A Method to Solve Cold-Start Problem in Recommendation System based on Social Network Sub-community and Ontology Decision Model

Meng Chen[1] • Cheng Yang[2] • Jiechao Chen[2] • Peng Yi[2]

Abstract. The paper presents a method to solve Cold-start problem in collaborative filtering recommendation system. Social sub-community is divided following analyzing exiting users' history data and mining relationship between each other. Then ontology decision model is built in the basis of sub-community and users' static information, which makes recommendation for new user based on his static ontology information. At last, the proposed method is used to recommenditems tonew users. In this paper, data simulation experiment is taken to test the technical method.

Keywords: Cold-start, recommendation system, ontology model, decision tree, collaborative filtering, social sub-community

1 Introduction

With the increased penetration of the Internet and the rise of the digital TV, users enjoy the rich resources and service， while they have to face the problem of "information fog". In this case, it is an urgent need that the private recommendation system is used in digital TV to provide users with personalized information filtering service.

Private recommended algorithm is including of content-based recommendation[1], collaborative filtering recommendation[2][3], association rules based recommendation[4] and hybrid recommendation[5], of which collaborative filtering

[1]Meng Chen (🖂)
Information Engineering School
Communication University of China ,Beijing, China
e-mail:chenmeng14@cuc.edu.cn
[2]Cheng Yang, Jiechao Chen, Peng Yi
Information Engineering School
Communication University of China ,Beijing, China

is the most popular method. It gets and analysis historical information to find user's similar neighbour ,then modelling user's preferences in order to make private recommendations initiatively. But when there is not enough user history behavioural information, the system is difficult to find user's similar neighbours, particularly being acute for new users who have little historical information. It is the cold start problem to be solved in this paper.

Currently cold start solutions are the following: 1)Statistical model-based approach [6]: the corresponding probability distribution statistics is made according to the user, project and initialize rates and high probability items are priority recommended; 2)Average approach [7]: the original rating matrix is filled using the average of all ratings of the item before collaborative filtering. 3）Modeapproach: The predict results of the user is the score which occurred in his rating most often. Howeverthere is still the problem of low precision in recommendations in these methods.

This article describes a method using social network sub-community and ontology decision mode, which takes into account the historical information and user ontology information. Analysing historical ratings of existing users, the user relationship network is established using social network theory, which is used to divide into multiple sub-communities based on the strength of relationship. According to the sub-community and ontology characteristics of the existing user, classification decision tree is used to train ontology decision-making model. With the new user's ontology information, the model adds him into a groped cluster to get his similar neighbours. Popular items in the sub-community arerecommended to the user.

The rest of this paper is organized as follow: Section 2 describes the model of cold-start problem and the system. Section3 details the solutions from the social network clustering model, which build a social network and cluster analysis by analysing the behaviour of the existing user preferences, user ontology decision model, and making decision and recommendation. In section 4, the proposed method is expired and compared. Finally, the contributions of this paper are summarized in Section 5.

2 Related Works

2.1 Collaborative Filtering

Private recommendation system [4] has three elements: items, users and recommendation algorithm. Let C be the set of all users and S means the set of all items; Utility function u () is used to calculate the recommended degrees of item s to the user c. What is the main problem of recommendation algorithm is finding the item s*which has the largest recommended degree.[3]Like:

$$\forall c \in C, s* = \arg\max_{s \in S} u(c,s) \qquad (1)$$

The idea of traditional collaborative filtering is: 1) calculating the similarity sim（c, c’）between the user c and others c’ with rating vectors, to find the preference neighbor of the user c; 2) u(c,s) means the result of weighted average with the ratings of user c* to item s and sim（c, c*）:

$$\mathrm{u(c,s)} = \mathrm{ave}(\sum \mathrm{u}(\mathrm{c}^*, \mathrm{s}) * \mathrm{sim}(\mathrm{c}, \mathrm{c}^*)) \ (\mathrm{ci} \in \mathrm{C}) \text{（2）}$$

Where

$$\mathrm{c}^* = \arg\max\nolimits_{\mathrm{c}' \in \mathrm{C}} \mathrm{sim}(\mathrm{c}, \mathrm{c}') \qquad (3)$$

The solution of cold-start proposed in this paper is to find a method in place of the equation (3).

