Recommendation System for E-commerce Based on Clustering and Association Rules

Qu Shixin

Economic and Management School, Tongji University, Shanghai 20092, China

Keywords: Personalized Recommendation Customers Clustering Association Rules

Abstract: **[Objective]**To generate recommendations based on customers' behavior history and to evaluate it preliminary.**[Methods]** This paper clusters customers and uses the result to generate association rules. Then it calculates customers' interests on each commodity. Finally the recommendation results are obtained. **[Results]**The performance of this method is pretty good when we use date from 'Taobao'. It indicates that this idea is correct and it is significant for us to improve it later.**[Limitations]** This paper did not make good use of all data and the result of clustering and association rules should be improved.**[Conclusions]**This method can generate desirable recommendation results and it is meaningful to improve it in later research.

1. Introduction

In recent years, Internet technology has developed rapidly. The speed of data generation has been higher and higher, which causes the problem that we call 'information overload'. It means that we are much more efficient at processing and using information than they are generated. In addition, the competition in e-commerce and online retail is being intenser. Under such circumstances, the concept of recommendation system was raised. A common definition of recommendation system is 'providing information and advice to customers, helping them decide which one to buy, simulating the process of salespeople's helping process. [1] There are four types of recommendation systems:[2] a) recommendation systems based on content, b) collaborative filtering, c) recommendation systems based on knowledge, d) the combination of all these three. All of these methods have their own advantages and disadvantages and complement each other. Related researches have been carried out in the fields of computer and information science, statistics, marketing, etc., mainly focusing on personalized recommendation algorithms and the evaluation of these algorithms[3]. Recommendation systems can provide personal recommendation services for customers and this technology has been a hot spot. It's of great significance to improve the performance of reommendation systems .

In the study of collaborative filtering, Ungar and Foster[4] apply clustering analysis. They clustered the customers based on their purchase history and generated recommendation with the result of clustering. Chen and George[5] collected readers' reading data on books, calculated customers' similarity and used Bayes formula to calculate missing values. The missing values can be viewed as customers' potential satisfaction on these books. Then they get recommendation based on the missing value. Sarwar[6] thinks a recommendation system consists of three parts: input data, neighborhood formation, and recommendation generation. He combines association rules, neighborhood formation with dimensionality reduction to generate high quality recommendation. Bresses[7] utilize Bayesian netword and decision tree, put forward an empirical algorithm for collaborative filtering recommendation system.

Different recommendation systems come from diverse thoughts. The process of implementation of them and the basic data required are also not exactly the same. In this paper, we take customers' behavior history as the basic data, mianly use the technology of clustering and association rules mining, propose a customer interest calculate model, and finally generate reliable recommendations for customers. We use the common indicator, precision rate, recall rate, and F1 to evaluate the result. The main purpose of this paper is to test whether a recommendation system combining clusering and association rules can have an acceptable recommendation result or not. If it does, we believe this method can do a good job on recommendation and it is meaningful for us to do more improvement on it.

2. Recommendation System

There is a more specific classification for recommendation system: recommendation system based on content, recommendation system based on customers' behavior(collaborative filetering), recommendation system based on knowledge, recommendation system based on customer-product network, recommendation system based on contextual awareness and the combination of them. At present, the most recognized method is collaborative filtering. The first step of this is to calculate similarity among customers based on their behavior. The similar customers are tend to have more similar interests and behavior. Based on this assumption, we can generate recommendations to a customer by observing other customers' purchasing history. Grundy's book recommendation system, Tapes-try's e-mail processing system, Ringo's music recommendation system and Phoaks's information recommendation system are all collaborative filetering examples which have a great performance.

The recommendation system based on customers' behavior can also be divided into three sorts, based on customers, based on items, and based on models[8]. Based on customers means user A and user B will have high similarity if they have a large amount of common interested items. Qualify the similarity such as using Pearson correlation coefficient, find a pair of customers who have the higest similarity, then use one of them to generate recommendation to the other. Based on customers means recommendations based on the same customer's purchasing history. It assumes that a customer's interest remain unchanged in a short period of time. Based on models means building and training models from data, using these models to recommendate such as the probability model, Bayesian netword and clustering etc.

