International Journal of Database Theory and Application
Vol.8, No.2 (2015), pp.1-10
http://dx.doi.org/10.14257/ijdta.2015.8.2.01

Improvements in Data Mining Association Rules Algorithm

Dai Li

Dept. of Computer Science
Yunyang Teachers' College, Shiyan 442000, China
lidai72@gmail.com

Abstract

Because of the traditional Apriori algorithm in data mining in the process of operation for a long time, and produce a large number of unrelated item sets, caused great waste of data space, so this article puts forward an improved Apriori algorithm based on SQL, increase degree by calculation method, for pruning association rules and is independent of item sets.

Keywords: *Apriori algorithm, SQL, Association rules, Ascension degree*

1. Introduction

Mining association rules in data mining as an important species, has now got very extensive application in many fields, is a hot research topic in data mining. Association rules are to specify the database and the connection between the items in the description of these descriptions can be from a large number of records in the mining valuable support relationship, to help decide things. In 1993, Agrawal R. and others take the lead in put forward in the transaction database can find implicit association rules, the second year Agrawal R. Apriori algorithm [1] is given in, this algorithm can find the transaction database of frequent item set and use of these frequent item sets generate association rules. At present, in view of the Apriori algorithm is improved very much, most is the use of enhance the efficiency of the mess of frequent item sets, the method of using two standard support and credibility to the output rules, but if only the two indicators to measure the value of the rules, some not correct rules. Considering the SQL is a kind of structured query language (SQL), which itself has a strong group and statistical functions, so the paper puts forward an improved algorithm based on SQL Apriori. The algorithm is divided into three steps, the first step to use SQL statements to query of relational tables, calculate meet minimum support frequent item sets; The second step calculation rules of interest degree, judge the positive correlation rules; Last step calculation of the reliability of each rule, thus generating meet the minimum support and credibility is association rules. The algorithm introduces the concept of ascension degree, the use of ascending degrees will mistakes pruning association rules; Using SQL query relational tables, computing support count, effectively reduce the complexity of the space, and easy to program.

With the continuous development of database technology and the wide application of database management system, the amount of data stored in the database increases sharply. In a large amount of data hidden behind a lot of important information, and this important information is a good way to support people's decisions. The current database system can only access to existing data in the database, through the data gained by the amount of information is only part of the entire database contains information, hidden in the data after the more important information about the overall characteristics of the data description and prediction about the trend, the information generated in the decision process is of important

ISSN: 2005-4270 IJDTA

reference value. So people also constantly improve, to the requirement of data processing technology of data need to be able to further processing, in order to get on the general characteristics of the data and the prediction of development trend.

For example: the supermarket operators want to often buy goods together at the same time, in order to increase sales; Insurance companies want to know to buy insurance customers generally have what characteristics; Medical researchers want to from already of tens of thousands of cases to find the common features of patients suffering from a disease, so as to provide some help to cure the disease. For the above problems, the traditional database management system is hard to do, and is used for analyzing the data processing tools are rarely. In fact, the data is only observe the objective world people of raw material, itself do not have much meaning, it is just describe what happened, does not constitute a reliable basis of decision-making; Through the analysis of data to find the relations, gives a sense, and associated data, which form the so-called information. Although information is given some of the data has certain significance, but it often and people need to complete the task of no direct contact, also cannot be used as the basis for decision making. The information processing again, more in-depth analysis, in order to obtain more useful information, that is knowledge. So from the data to information, to knowledge, analysis processing of refining process is needed. Explosive growth, however, the amount of data that the user now is hard to like once upon a time, a large number of calculation based on the experience and the command of the human brain to find out a more comprehensive knowledge about data artificially, many knowledge still hidden in the data and can't be found and utilization of data resources waste. As John Naisbett said, "we have been overwhelmed with information, but they are enduring the torment of lack of knowledge". In the 1980 s, the Data warehouse and Data Mining (Data Mining, and DM) information processing technology such as it is in order to solve this problem and developed rapidly.

