Marxist theory database query optimization based on improved ID3 algorithm

Fan Jiang [1,2] *

[1] *Northeast Forestry University, School of Marxism, Harbin 150040, Heilongjiang, China*

[2] *Harbin Party School of CPC, Department of The party's history and construction, Harbin, 150040, Heilongjiang, China*

Received 1 June 2014, www.tsi.lv

Abstract

Focused on the problem that the data query of Marxist theory database requires to be optimized, this paper proposes a database query optimization strategy based on improved ID3 algorithm. It firstly changes the measure of property selection in data set so as to decrease the computational expense and generation time, then adjusts the calculation of information gain to the calculation of residual value of information gain and selects the property with minimum residual value as a new standard to replace the original information gain. Simulation experiment shows that improved ID3 algorithm is superior to standard ID3 algorithm in accuracy and time consumption in the establishment of decision-making tree.

Keywords: Marxist Theory Database, Data Query Optimization, Decision-making Establishment Optimization, Stabilization Optimization, Improved ID3 Algorithm

1 Introduction

With the development of society, Marxist theory and ideological and political education have been granted as the state-level key disciplines and specialties by the Ministry of Education. The construction of Marxist theory database is the product of discipline development and professsionalization [1]. Through data processing, online full text browse and query makes more students and teachers quickly acquire the essence of Marxist theory [2]. Spreading Marxist theory information via Internet has a far-reaching significance in discipline construction and can help new Marxists grow rapidly. However, current Marxism theory database requires optimization of database query [3].

R* algorithm, a distributed successor of relational database system developed by IBM St. Joseph Laboratory, is query optimization algorithm based on direct connection operation with the purpose of cooperating with distributed database systems constructed by multiple independent sites which is also a relational database system[4-5]. The principle is to enumerate all the connections according to the query, distribute each possible site and finally select the best one based on the optimum principle [6]. Distributed INGRES is also an early algorithm based on direct connection, developed by UC-Berkeley based on INGRES [7]. It set the decomposition as an optimization strategy: firstly it decomposed the multiple relations query into the query with one relation; secondly it executed every single relation query, chose an initial executing plan with heuristic method and determined the query order through intermediate relationships. It is a dynamic interpretation algorithm.

C-POREL was designed cooperatively by Chinese Academic of Science (Institute of Mathematics), University of Science and Technology and East China Normal University [8], which supported level data fragment. LSZ is a heterogeneous distributed database system designed by Nanjing University realizing the local query optimization and multiple connection optimization [9]. SUNDDB is a distributed database management system designed by Southeast University with functions like multiple duplicates, level relation fragment and snapshot. Galaxy is a distributed heterogeneous data source integrated system based on CORBA designed also by Southeast University, and also realized the query optimization [10]. However, most of these query optimization algorithms have existed limitations to some extent. In practice, the difference of network structures and data structures makes it a NP problem and the query efficiency of some query optimazation algorithms changes greatly when the data size changes.

Based on the requirement of data query of Marxist theory database, this paper put forward a data query optimization strategy based on improved ID3 algorithm, and simulated with sample experiment to prove its excellent performance.

2 Standard ID3 algorithm

2.1 THE THOUGHTS OF ALGORITHM

The core of ID3 algorithm is, when we choose attributes at each node of decision making tree, we choose the information gain or mutual information as the measure of the split attribute. According to the definition of information gain, the split should minimize the required information for accurate classification. The specific steps include: detec-

* ***Corresponding author's*** e-mail: jiang9213fan@163.com

ting all the attributes and selecting the attribute of maximum information gain as the node of decision-making tree; building corresponding branches due to different values; building the branches of the decision-making tree nodes with the recursion of subsets of each branches until all the subsets only involve the data of identical attributes.

The information gain of attributes is calculated with following method. By comparing with each other, the attribute of maximum information gain is obtained.

Suppose S is set composed of $|S|$ data samples and category number attribute has n different values, defined as $C_i(i=1,2,...,n)$. The sample number of C_i is $|C_i|$, then the expected information of the given sample category is expressed as follows.

$$I(S_1, S_2, ...S_n) = \sum_{i=1}^{n} -p\log_2(p_i), \tag{1}$$

where, $p_i = |C_i| / |S|$ is the probability of a sample belonging to category i.