Classified by target, recommedation system has two types, to all customers and to a certain customer. It is obvious that the latter has a wider meaning. So almost all studies are about personalized recommendation. Generally, a complete recommendation system consist of three parts, background data, input data and algorthm[9]. Using data mining technology ,we can modeling on factors which determine a customer's behavior preference. For example, we can take an analysis on a customer's behavior history and combine the similarity among products to select items which this customer is likely to buy. These products are what we should recommendate to this customer[10, 11].

Recommendation system is not only a tool to filter information. It plays an important role in e-commerce field. When a customer don't have an clear demand or excessive information makes customer feel confused, a good recommendation system will give the most suitable products. And to sellers, there's no doubt that it will bring a sales growth. So on the one hand, sellers use recommendation system to develop potential markets. Great recommendations will improve customer satisfaction and loyalty, which lead to higher sales. On the other hand, customers will have a more pleasant shopping experience. They can easily get what they exactly want to buy. Especially for those who are extremely busy, it is very helpful. And for a recommendation system, recording more customers' behaviors can enrich its database, which will be useful to next recommendation[12].

3. Recommendation System Based on Clustering and Association Rules

Association rules mining is a pretty basic technology in data mining. Association rules can reveal potential connections among products. According to these connections and the products involved in a customer's existing behavior, we can generate reliable recommendations. Generally speaking, every customer has a distinct taste. So from his or her view, products' connection are also different from others. But the number of customers is large. To build one rule base for one customer is too slow. In addition, history data about a customer perhaps don't cover all the information about a customer's interest. That's why we use clustering in this paper. We believe that after clustered, customers in the same cluster will have the same purchase behavior characteristics. It means that we can generate association rules in every cluster. And for customers in the same culster, they share the same rules. Then we calculate customers' interests on every products. In this paper, we make recommendation on product categories not specfic products.

3.1 Customer Clustering

Clustering is aim to divide the data set into k subset which called cluster. Items in the same cluster should have high similarity and items in different clusters should have low similarity. Clustering is a unsupervised learning process. The classification of items can be achieved based on the potential distribution of features among items without mining these features.

K-means is the most common method in clustering. It is put forward by Stcinhaus in 1955, Lloyd in 1957, Ball & Hall in 1965 and McQuccn in 1967 in different field[13]. It has been more than 50 years, but it is still one of the most widely used clustering algorithms[14]. The process of this alorithm is:

1) select k initial cluster centers;
2) calculate distance between every item and every cluster;
3) for a single item, find the nearest center and calssify it into this cluser;
4) calculate the new k centers;
5) redo step b) to d) use the new k centers;
6) if the centers remain unchanged, return the result.

The advantage of K-means is its efficiency. It's easy to implement and suitable for analysing and processing of large data sets. But since each item is uniquely divided into a category, it is not easy to deal with the case of spherical shell clustering. This is because if the distribution of items is spherical, the local extremum of the objective function is easy to fall into local extreme points[15]. It is found that K-means clustering has a good performance for clusters with spherical shapes which have small differences in size, but clusters with arbitrary shapes and large differences in size cannot be found, and the clustering results are easily to be affected by noise data[16]. Also, the value of k will have an impact on the result.

To overcome some deficiency mentioned above and make it more suitable for our situation, we make some improvements on this algorithm.

First is the distance measurement, according to the characteristics of the data, we use Jaccard index rather than Euclidean Distance. Every customer has a set of products which he or she has browsed, marked, added into cart or purchased. We can say that these products have postive feedback to this customer. Similarity between two customers is defined as the size of the intersection of these two customers' products set divided by the size of union. The formula is shown below.

$$\mathrm{sim}(\mathrm{N}(\mathrm{i}),\mathrm{N}(\mathrm{j})) = \frac{\mathrm{N}(\mathrm{i}) \cap \mathrm{N}(\mathrm{j})}{\mathrm{N}(\mathrm{i}) \cup \mathrm{N}(\mathrm{j})}$$

N(i) in the set of products which have postive feedback to customer i. N(j) in the set of products which have postive feedback to customer j.