After more than ten years of development, based on statistics, artificial intelligence, such as theory and technical achievements has been successfully applied to the business processing and analysis. These applications to some extent for the emergence and development of data mining technology have played a great role in promoting. The core module technology and algorithm of data mining system is inseparable from the theory and technology support. In some sense, the theory itself development and application of data mining provides a valuable accumulation theory and application. Mathematical statistics is an applied mathematics discipline, the history of the development for hundreds of years, however, the combination of it and database technology research should be said that the recent more than ten years to be regarded. Before the application of mathematical statistics method were mostly through special process. And, most of the statistical analysis technique is based on strict mathematical theory and the application of the technique, which makes the general users are difficult to handle it gracefully. Data mining technology is actually an extension of the application of mathematical statistics analysis and development of probability theory and mathematical statistics for the data mining technology to provide a theoretical basis.

Artificial intelligence is the most computer science research in the controversial but still maintains a strong research field of life. Machine learning should be got fully research and development, inherited the machine learning and data mining technology to solve the problem. Expert System (Expert System) was thought to be artificial intelligence toward practical direction to develop the most promising technology, however, this technique also gradually show a large investment, strong subjectivity, narrow range of Achilles' heel. For example, knowledge acquisition is widely considered to be the bottleneck problem in the study of the expert system. In

addition, because of the expert system is subjective knowledge, so the mechanism of inevitable zone has bias and error. Data mining inherited the characteristics of expert system is highly practical, and with the data as the basic starting point, objectively mining knowledge. Say, therefore, data mining research in the inheritance of the existing related on the basis of research achievements in the field of artificial intelligence, out of the ivory tower before research model, really started to objectively found contain knowledge from data set. In particular, data mining can be seen as interdisciplinary database theory and machine learning, database technology focuses on the research on the efficient method of data processing, and machine learning is focused on designing a new method to extract knowledge from data. Data mining using database technology to the front-end processing of data, and using the machine learning method to extract useful knowledge from the processed data.

2. Related Works

Bar code technology and the development of shopping malls POS machine Settings made the super market store tens of thousands of number According to records, the detailed records for each customer each transaction time, commodity, quantity and price and other information, Provides a data basis for association rules mining. Association rule mining initially by R. Agrawal, T. Imielinski and A. Swami is put forward, applied in the transaction database, which is used to find users to buy in the supermarket

The implied relationship between commodity, *i.e.*, association rules, in order to provide basis for decision-making for the mall. These rules are to find out Customer purchase behavior patterns, such as bought a commodity to buy other goods. Decision makers can according to the level League rules provide information for the reasonable design goods shelves and arrange inventory to optimize the store layout (for example: Users often buy goods put together), do all kinds of promotional activities in sales and advertising, As well as the users are classified according to the buying patterns. Association rules is derived from the POS, but can be applied to many fields. The application of association rules also

Including the customers, shopping malls, product advertising, post analysis, network fault analysis, *etc*. Wal * Mart retail"" diapers and beer of the story is a successful typical case of association rule mining. Headquartered in the United States Ken color state Wal * Mart has the world's largest data warehouse system, which USES data mining tools for data. To analyze the original transaction data warehouse, got a surprise: buy most traders with diapers Product is beer. If not with the aid of data warehouse and data mining, the businessman can never find the hidden behind the fact that in the United States, some of the young father often after work to go to the supermarket to buy baby diaper, and he Were 30% ~ 40% of the people also buy some beer for himself. With this discovered, adjusted the supermarket goods Put, put nappies and beer together, and significantly increased the sales.

Association rule mining work one of the key problem is found in transaction set all meet user given minimum support of frequent item sets, this step focus all the amount of calculation. To solve the problem of association rules

The original algorithm is Rakesh Agrawal the AIS algorithm is put forward. For improving AIS algorithm, Heikki Mannila et OCD algorithm is proposed, the combination of OCD algorithm using the last search information to reduce the candidate items Set production quantity. Later, Rakesh Agrawal proposed the Apriori algorithm of association rules mining is one of the most famous and its variants AprioriTid and AprioriHybrid algorithm to find frequent itemsets. Since

then, many Scholars have put forward association rule discovery algorithm of frequent itemsets, but most algorithms are variant or its improved Apriori algorithm. Due to the Apriori algorithm is more than a trip to search algorithm, to huge amounts of data collection, each Search time, all want to read the storage time, peripheral I/O overhead. So, for most of the improved algorithm, and make an issue of how to reduce the search times. In fact, to make a real difference to frequent Apriori algorithm based on the classical association rules Item sets found the efficiency of the algorithm is to measure the level of project set and its problems, if the transaction data set contains the number of different projects for n, on the basis of the Apriori algorithm of frequent item sets found will meter 2 n item set. When n is large, will produce the combination explosion. In fact this is an np-hard problem.

