ISSN 1000-9825, CODEN RUXUEW
Journal of Software, Vol.20, Supplement, December 2009, pp.123−130

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

面向个性化推荐的两层混合图模型*

张少中[1+], 陈德人[2]

[1](浙江万里学院 电子信息学院,浙江 宁波 315100)

[2](浙江大学 软件学院,浙江 宁波 315100)

Hybrid Graph Model with Two Layers for Personalized Recommendation

ZHANG Shao-Zhong[1+], CHEN De-Ren[2]

[1](Institute of Electronics and Information, Zhejiang Wanli University, Ningbo 315100, China)

[2](College of Software Technology, Zhejiang University, Ningbo 315100, China)

+ Corresponding author: E-mail: dlut_z88@163.com

Zhang SZ, Chen DR. Hybrid graph model with two layers for personalized recommendation. *Journal of Software*, 2009,20(Suppl.):123−130. http://www.jos.org.cn/1000-9825/09015.htm

Abstract: A hybrid graph model for personalized recommendation, which is based on small world network and Bayesian network, is presented. Small world network has a good property in clustering and Bayesian network is compatible for probability inference. The hybrid graph model consists of two layers. One is user's layer for representing users or customers and the other is merchandise's layer for representing goods or products. Small world network describes the relationships among the nodes of users in lower layer. The implications among nodes of merchandises are represented by Bayesian network in higher layer. Directed arcs denote the tendency of nodes between user's layer and merchandise's layer. This paper also introduces several algorithms for clustering based on small world network, structure learning and parameter learning of Bayesian network, and recommended algorithm based this model. The experimentation shows that the model be accomplished to represent the relationships from user to user, merchandise to merchandise, and user to merchandise. The experimental results show that the hybrid graph model has a good performance in personalized recommendation.

Key words: small world network; Bayesian network; personalized recommendation; hybrid graph model

摘 要: 小世界网络在聚类应用中具有良好的性质,贝叶斯网络在概率推理中也得到了广泛的研究.将小世界网络和贝叶斯网络结合起来,形成一种混合图模型,并将该模型用于个性化推荐系统中.该混合图模型由两层组成,分别是用户层和商品层.其中小世界网络用于描述用户层内用户-用户结点间的关系,贝叶斯网络用于描述商品层内商品-商品结点以及层间用户-商品结点间的偏好关系.对小世界网络的用户聚类方法、贝叶斯网络结构和参数学习方法以及两层图模型的推荐算法进行描述,实验分析表明,该模型能够很好地表示用户-用户、商品-商品以及用户-商品间的关系,推荐结果具有良好的准确度.

关键词: 小世界网络;贝叶斯网络;个性化推荐;混合图模型

* Supported by the National Natural Science Foundation of China under Grant No.70671007 (国家自然科学基金); the Postdoctoral Science Foundation of China under Grant No.20060390391 (中国博士后基金)

Received 2009-03-05; Accepted 2009-04-03

With the development of information available on the Internet, the effects of intelligence technology on various commerce applications are widespread and exponential increase. As a new intelligent information service, Personalized recommendation is introduced into E-commerce[1,2].

Watts proposed a Small world model based on analysis of human social relationship in 1998[3]. It aimed at analyzing the inner relations and interactions among social groups to discovery their characteristics and behavior methods. Small world model is similar to networks of consumer's preference in E-commerce. Consumers can be divided into some kinds of clustering or groups according to their properties and purchasing behavior. Therefore, the small world networks model can be used to study personalized recommendation model. Batul proposed a graphic model using Jumping Connection for recommending method and adopted small world networks for users clustering. Jumping Connecting connects the similar preference[4]. Adriana proposed a files sharing model based on small world networks[5]. Kashif put forwards a method for grid resources discovery based on small world model[6]. He divided the grid resources into producer and consumer. Chen proposed a relation grid of social network[7]and Liu proposed a consensus spread model based on small world model[8].

The personalized recommendation system use some training sets to establish relevant recommendation model based on Bayesian network. Aggarwal built up a directed graphical model for coordinated filter system[9]. His basic thought is to show user by node in a graph, predict result by edge and he also gave the predict result by advice between users. Huang put forwards a kind of double layers graphical model[10]. He used different layer to stand for users and goods. And arcs between goods and users show the transaction information. Ji proposed a kind of recommended mechanism based on Bayesian network[11]. He divided the recommended process into two steps.

