Intelligent Mining on Purchase Information and Recommendation System for E-Commerce

Weikang Xue, Bopin Xiao, Lin Mu
Department of Reliability and Systems Engineering, Beihang University, Beijing, China
xue320320@126.com
xbp@buaa.edu.cn

Abstract - **As an important marketing tool, recommendation systems for e-commerce offer an opportunity for merchants to discovery potential consumption tendency. This paper puts forward a novel recommendation algorithm to make the recommendation system more accurate, personalized and intelligent. Firstly, we use intelligent mining on purchase information, and regress consumer preference rating on click behavior. Secondly, we use Bipartite Network Recommendation model based on resource allocation and improved collaborative filtering model; the former abstracts products and consumers into nodes in the graph, and finds the correlation of products that recommend to others using alternative relation; and the latter solves the problem, caused by sparse data, by compressing rating matrix and predicting null values. Finally, according to Alibaba e-commerce customers purchase data, we verify that Hybrid Recommendation Model optimizes the accuracy and coverage of the recommendation results.**

Keywords - **purchase information, collaborative filtering, sparse data, recommendation system**

I. INTRODUCTION

With the development of web 2.0 technology, electronic commerce has become one of the most important channels in the modern-way shopping. However, customers are often trapped in multifarious information, which make them harder to choose goods. Under this background, the intelligent data miming on e-commerce customers' purchase information and goods recommendations are showing their importance. On the one hand, it provides customers with fast and convenient way to trade and choose goods in the complex information space. On the other hand, it helps shops to more deeply understand the customers' demand and dig out the characteristics of shopping behaviors, which improves customer loyalty, excavates potential consumers and improves the ability of cross-selling.

Literature review

Various recommendation systems and algorithms have emerged in e-commerce field. In the mid-1990s, personalized recommendation system LIRA was first proposed in the American association of artificial intelligence, and collaborative filtering algorithm appeared at Bell Labs in the United States. In 2003, Google developed a function mode called AdWords, which pushes ads with high similarity with customers' interest through the analysis of the records of keywords that customers frequently search. Methods in this field mainly include content-based recommendation and collaborative filtering recommendation, knowledge-based recommendation, etc. [1] proposed a measurement called Item-to-Item correlation, which belongs to content-based recommendation. The collaborative filtering algorithm is by far most widely recommended algorithm. [2] used collaborative filtering algorithm to recommend, and the degree of accuracy was increased. Although it has the ability to recommend new information, it has difficulty to analyze characteristics of products and deal with data sparseness problem, so some improved collaborative filtering algorithms (in [5] and [6]) were put forward. [14] elucidated the current researches about e-commerce recommendation have mainly two problems: 1) The relevant information was not sufficiently mined 2) Model data extremely sparse. So this article combines with a variety of advanced technologies to provide customers with more efficient, real-time recommendation service.

In recent years, some people proposed combination model (hybrid recommendation model), for example [9], etc. The biggest advantage of this model is to avoid the disadvantage of single recommendation algorithm. Combination recommend model is generally through switching, series, weighting, features expanding, features combining, mixing, to present different mixed strategies. In [3], Ansari, Essegaier and other scholars put forward a new combination recommendation model in journal of "Internet recommendation systems" in Marketing Research. This model inputs features of customers and products to calculate the utility value. For example, object j's utility value for customers i, which is determined by customer attribute z, commodity attribute w and their interaction (such as selection) x factor. The three parameters of the three distributions are estimated by Monte Carlo method. Assume three variables follow normal distribution and represent randomness of system, customer properties and product properties. The expression as shown below:

$r_{ij} = x_{ij}\mu + z_i\gamma_j + w_j\lambda_i + e_{ij}$, Where $e_{ij} \sim N(0,\sigma^2)$, $\lambda_i \sim N(0,\Lambda)$, $\gamma_j \sim N(0,\tau)$