Attribute A has m different values $\{x_1, x_2, ..., x_m\}$, which can divide the S into m subsets $\{S_1, S_2, ..., S_m\}$, where the sample of S_j has the same value $x_j(j=1,2,...,m)$. In subset S_j, the sample number of category i is supposed to be $|S_{ij}|$, given by the entropy of information gain of subsets divided from A.

$$E(A) = \sum_{j=1}^{m} \frac{|S_{1j} + S_{2j} + ... + S_{nj}|}{|S|} I(S_{1j} + S_{2j} + ... + S_{nj}) \tag{2}$$

Then the information gain obtained on attribute A is,

$$Gain(A) = I(S_1, S_2, ..., S_n) - E(A) \tag{3}$$

From the equation above, the lower entropy is, the higher information gain will be.

2.2 THE DESCRIPTION OF ALGORITHM

1) Randomly select a subset, called as a window, with both positive sample and negative sample from training sample set.

2) Adopt ID3 algorithm to generate a decision-making tree with the subset above.

3) Judge the category of the tuple in the training sample set beyond the randomly selected subset with the decision-making tree, and find the misclassified tuple.

4) If there exists the misclassified tuple, then they are inserted into randomly selected subset, turn to step 2); otherwise, end the execution.

The algorithm flow is shown in Figure 1. The training sample set is composed of true example set TE and false example set FE. The subsets of TE are expressed with $TE1$ and $TE2$ while the subsets of FE are described by $FE1$ and $FE2$.

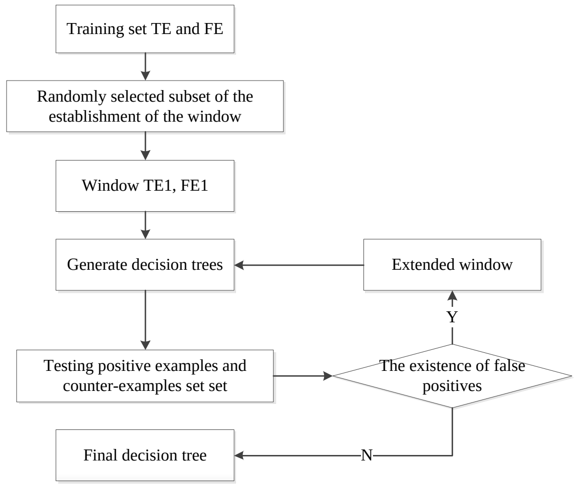

FIGURE 1 ID3 algorithm flowchart

The generated decision-making tree will be updated as soon as each loop is executed.

2.3 THE SHORTCOMINGS OF ID3 ALGORITHM

ID3 traverses the hypothesis space with a hill-climbing strategy from simple to complicated, starting from empty tree and then considering more complicated hypothesis. Through observation of search space and search strategy, we find it still exist shortcomings.

1) When traversing decision-making tree space, ID3 algorithm only sustains single current hypothesis, which loses the advantage of its representing all the consistency hypothesis. For instance, it cannot judge how many other decision-making trees are consistent with the training data available, or using new example query to optimally distinguish these competing hypotheses.

2) ID3 algorithm doesn't back track, and selects a attribute of some layer to test.

3) ID3 algorithm is a greedy algorithm. Because it doesn't accept training examples incrementally, the incremental learning task makes it give up primary decision-making tree and reconstruct one costly. Therefore, ID3 algorithm is not appropriate for incremental learning.

3 The improvement of ID3 algorithm

3.1 DECISION-MAKING TREE OPTIMIZATION

From the principle of decision-making tree, the construction of it is based on the information theory, mainly involving the equation of amount of information. Therefore, when choosing a split node, the algorithm must involve several times of logarithmic calculation. In large data volume calculation, it will apparently influence the efficiency of the establishment of decision-making tree. Therefore, considering the measure of property selection will decrease the calculation cost and achievement time.

Based on in-depth study of optimization theory, this paper uses convex function to adjust the information amount equation.

1) Suppose $f(x)$ is continuous in $[a,b]$, and has first-order and second-order derivatives in (a,b).

If in (a,b), $f''(x)>0$, then $f(x)$ has a concave shape in $[a,b]$; if $f''(x)<0$, then $f(x)$ has convex shape in $[a,b]$.

2) If $f(x)$ is a convex function in interval I, $\forall x_1, x_2 \in I$, $\lambda \in (0,1)$, then

$$\lambda f(x_1)+(1-\lambda)f(x_2) \le f[\lambda x_1+(1-\lambda)x_2] \tag{4}$$

In the function $\log_2 P$ used in information amount calculation, P represents the percentage of some record count in total record count with the domain of definition $(0,1]$.