We proposed a two-layers hybrid graph model of personal recommendation. The model is combined with two layers, one is based on small world network and the other is based on Bayesian network. The layer of small world network is used to denote users and the other one of Bayesian networks is for merchandises. In the users layer, the nodes means users and the arcs connecting users denote the similarity of these users. A small world networks is used to describe the relation of users. In the merchandises layer, a Bayesian network is used to describe the relation of these nodes in the merchandises layer. The nodes in merchandises layer denote goods and produce. The directed arcs, which connect form one good to others, denote the similarity or implication relationship of these goods.

1 Hybrid Graph Model Based on Small World Network and Bayesian Network

Small world network and Bayesian network are used to describe the different business variables in the hybrid graph model for electronic business personal recommendation. Small world network is used for consumers or users with undirected arcs and Bayesian network for merchandises with directed arcs. Directed arcs connect users and merchandises and then a two-layers hybrid graph model is structured.

Definition 1. $G=(V_{user},V_{produce},E)$ is a two-layers hybrid graph model of business recommendation. Where V_{user} means user's node, $V_{produce}$ is merchandise's node, E is arc's set of nodes, which is included directed and undirected arcs.

A two-layers hybrid graph model is shown in Fig. 1. Where, the bottom level means user's layer and the upper one means merchandise's layer.

The user's layer in this model is an undirected arcs graph, which describes the relation of the user's nodes by small world network. The undirected arcs inside the connected nodes of user's layer mean the similarity of the preference of users. These arcs are also called links. Links are weighted by relational strength. The link weight represents node's similarity or link's strength and intensity. Nodes in the same group are more similar to each other or more strongly connected. Users in a same group have the same or similar trendy of preferences.

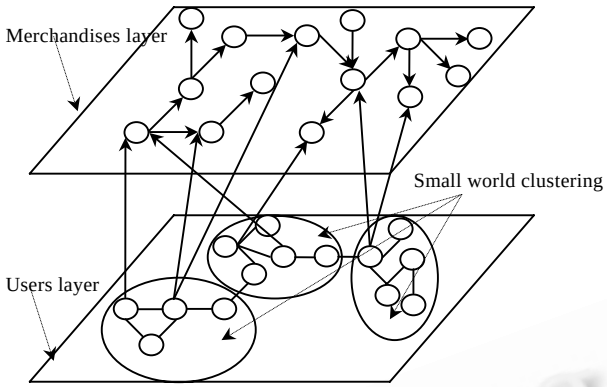

Fig.1 Two-Layers hybrid graph model for personalized recommendation

The merchandise's layer describes the relation of goods or produce to others. It is connected by directed links, which means an implicated definition among merchandises, a user that purchase certain merchandise also tends to purchase another. The properties and content of merchandise can be used to show the similarity of the merchandise.

The relations between user's layer and merchandise's layer are connected by directed links. The start node of the directed links is a user node in user's layer belonging to some node group, which is gained by small world network. The end node of links is the node of some merchandise of the merchandise's layer. The directed links between the user's layer and the merchandise's layer are connected based on trade information of users. The strength of the relation between users and merchandises can be denoted by the probability parameter. The probability parameter shows a possibility of some users selecting for some merchandises.

2 Construction of Hybrid Graph Model

2.1 Small world network for clustering of users

2.1.1 Characteristic paths and clustering coefficient

A small world network model contains two important characteristic parameters, namely characteristic path length L and clustering coefficient C.

Characteristic path length L is the mean of the shortest path lengths over all pairs of nodes, that is,

$$L = \frac{1}{n(n-2)/2} \sum_{1 \le i, j \le n} D(i, j) \tag{1}$$

Where, $D(i, j)$ denotes the links number of shortest path between discretional two connected nodes.

Clustering coefficient C is a parameter to measure the closely degree of the near neighbor nodes. C_v means the ratio of the number of actual sub-graph arcs and the most largest arcs. C means the mathematic expectation of C_v for all nodes and then it is the clustering coefficient.

$$C = \frac{\sum_{n=1}^{\infty} C_v}{n} \tag{2}$$

Small world network is a model that is between regular networks and random networks. It has a high clustering coefficient and a short average path length.