II. HYBRID RECOMMENDATION MODEL

Recommendation model has three modules, including input module, recommendation algorithm module and

output module. Input module uses implicit mining and explicit obtaining, mining data respectively through demographic characteristics, evaluation index and comments, customer purchase records, page retention time, history and so on. In recommendation algorithm module, collaborative filtering recommendation algorithm is combined with sparse data compression and null values estimation step. It first deals with data sets, solving the low recommendation accuracy of traditional algorithm due to the sparse data problem. It is also combined with network structure-based resource allocation recommendation method, and makes up for the shortcomings of the former. Output module uses the Top - N algorithm, recommending goods with high similarity of customers' preference but they never read before. The system structure is shown below:

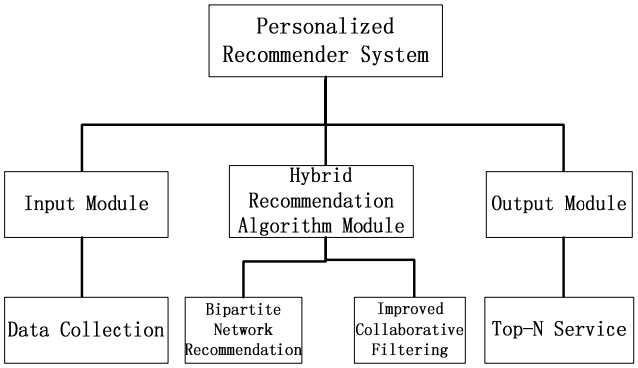

Fig. 1. The architecture of recommendation system

Bipartite Network Recommendation model based on resource allocation like association-rule focuses on relation among commodities. And improved collaborative filtering model focuses on similarity between customers. To combine the both in reasonable proportions, we put forward a back-fuse method, called Hybrid Recommendation Model, which determines the proportions by abundant experiments training. The Alibaba e-commerce website plans to offer every customer 12 recommendations, we can determine the number(x, y) of elements in A(Bipartite Network Recommendation set) and B(improved collaborative filtering recommendation set), according to average times(a, b) that real purchase appears in recommendation set A and B.

$$\text{S.t.}\begin{cases}\dfrac{x}{y}\approx\dfrac{a}{b}\\ x+y=12\\ x,y\in\{1,2,\cdots,12\}\end{cases}\qquad(1)$$

We hope that the recommendation is accurate and suitable for more customers and commodities, so Precision and Recall are applied as index. H_i denotes the number of intersections between recommendation list and purchase list for customer i.

$Precision=\dfrac{\sum_{i=1}^{N}H_i}{\sum_{i=1}^{N}L_i}$, N denotes the number of customers who we offer recommendation, L_i denotes the number of recommendation commodities for customer i.

$Recall=\dfrac{\sum_{i=1}^{M}H_i}{\sum_{i=1}^{M}K_i}$, M denotes the number of customers who have actual behavior, K_i denotes the purchase quantity of customer i

Finally, we use F to optimal fit Precision and Recall, that is $F=\dfrac{2\cdot P\cdot R}{P+R}$ (P denotes Precision, and R denotes Recall). Apparently, the F not only guarantees the probability that the recommendation is valid to chose, but also ensures the probability that the recommendation is in customer purchase. Thus, we can use F to evaluate the recommendation results.

III. OPTIMIZATION ALGORITHM

A. Bipartite Network Recommendation model based on resource allocation

Bipartite Network Recommendation model based on resource allocation abstracts customers and commodities to nodes in network. This method is similar with resource allocation and uses the information hidden in the choice relation between customer and commodities. This method focuses on the interaction among commodities, and improves the ability of personalized recommendation comparing with association-rule model. Assuming that an E-commerce website has m commodities and n customers, we abstract them into $m+n$ nodes. If an alternative relation between customer i and commodity j is existed, then an edge between node i and j is added, i.e. $a_{ji}=1(i=1,2,\cdots,n;\ j=1,2,\cdots,m)$; Otherwise $a_{ji}=0$. In fact, the commodities that have been chosen by customer i have the ability to recommend other commodity for i. k_l denotes the degree of commodity j, which is how many customers have chosen it ever. k_l denotes the degree of customer l, which is how many commodities have been chosen by him. w_{ij} denotes how many resource can commodity i get from commodity j; We can get the general expression as follow:

