Apriori Improved Algorithm and its Application in Tmall

Jianhua Xiao[1,a], Shaoyu Luo[1,b]

[1]School of Economics and Management, Wuyi University, Jiangmen, 520900, China

[a]email: 1325702977@qq.com, [b]email:352071385@qq.com

Keywords: Data mining; Apriori Algorithm; Marketing Strategy; Tmall

Abstract. In view of the large data volume and fast update feature of Tmall transaction database, the original Apriori algorithm is not suitable for Tmall transaction data mining due to its frequent scan of the database, so improved the search efficiency of original Apriori algorithm combining Apriori nature, and applied it to the association rules extraction of Tmall alcohol commodities trading data. Finally obtain association rules based on the algorithm and put forward targeted marketing strategy.

Introduction

In recent years, with the increasing popularization of online shopping, such as the huge daily trading volume of Tmall, Jingdong, Suning, Amazon and other big e-business platforms, a large number of transaction data has been produced, which has concealed important information such as consumer shopping preferences. Anyone who masters these information can grab chance in the increasingly fierce e-business war, and develop targeted marketing strategy, win clients, and achieve development. Therefore, it seems to be critical how to develop the useful information from the data.

Apriori algorithm, classic association rules mining, is an effective algorithm for analysis of transaction data, it can search frequent patterns, relevance, correlation or casual relationship among item sets in large and complex transaction data, such as commodity transaction amount, transaction type, transaction time, etc., and make the complex, disorderly trading data as effective basis of market analysis, business strategy, customer relationship management, so as to realize the true value of e-business activities. But the fatal weakness of Apriori algorithm is frequent scan of transaction database, which is unpractical for Tmall database with large base and fast update. Therefore, this article will improve the search efficiency of the Apriori algorithm with Apriori nature, and apply the improved Apriori algorithm for Tmall alcohol commodities data mining.

Deficiency of Apriori algorithm and improvement

Association rules are to describe the potential relationship between data items in the database, but association rules mining is to find item set with minimum support in the transaction database, namely the largest item set, and then generate association rules required with largest item set. It is obvious that to find the largest item set in the whole execution process of association rules mining is the core problem to extract association rules between things.

Basic concept. The Set $I=\{i_1,i_2,\cdots,i_n\}$ as collection of item. Set D related to task as collection of database transaction, in which each transaction T is a collection of item, making $T\subseteq I$. Each transaction has an identifier called TID. Set A as an item set, and transaction T includes A when and only when $A\subseteq T$.

Definition 1. Association rule is the expression as $A\Rightarrow B$, in which $A\subset T$, $B\subset T$, and $A\cap B=\varnothing$.

Definition 2. Support $(A\cup B)=P(A\cup B)$, namely the rules $A\Rightarrow B$ is workable in the transaction set D with support s, s is the percentage of containing $A\cup B$ in transaction D, written as $P(A\cup B)$.

Definition 3. Confidence $(A\Rightarrow B)=P(A\mid B)$, namely the rules $A\Rightarrow B$ has confidence c in the

transaction set D, and c is transaction of containing A in D, and at the same time also the percentage of containing B in D, written as $P(A|B)$.

Definition 4. The collection of item is called itemset, in which itemset containing k items is called k-itemset. Itemset frequency is the number of transactions containing itemset, referred to as the frequency of itemset. If the emergency of itemset is not less than minimum support, this itemset is called frequent itemset, frequent k-itemset is written as L_k. If the frequent itemset is not the subset of any other frequent itemset, the frequent itemset is the maximum frequent itemset.

Definition 5. Association rule in which both support and confidence are greater than the given value of the user is called strong association rule.

Aprior algorithm. In 1993, Rakesh Agrawal Rama and Krishman Skrikant put forward a classical frequent item set -- Apriori algorithm to mine boole association rules. Its basic idea is to find out which items in the database have the highest frequency of occurrence at the same time, and then look for association rules according to these items. Apriori algorithm is widely applied in the e-commerce filed because it's able to effectively mine related relations hidden in transaction data, and find the knowledge such as some customer buys both product A and product B in the single purchase from the transaction information database.

The basic principle for Apriori algorithm to look for the largest item set is to use the priori knowledge of frequent item set property, and then proceed multistep processing by means of one-by-one search iterative method. First of all, find the set of frequent 1-item sets, namely simply calculate the item sets where the frequency of occurrence of itemsets containing one element is no less than the minimum support, marking it L_1; and use L_1 to look for the set of frequent 2-item sets L_2, and then use L_2 to look for L_3, and so on. This process will not stop until the step k processing cannot generate the set of the frequent $k+1$ item set.