Because of this method to express similarity, the way to generate new centers should be changed. In the traditional K-means algoritms, take average value as the new center. But under Jaccard index, we can calculate average value. So we define that the new center in a cluster is the customer who has the highest average similarity with others.

Second, the way to select the initial centers. In the traditional K-means algorithm, just select initial centers randomly. But this probably results in a bad clustering result. In addition, use random initial center will take more time to achieve the process of clustering. So in this paper, we make some difference. We select the first customer in database as the first initial center. Then find the customer who has the smallest similarity with the first to be the second initial center. And then to find another customer who has the smallest average similarity with the existing initial center. By doing this, we can get k initial centers and ensure they have a strong difference.

That's all about clustering in this paper and the whole process is as shown in Figure 1.

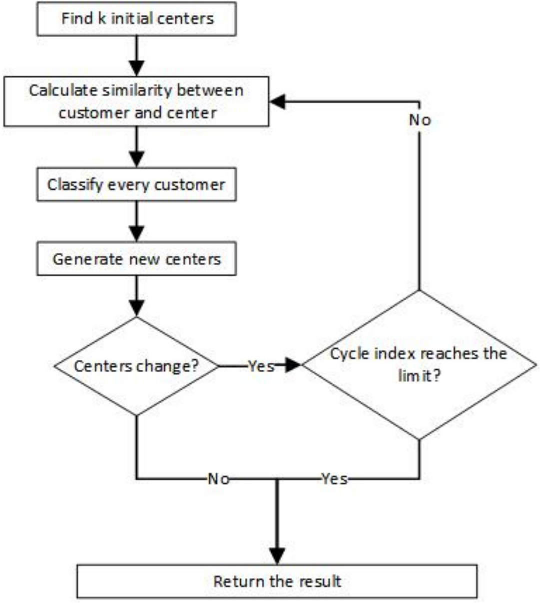

Figure 1. Clustering process.

First generate k initial centers using the mentioned method. Then calculate the Jaccard between every customer and center. For one customer, the cluster it belongs to is the one whose center is nearest to it. After every customer is allocated, recalculate centers. Then we can get k new cluster centers. Check if these k centers have any difference with previous or not. If not, then we are done. If yes, we should check if current cycle index has reached the limit. If not, repeat. Else, return even it is not an expected clustering result.

By clustering, customers are divided into groups according to their potential interest distribution. In the later, we will generate association rules respectively in every cluster.

3.2 Association Rules

Given an item set D, define a transaction ti to be a subset of items in D. In different situation, ti can have different practical meanings. Use T to represent the set of all ti. Association rules is the relation between an item set X and an item set Y. It is recorded as X⇒Y. The practical meaning is that under the restriction of specified support and confidence, if a transaction covers all items in X, it will cover the items in Y.

In a recommendation system, an item is a product and a transaction is a single piece of behavior history. Association rules can tell us the internal relation between items and this relation is manifested in the simultaneously purchasing of these items. So for a customer with association rule

X⇒Y, if he or she has purchased all items in X but not Y, we will recommend items in Y to him or her because there is great possibility for him or her to buy them.

The Apriori algorithm is a common one for association rules mining. The process is:

1) generate frequent itemset: A frequent itemset is an itemset whose support is bigger than the minsup;

2) generate association rules: A confident association rule is a rule whose confidence is bigger than minconf;

Minsup and minconf are all parameters which should be assigned manually.

The Apriori algorithm generate frequent itemset efficiently based on the downward closure property and select candidate sets of association rules. Then do pruning and merging on these candidate sets. There are two basic properties: the subset of frequent itemset is also frequent itemset, the superset of infrequent items is also an infrequent itemset. Because if a transaction covers an itemset X, it will cover all the subset of X. Its converse negative proposition is also true. The two properties can help us generate association rule more efficiently.

First, we generate rules whose size of Y is one(recorded as 1-rules),then use these rules and the generation method, get all 2-rules and so on. It is more efficiently compared to the method of exhaustion. The reason why we can do that is that if α is the Y, to the same X, all the subsets of Y can form association rules with X.

