Apriori algorithm for economic data mining in sports industry

Yaguang Xiang*

Sports Institute, West Anhui University, Liu'an, 237012,Anhui, China

Received 1 June 2014, www.cmnt.lv

Abstract

Current data mining model cannot meet the increasing requirement of economic data mining in sports industry. In view of it, this paper put forward a data mining model based on improved Apriori algorithm, firstly took Hash technology to remove unnecessary candidate item sets to improve the algorithm efficiency, and then translated the transactional databases to the form of matrix to decrease the memory consumption. The simulation experiments show that compared with standard Apriori algorithms, our proposed data mining model based on improved Apriori algorithm greatly shortens the running time, decreases the space consumption, and can be completely applied into the sports industry economic data mining.

Keywords: Sports industry economy, data mining, improved Apriori algorithm, Hash technology, memory consumption optimization

1 Introduction

Since the reform and opening up, sports industry has achieved rapid improvement in our country. The success of Beijing Olympic Games promotes it to develop constantly with significant economic benefits [1]. With the increase of household consumption rate and people's requirement of high quality life, sports industry will be a powerful force in Chinese economic development in the future [2]. Due to this background, more and more researchers pay attention to the data mining of sports industry economy [3].

The toughest task of the data mining is the association rules mining that can find the relationship rules among items or properties which cannot be searched with traditional artificial intelligence and statistical methods, satisfying the requirement of knowledge acquisition from large-scale data storage [4]. Some well-known research institutes in university and researches department in companies have paid much attention to it and made many achievements [5]. The Data Base System LAB in Stanford developed large number of commercial data mining system, especially DBMiner system. It involves many advanced mining algorithms, and searches out many types of knowledge including association rule, sequence model and classification. It also can be operated on variety of platforms and combined with some mainstream data base management system like SQLSever and Oracle, and at the same time, introduces the online analysis and mining technology so as to make use of the analysis advantage of data warehouse[6]. The QUEST project of IBM Almaden laboratory is also in the top class of data mining, involving the study of association rule, sequence pattern, classification and time sequence clustering, with the representative product, DB2 Intelligent Miner for Data in IBM DB2 platform [7]. There are a lot of researchers that foundationally contributes to the development in this area, such as the Jiawei Han in SimonFraser University, Mannila and Toivonnen in University of Helsinki [8]. Nowadays, the study on association rule mining becomes a hot topic in China, and we have made breakthrough in related algorithms and applications [9].

In view of immaturity of current economic data mining model in sports industry, this paper put forward the data mining model based on improved Apriori algorithm which optimized the traditional Apriori algorithm.

2 Apriori algorithm

2.1 THE BASIC IDEA OF APRIORI ALGORITHM

Apriori algorithm uses repeated iteration starting from $1-itemset$ and pruning it frequently according to the given support threshold value $\min sup$ until iterating to L_1. According to Apriori principle, if some item-set is frequent, then all of its subsets are frequent.

Therefore, the candidate $2-item$, named as C_2, can be generated with the frequent $1-itemset$ L_1. After the generation of $2-item$, frequent $2-item$ L_2 is obtained by pruning it again according to $\min sup$,and so on, until frequent item-set L_k with maximum items is generated. As previously mentioned, the realization of Apriori algorithm mining rule has two steps:

1) To find out all the frequent item sets L in the data set;
2) To extract strong rules from L.

Here, step1 is the key point of the Apriori algorithm, the key measure to evaluate the algorithm performance, while step 2 is relatively simple. Presently, the improvement method of Apriori algorithm focuses on step 1[10]. Step 1 can be subdivide into two operations: the first is to generate candidate item set C; the second step is to prune

* ***Corresponding author's*** e-mail: xygang1977@163.com

the candidate item set until the frequent item set L is found according to $\min sup$.

The candidate item sets can be generated in many ways, usually including brute force, $F_{k-1} \times F_1$ method and $F_{k-1} \times F_{k-1}$ method.

