Study on effective detection method for specific data of large database

LI Jin-feng

(Vocational College of DongYing, Shandong 257091, China)

Keywords: data detection; database; mapping;

Abstract: in the process of detecting specific data of large database, when the traditional detection method is utilized for detecting specific data, it is vulnerable for interference of mass information, which makes the specific data detection process time-consuming, and of low efficiency. For this, an effective detection method for specific data of large database is proposed based on improved TFIDF algorithm, the information entropy between the specific data features of large database and the information entropy within the features are viewed as the weighted factor for specific data detection, nonlinear mapping ability of neural network is adopted to achieve calculation of weights and fuzzification of TFIDF algorithm, thus solving the detection problem for specific data of large database. The experimental results show that, improved algorithm for effective detection of specific data in large databases, can effectively reduce time consumed for detection of specific data, ensure the detection quality of specific data to meet customer requirements.

1 Introduction

With the rapid increase of database management technology, specific data detection technology have been widely used in database management of various industries, and plays a more and more important role [1]. Therefore, how to process effective detection for specific data in large database according to the needs of users [2], has become the core problem to research in the field of database management [3]. The current stage, the main detection methods for specific data in large database includes detection method for specific data in large database based on improved support vector machine algorithm [4, 5], detection method for specific data in large database based on Gauss's model [6] and detection method for specific data in large database based on fuzzy clustering algorithm [7]. One of the most commonly used is detection method for specific data in large database based on improved support vector machine algorithm [8]. Because detection methods for specific data in large database play an irreplaceable role in the field of database management, therefore, has a broad prospect for developing [9, 10], and become the focus studied by many experts.

2 Principle of detection method for specific data in large database based on improved TFIDF algorithm

2.1 description of improved TFIDF algorithm

With the given probability distribution $P = (p_1, p_1, ..., p_n)$ of data in large database, the information entropy of distribution transfer is defined as:

$$H(P) = -\sum_{i=1}^{n} p_i \bullet \log_2 p_i \qquad (1)$$

In the specific data collection D of a large databases, according to data types are divided into k class, denoted as $C_1, C_2, ..., C_k$, defines the probability distribution containing features t as $P = (n_1 / N, n_2 / N, ..., n_k / N)$.

In the detection process of specific data in large database, higher specific data distribution uniformity index of a characteristic item contained in large database, shows that the distribution entropy H_{ac} between each class after specific data detection is bigger, namely the system

contribution is small. Therefore, the improved TFIDF method, based on the traditional method, combines the distribution entropy of each class and separate information entropy within each class, to form the new influence parameter, new information entropy factor between classes is defined as:

$$\begin{cases} a(H_{ac}) = 1 - \dfrac{H_{ac}}{\max(H_{ac}) + l} \\ \max(H_{ac}) = \log_2 k \end{cases} \qquad (2)$$

Wherein: $\max(H_{ac})$ is the maximum of each feature class information entropy of correlated feature item after extracting the feature of specific data, k is the class number of data in large database, l is a constant coefficient.

Thus, based on the information entropy distribution of specific data feature items between class and within class in large-scale database, TFIDF weighting method is defined as:

$$\begin{cases} W_{ik}(d) = \dfrac{IDF_1}{IDF_{const}} \times a(H_{ac}) \\ IDF_1 = tf_{ik}(d) \times \log(\dfrac{N}{n_k} + 0.01) \\ IDF_{const} = \sqrt{\sum_{i=1}^{n} (tf_{ik}(d))^2 \times [\log(\dfrac{N}{n_k} + 0.01)]^2} \end{cases} \qquad (3)$$

The definition is optimized with the above formula, contribution of classifying of characteristics of specific data in large database can be shown obviously.

2.2 weight calculation based on neural network

The improved TFIDF method after weighting works better for considering detailed distribution situation of each specific data characteristics in large database, when specific data quantity is very low, the manual way is adopted to calculate weights of each specific data set, while, when specific data quantity in large databases is enormous, manual calculation is not realistic, thus, it is necessary to adopt a new data fusion processing method for large amount of data to calculate the weight.

BP neural network is a multilayer feedforward neural network according to the error back-propagation algorithm training, is one of the most widely used neural network models. BP network can learn and store a lot of input and output scheme mapping relation, without revealing the mathematical equations to describe this mapping relation in advance. Its learning rule is based on the method of steepest descent, constantly adjust the network weight of network. BP neural network model topology structure includes input layer, hidden layer and output layer.

The basic idea of BP neural network is to obtain a set of optimal weights after training with a large number of data, for utilization afterwards, for a set of specific information, according to the optimal weights which are trained before, the output information of prediction is obtained. BP neural network have great input and output nonlinear mapping relationship, can achieve self-learning and update, which isn't applicable for the probability and statistics method.

The relationship between input and output is defined as:

$$y_j = \sum_{i=1}^{n} x_i * a_{ij} + \delta 1 \qquad (4)$$

$$z_k = \sum_{j=1}^{m} y_j * b_{jk} + \delta 2 \qquad (5)$$

Among them:

$x_i (i = 1, 2, , n)$ —The element of input layer, n is the number of elements of output layer;

$a_{ij} (i = 1, 2, , n; j = 1, 2, , m)$ —The weighted value of input layer to the hidden layer;

m —the number of elements of hidden layer;

$\delta 1$—The threshold value of input layer to the hidden layer;

$y_j (j = 1, 2, , m)$ —The element value of hidden layer;

$b_{jk}(j=1,2,,m;k=1,2,,p)$—The weighted value of hidden layer to the output layer;

p—Numbers of hidden layer elements;

$z_k(k=1,2,,p)$—The element value of output;

$\delta 2$—The threshold value of hidden layer to the output layer.

