Optimization for massive data query method in database

Xie xiaodong, zou jinpin

Jiangxi Biotech Vocational College 330200

Jiangxi Biotech Vocational Collegeg330200

Jiangxi College of Foreign StudiesHuang xi 330099

Keywords: Database; inquiry; correlation clustering;

Abstract: The massive data query methods in database is studied to improve accuracy of query. In the process of data querying in the database, once the data volume is overflow and thetype of data becomes complex, the query requires a lot of restrictions, resulting in time-consuming and low query accuracy rate for data query. To this end, optimization for massive data query method in database based oncorrelation clustering algorithm is proposed. Correlation clustering is processed to all the data in the database to obtain the correlation between different data. Query for the specified categories of data to achieve query optimization of massive data in database. Experimental results show that the proposed method for mass data query in database can improve the accuracy and efficiency of the query, and reduce time for querying.

1 Introduction

With the arriving of the modern computer information era, databases is widely applied in people's production and life [1]. In the application process of the database, how to improve query efficiency and accuracy of the database, has become a hot issue in the field of the database research [2]. Therefore, the data query methods in the database, has become one of the key research topics in the field [3.4]. Currently, the mainstream data query method in the database includes data query method in the database based on decision tree algorithm [5], data query method in the database based on support vector machine algorithm [6] and data query method in the database based on artificial neural network algorithm. Among them, the most commonly used data query method in the database is on the basis of decision tree algorithm. As the data query method in the database plays a vital role in improving the accuracy and efficiency of data query, it has much room for development, which has attracted more attention from the relevant experts.

2 Principle of data query in the database

2.1 Correlated data clustering process

The sequence composed of massive data in database can be represented by $V=\{v_1,v_2,\cdots,v_q\}$, where, v_l represents l-th data among the sequence, the corresponding characteristics of the above date can be expressed by $K=\{k_1,k_2,\cdots,k_p\}$. According to the related theory of fuzzy clustering can classify all the data.

On the basis of the following formula to calculate fuzzy clustering objectives of the massive data in the database:

$$L_p(Y,B)=\sum_{k=1}^{q}\sum_{l=1}^{e}y_l^p f_{kl}^2(z_k,b_l) \qquad (1)$$

State parameters of massive data in database can be denoted by $e,q,p,d=1$, the cluster center is denoted by $B_{(d)}=(b_1,b_2,\cdots,b_e)$, andthe following formula can be carried out to update all fuzzy clustering center timely:

$$y_{kl} = \frac{1}{\sum_{m=1}^{e}\left[\frac{f_{kl}}{f_{km}}\right]^{\frac{2}{p-1}}} \text{ when } f_{kl} \neq 1$$

$$y_{kl} = 0 \text{ 当} e_{kl} = 0, l \neq m$$

$$y_{kl} = 1 \text{ 当} e_{kl} = 0 \qquad (2)$$

According to the following formula to calculate the mean of all data in the database:

$$b_l = \frac{\sum_{k=1}^{q} y_{kl}}{\sum_{k=1}^{q} y_{lm}} \qquad (3)$$

b_d iscompared with $b_{(d+1)}$, if these data can satisfy the following constraints, it is possible to achieve fuzzy clustering processing of massive data:

$$\left|b_d - b_{(d+1)}\right| \leq \varphi \qquad (4)$$

In the database, the value of the objective function of fuzzy clustering is dwindling. In the process of fuzzy clustering of massive data, can effectively avoid getting into the defect of local minimum in iterative processing, resulting in obtaining massive data clustering structure.

2.2 optimization processing for massive data query

The collection composed of massive data in database is $Z = \{(z_k, a_k) \mid k = 1, 2, \cdots, total\}$, the k-th element in the collection can be expressed by $z_k = (z_{k1}, z_{k2}, \cdots, z_{kf})$. According to the following formula to calculate the expected value of massive data:

$$K(q_1, q_2, \cdots, q_p) = -\sum_{l=1}^{p} \frac{q_l}{total} \log_2 (\frac{q_l}{total}) \quad (5)$$

Among the data collection constituted by massive data, the data set of all attributes is $C_h (h = 1, 2, \cdots, f)$, the following equation can be utilized to decompose attributes of data:

$$G(C_h) = \sum_{u=1}^{s} \frac{q_{1u} + \cdots + q_{pu}}{total} K(q_{1u}, q_{2u}, \cdots, q_{pu}) \quad (6)$$