2.1.2 Clustering with small world network for users

Definition 2. For a personalized recommendation, a sub-graph is a clustering of users of small world networks, which is obtained by deleting m arcs from the regular network so that maximizing $f = aL + bC|_m$, when a regular

network $G_{user}=(V_{user},E_{user})$ with k degree is given. In which V_{user} is users nodes set and E_{user} is arcs or links set. In conditional parameter $f=aL+bC|_m$, a and b are constants, m is integer, L and C is characteristic path length and clustering coefficient respectively.

Solving the optimal connectivity problems of the nodes in a small world network is a NP problem. We propose an algorithm for the optimization as follows.

Algorithm 1. Users clustering.

Step 1. Repeat cut an arc, which could maximize f, until m arcs are moved.

Step 2. Join an arc that could maximize f and make a judgment. If the joined arc is the same as the cut one then the algorithm will end.

Step 3. Cutting an arc, which can maximize f and then jump to Step 2.

The parameters L and C that satisfy the maximal $f_{\max}$ are the clustering about the users nodes, in which the clustering group can be expressed as.

$$V'_{user}=\{v_i \mid v_i \in V_{user} \wedge D(i,j) \le L\} \tag{3}$$

2.1.3 Analysis of new user interest

New user interest analysis is to judge which clustering group is the best match by calculating the distance of the new user node to the others user nodes. Given the new user node x, the calculating algorithm is as following:

Algorithm 2. Clustering of a new user.

Step 1. Selecting the best matching clustering unit.

Calculating the path length $D_i(x,v_i)$, which is from user node x to others user node. If x has a shortest path length D_a with the node a in the network, the group that node a located in is the best clustering match. The process can be expressed as:

$$\| x-D_a \| = \min_{i \in V_{user}} \| x-D_i \| \tag{4}$$

Step 2. Adjustment of the path length of D and the direct neighbors topology of a.

$$D_i=\begin{cases} D_i+\varepsilon_a(x-D_i), & \text{if } i=a \\ D_i+\varepsilon_n(x-D_i), & \text{if } i \in N_a \\ D_i, & \text{otherwise} \end{cases} \tag{5}$$

Where N_a denotes the direct neighbor node of a, as shown in Fig.2. Nodes b_1, b_2, b_3 are direct neighbors of a, parameter ε_a and ε_n mean the learning rate of a and its direct neighbor respectively. The two rates are smaller constants. Generally, the range of ε_a is at 0.05 to 0.1 and ε_n is 0.002 to 0.01.

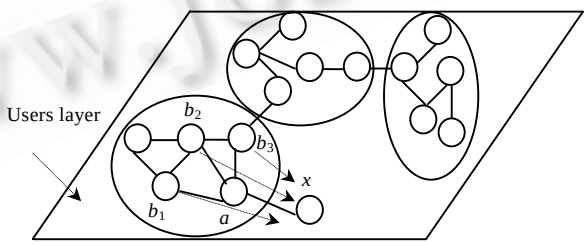

Fig.2 Clustering of new user and adjustment of structure

Step 3. Repeat Step 1 and Step 2 until all nodes are treated.

2.2 Bayesian network for causality of merchandises and users

2.2.1 Structure learning

We adopt the maximal mutual information principle to restrict complexity based on degree of Bayesian

network.

Definition 3. The complexity of structure is defined as $B(G_B)$. The complexity is closely with the number of parent nodes of a node, which is the number of directed links, it can be calculated as follows.

$$B(G_B)=\sum_{i=1}^{n} o_i \log_2(n) \tag{6}$$

o_i is the number of the parent nodes of node v_i, G_B is directed graph structure of whole nodes of users and merchandises. Here, we have no consideration about the undirected arcs in the users layer.

Definition 4. The maximal mutual information entropy score function with restriction, which is used for Bayesian network structure learning of the user-merchandise and merchandise-merchandise, is defined as:

$$MMI=\sum_{i=1}^{n} MI(v_i, Pa(v_i)) - B(G_B) \tag{7}$$

Where MI is the mutual information among the nodes, $Pa(v_i)$ means the parent nodes set of node v_i.

