct Chinese word segmentation and part of speech tagging on the acquisition of the micro-blog network advertising and users' micro-blog text. Removing some of the stop words.

C. Data Feature Extraction

Document frequency method is adopted to carry on the selection of feature words, putting off words whose word frequency is less than 3 and whose emergence rate is more than 95% in order to get rid of words whose word frequency is too small or too large, then the rest will be taken as feature words. The extracted features are divided into nine categories: fitness, fashion, examination, entertainment, finance, life, science, technology, and tourism. The extraction results of users' micro-blog texts are shown in figure 1.

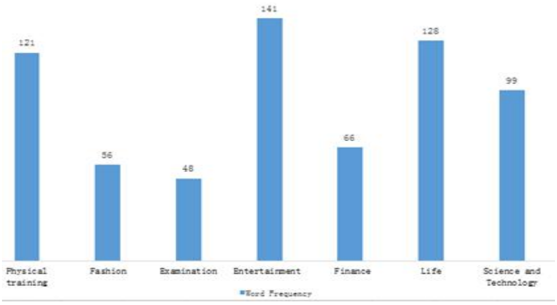


Figure 1. Word frequency statistics of feature word categories

D. Bayesian Classification

After obtaining the training samples, then according to the feature item, calculating the probability of each category, so as to achieve classification. Using naive Bayes to achieve a rough classification matching of micro-blog Internet advertising and users.

E. Establishment of User - project Matrix

According to the results of the naive Bayesian classification, micro-blog network advertising after rough classification matching is recommended to corresponding users, the users immediately thumb up, comment, or retweet the recommended micro-blog advertising, and then according to this kind of behavior feedback of users, setting up a corresponding user - project evaluation matrix. Taking five ordinary users' behavior feedback to ten micro-blog advertising recommended by them as an example.

F. Recommendation Algorithm Based on User Clustering

This algorithm is conducted on the basis of user clustering, and it needed to find the user's nearest neighbors, thus forming a cluster. To recommend the user the former N advertising micro-blogs that he may be interested in, here we set *N* to 8.

The main process of calculating the similarity function between the users is shown in figure 2.

The main flow of the matrix ranking function of users' similarity is shown in figure 3.

The main process of how the user *i* predict the function of users' interest degree towards item *j* is shown in figure 4.

Operation result of collaborative filtering algorithm which is based on user clustering is shown in figure 5.

Ⅳ SUMMARY

In this paper, firstly,using naive Bayesian classification algorithm to roughly classify micro-blog network advertising and micro-blog users, advertising that is similar to users' preferences will be recommended to them. Then, on the basis of matching results of rough classification, this paper proposes a collaborative filtering recommendation algorithm based on user clustering, according to customer's feedback to recommended results, establishing user - project evaluation matrix, predicting unknown project evaluation, recommendation matrix being obtained. Finally, the combination of the simple Bayesian classification algorithm and the collaborative filtering algorithm are combined, and a hybrid recommendation algorithm is proposed, which makes the results more accurate. This hybrid strategy can effectively reduce the amount of computation of collaborative filtering algorithm, and improve the overall efficiency of the algorithm, and to a certain extent, it can solve the problem of data sparsity and cold boot .The hybrid recommendation algorithm can help enterprises more accurately find the target users, thus making recommendation results more precise in order to enhance using experience.

ACKNOWLEDGMENT

The work in this paper is partially supported by the Humanities and social science Research Plan Project of Education Department of Hubei of China under Grant No. 16G250.

REFERENCES

[1] S. Huang, J. Sun, X. Wang, H, Zeng and Z, Chen. Subjectivity Categorization in Weblog Space using Part-Of-Speech based Smoothing. In Proceedings of 6th Inernational Conference on Data Mining.

[2] R. Kumar,J. Novak,P. Raghavan,and A. Tomkins. Structure and Evolution of Blogspace. Commun. ACM, 47(12):35-39, 2010.

[3] J.D. Lasica,Weblogs:A New Source of Information. In We've got blog: How weblogs are changing our culture, John Rodzvilla (ed). Perseus Publishing,Cambridge,MA,2012.

[4] F. Sebastiani, "Machine Learning in Automated Text Categorization", ACM Computing Surveys, Vol.34, No.1, p1-p47, March 2012.

[5] J. Bar-llan. An Outsider's View on "Topic-oriented" Blogging. In Proceedings of the Alt. Papers Track of the 13th International Conference on World Wide Web, papers 28-34,May,2013

[6] K.T. Durant and M.D. Smith. Mining Sentiment Classification from Political Web Logs. In Proceedings of Workshop on Web Mining and Web Usage Analysis of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (WebKDD-2012). August, 2012.