3. System Model

3.1. Improve Degree's Concept

Data mining is to a large database of implicit knowledge and the process of potentially useful information is extracted. In the field of data mining, the research of association rule mining to conduct more in-depth.

Association rules is a set of transaction database has a certain relationship between rules. Target of association rule mining is the transaction database, association rule mining is to project (refer to the contents of the transaction) whether there is a relationship between for identification. Association rules can be said, for example, 80% of those who bought A and B have bought C and D again at the same time.

Association rule mining problem description is as follows: set the $I =' i_1, i_2, \cdots, i_m '$ is a collection of different projects, the number is m, D is a collection of trading on the I, every transaction contains several project $i_1, i_2, \cdots, i_k$. Association rules represented by $X \Rightarrow Y$, $X, Y \subset I$ and $X \bigcap Y = \phi$. X stands for the premise, Y stands for the results. Itemset is a collection of several projects, namely the item sets. Itemset is a statistical measure for Support: for X<I, if the collection contains the number of X for s, D so Support(X) = s. Measure of rule is Confidence, defined as follows

$$\text{confidence}(X \Rightarrow Y) = \frac{\text{Support}(X \cup Y)}{\text{Support}(X)} \quad (1)$$

Find these rules is to solve the main problem of mining association rules, the user to specify the minimum Support and minimum Confidence limit were less than their Support and Confidence.

$$\text{Ascension degree} = \frac{\text{The probability of events } A \text{ and } B \text{ actually occurs at the same time}}{\text{The probability of expectation occur simultaneously}} \quad (2)$$

Formula is expressed as

$$\text{Ascension degree} = \frac{P(A \cup B)}{P(A)P(B)} = \frac{P(B \mid A)}{P(B)} = \frac{P(A \mid B)}{P(A)} \quad (3)$$

A rule $A \Rightarrow B$ interest degree is also called the correlation, with represented with $\frac{P(B|A)}{P(B)}$, is the ratio of the real strength and expected strength, calculate the ratio is based on the premise of statistical independence. When events A and B are related, get rule $A \Rightarrow B$; When the event A negative is negatively related to the event B, get rule $A \Rightarrow -B$. Among them, the quantitative correlation index is called the degree of ascension, value range is the value of interval $[0, \infty]$, the two events are independent of each other things improve degree of close to 1, and greater than 1 when the event is related, and negatively related to the event is less than 1. In order to more timely and accurately determine correlation between item sets, need to introduce the concept of ascension degrees in the algorithm.

3.2 The Problem of Apriori Algorithm

The basic ideas of Apriori algorithm is the simplest form of association rule mining method, the association rules is one-dimensional, single-layer, Boolean association rules. Apriori algorithm is one of the most influential of the basic algorithm of mining frequent item sets, gets its name from the algorithm using the prior knowledge of the nature of frequent item sets. Apriori algorithm is a kind of broadband priority algorithm, through many times of database D scanning to find the entire frequent item sets, have the same number is taken into account in each scan (the same length) project. Apriori algorithm scans the database for the first time, all the data set D all individual in support of the project, also is to find frequent item sets 1. After every time before scanning the database, first of all, according to the frequent (k-1) item sets to generate new candidate item sets, and then by scanning the database statistics and their respective support delete support is lower than the minimum support threshold of candidate item sets, resulting in frequent k-item sets. Repeat the process until there is no number so far more frequent item sets.

Apriori algorithm, using the iterative step by step to find frequent item sets process description is as follows:

Input: D transaction databases, and the minimum support threshold minus.

Output: D the frequent item sets in L.