$$w_{ij}=\frac{1}{k_j}\sum_{l=1}^{n}\frac{a_{il}a_{jl}}{k_l}\qquad(2)$$

For customer i, the original resource of the commodities that have been chosen by him ever is 1, and the original resource of other that have not been chosen is 0. In this way, we can get a vector $\mathbf{P}_i$ composed by 0/1 whose dimension is m, representing the original

preference vector of the customer. According to the resource allocation process above, we can get the final preference vector of the customer as following:

$$\mathbf{P}_i^{'} = \mathbf{P}_i\mathbf{W} = \mathbf{P}_i \bullet \begin{pmatrix} w_{11} & w_{21} & \cdots \\ w_{12} & w_{22} & \cdots \\ \vdots & \vdots & \ddots \end{pmatrix} \quad (3)$$

We order the vector P corresponding numerical value s of component elements (the larger the value means the more customer likes), and then delete the goods that the t arget customer has browsed. For every customer in training set, we select top 10 elements that have not been chosen from $\mathbf{P}_i^{'} = \{p_{\mu 1}, p_{\mu 2}, p_{\mu 3}, p_{\mu 4}, p_{\mu 5}, p_{\mu 6}, \cdots\}$, that is $\{\mu_1, \mu_2, \cdots, \mu_{10}\}$, and put it into the recommendation set of the customer.

B. Improved collaborative filtering based on BP-Neural Network

We define the customer-commodity preference score matrix as follow, $U_i (1 \le i \le n)$ denotes preference vector of the customer i, $P_j (1 \le \mathrm{j} \le \mathrm{m})$ denotes score vector of the commodity j

$$\mathbf{UP} = \begin{pmatrix} u_1p_1 & u_1p_2 & \cdots & u_1p_m \\ u_2p_1 & u_2p_2 & \cdots & u_2p_m \\ \vdots & \vdots & \ddots & \vdots \\ u_np_1 & u_np_2 & \cdots & u_np_m \end{pmatrix} = \begin{pmatrix} \mathbf{U}_1 \\ \mathbf{U}_2 \\ \vdots \\ \mathbf{U}_n \end{pmatrix} = \begin{pmatrix} \mathbf{P}_1 & \mathbf{P}_2 & \cdots & \mathbf{P}_m \end{pmatrix}$$

We define the sparsity of the preference score matrix $\beta = 1 - \frac{\text{non-null matrix elements}}{\text{Total number of matrix elements}}$. Since the $\mathbf{UP}$ matrix is extremely sparse, we improve the method to search the nearest neighbors of $\mathbf{U_i}$ accurately. (In this study, we define the nearest neighbors are those who have higher similarity.) We manipulate the data compression by following steps:

α denotes the number of elements in nearest neighbor candidate set, $\beta_{\min}$ denotes the minimum sparsity of preference score matrix in candidate set, γ denotes the number of nearest neighbors of $\mathbf{U_i}$. After the preference vector of customer i has been inputted, the nearest neighbors set $\mathbf{NU}_{\mathbf{i}}$ of $\mathbf{U_i}$ is outputted.

1) $\mathbf{U}_i^{'} = u_i \bullet \theta(\mathbf{U}_i), \mathbf{S} = (\ ), \mathbf{S}^{'} = (\ ), \mathbf{S} = \mathbf{S} \cup \mathbf{U}_i, \mathbf{S}^{'} = \mathbf{S}^{'} \cup \mathbf{U}_i^{'}, \mathbf{V} = \mathbf{U}_i$;

Operator functions $\theta(\mathbf{U}_i)$ pick up all non-null elements into the new vector from the preference vector of the customer i, for example, $\theta(\mathbf{U}_i) = (p_3 \quad p_7 \quad \cdots \quad p_{m-3})$.

