Research on the effective detection method of specific data in large database

LI Jin-feng

(Vocational College of DongYing, Shandong 257091, China)

Keywords: large database; specific data; frequent sets;

Abstract. the effective detection of specific data for large databases can ensure the security of large data. Because of the strong subjectivity of the traditional detection method, the accuracy of detection and the detection efficiency are decreased. To this end, a specific data detection method based on Apriori algorithm for large databases is proposed. The characteristics of specific data nodes is calculated to provide data base for the detection of specific data. To find out the frequent sets of all the specific data, the frequent sets are obtained strong association rules meet the minimum support degree and the minimum reliability, and then the detection of the specific data in large database is completed. Experimental results show that the detection of specific data of large database can be effectively improved by using the improved algorithm, the accuracy and detection efficiency of the detection are increased as well.

1 Introduction

Specific data detection is also known as the knowledge discovery in database [1]. With the continuous expansion of the scale of the Internet, the variety of specific data in the large database is becoming more and more complex, and the scale is also more and more huge [2]. In large databases, specific data is often hidden in complex data, so that the detection process of specific data becomes very difficult. Therefore, how to detect the specific data in the large database is of great practical significance, and it becomes a hot research hotspot in the current database field [3-6].

At present, the detection of specific data in large databases is mainly based on association rules [7], fuzzy clustering [8] and detection method based on support vector machine (SVM) [9]. The application of the specific data detection method based on SVM is most popular. However, in the process of using traditional algorithm for large databases specific data detection, it is difficult to establish effective detection rules in the detection process, resulting in the detection process takes longer, an increase of invalid results, low detection efficiency [10].

In view of the defects of the traditional algorithm, a large database specific data detection method based on Apriori algorithm is proposed, which provides a new method for the effective detection of large database specific data.

2 specific data detection method principle based on Apriori algorithm

Using traditional algorithm for specific data detection in large databases, it is difficult to establish effective detection rules in the detection process, resulting in the detection process takes longer, an increase of invalid results, low detection efficiency. To this end, a specific data detection method based on Apriori algorithm for large databases is proposed.

2.1 accurate classification of specific data nodes

Because the specific data distribution is stored in different data nodes, it is difficult to classify the nodes when detecting with traditional algorithm, due to the characteristics of the sub nodes is redundant. Subdata nodes has some commons, therefore, through the calculation of the correlation between subdata nodes, the subdata nodes is classified effectively, so as to ensure the accuracy of the detection of specific data.

Assuming the number of data nodes is p, e_k can describe the attribute characteristic of the k-th $(k = 0,1,\cdots, p)$ data, assuming these attribute characteristics can be divided into L classes,

sub-attribute feature is $(w_1, w_2, ..., w_n)$, n is the number of sub-attributes, each sub attribute can be said $e_k(w_{k1}, w_{k2}, ..., w_{kn})$, $k = 0,1,\cdots, n$, all the nodes of attributes' suspected feature data can be expressed in terms of $e_k = (w_{k1}, w_{k2}, ..., w_{kn})$, the association decision probability algorithm is to sort sub e_k according to the different characteristics, which can be divided into normal nodes and specific data nodes. The association degree probability of a sub-attribute c_j belonging to the category f_k is calculated as $q(f_k \cdot c_j)$, and the probability decision can be used for calculation:

$$q(f_k \cdot c_j) = (q(f_k) + q(f_k \cdot \chi_\upsilon)) / q(c_j) \qquad k = 0,1,\cdots, p \qquad (1)$$