It can be seen from the basic principle of Apriori algorithm that it mainly has three defects: the frequency of scanning database is too high, and it's not applicable for association rules mining of dense sets, as well as it may generate much too many association rules. Of these three defects, the first one is the greatest disadvantage.

Improvement of Aprior algorithm. Aiming at the defect of frequent scanning on database of classical Aprior algorithm, we have some optimizing strategies for Aprior algorithm, such as reverse operation hash tables, transaction reduction, reconfiguration of database, significance of item sets and "separation-integration". However, how to improve the search efficiency of Aprior algorithm in the transaction database that is provided with huge data size and in constant update, still remains the problem needing to be solved in the data mining of Tmall. Accordingly, through the literature and empirical evidence, an important property called Apriori property is utilized to improve the efficiency of search frequency mode of Aprior algorithm. This property defines that if one item set is a frequent item set, then all the nonvoid subsets of this item set are frequent. Based on this definition, if item set I does not meet the threshold value of minimum support, namely $P(I) < \min\_\sup$, then I is not a frequent item set. If add the item M into the item set I, then the new item set $(I \cup A)$ is impossible to have a higher frequency of occurrence than I does. Consequently, $(I \cup A)$ is also not frequent, namely $P(I \cup A) < \min\_\sup$.

Application of improved Aprior algorithm in Tmall

Tmall is the comprehensive B2C shopping platform separated from Taobao. With its advantages of integration of thousands of brand flagship stores, 100% authentic product guarantee and changing or returning in 7 days without any reason, it develops rapidly into the largest e-commerce shopping platform in China and even in Asia within a few years, so that it creates more than 90% volume of business of online shopping in China, turning out the miracle of the development of internet enterprises. Therefore, it is considerable for sellers to find helpful marketing points or promotion means among a huge amount of transaction data of Tmall, in order to make themselves stand out and expand potential client bases.

Synthesizing the characteristics of transaction data in Tmall, this paper selects the transaction data of liquor commodities from January, 2014 to December, 2014 in Tmall as the object of study.

Data processing.

Selection of data variables. Through the analysis of key features of liquor commodities, choose the year, odor type, degree of alcohol, producing area, price, transaction time and level of buyers' transaction evaluation as the data variables of liquor commodities, in order to establish the transaction database of liquor commodities from January, 2014 to December, 2014 as shown in Table 1.

Table 1 Original record table of liquor commodity transaction in Tmall (omitted)

Variable / Record	Year		Type		Degree		Place		Price		Time		Level	
Record1	5	A1	Maotai-flavor	B1	39	C2	Sichuan	D1	128	E2	May	F2	Five star	G3
Record 2	15	A2	Fen-flavor	B2	45	C3	Shanxi	D2	158	E2	September	F2	Three star	G1
Record 3	10	A1	Maotai-flavor	B1	53	C4	Guizhou	D3	59	E1	September	F1	Four star	G2
Record 4	30	A3	Chicken-flavor	B3	52	C4	Shanxi	D4	388	E4	March	F4	Three star	G1
Record 5	5	A1	Special flavor	B4	53	C4	Hubei	D5	168	E2	February	F2	Five star	G3
Record 6	10	A1	Mixed- flavour	B5	45	C3	Jiangxi	D6	168	E2	March	F2	Four star	G2
Record 7	8	A1	Strong-flavor	B6	52	C4	Sichuan	D1	378	E4	December	F4	Four star	G2
Record 8	30	A3	Dry white	B7	46	C3	Beijing	D7	199	E2	November	F2	Three star	G1
……	……													

Data pre-processing. It can be seen from table 1 that the collected transaction records of liquor commodities are not suitable for data mining, thus the normalized method of narrowing down the attribute data into specific sections in proportion is used, in order to normalize and gather the original data. Specific operations are as following:

The year when liquor is made has great difference. If a decade is one section, then liquor during 0 to 10 years is A1; liquor during 11 to 20 years is A2; liquor during 21 to 30 years is A3; liquor during 30 to 40 years is A4, the rest is A5. If the liquors are classified by odor type, then the maotai-flavor is B1; the fen-flavor is B2; the feng-flavor is B3; the special flavor is B4; the mixed-flavor is B5; the strong-flavor is B6; the dry white is B7, and the rest are B8. If every 10 degree of alcohol is one section, then the liquors lower than 30 degree is C1; liquors between 30 to 40 degree is C2; liquors between 40 to 50 degree is C2; liquors between 40 to 50 degree is C3; liquors above 50 degree is C4. If the liquors are classified by production area, then Sichuan is D1; Shanxi is D2; Jiangxi is D3; Guizhou is D4; Shanxi is D5; Hubei is D6, and the other places are D7. If the liquors are classified by prices, then 0 to 100 yuan is E1; 100 to 200 yuan is E2; 200 to 300 yuan is E3; 300 to 400 yuan is E4, more than 400 yuan is E5. If the liquors are classified by transaction time, then firstly, proceed generalized processing. Generalize the primary concept "month" to high-level concept "season", so that the data is simple: January to March is spring (F1); April to June is summer (F2); July to September is autumn (F3) and the October to December is winter (F4). If the liquors are classified by levels of transaction evaluation, then those with less than three stars are G1; those with four stars are G2; those with five stars are G3. Finally, the transformation of original data is as shown in Table 1.