Begin

$L_1 = (\text{Large 1-}itemset)$;

For $(k = 2; L_{k-1} \neq \phi; k++)$ Do

Begin

$C_k = \text{apriori\_gen}(L_{k-1})$;

For all *transactions* $t \in D$ Do

Begin

$C_t = \text{subset}(C_k, t)$;

For all *Candidate* $c \in C_t$ Do

$c.count++$;

End;

$L_k = \{c \in C_t \mid c.count \geq minsup\}$

End;

Answer=$\bigcup_k L_k$;

End

Generate candidate item sets process description is as follows:

Connect steps: $\text{Apriori}-\text{gen}(L_{k-1})$

Begin

insert into C_k

Select $p.item_1, p.item_2 \cdots, p.item_{k-1}, q.item_{k-1}$

From $L_{k-1}p, L_{k-1}q$

Where $p.item_1 = q.item_1, p.item_2 = q.item_2 \cdots, p.item_{k-2} = q.item_{k-2}, p.item_{k-1} < q.item_{k-1}$

End;

Pruning steps: $\text{Apriori}-\text{gen}$

For all *itemsets* $c \in C_k$ Do

For all$(k-1)-$*subsets* s of c Do

If$(s \notin L_{k-1})$then

delete c from C_k;

problem posing

See from the above steps, the collection of C_k after cycle is required to produce, C_k every item of the set is on two belong to only one item different L_{k-1} set of frequency do $k-2$-connection to produce. L_k frequent item sets is a C_k a subset of the items that are included in the focus. All item sets of C_k needs to be validated in the database, so as to decide whether to join the frequent item sets L_k, several scans may be large database is the bottleneck of the method.

This method will cause when the database is large, the mining efficiency is very low. And algorithm simply dug up all the possible rules, lack the necessary human-computer interaction, users are not interested in a lot of rules, these all belong to the redundant rules, affect the mining efficiency. So can be judged according to improve degree of rules, classifying the original project, narrowing the scope of mining, improve the mining efficiency.

4. The Proposed Scheme

In this paper, using Apriori algorithm of scanning database and pattern matching calculation of candidate set support, improvement for using SQL statements to query of relational tables to calculate the candidate set support, n - frequent item sets and their support stored in relational tables tb_item_n (item 1 , item 2 $\cdots$ item n , support), calculating the ascension of the rules first, generating association rules, and then judge whether credibility is greater than the set minimum confidence threshold.

In this process of the excavation, the use of SQL statements to query of relational tables, first calculate the meet the minimum support count of frequent item sets; Then, correlation calculation rules, find out all the positive correlation rules, has nothing to do delete negative association rules and rules; Finally, calculation rules of credibility, generating all meet the minimum support and credibility is association rules. The algorithm takes advantage of the correlation pruning

redundant rules; Using SQL query relational tables, computing ascension, greatly reduced the space complexity.

4.1. Determine the Frequent l-Item Sets

Determination of frequent 1-item sets [4-5] flow chart shown in Figure 1.

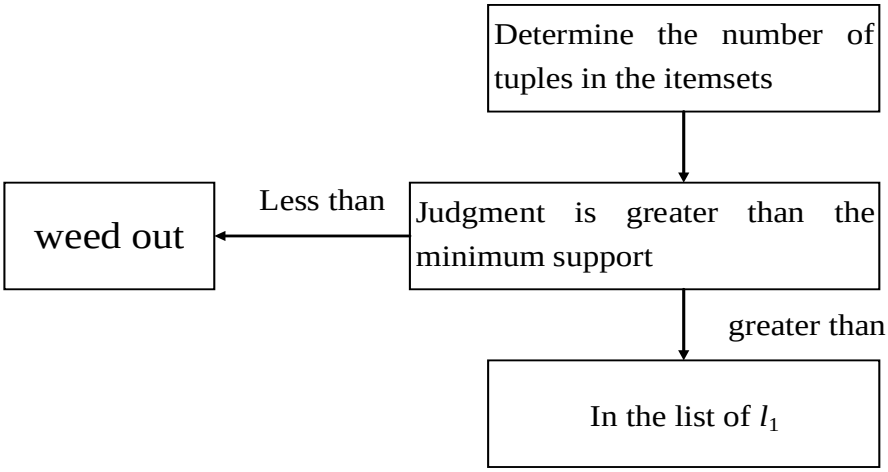

Figure 1. Determine the 1-Frequent Item Set Flow Chart

4.2. Determine the Frequent n-Item Sets

Determination of frequent n-item sets [4-5] flow chart shown in Figure 2.