BP neural network is used for weight calculation, through input large amount of training data of large database sample distribution to the neural network to obtain the collection of weight feature. Then specific data feature database is established, afterwards, through the weighted calculation for new data, the results is compared to specific data feature database, if it is consistent with the characteristic, appropriate weight will be given.

In this paper, BP neural network is used to calculate weight, with the following advantages:

(1) the algorithm has strong anti-jamming ability;

(2) self-adaptive to various environments;

(3) do not need tedious statistical model.

3 Experiment results and analysis

In order to verify the effectiveness of improved algorithms, there is the need for an experiment, the experimental environment is Visual C++6.0. For the experiment conducted with large database, the number of all data contained is P, the species number of all special data is p, all special data collection is $\{b_1,b_2,\cdots,b_p\}$, the data set constituted of all special data attributes is $\{c_1,c_2,\cdots,c_p\}$, the probability of specific data b_j belonging to attribute c_k is λ.

The following formula was employed to calculate the time consumed by specific data detection in large database:

$$\varphi=\frac{\sqrt{b_j-b_{j-1}^2}}{|c_k-1|} \qquad (6)$$

Specific data detection consuming time in large database, is an important index to measure the specific data detection method.

The number of all data samples in large database was 1000, all species of data is 15.

The data sample data and attribute of the large database is organized and analyzed, so as to obtain the table as shown below:

Table 1 experimental samples data

NO.	Attribute	Quantity of data(million)
1	Residential building	1.5
2	Dwelling area	2.2
3	Property	3.4
4	Warm oneself	2.8
5	Water supply	4.2
6	Power supply	1.9
7	Gas	2.8
8	Pension	2.7
9	Medical care	3.4
10	Community	3.6
11	Books	4.4
12	Library	5.8
13	Supermarket	4.1
14	Shopping	3.3
15	Tourism	5.7

When the sample complexity was high, the traditional algorithm and the improved algorithm were adopted separately to detect specific data in large database, the detection results can be

described in the following table:

Table 2 experimental results when sample complexity is high

The number of experiments	**time consuming of the traditional algorithm（ms）**	**time consuming of the improved algorithm（ms）**
1	65	33
2	66	34
3	67	33
4	66	36
5	54	41
6	66	44
7	63	39
8	58	38
9	69	45
10	71	42

According to above table, it can be learnt that by using the improved algorithm for specific data detection in large database, can avoid the defects of the traditional algorithm, therefore, improves the accuracy of detection for specific data in large database.

4 Conclusion

Aiming at the problem happened in the process of detecting specific data of large database, when the traditional detection method is utilized for detecting specific data, it is vulnerable for interference of mass information, which makes the specific data detection process time-consuming, and of low efficiency. For this, an effective detection method for specific data of large database is proposed based on improved TFIDF algorithm, the information entropy between the specific data features of large database and the information entropy within the features are viewed as the weighted factor for specific data detection, nonlinear mapping ability of neural network is adopted to achieve calculation of weights and fuzzification of TFIDF algorithm, thus solving the detection problem for specific data of large database. The experimental results show that, improved algorithm for effective detection of specific data in large databases, can effectively reduce time consumed for detection of specific data, ensure the detection quality of specific data to meet customer requirements.

References

[1] Ren Li’an, He Qing. A new classification method of mass data [J]. Computer engineering and applications, 2002.14:58-60.

[2] Fan Bo, Li Haigang, Guo Qiong. Research on spatial data classification for customer segmentation [J]. Journal of systems engineering, 2008.1:87-95.

[3] Wu Fei, [J]. Classification method of real network data and anonymous address. Journal of Changjiang University natural sciences: Polytechnic volume, based on 2008.1:244-246.

[4] Liu Hongyan, Chen Jian, Chen Guoqing. Review on data classification algorithm in data mining [J]. Journal of Tsinghua University: Natural Science Edition, 2002.6:727-730.

[5] Ni Xianjun. Research on the teaching method based on data mining classification technology [J]. Science technology and engineering, 2006.4: 390-393.

[6] Li Guangda, Chang Chun, Zheng Huaiguo, Tan Cuiping, Zhao Jingjuan. Study on the classification and search method of agricultural scientific data of foreign network [J]. Anhui Agricultural Sciences, 2010.20:10998-10999.

[7] Gao Hongbing, Li Fengbin, Wang Jin, Liu Yu. SCS data classification conversion into Shape document based on VBA [J]. Modern surveying and mapping, 2007.4:37-39.

[8] Wang Defen, Gao Jianqiang, Li Li. Research on [J]. Data classification based on median PCA and weighted PCA, information technology, 2014.2:14-18.

[9] Zhang Haifeng. The input method of effectiveness sequence the multi-level classification data in

Excel [J]. Application of computer system, 2006.8:68-70.
[10] Hu Zhaoling, Li Haiquan. Extraction and classification of SAR image texture feature [J]. Journal of China University of Mining and Technology, 2009, 38 (3): 422-427.