2) To make the $\frac{|\theta(\mathbf{V}) \cap \theta(\mathbf{U}_j)|}{|\theta(\mathbf{V}) \cup \theta(\mathbf{U}_j)|}$ maximum, we pick up $\mathbf{U}_j$ from $\mathbf{UP} - \mathbf{S}$ $\mathbf{U}_j^{'} = \theta(\mathbf{U}_j), \mathbf{S} = \mathbf{S} \cup \mathbf{U}_j, \mathbf{S}^{'} = \mathbf{S}^{'} \cup \mathbf{U}_j^{'}$;

The purpose of this step is to find out nearest neighbor candidate set whose preference vectors have more intersection.

3) If $|\mathbf{S}| \ge \alpha$, then go to step 5; otherwise continue;

4) Let V = the element $\mathbf{U}_\mathrm{j}$ that added into S newly, then go to step 2;

5) Let $\mathbf{S}^{'} = \sigma(\mathbf{S}^{'})$; $\beta \le \beta_{\min}$, σ is aligning and adding null value for $\mathbf{S}^{'}$ matrix, i.e. $\sigma(\mathbf{V}_1, \mathbf{V}_2, .., \mathbf{V}_n) = (\mathbf{V}_1^{'}, \mathbf{V}_2^{'}, .., \mathbf{V}_n^{'})$,

$\mathbf{V}_i^{'} = \{v | v \in \theta(\mathbf{V}_i)\} \cup$

$\{v = 0 | v \notin \theta(\mathbf{V}_i) \text{ and } v \in \theta(\mathbf{V}_1) \cup .. \cup \theta(\mathbf{V}_{i-1}) \cup \theta(\mathbf{V}_{i+1}) \cup .. \cup \theta(\mathbf{V}_n)\}$

6) Using BP(Back Propagation) neural network to evaluate the null value in $\mathbf{S}^{'}$ as following table1, in which $u_2 \sim u_5$ is the training set, and activation function $f(x) = \frac{1}{1 + e^{-ax}}$ $(0 < f(x) < 1)$

Table1

Using BP- neural network to evaluate u_1p_5 value

User	In					Out
	p_1	p_2	p_3	p_4	p_6	p_5
u_1	1	0	2	1	3	1
u_3	4	5	1	5	1	3
u_4	1	2	3	0	1	4
u_5	0	2	4	1	0	5
u_1	2	3	0	4	1	?

Until $\beta \le \beta_{\min}$, then go to the next step;

7) In $\mathbf{S}^{'}$, we obtain the nearest neighbors set $\mathbf{NU}_{\mathrm{i}} = \{\mathbf{U}_\alpha, \mathbf{U}_\beta, \mathbf{U}_\gamma, \cdots\}$ by revised cosine similarity and then order in ascending sort;

Similarity of u and v

$$\mathrm{sim}(\mathrm{u}, \mathrm{v}) = \frac{\sum_{i \in \phi(\mathrm{u},\mathrm{v})} (\mathrm{R}_{u,i} - \mathrm{R}_u)(\mathrm{R}_{v,i} - \mathrm{R}_v)}{\sqrt{\sum_{i \in \theta(\mathrm{u})} (R_{u,i} - R_u)^2} \sqrt{\sum_{i \in \theta(\mathrm{v})} (R_{v,i} - R_v)^2}} \quad (4)$$

In formula (4), $\theta(\mathrm{u})$ denotes preference vector of the customer u, $\phi(\mathrm{u}, \mathrm{v})$ denotes the commodity intersection that customer u and v both evaluate, and R_u denotes the customer u arithmetic mean of preference score in history, and $\mathrm{R}_{v,i}$ denotes the customer v preference score for commodity i, and $\theta(\mathrm{v})$, R_v, $\mathrm{R}_{\mathrm{u},i}$ by this analogy.

