Research on Apriori algorithm and its application in electronic commerce system

Yiyong YE
Wuyi University, School of Economics & Management, Jiangmen Guangdong , China
30117406@qq.com

Keywords：Apriori algorithm; data mining; association rules;electronic commerce

Abstract：For large amounts of data generated by e-commerce platform, this paper combined with the actual needs of e-commerce recommendation system, study a common technique of the correlation analysis which call association rules that oriented e-commerce web mining, then introduces the Apriori algorithm of association rules in detail, and discusses the concrete application of the Apriori algorithm through an example. Finally, it points out the deficiency of the classical Apriori algorithm, and gives the direction of improvement.

1 Introduction

With the rapid development of internet technology and internet economy, the Chinese e-commerce market transaction scale continues to expand, large amount of data were produce and accumulate by all kinds of electric business platform. Behind the explosive growth of data ,which hiding many important and valuable information, and the managers hope for higher level analysis of these data and to make use of. But the existing Network data management technology can only to complete the basic management functions of data , and still has some deficiency in data mining, knowledge discovery, leading to the emergence of "data explosion but lack of knowledge". An effective way to solve these problems, it is the traditional data mining technology combined with the web, to make the web data mining.

2 The Web data mining system based on E-commerce

Web data mining is first proposed by Oren Etzioni in 1996, the general definition is defined that found the mode P implied from the Web document structure and a set of C. If the C as input, P as output, then the Web mining is a process of mapping from input to output[1].

Web data mining is a process of a number of comprehensive use of technology, it is based on the Internet resources, discovering knowledge by extracting contains, unknown and potentially valuable patterns, and used to guide practice activities. It is the integrated use of a variety of data mining algorithms to find patterns and trends from sample data. It is the traditional data mining technology and theory which applied to a new research field of web resources mining[2]. Web mining research covers including database technology, information acquisition technology, statistics and artificial intelligence in machine learning and neural networks etc..

For the application of electronic commerce, web data mining is mainly through the analysis of users to visit a site, including the access preference page, columns, visit order, and make analysis of its activity rule, after that, make a prediction to the user's access behavior for the next step, and provide personalized recommendation service based on user habits to attract and retain customers.

Because the topological structure of e-commerce web site is relatively stable, although the time access and browsing mode for different users have differences, but in the long run, this habit is relatively stable, therefore, through the analysis of the visit of customers in a certain period of time, we can also find the potential clients, understand the relevant customer preference page browsing, so as to make customer clustering for targeted product promotion and recommendation. At present, the data mining in e-business process, mainly through the association rules, sequential pattern analysis, clustering and classification of these four kinds of technology to achieve.

3 Mining association rules and Apriori algorithm

3.1 Association rules overview

The concept of association rule was first proposed by Agrawal, Imielinski, Swami, it is a simple and practical rules in data mining[3]. In the knowledge discovery in database, association rules is a kind of knowledge model, it describes the rules that appear at the same time between different objects in a transaction; simple said, is to determine the impact of the object A's appearance for object B's appearance by the mathematical relationship.

Association rules is a common technique for e-commerce of web mining association analysis, it can find association knowledge do not know in advance hidden between different goods or services, through accessing to these rules, which can help site managers make better product promotion, marketing programs, and provide users with a more targeted recommendation service. At present, in the field of electronic commerce, the main content of the research on association rules include two aspects, one is how to quickly and efficiently generated frequent item sets, the other is based on frequent item sets generating management rules. The second step is relatively simple to operate, and the difficulty lies in the generation of frequent item sets, therefore, many scholars make research focused on how to generate the frequent item sets.

3.2 The basic definition of association rules mining[4]

Definition 1: Set $I=\{ i_1,i_2, \cdots\cdots i_m \}$is a set which has the collection of *m* different project, each i_k called a project. The collection of items called *I* item set. The number of its elements is known as the item sets of length, the length of the *k* item set called *k*- set.