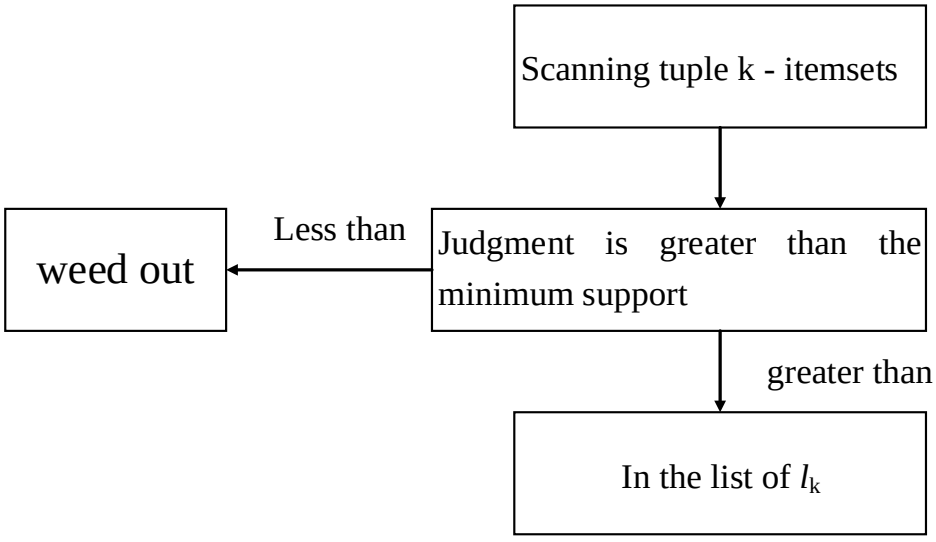

Figure 2. Determine the n-Frequent Item Sets Flow Chart

4.3 Generate Rules

First calculating the ascension of the rules, when events A and B are related, conclude rules $A \Rightarrow B$; When events A and B negative correlation, conclude rules $A \Rightarrow -B$; When independent events A and B, no rules. Generate rules of flow chart shown in Figure 3.

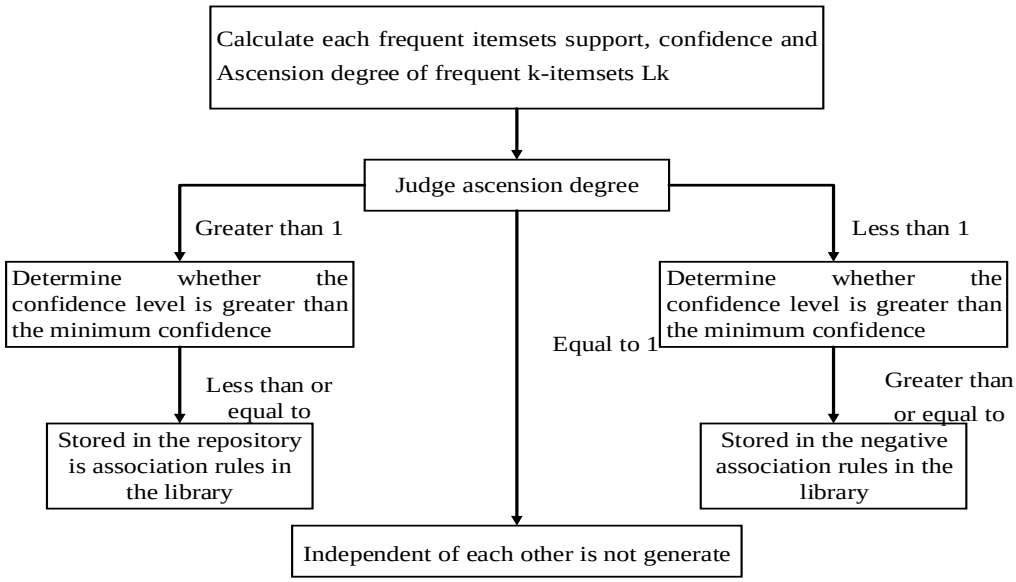

Figure 3. The Flow Chart of Generate the Rule

Generate rules of procedure is as follows:

for(k=2;$l_k \neq \phi$; k++)

for Frequent k - itemsets lk each of the frequent itemsets P

{

/*Calculation the support, improve the credibility of degrees of the rule e $\Rightarrow$ p-e*/

Confidence= $\frac{\text{p.count}}{\text{e.count}}$;

Lift= $\frac{\text{confidence}}{\frac{\text{(p-e).count}}{n}}$;

Support= $\frac{\text{p.count}}{\text{n}}$;

if Lift>1 then

{

if confidence min_conf} then

R_S=R_S$\cup${e $\Rightarrow$ p-e};