3.3 The Combination of Clustering and Association Rules

After the clustering is completed, a rule base is established in each cluster to generate association rules.

In the traditional method, the association rules is existing among products. Pay attention to this that if a customer's behavior history dosen't cover any association rule. The recommendation based on association rules fails. In this paper, we use product category rather specfic product to minimize the emergence of this problem. The category is a product's property. We can use it directly. And for a customer, we can get a transaction from all his or her behavior history. The process of association rules mining is:

1) find frequent itemsets;

2) find association rules. This step is achieved in this way. For every frequent itemset, find itemsets A and B which satisfy the follwing conditions: $A \cup B = C, A \cap B = \emptyset$ and the confidence of rule $A \Rightarrow B$ is bigger than the minconf.

When we want to generate recommedations for a customer. First we should determine which cluster the customer should be classified to based on the result of clustering, which means we will determine which association rule base we should use. Then check the existing products that have feedback to this customer, filter them with association rules. We can end with a set of products which should be coverd in this customer's feedback set according to the rules but not yet. Then based on the behavior types, calculate customer’s possible interests on every product. Then we sort the products by customer interest and get the final recommendations.

3.4 Calculate Customer Interest

Customers’ behavior has four types: browse, mark, put into cart and purchase. We don’t distinguish which type of behavior in the process of association rules mining. But we believe that different type of behavior have different importance when we calculate customer’s interest.

According the behavior type, we divide a customer’s behavior history into four subsets: product set X a customer has browsed, product set Y a customer has marked, product set Z a customer has put into cart and product set W a customer has purchased.

First we define the importance of one association rule. Assume R is the rule base, $|R|$ is the number of rules. Ri is a single association rule. The importance Imp(Ri) can be calculated as follwing:

$$\mathrm{Imp(R_i)} = |R| * (\frac{\mathrm{support}(R_i)}{\sum_{i=1}^{|R|}\mathrm{support}(R_i)} + \frac{\mathrm{confidence(R_i)}}{\sum_{i=1}^{|R|}\mathrm{confidence(R_i)}})$$

Take browsing-interest as an example, for a given pruduct z, the browsing-interest L(z,X) is defined as:

$$\mathrm{L(z,X)} = \sum_{i=1}^{|R|} y_i * \mathrm{Imp(R_i)}$$

yi indicates whether we can get z by going through the association rules with X. If we can, yi will be 1. or else, set yi to be 0.

Finally, we set weight w to every type of behavior manually. And the comprehensive interest is calculated as:

$$\mathrm{L(z)} = \sum_{type} L(z, type) * w$$

Then we sort all the products by L(z), find the top N products to recommend to this customer.

4. Experiment

We use the data from ‘https://tianchi.aliyun.com/getStart/introduction.htm?spm=5176.100066.0.0.7cabd780CcLV6e&raceId=231522’ to complete the recommendation and test its per-formance.

4.1 Data Description and Data Preprocessing

The whole data set contains 20,000 customers behavior history. Each piece of history has these information: customer id, product id, behavior type, time and product category.

Considering the cold start problem of this approach, we delete some customers’ data from the database. These customers only have a small number of behaviors. Finally we just use 5,000 customers’ data to training the model and 1,427 customers’ data to test the performance.

4.2 Customer Clustering

Set the number of initial center to 10. In order to decrease the influence of hyperparameters. We generate a loop to choose the optimal k. Every time increase k's value by 5 and use F1 to evaluate this k. After a whole loop, we find that when we set k to 80, we get a best recommendation result. The different performance of k is partly shown in Figure 2.