Structure learning uses maximal mutual information score function with restriction as the estimating score of network structure. Its aim is to make the maximal score and then get the Bayesian network topologic structure of merchandise. Merchandises nodes layer is located in the upper of the two-layers hybrid model. Its internal nodes mean the different types of goods or produces. They are connected by directed links, which show the implication relation of these merchandises. The users and customers who tend to buy some other merchandises when purchase some merchandises have some causality relations. The strength of this trendy is identified by the probability distribution. The directed acyclic graph expressed by the probability distribution is the Bayesian network. In order to analysis the implication relation between goods, we should build Bayesian network first, and this building process is Bayesian network learning.

2.2.2 Parameter learning

The density of probability distribution of the users-merchandises and the merchandises-merchandises can be received by parameter learning. Assuming there is a fixed unknown parameter θ, considering all the possible value of parameter θ when the topology structure G_B is given. We can find the maximum posterior probability distribution when topology structure G_B and training sample set S are given. Based on Bayesian rules, there is:

$$P(\theta \mid S, G_B)=\frac{p(S \mid \theta, G_B)P(\theta \mid G_B)}{p(S \mid G_B)} \tag{8}$$

$P(\theta \mid G_B)$ is the priori probability distribution of the parameter θ of structure G_B and $P(S \mid G_B)$ has nothing with the actual value of the parameter. In some references about Bayesian network, the most used priori probability distribution is *Dirichlet*. Considering the parameter of polynomial are $\theta_1,...,\theta_k$ and $\sum_i \theta_i = 1$, *Dirichlet* distribution is determined by a set of hyper-parameter $\alpha_1,...,\alpha_k$. The priori probability of θ is as follows when $P(\theta \mid S, G_B)$ approves of *Dirichlet*.

$$P(\theta \mid S, G_B)=Dir(\theta \mid \alpha_1,...,\alpha_k)=\frac{T(\alpha)}{\prod_i T(\alpha_i)}\prod_i \theta^{\alpha_i-1} \tag{9}$$

So there is $P(v_i)=\frac{\alpha_i}{\sum_j \alpha_j}$ and for sample dataset S, the statistic values are: $N_1,...,N_k$, and then:

$$P(\theta \mid S, G_B)=Dir(\theta \mid \alpha_1+N_1,...,\alpha_k+N_k) \tag{10}$$

We can expand this result for Bayesian network. Defining that event is $V=v$ and the statistic value of the parents' node is $Pa(v)=u$, which is can be written as $N(v,u)$. In the algorithm of Maximum Likelihood Estimate, the estimating of parameter can be calculated by the following formula:

$$\hat{\theta}_{v|u}=\frac{N(v,u)}{N(u)} \tag{11}$$

In the algorithm of Bayesian rules, the estimating of parameter can calculated by the following formula:

$$\hat{\theta}_{v|u} = Dir(\alpha_1 + N(v_1,u),...,\alpha_{vk} + N(v_k,u)) \tag{12}$$

By the formula (11) or (12), we can estimate the conditional probability table of each node and we can get a complete graph model G_B which shows the behavior characteristics and complication relationship of nodes of users-merchandises and merchandises-merchandises.

And then, we can get a hybrid graph model structure G and it is:

$$G = G_{user} \cup G_B \tag{13}$$

3 Algorithm of Recommendation

The basic object of recommending system is to get the recommendation sets by recommending algorithm. The two-layers hybrid graph model in this paper is composed of the small world network for representing users and the Bayesian network for representing merchandises. Directed links between layers is from users to merchandise. And it is also described by Bayesian network. So the essentially method is clustering analysis by use of small world network for users and inference by use of Bayesian network.

The initialized inputs can utilize some users information for recommending algorithm including the attributes and browsing process of a user. A proper user-clustering group will be gained by clustering matching with other users in small world network based on this information. Then all the other users nodes, which connect to this user, are selected based on a threshold of path length in the clustering. The recommended merchandise set of these users will be obtained by Bayesian network inference using these nodes as proofs. Finally, a set of recommendation of merchandise is presented for user according to their order of probability distribution.