}

Else if Lift<1 then

{

confidence=1- $\frac{\text{p.count}}{\text{n}}$;

support= $\frac{\text{e.count-p.count}}{\text{n}}$;

$$\text{Lift}=\frac{\text{confidence}}{1-\frac{\text{(p-e).count}}{\text{n}}};$$

if confidence min_conf then

$R\_S=R\_S\cup\{e\Rightarrow p\text{-}e\};$

}

else

e and (p-e) are independent of each other, do not generate rules;

}

5. The Experiment and Test Results

In this paper, the test object as a metro company running a database, the subway accident form a daily with 920 transaction database, a total of 82 sets, using the improved algorithm to test the effectiveness of the improved. First by Apriori algorithm to overall scanning database, find out contains equipment damaged item set, a total of 3912 rules meet the threshold set, takes 3.3s; After using the improved algorithm, and then to a scan of the database, according to the communication by the user supplies the retrieve information, the data are pruned, to retrieve all related consumable items set of communication; Finally, for these projects, the excavation in this step consider minimum support and minimum confidence and improve degree of three elements, so dig up meet the requirements of correct rules only 1937, takes 0.8s, remove redundant rules, 1930. By comparison with the example of two algorithms, the query can be seen that the improved algorithm to search out the meet the rule conditions is less than that of the original Apriori algorithm, and takes shorter, also proved in this paper, the Apriori algorithm is feasible. By this algorithm can improve the query efficiency, avoid error rule, and decrease the space complexity.

6. Conclusion

For large and medium-sized enterprises, if a lot of project records, it is difficult to implement effectively mining Apriori algorithm. This article on how to improve the algorithm efficiency, reduce the irrelevant rules are studied, Apriori improved algorithm is proposed based on SQL. A little of this algorithm is to save storage space and reduce I/O load; avoiding repeated scanning database has nothing to do a lot of problems. Reading the database only twice, can significantly improve the mining efficiency, and with the fastest speed output query information effectively, the actual test confirmed the paper have put forward the feasibility of the improved algorithm Apriori.

References

[1] Z. Xiang, Z. Wei, “A large database of efficient sequential patterns incremental updating algorithm”, Journal of nanjing university, natural science edition, vol. 33, no. 2, **(2003)**, pp. 165-171.
[2] Z. Haifeng, X. YongKang, Yang Huali, *etc.*, “A is used for mining, negative association rules Apriori algorithm”, Journal of computer science, **(2007)**, pp. 242-244.
[3] Z. Yuquan, C. Geng and Y. Hebiao, “Positive and negative association rules mining algorithm study”, Journal of computer science, no. 3, **(2006)**, pp. 188-190.
[4] Innovation, “The king in the association rules extraction of an improved Apriori algorithm”, Computer engineering and application, vol. 40, no. 34, **(2005)**, pp. 183-185.
[5] Z. Xiaoyu, W. L. Winter and W. Guangyang, “An improved Apriori algorithm for mining association rules,” Computer technology and development, no. 12, **(2007)**, pp. 89-90.

[6] J. Liu, Y. Pan, K. Wang, and J. Han, "Mining frequent item sets by opportunistic projection," Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD'02), ACM Press, **(2002)**, pp. 229-238.
[7] M. J. Zaki and K. Gouda, "Fast vertical mining using daffiest" Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD'03), ACM Press, **(2003)**, pp. 326-335.
[8] R. Srikant and R. Agrawal, "Mining quantitative association rules in large relational tables", Proceedings of the 1996 ACM SIGMOD International Conference on Management of Data (SIGMOD'96), **(1996)**, pp. 1-12.
[9] R. Srikant, Q. Vu, and R. Agrawal, "Mining association rules with item constraints", Proc. 3rd Int. Conf. Knowledge Discovery and Data Mining (KDD'97), AAAI Press, **(1997)**, pp. 14-17.
[10] E. Baralis and G. Psaila, "Designing templates for mining association rules", Journal of Intelligent Information Systems, Kluwer Academic Publishers, **(1997)**, vol. 9, no. 1, pp. 7-32.

Author

Dai Li, born in 1972, an associate professor of Yunyang Teachers' College in China. He got a Master's degree in Computer Science. He is mainly researching on computer network, computer science education and Data Mining etc.