Two points P_1 and P_2 arbitrary in $(0,1]$ meet the condition that when $P_1-P_2=\Delta P \to 0$, $\log_2 P$ is continuous. According to (1), we judge the concavity and convexity of the function $\log_2 P$.

$$(\log_2 P)' = \frac{1}{P \times \ln 2} \tag{5}$$

$$(\log_2 P)'' = -\frac{1}{P^2 \times \ln 2} < 0 \tag{6}$$

Therefore, the function $\log_2 P$ shows a convex morphology.

3) If $f(x)$ is convex function in I, and: $\forall x_1, x_2, ..., x_n \in I$, $\lambda_1, \lambda_2, ..., \lambda_n > 0$ and $\lambda_1+\lambda_2+...+\lambda_n=1$, then,

$$\lambda_1 f(x_1)+...+\lambda_n f(x_n) \le f(\lambda_1 x_1+...+\lambda_n x_n) \tag{7}$$

This paper adjusts the information amount equation into (8),

$$I(S_1, S_2, ..., S_m)' = -\log_2 \sum_{i=1}^{m} P_i^2 \tag{8}$$

Apparently, the accuracy of the decision-making tree classification will be influenced by this improved information content equation. But this influence is very small to the whole performance of data classification. The change of information content will contribute to the change of information entropy.

From equation (2), the improved information entropy equation is,

$$E(A)' = \sum_{j=1}^{m} \frac{|S_{1j}+S_{2j}+...+S_{nj}|}{|S|} \cdot (-\log_2 \sum_{j=1}^{m} P_{1j}^2 + P_{2j}^2 + ... + P_{nj}^2), \tag{9}$$

where, S_{ij} is the sample set of subset S_j belonging to category C_i; $\frac{|S_{1j}+S_{2j}+...+S_{nj}|}{|S|}$ represents the weight of the j subset.

3.2 STABILITY OPTIMIZATION

ID3 decision-making tree changes when training set increases. During the tree establishment, mutual information of each characteristic will change with the examples together with the decision-making tree. This kind of changeable data set is not appropriate for learning.

In a common decision-making tree, amount of information is utilized as the measure for testing attributes. ID3 algorithm uses $Gini$ index to replace the information gain with better performance. For a data set S with n categories, $Gini(S)$ is defined as,

$$gini(s) = 1 - \sum p_j \cdot p_j, \tag{10}$$

where, p_j is the frequency of j category data in S. The smaller $Gini$ is, the larger information gain will be.

To solve the instability of the decision-making tree in ID3 algorithm, this paper made corresponding adjustment to the calculation of information gain of ID3. Different with $Gini$, ID3 algorithm chooses the maximum information gain as the measure of test attributes while $Gini$ chooses minimum $Gini$ as the index. This paper chooses the improvement based on $Gini$ split index, thus this paper adjusts the calculation of information gain to its residual value, and then chooses minimum residual value as the new standard to replace primary information gain.

The improved information gain has the following expression.

$$\begin{aligned} gain_{left} &= 1 - gain(A)' \\ &= 1 - \frac{I(S_1, S_2, ..., S_m) - aE(A)}{m}, \end{aligned} \tag{11}$$

where, a is the attribute priority value, $(0,1]$, and m is the number of values of attribute A.

Choosing the minimum $gain_{left}$ as the new measure, not only overcomes the shortcoming of ID3 that easily chooses property with much more values, but also offsets the error from convex function, improves the classification efficiency of decision-making tree. At the same time, it solves the instability problem of the decision-making tree.

4 Simulation research

To verify the validity of the proposed algorithm, this paper tested original ID3 algorithm and improved algorithm, selecting four data sets from Marxist theory database, and comparing with each other from aspects of rules number and tree establishment time. Every group of data set is experimented 20 times and their mean value is calculated to make the experiment generalized.