2.1.1 Brute force method

The brute force method is to arrange all the $1-$ item sets, and list all the possible candidate item sets. If there are n $1-$ item sets, then C_n^k candidate item sets are generated, and some unnecessary candidates are cut off. This method is simple in candidate generation, but complicated in pruning these large number of candidates.

2.1.2 $F_{k-1} \times F_1$ method

This method adopts the combination of L_{k-1} and L_1 to generate the $K-$ items set C_{k-1}. Figure 1 shows the process of this method to generate the $3-$ items set.

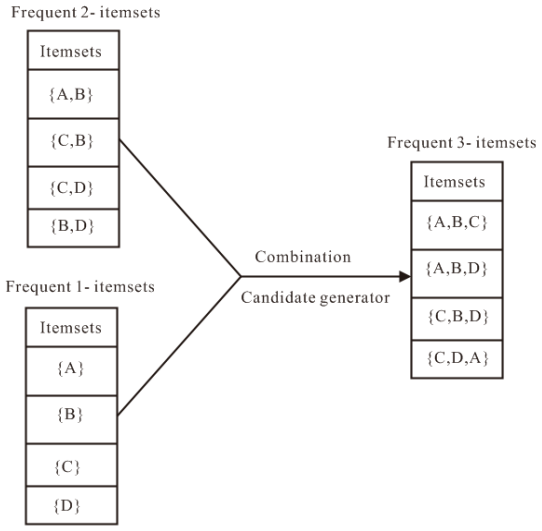

FIGURE 1 The process of candidate 3-itemsets generation

But this method will bring repeated candidate item sets due to this combination way.

2.1.3 $F_{k-1} \times F_{k-1}$ method

In this method, candidate $k-$ item set is obtained by combination of a couple of frequent $(k-1)-$ item sets meeting the requirement of same $k-2$ items. That is to say,

$$A = \{a_1, a_2, ..., a_{k-1}\} , \quad (1)$$

$$B = \{b_1, b_2, ..., b_{k-1}\} . \quad (2)$$

When they meet the following conditions, A and B are combined.

$$a_i = b_i (i = 1, 2, 3, ..., k-2), a_{k-1} \neq b_{k-1} , \quad (3)$$

Figure 2 shows the process to generate the candidate $3-$ itemset with this method.

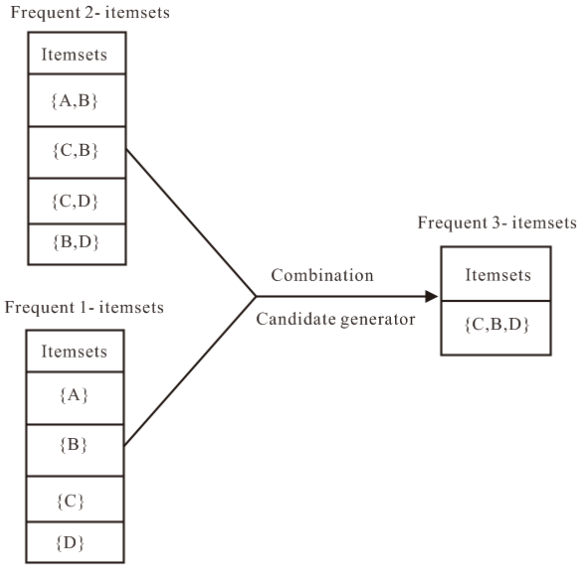

FIGURE 2 The process of generating the candidate set

Because this method merges a couple of frequent $(k-1)-$ item set to get the candidate $k-$ item set, it requires to add a step to ensure the $(k-2)$ items same before frequent $(k-1)-$ item set.

The frequent item set generation of Apriori algorithm has two important characteristics:

Firstly, it is a *level – wise* process, namely going through the transaction set each time from frequent $1-$ itemset to the frequent item sets with the maximum items.

Secondly, it uses generation-pruning rule to generate frequent item sets. The frequent item sets discovered last time is used every time when new candidate frequent items are generated, then the support count is calculated and compared with the given one so as to delete the candidate sets with the support value less than threshold value.