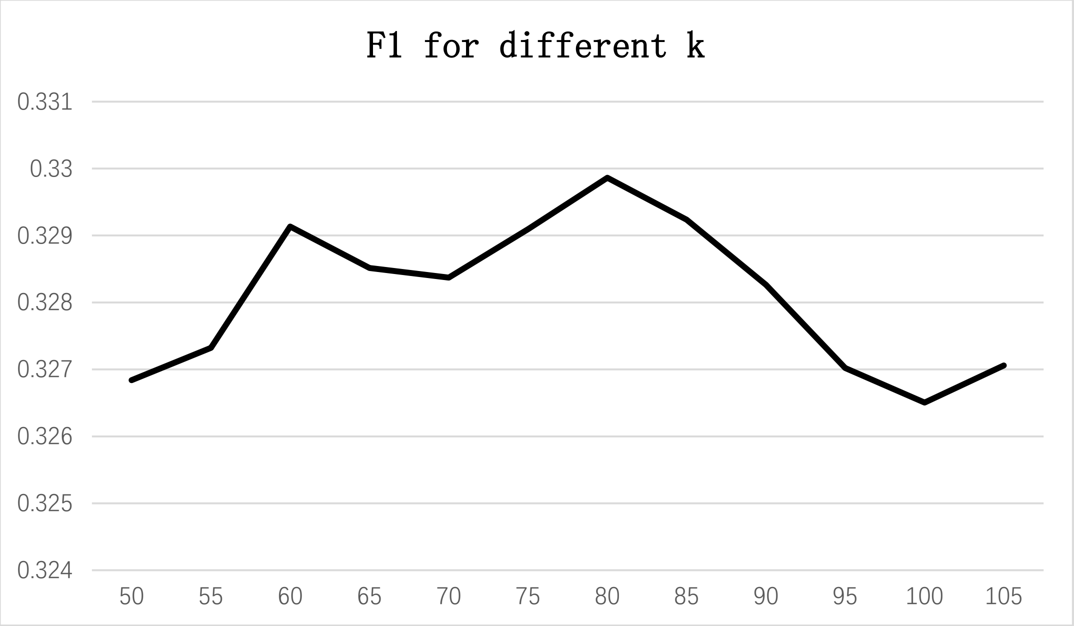

Figure 2. F1 for different k

So in the process of clustering, we set k to 80, evaluate customers' similarity by Jaccard index and after 45 times iterations, the cluster center remain unchanged. We believe that we have got a stable result. Customers in the same cluster will have interest distribution.

4.3 Association Rules Generating

Mining association rules in every cluster respectively. In our experiment, we set minsup to 0.3 and minconf to 0.45. We have not made some improvements to these two parameters. But it's obvious that a suitable choice on these two parameters will improve the preformance and we will do this job in later studies.

4.4 Result

4.4.1 Index

We use the most common indexes: precision, recall and F1 to evaluate our result. We haven't chosen some more complicated indexed because we just want to test this approach preliminarily. If it gives a pretty great result, we will study more on this method. And it fails at the beginning, we will drop this method.

The formula to calculate these three indexes are:

$$\text{Precision} = \frac{Precisionset \cap Referenceset}{|Precisionset|}$$

$$\text{Recall} = \frac{Precisionset \cap Referenceset}{|Referenceset|}$$

$$\text{F1} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Both precision and recall are between 0 and 1. Bigger values indicate a more recommendation result.

4.4.2 Customer Interest Calculation

We use customers' behavior history from November 18th to December 18th to train the model and data from December 9th to 18th to evaluate the preformance. The indexes are as follows:

$$\text{Precision} = 0.34$$

$$\text{Recall} = 0.32$$

$$\text{F1} = 0.33$$

It indicates that the commendation method in this paper has a pretty good performance. Its precision and recall can both reach the general level. So using this method to generate recommendation is completely feasible. It is meaningful for us to improve this method in the later studies.

5. Conclusion

The recommendation system is a hot topic in the current data mining field. How to accurately provide personalized recommendation services for customers has not yet been a very mature technology. Starting from this problem, this paper designs a recommendation system that combines clustering and association rules mining. First clustering these customers, treat customers in the same cluster as a whole. Then generate association rules based on their behavior history. For a customer, when we want to recommend products to him or her. We first determine which cluster he or she belongs to. Then we can know which association rules we can apply. With these rules and the behavior history, we can select products a customer will be interested in. To do a qualitative analysis, we calculate the customer's interest. The customer interest is more comparable and then we can generate recommendation.