2.2 Social Network

Social network [8] is a collection of social actors and the relationship, including nodes (social actors), the edge between the nodes (the actors’ association) and the weights of the edges (the impact between the actors). Each node, not independent individuals, is interdependent by sides. Sides, the channels of resource flowing, provide guidance for individual actions, greater weight bring stronger guidance.

Sub-community whose points have strong relationship with each other means nodes’ impact to others is larger than those outside. When multiple edges between the nodes, it can be changed to the matrix for data analysis and refining side weight in order to simplify the network.

2.3 Ontology Modelling

According to Studer’s definition putting forward in 1998, the ontology [9] is a shared conceptual model explicit formal specification. The goal of ontology is to capture the knowledge of related fields, to provide a common understanding of the domain knowledge to determine the terms of mutual recognition in the field, and give a clear definition of the mutual relations between these terms and terminology from the different levels of formalization model.

2.4 Decision Tree

Decision tree is a supervised learning, widely used in business decision-making, risk analysis and other fields. Classification and Regression Tree[10] (Classification And Regression Tree, CART) proposed by Breiman in 1984 creates a simple binary tree from the top down based on the training set for classification of the training set. CART decision tree is a high-level overview of all the sample data, which not only can accurately identify all categories of the sample, but also can effectively identify the class of the new sample [11].

3 Algorithm Analyses

The proposed algorithm structure, which is designed based on social network ideas, is divided into two parts: existing users' data mining and prediction for new users. As shown in Figure.1, the first part returns user sub-communities through clustering after making up social network topology map with existing users' dynamic information. Then ontology decision model is established by analyzing sub-communities and users' static information. The second part is for a new user, whose static information is used to input decision model to find his sub-community. At last, results are recommended to the new user followed sub-community average recommendation.

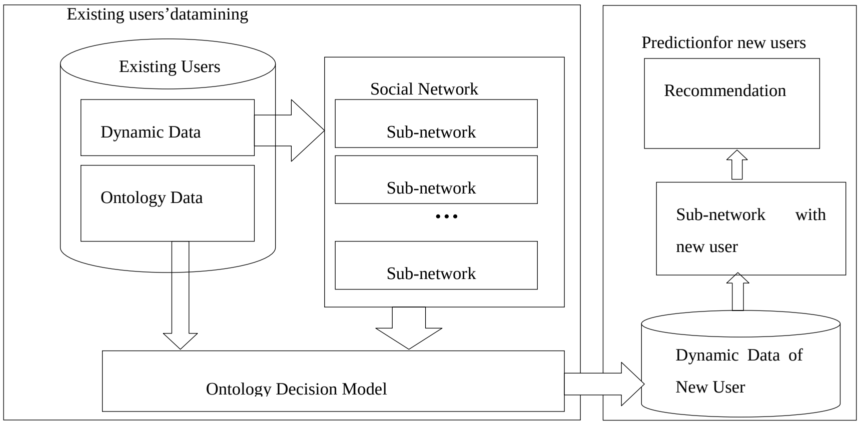

Fig. 1 proposed algorithm architecture

3.1 Social Network Clustering

A social network needs three elements: nodes, edges and weight. Let user as the nodes, relationship as rides and the degree of similarity as weight. According to existing users' dynamic data, user-item matrix is described as Table.1. With the ratings of items those are both watched by some two users, Pearson similarity is computed by formula (4).

$$r_{xy} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{[n\sum x_i^2 - (\sum x_i)^2][n\sum y_i^2 - (\sum y_i)^2]}} \quad (4)$$

Where x and y means some two users, x_i and$y_i$means the ratings of items from User x and User respectively, and n is the number of items. The results are between -1 and 1, where only r_{xy} up to 0 indicts there is a side between them. The Pearson weight larger means the relationship stronger. So the network is established.