2.2 PERFORMANCE ANALYSIS

Apriori algorithm doesn't have heavy and complicated formula derivation process. It is easily realized. At the same time, the generation process of candidate item sets has included a part of pruning which decreases some unnecessary candidate item sets and the pruning work later. However, this algorithm shows a little powerless when dealing with large number of item sets with longer length and small support threshold value.

1) In the rule generation process, algorithm must scan transaction base repeatedly, especially when the item sets are too long. Algorithm must scan the subsets of the $k-$ candidate item set C_k one by one, and check whether the subset belongs to L_{k-1}. If one subset doesn't belong to L_{k-1}, then it must be pruned; otherwise, it will cause huge amount of calculation, give high press on the I/O load and increase the running time.

2) Many candidate item sets are generated in the combination of L_{k-1} to C_k, increasing fast with initial $1-$ itemset. For example, when there are five frequent $1-$ itemsets, the combined C_2 will include 10 items; when the number of frequent $1-$ itemsets increases to 1000, then the number of C_2 will reach to C_{1000}^2. It demonstrates that the computational amount is huge when the number of item sets of the candidate item set is too much.
3) Apriori algorithm uses Apriori-gen function, namely a couple of L_{k-1} is combined to generate C_k, meeting the requirement that $k-2$ items before L_{k-1} must be same and the last item is different. Therefore, each item before L_{k-1} should be compared to ensure their difference, which increases the time consumption and decreases algorithm efficiency. This is a bottleneck problem requiring to be solved.

3 Improvement of apriori algorithm

3.1 OPERATING EFFICIENCY OPTIMIZATION BASED ON HASH

According to Apriori algorithm, before generating k item set in each step, the candidate set C_k of frequent $k-1$ item sets is firstly required, then the support of each candidate item set is calculated from the data base, which costs much in time and space. A small candidate item set plays a vital role in improving the efficiency of finding frequent item set. However, in Apriori algorithm, C_k is generated from L_{k-1} with high potential. Through Hass technology, unnecessary candidates are removed to decrease the potential of C_k, so as to decrease the cost of time and space and improve the algorithm efficiency.

When the transaction in D and item number of I is too much, the item number m of L_1 will be large, then the number of C_2 is $\frac{m(m-1)}{2}$. At this time, the support calculation of the elements in C_2 is a huge number.

Hash technology is mainly used to solve the conflict problem. This paper gives a two dimensional hash function to avoid this conflict.

Suppose the item set $I=\{I_1, I_2, ..., I_k, ..., I_m\}$, where $I_k(k=1,2,...,m)$ is the item valued by the sequential value $1,2,...,k,...,m$.

$order(x)$ and $order(y)$ represents the sequential value of item x and y. The new Hash function is derived,

$$h_1(x,y)=(|L|\times order(x)+order(y)-\\ -\frac{x(x-1)}{2})\bmod p_1 \quad , \tag{4}$$

$$h_2(x,y)=(|L|\times order(x)+order(y)-\\ -\frac{x(x-1)}{2})\bmod p_2 \quad . \tag{5}$$

Here,

$$p_1\times p_2\geq(1-\min\sup)C_{|L|}^2 . \tag{6}$$

And $p_1\neq p_2$, and they are relatively primes. If they are valued by a large number, then Hash table will occupy much space, while if it is small number, then it is easy to have a conflict. The value of p_1 and p_2 is adjusted with the item number and minimum support set by users.

$$H(x,y)=H(h_1(x,y),h_2(x,y)) . \tag{7}$$

Where, $h_1(x,y)$ and $h_2(x,y)$ represent the subscripts of $H(x,y)$. When the Hash value is projected to a unit, then the count value is plus one. $|L|$ is the number of items, with the single function:

$|L|\times order(x)+order(y)-\frac{x(x-1)}{2}$ and

two dimensional Hash table, which greatly decreases the Hash conflicts.