We calculate mean absolute error (MAE) that is deviation between real preference score and predicted preference score and then determine α that is the number of elements in nearest neighbor candidate set. The value of MAE is bigger, the prediction accuracy is worse. Assuming the predicted customer preference score set is $\{p_1, p_2, \cdots, p_N\}$, and the real customer preference score set is $\{q_1, q_2, \cdots, q_N\}$, and then MAE is as formula (5).

$$E_{MAE} = \frac{\sum_{i=1}^{N} | p_i - q_i |}{N} . \quad (5)$$

IV. APPLICATION CASE

68751 items of Alibaba E-commerce purchase have been used in experiment, including 2793 kinds of commodity and 368 customers. (The data was from Tianchi Recommendation Algorithm competition, launched by Alibaba.) The training data set, accounting for 70% of total, is used to adjust the parameters of two algorithms and weigh the proportion of two algorithms in final recommendation. The test data set, accounting for 30% of total, is used to testify the effect of Hybrid Recommendation Model and then compare with traditional collaborative filtering model and improved collaborative filtering model.

According to the experiment result, the curve graph that MAE value influenced by the number of nearest neighbors is drawn. From the graph 2, we can find that improved collaborative filtering model is more accurate after sparse data is dealt by BP-neutral network. We can also find that the MAE value tends to be stable when the number of nearest neighbors is about 20.

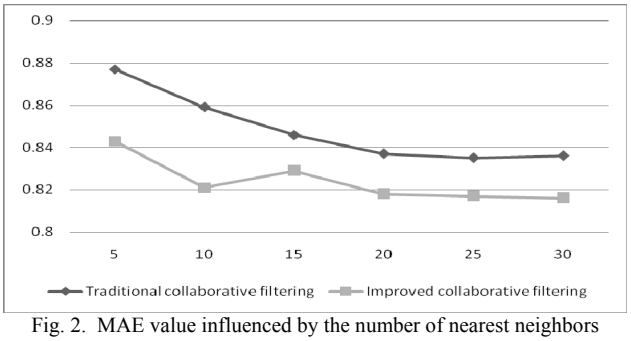

Fig. 2. MAE value influenced by the number of nearest neighbors

Finally we select the highest preference score commodity from the top 20 neighbors $\mathbf{U}_{\beta_1}, \mathbf{U}_{\beta_2}, \cdots \mathbf{U}_{\beta_\gamma}$, so that $B = \{\gamma_1, \gamma_2, \cdots \gamma_{20}\}$ is added into customer i recommendation set. On the other hand, Bipartite Network Recommendation model based on resource allocation recommends $A = \{\mu_1, \mu_2, \cdots\}$ for customer i. According to formula (1), we know x=4，y=8, and the final recommendation set of customer i is $C = \{\mu_1, \mu_2, \cdots, \gamma_1, \gamma_2, \cdots\}$. With the increasing number of test samples, Hybrid Recommendation Model's Precision is about 0.031, Recall is about 0.104, and F is about 0.0478.

Table 2
Contrastive analysis about the experiment results

	Precision	Recall	F
collaborative filtering model	0.027	0.065	0.0382
improved collaborative filtering model	0.032	0.086	0.0466
Hybrid Recommendation Model	0.031	0.104	0.0478

V. CONCLUSION

Based on the application case above, the improved collaborative filtering recommendation model is more accurate than the traditional one. In addition, the combination of the improved collaborative filtering and Bipartite Network Recommendation model, the F value improved significantly comparing with other recommendation algorithm. (Hybrid model (F=0.0478) compared to improve collaborative filtering model (F=0.0466) and collaborative filtering model (F=0.0382).) Thus, the feasibility and advantages of Hybrid Recommendation Model are fully validated.

Bipartite Network Recommendation model based on resource allocation focuses on interaction among commodities. And the improved collaborative filtering model focuses on similarities of customers and has advantages to deal with sparse data. The combination of both makes up for their each defect, in case that recommendation results fall into narrow space. Application of this method is expected to improve the accuracy of the e-commerce personalized recommendation and customer satisfaction.