In the above equation, $q(c_j)$ is the prior probability of the characteristic data detection, and $q(f_k \cdot c_j)$ is the prior conditional probability of the node partition. The detection results of $q(f_k)$ are invariant for the data nodes sub-attribute of same features. If there is no relation between the features of the sub- attributes, the following relationship can be established:

$$q(f_k \cdot c_j) = q(w_1 \cdot c_j) + q(w_2 \cdot c_j) + \cdots + q(w_m \cdot c_j) = \sum_{k=2} q(w_k \cdot c_j) \quad (2)$$

$$q(c_j) = \sum_{k=2}^{p} q(f_k) / q(c_j \cdot f_k) \quad (3)$$

After the feature of the above parameters is collected, the correlation probability values of different data node attribute characteristics is able to be calculated, the method is as follows:

$$q(c_j \cdot f_k) = T(p(c_j) + f_k) / T_{ek}^{\ 2} \quad (4)$$

$T(p(c_j) + f_k)$ is the number of data node sub-attributes that appear in the f_k class, and T_{ek} is the number of nodes that belong to the data node sub-attributes category in the samples for detection. If $q \sin\beta / \cos\beta < a$, it can be judged that belonging to specific data node attribute class, otherwise normal nodes.

According to the above method, feature sub attributes of the data nodes can be classified effectively, and can screen specific data in large databases, so as to provide a basis for the detection of specific data in large databases.

2.2 effective detection of specific data

According to Apriori algorithm related principle, the detection of specific data in large database can be accomplished. All data and relevant data of specific data in large database were counted, so as to obtain the number of specific data in large data, by calculating to acquire the specific data of which gain support reaches the minimum value. The obtained data collection of frequency 1 is N_1, using N_1 to calculate the collection N_2 composed of all elements of frequency 2, continuous calculating until obtain the collection of all the frequency M. The specific method is as follows:

(1) data connection processing: in the course of the data connection, all data need to be arranged in accordance with the rules of the descending, the sorting method is as follows:

The total element N_{m-1} of the large database can be self connected, so that the specific data frequency item E_m. In the collection, two elements n_1 and n_2 are selected, and the l-th item of the k-th element in the collection is obtained. The element relation of the large database is as follows:

$$n_k[1] < n_k[2] < \cdots < n_k[m-1] \quad (5)$$

In a large database, setting the specific data of the first $m-2$ items has the same attribute, and it can be self-connected for element N_{m-1}. The detection of specific data in large database needs to meet the following conditions:

$$(n_1[1] = n_2[1]) \wedge (n_1[2] = n_2[2]) \wedge \ldots \wedge (n_1[m-2] = n_2[m-2]) \wedge (n_1[m-1]) < n_2[m-1] \quad (6)$$

The following formula can be used for self-connection transformation:

$$n_1[1], n_1[2], \cdots, n_1[m-1], n_2[m-1] \quad (7)$$

(2) data pruning. Setting specific data E_m is a superset of N_m, in the specific data set, according to the values of candidate elements can calculate N_{m-1}. If the m-th specific data is not in the collection N_{m-1}, then the element is removed, so that all specific data can be obtained, and the detection of the specific data in the large database is finished. Firstly, find all the itemsets, the frequency of item appear is at least same to the predefined minimum support. Then, a strong association rule is generated based on itemsets, which must satisfy the minimum support degree and the minimum reliability.

According to the method described above, the characteristics of specific data nodes is calculated to provide data base for the detection of specific data. To find out the frequent sets of all the specific data, the frequent sets are obtained strong association rules meet the minimum support degree and the minimum reliability, and then the detection of the specific data in large database is completed.

3 experimental results and analysis

3.1 experimental environment setting

In order to verify the effectiveness of the improved algorithm, an experiment is needed. In the process of experiment, the program was written in Visual C++6.0 programming language. Experiment of detecting specific data of large database was conducted with different algorithms. Setting the number of data nodes in large databases is n and the number of specific data node is p, the set composed of feature vectors of specific data nodes is $\{g_1, g_2, \cdots, g_m\}$, equilibrium distribution coefficient of specific data nodes in a large database is μ, using the following formula to calculate detection accuracy of different algorithms.

$$\psi = \frac{\sqrt{n-p}}{g_i^2 - \mu} \times 100\% \quad (8)$$

Using the above formula, the detection accuracy of different algorithms can be calculated and the performance of the detection algorithm is described.