Definition 2: Each transaction *T* is a subset of a set of *I*. Corresponding to each transaction has a unique identifier transaction number, denoted as *TID*. All trading constitute the transaction database *D*, |*D*| is equal to the number of *D* transactions.

Definition 3: For a set *X*, set count($X\subseteq T$) as the transactions number in *D* set which contains the *X* , the degree of support for the item sets of *X*: *support (X) =count (X T) /*|*D*|.

Definition 4: The minimum support degree is a set of minimum support threshold, denoted as *sup_min*, which represents the lowest importance of the association rules that user care about. Support of no less than *sup_min* set called the frequent item sets, the length of the *k* frequent item sets called *k*- frequent item sets.

Definition 5: Association rule is an implication: $R:X\Rightarrow Y$, $X\subset I$, $Y\subset I$, and $X\cap Y=$ □, which indicate that when set *X* appear in a transaction, it will lead to *Y* with a certain probability will appear. Association rules of interest to the user, which can use two standards to measured : support and confidence.

Definition 6: Support association rules of *R* is the number of transactions contains both the *X* and *Y* and the ratio of |*D*|. Namely: $support(X\Rightarrow Y)=count(X\cup Y)/|D|$, support degree reflects the probability *X, Y* appeared at the same time. Association rule's support is equal to the frequent item set support.

Definition 7: For association rules *R*, reliability refers to the number of transaction which contains *X* and *Y* and the ratio of the number of the transaction that contains a *X*. That is:

confidence$(X \Rightarrow Y)=$*support*$(X \Rightarrow Y)/$*support*(X), credibility reflects that: if the transaction contains *X*, the probability of transaction contains *Y*. In general, only the support and confidence of higher association rules is of interest to the user.

Definition 8: Frequent item sets, set $U=\{ u_1,u_2, \cdots\cdots u_n\}$ as the collection of items, and $U \in I$，$U \neq$ □, for a given minimum support degree *min_sup*, if the support of item sets *support*$(U) \geq$ *min_sup*, then *U* is said to be a frequent item sets, otherwise, *U* is the non frequent item sets.

3.3 Association rules mining algorithm Apriori

(1)The basic idea of the Apriori algorithm

Apriori algorithm is one of the most effect algorithm on mining Boolean association rule frequent item sets. Its core is the recursive algorithm based on the idea of the two phase frequency set. The association rule belongs to a single dimension, single, Boolean association rule in classification[5]. The algorithm will find the process of association rules which is divided into two steps: the first step by an iterative, retrieve the affairs of all the frequent item sets in a database, which support the degree not lower than the user set the threshold set; the second step, using the frequent item set to construct the rules to meet the users the minimum confidence. Specific approach is: first find frequent *1*-item sets, denoted as L_1; and then use L_1 to generate candidate item sets C_2, the items in the C_2 to determine the mining L_2, namely frequent *2*-item sets; continuously so the cycle continues until is unable to find the frequent *k*-item sets more far. While mining on each layer L_k, requires scanning the entire database again, the execution of the algorithm is mainly the use of a Apriori properties: all nonempty any frequent item set must also be frequent.

(2)The generating process of frequent item sets by Apriori algorithm

To generate frequent item sets process consists of two steps : connection and pruning[5]:

①Connection: For L_k, through L_{k-1} and its connected to generate candidate *k*-item sets in C_k . set l_1 and l_2 be the set of items in L_{K-1}. Set the $l_i[j]$ representation of L_i article *j* items. Apriori assumes that the transaction or the item order by the dictionary sort. For *(k-1)* set l_i, that will sort the items to be like this:$l_i[1]<l_i[2]<\ldots\ldots<l_i[k-1]$. For the elements l_1 and l_2 of L_{k-1},If l_1 and l_2's former *(K-2)* corresponding terms are equal, then l_1 and l_2 can be connected. That is, if:

$(l_1[1]=l_2[1]) \cap (l_1[2]=l_2[2]) \cap \ldots \cap (l_1[k-2]=l_2[k-2]) \cap (l_1[k-1]<l_2[k-1])$

$l_1$and l_2 can be connected. Condition $l_1[k-1]<l_2[k-1]$ is only guarantees of non repetition. The connection of $l_1$and l_2 generate the resulting item set for:$l_1[1], l_1[2], \ldots\ldots l_1[k-1], l_2[k-1]$.