This method takes a balance between the performance and the workload. And the participation of behavior type are also beneficial to generating great recommendation. This paper makes a detailed introduction to the proposed method and makes a preliminary evaluation on its performance. The result indicates that this method can generate pretty great recommendations, which means it indeed has practical meanings to improve it in our later studies.

6. Shortcomings

As mentioned above, this is just the start of this method. There are still many things we can do the improve the recommendation performance.

1) We just recommend products from the view of a category not a specific product. In order to make the recommendation process more efficient, we use the product category to distinct from every product. But recommending a specific product has a wider application. To solve this problem, in the later studies, we will mine more information from a customer's behavior history such as the

price level he or she can accept, the brand he or she is likely to pick to achieve a more specific recommendation.

2) We just assume a customer's interest remain unchanged over time. It has no problem when we process data in a short period time. But it is obvious that when we expand to a large time range, the result will not always be great. It is also an issue that we will focus on in the later studies.

3) The cold start problem remain unsolved. It is a extreme hard problem exists in most recommendation systems, which we will try our best to overcome.

This paper is just a preliminary evaluation on this method. We haven't committed to the details of this method. The test result indicates that this method is practical feasible and can generate good recommendations. It's true that there are lots of problems but it is also true that it is promising and meaningful for us to improve this method in later studies.

References

[1] RESNICK P, VARIAN H R. Recommender systems [M]. ACM, 1997.

[2] Xu Hailin, Wu Xiao, Li Xiaodong, et al. Comparison Study of Internet Recommendation System [J]. Journal of Software, 2009, 20(02): 350-62

[3] Sun Luping, Zhang Linjun, Wang Ping. Review and Prospect of personalized recommendation research on the Internet [J]. Foreign Economics & Management, 2016, 38(06):82-99

[4] UNGAR L H, FOSTER D P. Clustering Methods for Collaborative Filtering [J]. Aaai Workshop on Recommendation Systems, 1998,

[5] CHIEN Y H. A Bayesian model for collaborative filtering; proceedings of the Proc of the Int Workshop on Artificial Intelligence and Situations, F, 2002 [C].

[6] SARWAR B, KARYPIS G, KONSTAN J, et al. Analysis of recommendation algorithms for e-commerce; proceedings of the ACM Conference on Electronic Commerce, F, 2000 [C].

[7] BREESE J S, HECKERMAN D, KADIE C. Empirical analysis of predictive algorithms for collaborative filtering; proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence, F, 1998 [C].

[8] Zhu Yangyong, Sun Jing. Recommender System:Up to Now [J]. Journal of Frontiers of Computer Science and Technology, 2015, 9(5): 513-25.

[9] Tian Ye, Zhu Zhongming, Liu Shudong. Review of Recommendation System Based on Linked Data [J]. New Technology of Library and Information Service, 2013, 29(10): 1-7.

[10] Li Wenhai, Xu Shuren. Design and Implementation of recommendation system for E-commerce on Hadoop [J]. Computer Engineering and Design, 2014, 35(1): 130-6.

[11] Liu Jianguo, Zhou Tao, Guo Qiang, et al. Overview of the Evaluated Algorithms for the Personal Recommendation System [J]. Complex System and Complexity Science, 2009, 6(3): 1-10

[12] Liu Qingwen. Reaserch on Recommendation Algorithms based on Collaborative Filtering [D]; University of Science and Technology of China, 2013.

[13] Wang Qian, Wang Cheng, Feng Zhenyuan, et al. Review on K-means clustering algorithm [J]. Electronic Design Engineering, 2012, 20(7): 21-4.

[14] JAIN A K. Data Clustering: 50 Years Beyond K-means [J]. Pattern Recognition Letters, 2008, 5211(8): 3-4.

[15] ZHOU Tao, LU Huiling. Clustering algorithm research advances on data mining [J]. Computer Engineering and Applications, 2012, 48 （12）：100-111.

[16] Tang Dongming. Research on Clustering Analysis and its Applications [D]; University of Electronic Science and Technology of China, 2010.