Table 1 Users-video matrix

Item / User	Item1	Item2	……	Itemn
User1	R_{11}	R_{12}	……	R_{1n}
User2	R_{21}	R_{22}	……	R_{2n}
……	……	……	……	……
Userm	R_{m1}	R_{m2}	……	R_{mn}

K-means [12]which is one of unsupervised learning algorithms in data mining, is used to divided into sub-communities. A set of means C={C1,C2,⋯Ck} is got by classifying a set of nodes X={x1， x2， ⋯xn} which have n nodes, into k sub-communities. In first, the k means in C are initialized randomly. Secondly, each node xiin X gets its nearest mean c_j in C by computing. Like:

$$Y = \sum_{i=1}^{n} \min_{1 \le j \le k} (x_i - c_j) \quad (5)$$

Let Y exists. Then all nodes in X are grouped into the sub-community Cj （j=1， 2， ⋯， k）of the largest similarity , followed the central means of each sub-community as the new means. Above two steps are repeated until convergence is achieved or reach the maximum number of iterations.

3.2 User Ontology Decision Model

The algorithm structure is proposed for the new user without dynamic information, so the ontology decision model is making up by user ontology characteristics and results of sub-comities, which are regarded as attributes and classes, respectively. According to the theory of ontology, the user's age, gender and occupation are selected to be attribute items in video recommendation system, which have greater impacts on the viewing behavior. The ontology model format is as follows: ['age', 'gender', 'occupation', 'class'].

CART decision tree algorithm based on information theory, whose root node is the greatest attribute of the data set information entropy, whose intermediate nodes' entropy attribute decrease and whose leaf node is the resultant categories. The formation of the decision tree is classification recursively for each node in the dataset until the node belonging to the same class or no extra attributes to divide the training sample set. Using the values of attributes for classification, the algorithm calculates classified information gain and selects the one with the greatest attribute value as the root node. A character attribute comparison is divided into left or right based on whether it is equal to the selected property value; and numeric attribute is thought about if more than the value of the selected property. Each branch of the tree use CART building algorithm until no further increase in information gain or branch is classified directly to the class label. This is a user ontology decision tree model. After the merger of a group of nodes with the same parent node, if the sum of entropy increment can be ignored, do it, which called prune.

3.3 Recommendation

First, new users are divided into sub-communities using decision tree: the static information of a new user to be tested through the decision tree, starting from the root of the tree, and gradually down along the decision tree until it reaches the leaf nodes of the tree. The leaf nodes represented category is the new user categories and new users and the class of users have similar interest preferences. The decision tree builds the bridge between the new and existing users and a new user' interests preferences can be got based on the common interest of the users in the sub-community, which visible the ideas of algorithm for cold-start and collaborative filtering recommendation algorithm are the same.

4 Experiments and Analysis

The proposed algorithm for cold-start was implemented using Python 2.6. Movies lens database with 1000users（including age, occupation and gender）, 1682 movies and 100 000 ratings, where each user has given at least 20 ratings.10% of data is using as testing data and 90% is using to train.

In this paper, the average absolute error（MAE）, which measures the accuracy of prediction from calculating the deviation between the user's predicted ratings and the actual ratings, is used as a metric. The algorithm predicted user ratings collection means as {p1, p2 ... pi} and the user corresponding to the actual ratings collection is {q1, q2 ...qi}, so

$$\mathrm{MAE} = \frac{\sum_{i=1}^{N} |p_i - q_i|}{N} \quad (7)$$

where N is the number of ratings. As the MAE value decreases, the predictive value of the recommendation algorithm is closer to the actual rating and the quality of recommendation is increasing improved.

The value of the parameter k is the number of clusters. If the value of k is too small, cannot effectively distinguish between the different interests of the user’s preference groups; if the value of k is too large, the computational overhead is very large. According to reference [13] method and Pearson similarity principle, the value of k is selected as the optimal number of clusters, so that the average distance of without class is biggest and the average within-class distance is smallest. Let cluster space is K={X,R},where X={x1,x2,…,xn}.If n points are clustered into c groups,

$$\mathrm{b} = \frac{1}{n}\sum_{j-1}^{k}\sum_{i=1}^{n_j} \min\left(\frac{1}{n}\sum_{p=1}^{n_k} r_{x_p^{(k)} x_i^{(j)}}\right) \qquad (8)$$

$$\mathrm{w} = \frac{1}{n}\sum_{j-1}^{k}\sum_{i=1}^{n_j}\left(\frac{1}{n_j-1}\sum_{q=1}^{n_j} r_{x_q^{(j)} x_i^{(j)}}\right) \qquad (9)$$