4.1 CONTRAST OF THE NUMBER OF RULES

TABLE 1 Contrast of the number of rules

Dataset	Record number n	ID3 number of rules	Improved ID3 number of rules
1	512	51	42
2	721	74	45
3	892	92	62
4	1053	112	91

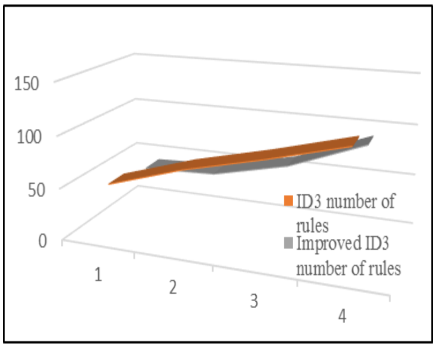

FIGURE 2 The contrast of number of rules

Seen from above examples, the number of decision-making tree rules established by improved algorithm is far less than that by ID3 algorithm. The number of rules corresponds to the number of leaf node. The fewer node the leaf has, the fewer number of rules is. And this attribute is more apparent when the examples set size is larger and the attributes sets are more.

4.2 ACHIEVEMENT TIME

TABLE 2 Achievement time comparison

Dataset	Record number n	ID3 elapsed time	Improved ID3 elapsed time
1	512	153.4	112.4
2	721	295.3	212.2
3	892	419.2	291.3
4	1053	562.7	342.8

From Figure 3, it is evident that improved ID3 algorithm consumes less time than original ID3 algorithm. In addition, with the increasing data size, time difference increases linearly. Certainly, it has something to do with the characteristics distribution of the data set. All of these prove that the improved ID3 algorithm is superior to the original ID3 algorithm both in efficiency and performance when dealing with the decision tree establishment problem of large size data set.

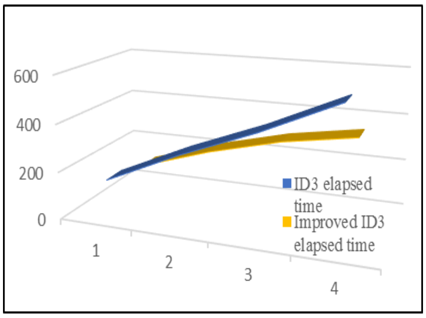

FIGURE 3 Achievement time comparison

5 Conclusions

The construction of Marxist theory database is beneficial to the insight study of Marxism and the insistence of scientific development. In view of the condition that the data query of Marxist theory database doesn't have a high performance, this paper put forward a database query optimization strategy based on improved ID3 algorithm. Experiment results show that compared with original ID3 algorithm, improved ID3 algorithm can construct simpler decision-making tree classification model, and is superior in the accuracy and time consumption.

Acknowledgments

This work was supported by Philosophy and Social Science Planning Project in Heilongjiang "Fusion Research about Marxism Popularization and Chinese Traditional culture" (No. 11B079).

References

[1] Bo Shanhua, Sun Chaojuan. (2014) Study on network communication mode for the population of Marxism. *The Journal of Xinzhou Teachers University*, 30, 107-109.
[2] Wen Hong. (2014) The application of new-developing carrier on the population of Marxism. *Journal of Social Science of Jiamusi University*, 32, 25-27.
[3] Pang Chengzhi. (2014) A philosophical thought on the Marxism belief education dilemma of college students today. *Motherland*, 2.
[4] Meng Xiangfu. (2012) A Categorization Approach Based on Adapted Decision Tree Algorithm for Web Databases Query Results. *Journal of Computer Research and Development*, 49, 2656-2670.
[5] Lin Guiya. (2012) Optimization of database query application research based on particle swarm algorithm. *Application Research of Computers*, 29, 947-949.
[6] Li Fangfang, Tian Zhijun. (2012) Research and Application of Fast Database Query Methods. *Microelectronics & Computer*, 39, 163-166.
[7] Zhang Xin. (2011) Relationship Database Query Optimization Based on SQL Standards. *Coal Technology*, 30, 284-286.
[8] Ye Rujun. (2011) Database Query Optimization and Simulation Framework. *Microelectronics & Computer*, 28, 195-197.
[9] Li Zhiwei. (2010) Study on distributed database query optimization based on greedy strategy. *Computer Engineering and Design*, 3838-3840.
[10] Xu Xinhua, Tang Shengqun. (2009) Review on latest parallel database query optimization technology. *Computer Engineering and Design*, 3814-3819.

Authors

Fan Jiang, 20. 07. 1980, China

Currently, Fan Jiang, is a PHD of College of Marxism, Northeast Forestry University, China. And she is a researcher at Party School of Harbin Municipal of C.P.C, China. Her research interests include Marxist theory, china traditional culture and database research.