②Pruning: According to the properties of the Apriori algorithm, frequent *k*-item sets any subset must be frequent item sets. So the generation set of C_k by connecting need to be verified, so as to remove non frequent *k*-item sets that does not meet the support.

(3)The main steps of Apriori algorithm[6]

①Scanning all data, and generate candidate *1*-item sets in C_1;

②According to the minimum support degree, generated frequent *1*-item sets L_1 from candidate *1*-item sets ;

③On the $k>1$, repeat *steps 4*, *5* and *6*;

④Connection and pruning operation was executed by L_k, which to generate a candidate *(k+1)*-item sets C_{k+1};

⑤Based on the minimum support degree, by the candidate *(k+1)*-item sets C_{k+1}, generating frequent *(k+1)*-item sets L_{k+1};

⑥if $L \neq \Phi$, then $k=k+1$, skip to *step 4*; otherwise, skip to *step 7*;

⑦According to the minimum confidence, generate strong association rules from frequent item sets, end.

3.4 Application and analysis of Apriori algorithm

According to the above algorithm process, combined with a transactional database (such as shopping cart) to carry on the example analysis to the Apriori algorithm, and to find out the frequent item sets. The specific transaction data such as shown in Table 1, first define the minimum support threshold *min_sup*=3, and use the support count directly to express support.

Table 1 The transaction database *D*

Transaction ID	1	2	3	4	5	6	7	8	9	10
Items	ACDE	ABD	ACE	ABC	C	BCD	AB	ABCE	ABC	BCE

Scanning the database for the first time, firstly calculate the transaction database candidate 1-item sets, and then get support count for each option in Items, as shown in Table 2.

Table 2 Candidate *1*-item sets C_1

Item set	{A}	{B}	{C}	{D}	{E}
Support	7	7	8	3	4

As we can see from Table 2, because the minimum support threshold is set to 3, so each support count all meet the requirements, so candidate *1*-item sets is the frequent *1*-item sets. At this time $L_1=C_1$.

Second scans of the database, in order to discover frequent 2-item sets L_2, algorithm using $L_1 \otimes L_1$ to generate candidate *2*-item sets in C_2, separately carries on the support count for each candidate, as shown in Table 3.

Table 3 Candidate *2*-item sets C_2

Item set	{AB}	{AC}	{AD}	{AE}	{BC}	{BD}	{BE}	{CD}	{CE}	{DE}
Support	5	5	2	3	5	2	2	2	4	1

From the Table 3, *{AD}, {BD}, {BE}, {CD}, {DE}* five item's support count is less than the minimum support threshold 3, so delete it from the table (pruning), the remaining parts of frequent *2*-item sets, such as shown in Table 4:

Table 4 Frequent *2*-item sets L_2

Item set	{AB}	{AC}	{AE}	{BC}	{CE}
Support	5	5	3	5	4

Third scans of the database, to solve the frequent *3*-item sets, connect L_2 with itself, namely $L_2 \otimes L_2$, and generating candidate *3*-item sets, separately calculate for each candidate support count, as shown in Table 5:

Table 5 Candidate *3*-item sets C_3

Item set	{ABC}	{ABE}	{ACE}	{BCE}
Support	3	1	3	2

It can be seen from Table 5, *{ABE}, {BCE}* two item's support count is smaller than the minimum support threshold 3, so delete it from the list, the remaining parts of frequent *3*-item sets, such as shown in Table 6:

Table 6 Frequent *3*-item sets L_3

Item set	{ABC}	{ACE}
Support	3	3

Fourth scans of the database, to solve the frequent *4*-item sets, connect L_3 with itself, namely $C_4=L_3 \otimes L_3=\{ABCE\}$, and generate the candidate *4*-item sets. According to the properties of the Apriori algorithm, because the *{ABE}* subset it is not frequent, so this item set is deleted. So $C_4= \Phi$, and the algorithm terminates. Thus all frequent item sets have been found out, the frequent item sets finally identified as $L=L_1 \cup L_2 \cup L_3$.