Based on the data set, k∈ [5,11],because when k=12,there will be some group in none.According to the table 2, 10 is the best k

Table 2. Result of test about k

k	5	6	7	8	9	10	11
b	0.004	0.012	0.015	0.067	0.097	0.129	0.139
w	0.203	0.216	0.222	0.214	0.229	0.242	0.231

In order to verify the effectiveness of the algorithm, comparative experiments among the proposal method (SSODM),mode method and average method is made. 10% of the users are taken as new users, respectively the three methods for experimental testing to compare the MAE values. Visible, SSODM have a relatively the smallest MAE.

Table 3. Comparison of SSODM, MODE and AVERAGE

Approach	SSODM	MODE	AVERAGE
MAE	0.7378	0.8014	0.7978

5 Conclusion

The paper presents a method combining social sub-community division and ontology decision model to solve the new user cold-start problem in collaborative filtering algorithm, which builds relationships between user static information and dynamic preferences by learning. Experience proves the function of the method and evaluation parameters. There is several points need for further research. For example, the method needs to be improved to make it apply to both new users and ordinary users and determine the optimal solution of parameters in the mathematical theory support to all data structures. In addition, user privacy and security also need to be research to get a better user experience.

Acknowledgments(1) The National Key Technology R&D Program of China (2008BAH28B06－05)(2) The National Nature Science Foundation of China (60902061)

References

1 A. I. Kovács, H. Ueno.Recommending in Context: A Spreading Activation Model that is Independent of the Type of Recommender System and Its Contents[C].Data Engineering Workshop.2007 IEEE 23rd International Conference,200704:871-878.

2 Schafer, B., Frankowski, D., Herlocker, J., Sen, S..Collaborative Filtering Recommender Systems[J]. The Adaptive Web: Methods and Strategies of Web Personalization. Lecture Notes in Computer Science, 2007,Vol. 4321

3 B. M. Sarwar, G. Karypis, J. A. Konstan, J. T.Riedl.Item-based collaborative _lteringrecommendation algorithms[C].Proceedings of theTenth International World Wide Web Conference, 2001: 285-295

4 Mican, D., Tomai, N.: Association-Rules-Based Recommender System for Personalization in Adaptive Web-Based Applications. In: Daniel, F., Facca, F.M. (eds.) ICWE 2010. LNCS, vol. 6385, pp. 85–90. Springer, Heidelberg (2010)

5 Meng C, Cheng Y, Huan W, etal. Research of Personalized Recommendation System in Interactive Digital TV[J]. Video Engineering, 201207(14):37-40.

6 X. N. Lam, T. Vu, T. D. Le, A. D. Duong.Addressing Cold-Start Problem in Recommendation Systems. Proceedings of the 2nd International Conference on Ubiquitous Information Management and Communication (ICUIMC'08), New York, USA, 2008: 208-211.

7 Y. H. Guo . On Collaborative Filtering Algorithm and Applications of Recommender Systems [D]. Dalian University of Technology, 2008.

8 Liu Jun, Social Network Analysis Introduction to [M] Beijing: Social Sciences Documentation Publishing House, 2004.

9 Stefan Decker , Michael Erdmann , Dieter Fensel and Rudi Studer, "Onto broker: Ontology based access to distributed and semi-structured information",1998

10 Breiman L, Friedman J, Stone C J, et al. Classification and regression trees[M]. Chapman & Hall/CRC, 1984.

11 J. R. Quinlan. Induction of Decision Trees. Machine learning, 1986, vol. 1, no. 1,81-106.

12 Kim, K.J., Ahn, H.: A recommendation system using GA K-means clustering in an online shopping market. International Journal of Expert System with Applied computing, pp. 1187-1191(2008)

13 ZHOU, Shi-bing, XU, Zhen-yuan, TANG, Xu-qing. Method for determining optimal number of clusters in K-means clustering algorithm [J]. Journal of Computer Applications, 2010, 30(08): pp.1995-19