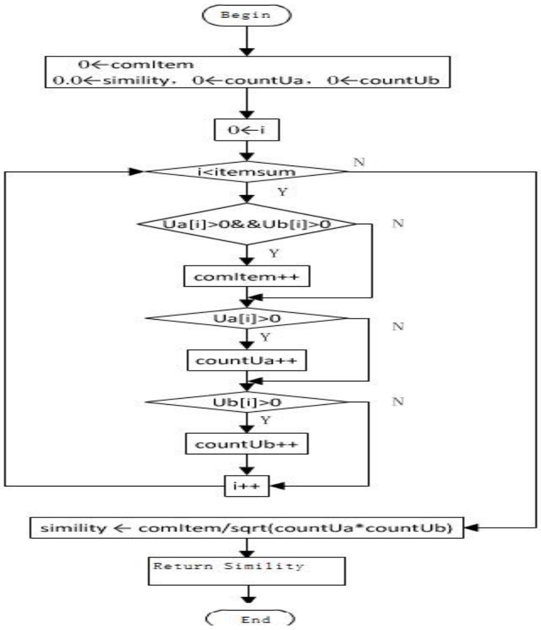


Figure 2. Flow chart of similarity function

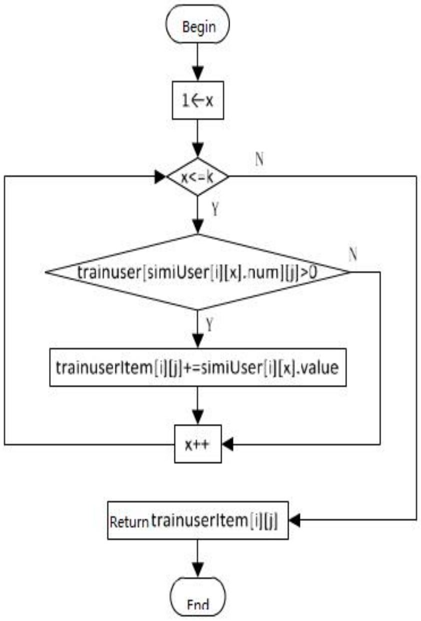

Begin
1←x
x<=k
N
Y
trainuser[simiUser[i][x].num][j]>0
N
Y
trainuserItem[i][j]+=simiUser[i][x].value
x++
Return trainuserItem[i][j]
End

Figure 3. Flow chart of similarity matrix

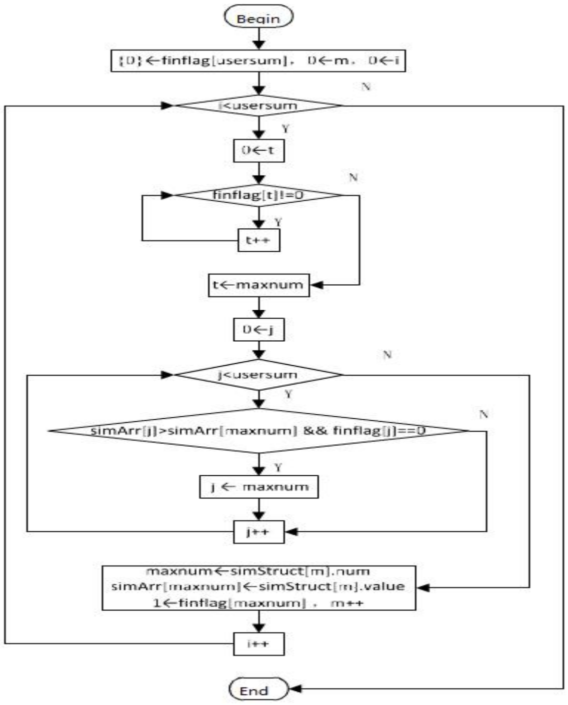
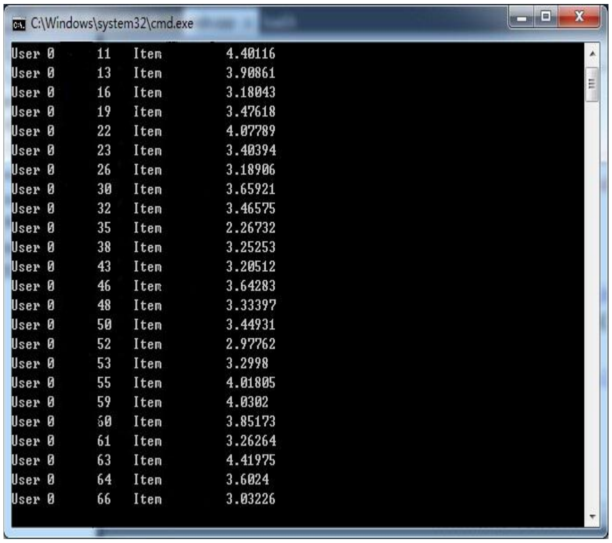


Figure 4. Flow chart of predicting the degree of user's interest



```
User 0     11   Item       4.40116
User 0     13   Item       3.90861
User 0     16   Item       3.18043
User 0     19   Item       3.47618
User 0     22   Item       4.07789
User 0     23   Item       3.40394
User 0     26   Item       3.18906
User 0     30   Item       3.65921
User 0     32   Item       3.46575
User 0     35   Item       2.26732
User 0     38   Item       3.25253
User 0     43   Item       3.20512
User 0     46   Item       3.64283
User 0     48   Item       3.33397
User 0     50   Item       3.44931
User 0     52   Item       2.97762
User 0     53   Item       3.2998
User 0     55   Item       4.01805
User 0     59   Item       4.0302
User 0     60   Item       3.85173
User 0     61   Item       3.26264
User 0     63   Item       4.41975
User 0     64   Item       3.6024
User 0     66   Item       3.03226
```

Figure 5. Collaborative filtering based on user clustering