For a two-layers graph model, the small world network of users layer, the Bayesian network of merchandise layer and links between users and merchandise is built up offline by training sample sets. Clustering matching of new user, re-conjunction between new user and primary users in small world network and inference with Bayesian network are processed online. So the recommending algorithm has a very good efficiency. The recommending algorithm is designed as following:

Algorithm 3. Recommending for a new user.

Input: A two-layers hybrid graph model $G = (V_{user}, V_{produce}, E, \theta)$. And V_{user} means nodes set of user and $V_{produce}$ means nodes set of merchandise. G is a two-layer hybrid graph model and θ is parameter. E is the set of arcs and x is the set of new user and its properties.

Output: A recommended set of merchandise for x.

Step 1. Clustering and re-constructing of new user based on algorithm 2 in chapter 2.1.3, and the set of user clustering and path length can be gained, they are:

$$V''_{user} \leftarrow \{v_i \mid v_i \in V'_{user} \wedge \| x - D_a \| = \min_{i \in V_{user}} \| x - D_i \| \}.$$

and D_i respectively.

Step 2. Calculating of $\min \| x - D_i \|$ for the path length set $D = \{D_1, D_2, ...D_i\}$ from user nodes in the group to new user node x.

Step 3. If $\min \| x - D_i \| \le \eta$, then add the node into the set of evidence set:

$$F \leftarrow \{v_i \mid v_i \in V'_{user} \wedge \min \| x - D_i \le \eta \| \}.$$

Step 4. Calculating the conditional probability distribution of which these nodes reaching all merchandises when they are as evidence nodes.

$$P(v_i \mid F) = \prod_{v_i \in V_{produce}} P(v_i \mid Pa(v_i))|_F .$$

Step 5. Combination of all the merchandises nodes that are satisfied a certain probability threshold, that is, the goods recommending set of a new user. And the process can be presented by:

$$\text{if} \quad P(v_i \mid F) > \mu \quad \text{then let} \quad R = \{v_i\} \cup R .$$

Step 6. Repeat Step 4 and Step 5 until all the merchandises nodes are processed.

4 Experimentations and Analysis

For evaluating the accuracy and real-time of the two-layer hybrid recommending model, we use the mean absolute error to evaluate the model.

Mean absolute error is defined as:

$$MAE = \frac{\sum_{x \in X} \sum_{v \in V_{produce}} | PreR_{x,v} - R_{x,v} |}{n} \tag{14}$$

Where, $PreR_{x,v}$ and $R_{x,v}$ mean the predicting value and the actual value of the user preference respectively. X and $V_{produce}$ mean the known sample set and merchandises node set. Mean absolute error directly shows the accuracy of recommending. It comes from the dispersion, which is between the predicting value and actual value of user preference. That is a different value of the comments score of production or merchandises from users. And then, mean absolute error can be obtained by adding these dispersions together and gets the mean value. The smaller the mean absolute error, the better the recommending performance, on the contrary, the worse the recommending capability.

We use MovieLens[12] database to evaluate the model and algorithm proposed. There are three tables in the database and they are Person, Movie and Votes. A statistics of the set is as shown in Table 1.

Table 1 Samples set for MovieLens of users

Dataset	Number of users	Number of movies	Number of votes
Movielens	943	1 682	100 000

In the experiment, we select training samples randomly from the user votes dataset according to a proportion of 10%, 30%, 50%, 70% of all the users and then take all these votes score as training set. Testing data set can be selected outside the training set. The result of the recommendation comparing with the other recommending method[13,14] is as shown in Fig.3.

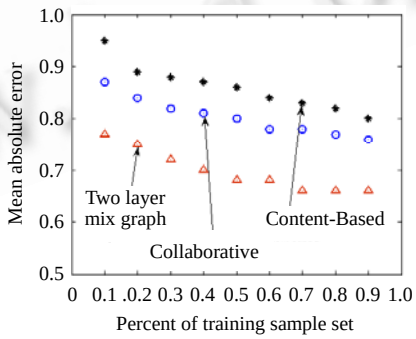

Fig.3 Comparable result of mean absolute error of three methods

When the votes score dataset of users, which is the known score dataset, is sparse. That is, the known votes dataset is small. The mean absolute error of personalized recommendation model with two-layers are obviously better than the model based on collaborative method and a little better than the model based on content-based. When there are more votes score data, the discrepancy of mean absolute error of the three models get smaller obviously.