This research has some shortcomings. E-commerce transactions data used in the experiment is limited, lacking of demography and other commodities attributes, so the recommendation algorithm selection is relatively limited. In future, we will consider the emotion mining to evaluate the customer preferences. And environmental factors such as other dimensions variable will be considered comprehensively, to carry out more appropriate personalized recommendation.

REFERENCES

[1] Schafer, J. B., Konstan, J., & Riedl, J. (1999). Recommender systems in e-commerce. *Proceedings of the 1st ACM conference on Electronic commerce* (pp. 158-166). ACM.

[2] Adomavicius, G., & Tuzhilin, A. (2005). Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *Knowledge and Data Engineering, IEEE Transactions on, 17(6),* 734-749.

[3] Ansari, A., Essegaier, S., & Kohli, R. (2000). Internet recommendation systems. *Journal of Marketing research*, *37*(3), 363-375.

[4] Adomavicius, G., & Tuzhilin, A. (2001). Multidimensional recommender systems: a data warehousing approach. *Electronic commerce* (pp. 180-192). Springer Berlin Heidelberg.

[5] Sandvig, J. J., Mobasher, B., & Burke, R. (2007). Robustness of collaborative recommendation based on association rule mining. *Proceedings of the 2007 ACM conference on Recommender systems* (pp. 105-112). ACM.

[6] Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P., & Riedl, J. (1994). GroupLens: an open architecture for collaborative filtering of netnews. *Proceedings of the 1994 ACM conference on Computer supported cooperative work* (pp. 175-186). ACM.

[7] Hill, W., Stead, L., Rosenstein, M., & Furnas, G. (1995). Recommending and evaluating choices in a virtual community of use. *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 194-201). ACM Press/Addison-Wesley Publishing Co..

[8] Adomavicius, G., & Tuzhilin, A. (2001). Expert-driven validation of rule-based user models in personalization applications. *Data Mining and Knowledge Discovery*, *5*(1-2), 33-58.

[9] Kim, B. M., Li, Q., Park, C. S., Kim, S. G., & Kim, J. Y. (2006). A new approach for combining content-based and collaborative filters. *Journal of Intelligent Information Systems*, *27*(1), 79-91.

[10] Lawrence, R. D., Almasi, G. S., Kotlyar, V., Viveros, M., & Duri, S. S. (2001). *Personalization of supermarket product recommendations* (pp. 11-32). Springer US.

[11] Montgomery, A. L., Li, S., Srinivasan, K., & Liechty, J. C. (2004). Predicting online purchase conversion using web path analysis. *Marketing Science*, *23*(4), 579-595.

[12] Zhou, T., Kuscsik, Z., Liu, J. G., Medo, M., Wakeling, J. R., & Zhang, Y. C. (2010). Solving the apparent diversity-accuracy dilemma of recommender systems. *Proceedings of the National Academy of Sciences*, *107*(10), 4511-4515.

[13] Huang, Y. P., Chuang, W. P., Ke, Y. H., & Sandnes, F. E. (2008). Using back-propagation to learn association rules for service personalization. *Expert Systems with Applications*, *35*(1), 245-253.

[14] Yeh, I., Lien, C. H., Ting, T. M., Wang, Y. Y., & Tu, C. M. (2010). Cosmetics purchasing behavior–An analysis using association reasoning neural networks.Expert Systems with Applications, 37(10), 7219-7226.

[15] Yen, S. J., & Lee, Y. S. (2006). An efficient data mining approach for discovering interesting knowledge from customer transactions. Expert Systems with Applications, 30(4), 650-657.

[16] Van den Poel, D., & Buckinx, W. (2005). Predicting online-purchasing behaviour. European Journal of Operational Research, 166(2), 557-575.

[17] Verheijden, R. (2012). Predicting purchasing behavior throughout the clickstream.