Apriori algorithm data mining. Use Java to run Apriori algorithm of relate rule mining, of which *minsup* of Double type is used for receiving users' defined minimum support; objective of ConnectDB is objectivces of datebase operation; TreeSet's objective *freq_set* is used for store the frequent item sets, *max-freq* is used for maximum frequent itemsets and *max_Association* is used for store all the association rules. Run Apriori algorithm written by Java, select wine product database after transformed, set minimum support of association rules as 20%, minimum confidence as 60%, we get the liquor products association rules of Tmall shown in Table 2.

Table 2 Apriori algorithm mining results

Serial number	Association rules			Explanation
（1）	Al B3 C3→	D5 G2	Confidence 65.4%	10 year, feng-flavour, 40~50 degree→Shanxi, five star
（2）	A2 B2→	D3 G2	confidence 71.5%	20 year, fen-flavor→Jiangxi, four star
（3）	A2 C3 →	D2 F3	confidence 62.3%	20 year, 40~50 degree→Jiangxi, autumn

From the Table 2, we can see that confidence coefficient of three association rules are greater than minimum confidence level of 60%, so that they are strong rules with reference value. From the association rules 1, generally when customers choose 0~10 years, feng-flavour 40~50 degrees wine, they often choose wine from Shanxi and give high consumption evaluation; from association rules 2, when customers choose 20 years, fen-flavor wine, they often choose wine from Jiangxi also give good evaluation; from association rules 3, sellers will choose wine produced by Shanxi when selling 20 years, 40~50 degrees wine and often purchase in autumn.

Analysis on Apriori algorithm mining results

Combine Apriori algorithm data mining results with actual situation, we can get following conclusions:

Bundle purchase. According to the results of data mining, during the purchase of Tmall wine merchants, if they purchase 40~50 degrees wine, they can purchase 10 years feng-flaovr from shanxi or 20 years all flavor types from Shanxi, in autumn, it should increase purchase quantity of 20 years wine with all flavor types and different prices; if purchase ten years feng-flaovr 40~50 degrees wine, they should purchase all with different prices; if purchase 20 years fen-flavor wine from Jiangxi, they should purchase all with different degrees and prices.

Chain sales. According to the results of data mining, when Tmall wine dealers sell products, they can put 10 years feng-flavor 40~50 degrees wine from shanxi together with 20 years feng-flavor 40~50 degrees wine from Shanxi (same flavor and degree), or put 20 years fen-flavor 40~50 degrees wine from Jiangxi with 20 years fen-flavor 40~50 degrees wine from Shanxi (same year, falvor and degree), or take 10 years feng-flavor 40~50 degrees wine and 20 years fen-flavor 40~50 degrees wine from Jiangxi as a package, they can sell through appropriate price cut.

Customized. First, determine characteristics of target customers, take association rules (20 years, fen-flavor→Jiangxi, four star) as example, properties of target customers include 20 years and fen-flavor, we may find customers who satisfied with these two properties in transaction record and locate them as potential customers. Second, determine the goods potential customers may interested in. Find the potential goods according to association rules, we still take association rules (20 years, fen-flavor→Jiangxi, four star) as example, if the customer's characteristic satisfy the standard (20 years, fen-flavor), then they may be interested in wine from Jiangxi, and the deal may get high evaluation at the end.

Conclusions

It may be helpful for merchants to get marketing strategy information by using Aprior to improve the efficiency of digging Tmall transaction database with Aprior algorithm. But Tmall transaction database is too large and updating quickly, the future research direction may lies in improving mining technology to better adapt information mining in transaction database.

References

[1] Jun Zhang: submitted to Journal of Software(2009),p.87-88
[2] Zhangyan Xu, Meiling Liu, Zhang Shichao, Lu Jingli and Ou Yuming: submitted to Journal of Computer Engineering and Applications(2004),p.190-192+202
[3] Hengjie Wang: Research on Application of association rules mining in the Tmall(2013)
[4] Song Zhao: Improvement and Application of Apriori Algorithm(2006)
[5] Xue-fei Xu, Peng-tao Qing: submitted to Journal of Modern Computer(2008),p.80-82+102