The most excellent is still the two-layers model. So we can conclude that the personalized recommendation based on two-layers model has the best accuracy and it can also resolve data sparse effectively.

5 Conclusion

The hybrid graph model with Two Layers for personalized recommendation combine the small world model and the Bayesian network. It makes use of the clustering property of the small world network and the probability inference characteristic of Bayesian network. The small world network is used to reflect the implication relations of users-users and the Bayesian network is for users-merchandises. The result of experiment shows that the two-layer hybrid graph model has a good accuracy in the personalized recommendation.

References:

[1] Sarwar B, Karypis G, Konstan J, Riedl J. Analysis of recommendation algorithms for e-commerce. In: Proc. of the 2nd ACM Conf. on Electronic Commerce. 2000. 158−167.

[2] Zhao L, Hu NJ, Zhang SZ. Algorithm design for personalization recommendation systems. Journal of Computer Research and Development, 2002,39(8):986−991 (in Chinese with English abstract).

[3] Watts DJ, Strogatz SH. Collective dynamics of small-world networks. Nature, 1998,393:440−442.

[4] Mirza BJ, Keller BJ, Ramakrishnan N. Studying recommendation algorithms by graph analysis. Journal of Intelligent Information Systems, 2003,20(2):131−160.

[5] Iamnitchi A, Ripeanu M, Toster I. Small-World file-sharing communities. In: Proc. of the INFOCOM. 2004.

[6] Ali K, Datta S, Datta S, Aboelaze M. Grid resource discovery using small world overlay graphs. In: Proc. of the 18th IEEE Canadian Conf. on Electrical and Computer Engineering. 2005.

[7] Chen SY, Song JX, Liu WD, Wang C. Relation grid: Small world based social relationship network. Application Research of Computers, 2006,23(5):194−197 (in Chinese with English abstract).

[8] Liu CY, Hu XF, Si GY, Luo P. Public opinion propagation model based on small world networks. Journal of System Simulation, 2006,18(12):3608−3610 (in Chinese with English abstract).

[9] Aggarwal CC, Wolf JL, Wu KL, Yu PS. Horting hatches an egg: A new graph-theoretic approach to collaborative filtering. In: Proc. of the KDD 1999. 1999. 201−212.

[10] Huang Z, Chung WY, Chen HC. A graph model for e-commerce recommender systems. Journal of the American Society for Information Science and Technology, 2004,55(3):259−274.

[11] Ji JZ, Sha ZQ, Liu CN. A method of commodity recommendation based on customer shopping model of Bayesian network. Application Research of Computer, 2005,22(4):65−69 (in Chinese with English abstract).

[12] http://www.grouplens.org/node/73

[13] Balabanovic M, Shoham YF. Content-Based, collaborative recommendation. Communications of the ACM, 1997,40(3):66−72.

[14] Lin W, Alvarez SA, Ruiz C. Efficient adaptive-support association rule mining for recommender systems. Data Mining and Knowledge Discovery, 2002,6:83−105.

附中文参考文献:

[2] 赵亮,胡乃静,张守志.个性化推荐算法设计.计算机研究与发展,2002,39(8):986−991.

[7] 陈绍宇,宋佳兴,刘卫东,王诚.关系网格:一种基于小世界模型的社会关系网络.计算机应用研究,2006,23(5):194−197.

[8] 刘常昱,胡晓峰,司光亚,罗批.基于小世界网络的舆论传播模型研究.系统仿真学报,2006,23(5):194−197.

[11] 冀俊忠,沙志强,刘椿年.一种基于贝叶斯网客户购物模型的商品推荐方法.计算机应用研究,2004,55(3):259−274.

ZHANG Shao-Zhong was born in 1969. He is a Ph.D. candidate and a professor at Institute of Electronics and Information, the Zhejiang Wanli University. His current research areas include data mining theory and application.

CHEN De-Ren was born in 1961. He is a professor and doctoral supervisor at the College of Software Technology, the Zhejiang University. His research areas are e-commerce theory and application.