According to the following formula to calculate the information gain ratio of properties of massive data in database:

$$E(C_h) = K(q_1, q_2, \cdots, q_h) - G(C_h)$$

$$u(C_h) = -\sum_{u=1}^{s} \frac{q_u}{total} \times nd\left(\frac{q_u}{total}\right)$$

$$I\_t(C_h) = \frac{I(C_h)}{u(C_h)} \qquad (7)$$

According to the following formula to build the decision tree of massive data query:

$$C_k = MIN + \frac{MAX - MIN}{Q} \times k \qquad (8)$$

Among them, $k = 1, 2, \cdots, Q$

According to the above-described methods, obtained decision tree structure can be expressed by the following diagram:

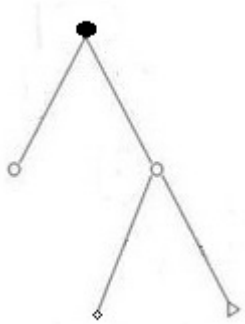

Figure 1 Structureof decision tree

The data of maximum information gain ratio in the database is regarded as a branch of the decision tree, to build a decision tree for massive data query in database.

3 Experimental results and analysis

In order to verify the validity of data query method in the database based on correlated clustering algorithm, there is the need for an experiment.

In this paper, both traditional algorithm and proposed algorithm are adopted for massive data query in database, query error results obtained can be expressed by the following Figure 2:

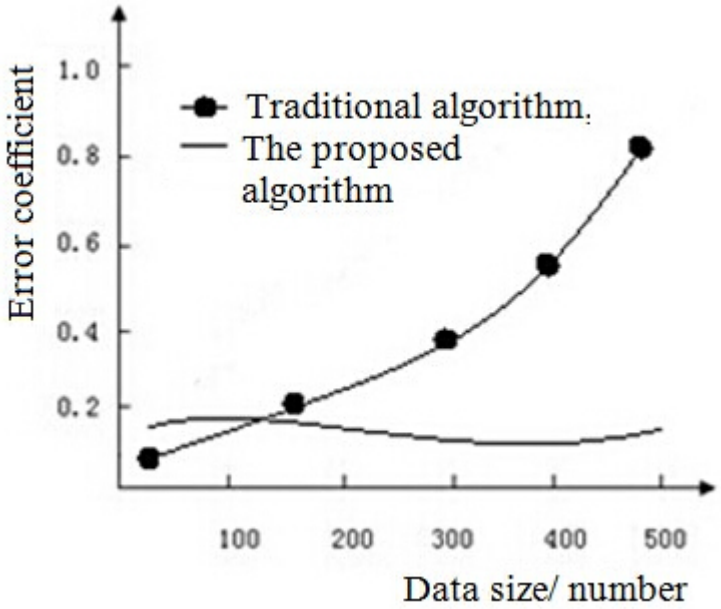

Figure 2 query error of different algorithms

According to the experimental results, it can be learned that by using this algorithm for massive data query in database, the query accuracy can be increased.

4 Conclusions

Optimization for massive data query method in database based on correlation clustering algorithm is proposed in this paper. Correlation clustering is processed to all the data in the database to obtain the correlation between different data. Query for the specified categories of data to achieve query optimization of massive data in database. Experimental results show that the proposed method for mass data query in database can improve the accuracy and efficiency of the query.

References:

[1] Chen Tao.On cloud computing theory and its technique [J].Journal of Chongqing Jiaotong University (Social Sciences Edition),2009.8:104-106.

[2] QiaoWa, XieYuebo.The establishment and application of China's storm data query database. China rural water and hydropower.2013 (7)78-80.

[3] Wu Chao, Shen Weiqun, Pan Shunliang, et al. Design and Realization of an Engineering Helicopter Simulator Control Center [J]. Computer simulation,2006,23(9）: 294-297.

[4] Sun Yong WU, Guanxiang. Design and Implementation of mobile GIS application based on Android [J].International Journal of Technology Management.2014 (4): 61-63.

[5] Tang Jian.Cloud computing Database research and its application in distance education in [J].Journal of ChifengUniversity,2009.11:35-36

[6] Liu Zhongbo. Performance comparison study ofNET data query and delete based on three-tier ObjectDatasource and two-tier SqIDataSource[J]. Electronic world,2013(13): 99-100.