Meanwhile, all the $2-$ items of each transaction is counted with two dimensional Hash. We can not only get L_1 but also a two dimensional Hash table where each unit value is a sum of the counts. Comparing the count value with $\min\sup$, if the count value is not less than $\min\sup$, then this $2-$ item group belongs to L_2; otherwise, it is not the frequent $2-$ item.

3.2 MEMORY CONSUMPTION OPTIMIZATION

If we change the database of the transactions into the form of matrix, then all the transactions are switched to two possibilities of 0 and 1, and only one time scanning can meet the requirement to greatly decrease the memory consumption. The connection step and pruning step is translated to the deletion of Boolean matrix, which effectively improves the algorithm efficiency.

1) To translate the database D into the form of Boolean matrix, with items $I=\{I_1, I_2, ..., I_m\}$ and transactions $T=\{T_1, T_2, ..., T_n\}$. If R is defined as the binary relation from I to T, denoted as $r_{ij}=R(I_i,T_j)$, $R=(r_{ij})_{m\times n}$ /

$$r_{ij}=\begin{cases}1, I_i\in T_j\\ 0, I_i\notin T_j\end{cases}, i=1,2,...,m;\ j=1,2,...,n . \tag{8}$$

2) To define the row vector of each item as

$I_i=\{r_{i1}, r_{i2}, ..., r_{in}\}$, where $r_{ij}=\begin{cases}1, I_i\in T_j\\ 0, I_i\notin T_j\end{cases}, i=1,2,...,m;$

$j=1,2,...,n$, with support count I_i,

$$support\_count\{I_i\}=\sum_{j=1}^{n}(r_{ij}) . \tag{9}$$

3) The support count of k item sets $\{I_1, I_2, ..., I_k\}$ is written as,

$$support\_count\{I_1, I_2, ..., I_k\} = \sum_{j=1}^{n}(r_{1j} \wedge r_{2j} \wedge ... \wedge r_{kj}) \quad , \tag{10}$$

$\wedge$ represents the ' and' operation in vector, meaning the value is 1 only when all the items equal to 1, otherwise equals to zero.

4) In the Boolean matrix R, if there is a row of item count less than k, then this row is deleted when calculating the k dimensional support.

This property indicates that when calculating L_k, all the data less than k dimensional is useless. In order to speed calculation, this part of data should be ignored or deleted.

5) If it exists an item $I_m \in X$ in k item set X, and the number of I_m in L_k is less than k, then the row of the I_m is deleted when frequent k items set generates the frequent $k+1$ items set.

If X is the frequent $k+1$ items set, then its $k+1$ $k-$ subsets are frequent k items set, which means each item appears k times in the $k-$ subsets. $\forall I_m \in X$, the number of I_m is larger than k. Therefore, if the time of I_m is less than k, then the $k+1$ items set is not generated any more.

4 Simulation research

To test the validity of the improved algorithm, we conducted the simulation experiments. The item number in the economic data set in sports industry is 1598293 with confidence coefficient 0.2, and changeable minimum support threshold value (0.05, 0.1, 0.15, 0.2). The operation condition of the proposed algorithm is shown in following figure.

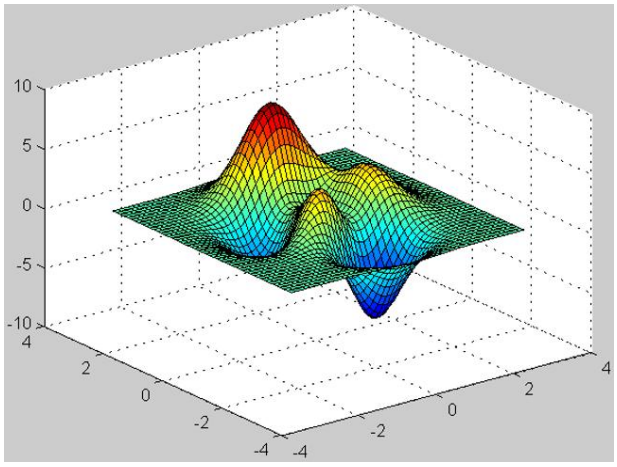

FIGURE 3 The Curve of Running Time with the Support

From the picture, it is seen that with the increasing minimum support threshold value, the running time decreases correspondingly, which indicates the improved Apriori algorithm is stable and convergent. When the minimum support threshold value is very small, the running time becomes long, which is similar to traditional Apriori algorithm.