During the application of the electronic commerce, we are more recommended commodities or services that are of interest to the user, if the transaction database as the shopping cart, then *A, B, C* and *A, B, E* these 3 kinds of products are respectively often appear together in the shopping cart, according to the conclusions, we will make corresponding adjustment in product placement, product combination or product links. On this basis, we make further analysis of the relationship between other goods and *A, B, C, E* four kinds of goods, and to identify more association rules, which can provide powerful basis for marketing plan formulation products, commodity promotion, and enhancing competition ability of electronic commerce website.

4 The improved Apriori algorithm

The characteristics of Apriori algorithm is simple and easy to use, it is based on the recursive statistics, gradually generated frequent item sets, but in practical application, it also has some problems, for example: It has to scan the transaction database repeatedly, when capacity is relatively large, it may use much time; second, the candidate item set created quantity is too large, resulting in too long operation; the minimum support and confidence set is fixed only, cannot reflect the difference between the set of important degree; mining algorithm itself is only applicable to single dimensional Boolean association rules, for the multi-dimension and multi-layer situation, we need to improve the algorithm.

In order to improve the performance of Apriori algorithm, there have been many scholars to conduct the improvement and perfection on Apriori algorithm from different angles [7] , for example:

①To reduce the times of scanning database to improve performance of I/O.

②To improve the computational performance of generating frequent item sets.

③The search for effective parallel association rules algorithm.

④The introduction of sampling technology to improve the generated frequent item sets I/O and computational performance.

One of the more effective is: Chenetal proposed DHP (Direct Hashing and Pruning) algorithm, the main idea is to cut unnecessary items set to improve the efficiency when relational rule mining, DHP algorithm is based on Apriori algorithm, also joined the hash table architecture, which improve the efficiency of further[8]. In addition, proposed by Savasere et al division algorithm based on the data set is divided into several parts, each part can be accommodated in the main memory, each part independently generated frequent item sets, this method is highly parallelizable, only will each part of distribution on each processor can be used to generate association rules[7].

5 Conclusion

The rapid development of electronic commerce industry has brought huge benefits to the society, but also lead to industry competition between individuals, how to provide better service for the user, will become an important manifestation of the core competitiveness of enterprises. In this paper, through the introduction of Apriori mining algorithm, and its application in electric business case, it provides a feasible way for managers to analysis problem and make decision.

As for managers, not only to master the technology of data mining tools, should also be aware of the background object, so that we can make reasonable judgment according to the key factors in the actual situation of the data mining, and get the information the decision makers need, thus to support marketing and decision-making, and achieve the purpose of improving the economic benefit.

Reference

[1] Yu Zhen Wang. Analysis and Discussion of Web Data Mining.Development and application of computer.Vol 16:72-74(2003).
[2] Zi Rong Yang. Study of Fields-oriented High Quality Information Retrieval Based on Web Data Mining. Guizhou University Master Thesis.2008.
[3] Mao Sheng Dou. Research and application on association rules based data mining.Changchun University of Science and Technology Master Thesis.2009.
[4] Ming Gao. The Research and Application on the Algorithms of Mining Association Rules.Shandong Normal University Master Thesis.2006.
[5] Shao Chun Chang. Efficient frequent item set discovery methods and improved Apriori.Jiangsu University of Science and Technology Master Thesis.2011.
[6] A Fang Feng. An optimization algorithm of association rules Apriori.Consumer guide.Vol 25:265-266(2010)
[7] Ji Ming Hu. Research and Improvement on Apriori' s Algorithm in Mining with Association Rules.Computer technology and development.Vol 16:99-102(2006)
[8] Zhen Zhou. The research of Web data mining system based on E-commerce.Journal of Hunan Industry Polytechnic.Vol 9:58-59(2009)