3.2 Experimental results comparison analysis

Using different algorithms to detect the specific data of large databases, the experimental results can be described with the following Figure1 and 2:

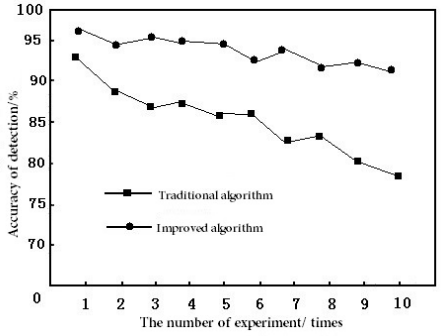

Figure 1 Detection accuracy comparison of different algorithms

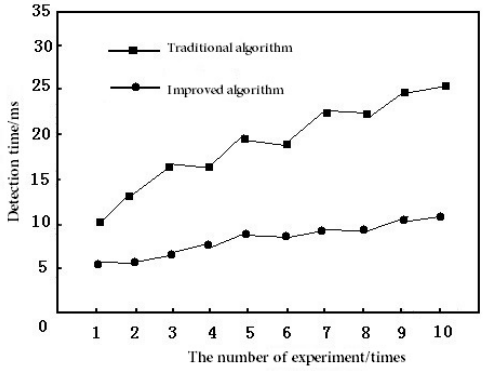

Fig. 2 Time consuming comparison of different algorithms

According to the experimental results it can be known that, using the improved algorithm for specific data detection of large database, obtained specific data detection accuracy was significantly higher than that of traditional algorithm, detection time was significantly lower than that of the traditional algorithm. This is because the proposed algorithm fully considers the characteristics of the specific data in large database, in the process of detection, first specific data nodes are classified effectively, to reduce the complexity of detection, so as to improve the detection accuracy and detection efficiency.

4 Conclusion

Aiming at the defects of the traditional algorithm, a specific data detection method based on Apriori algorithm for large databases is proposed. The characteristics of specific data nodes is calculated to provide data base for the detection of specific data. To find out the frequent sets of all the specific data, the frequent sets are obtained strong association rules meet the minimum support degree and the minimum reliability, and then the detection of the specific data in large database is completed. Experimental results show that the detection of specific data of large database can be effectively improved by using the improved algorithm, the accuracy and detection efficiency of the detection are increased as well, and the effect is satisfactory.

References:

[1] Li Yanqing, et al. Several Methods of Accessing Database in Labview [J]. Control and automation, 2006, 22 (1): 131-134.

[2] Xiao Ling, Liu Jihong, Yao Jianchu. The Research and Application on the Distributed Database System [J]. Computer Engineering, 27, 2001 (1): 33-35.

[3] Liu Weiwei, Xu Cheng, Li Renfa. The Mechanism and Applications of Berkley DB [J]. Science technology and engineering, 2005, 5 (2): 86-90.

[4] Yin Renping, Liu Gang, Wang Lixin, et al. Access to access database in LabVIEW [J] Electronic measurement technology, 2006, 29 (3): 51-52.

[5] Wang Guoqing, Tang Yanxia, Sun Xiaoli, et al. Process analysis and endpoint determination of the procedure for prepared radix rehmanniae based on special spectral dataprocessing [J]. Journal of Zhengzhou Institute of light industry: Natural Science Edition, 2009, 24 (5): 1-4.

[6] Wang Hua, Hu Xuegang, Tian Weidong. Mining Algorithm of Maximal Frequent Itemsets Suitable to Specific Database [J]. Computer Engineering, 34, 2008 (14): 63-65.

[7] Liu Yunsheng, Deng Huafeng, Dai Yichen, et al. A generic framework for storing specific streaming data [J]. Journal of Huazhong University of science and technology (Natural Science), 2005, 33:253-256.

[8] Liu Mingji, Wang Xiufeng, Huang Yalou. Data Preprocessing in Data Mining [J]. Computer science, 2000, 27 (4): 54-57.

[9] Luo Ke, Cai Biye, Bu Shengxian, et al. Research of Data Mining and Its Development [J]. Computer engineering and applications, 38, 2002 (14): 182-184.

[10] Bi Fangming, Zhang Yongping. Research of data mining technology [J]. Computer engineering and design, 2004, 25 (12): 2242-2244.