Then, the support is set as 0.15 and confidence coefficient 0.2, and the transaction number in the economic data set is changed. This paper takes a way of superimposing data in original data set repeatedly in order to generate high-volume new data set, taking the transaction number from 1 to 8 million, respectively, as shown in Figure 4.

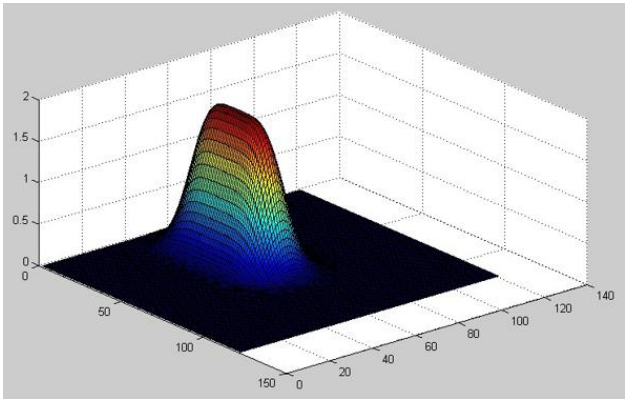

FIGURE 4 The Curve of Running Time with the Number of Transactions

From the figure, the running time curve is seem to linearly increase with the number of transactions, which indicates Apriori algorithm has better scalability and would not increase exponent-explosively. It proved the correctness of our algorithm.

5 Conclusions

Sports industry has a bright prospect in our country. Whether it can develop healthily depends on the whether we can dig out the relations and laws among the sports industry economy. This paper puts forward a data mining model of sports industry economy based on improved Apriori algorithm. The simulation experiments show that proposed data mining model appears good performance.

Acknowledgments

This work was supported by:
1. College foundation of excellent intellectual in Anhui province (NO. 2011SQRW136);
2. National foundation of philosophical and social sciences (10CTY014)

References

[1] Xiao Jianbo. (2009) Study over problems and countermeasures of current sports industry economy. *Henan Social Sciences*, 17, 15-16.
[2] Li Xiaohui. (2009) Analysis of the influence of sports industry on our modern economic development. *Enterprise Economy*, 6, 150-152.
[3] Zhang Hongsheng. (2013) Theoretical system architecture study of sports industrial economics. *The Journal of Xi'an Physical Education University*, 20, 17-19.
[4] Huang Zaixing. (2014) Correlated Rules Based Associative Classification for Imbalanced Datasets. *Computer Science*, 41, 111-113.
[5] Zeng Ziwei. (2014) Spatial Data Mining Method Based on Compact Dependences in Concept Lattice. *Computer Application and Software*, 31, 33-36.
[6] Dai Qian. (2010) The establishment of chain supermarket data mining model. *Population & Economics*, 8, 152-153.
[7] Zhao Lindu. (2012) The application of data mining technology in HIS. *Journal of Southeast University (Philosophy and Social Science)*, 7, 80-84.
[8] Wang Fudong. (2012) Clients relations analysis evaluation system based on data mining. *Journal of Southeast University (Philosophy and Social Science)*, 4, 99-102.
[9] Ding li. (2013) The Research about Data Mining of User's Behaviors Based on Apriori Algorithm. *Bulletin of Science and Technology*, 29, 214-217.
[10] Zhang Haitao. (2013) Research advances on privacy-preserving data mining. *Application Research of Computers*, 30, 3529-3535.

Authors

Yaguang Xiang, 29.8.1977, China

Current position, grades: an associate professor from West Anhui University.
Scientific interest: P.E. Economics, P.E. and computer, humanities and social science of P.E.
Publications: many high-quality articles in these fields.