Multilevel Index Algorithm Based on Improved Association Rule Mining

J. H. DUAN and M. YUAN

ABSTRACT

Association rule mining is an important research content in data analysis. At the same time it is also the key technology of database index, proposed an improved association rule mining algorithm. Using multi index database frequent item sets of association rules produce the desired set of features, the feature sub space narrowing of candidate item sets in scale compression method. Based on the repeated scanning of index database, index node according to the minimum support threshold filtering rules and constraints, depth first traversal of the index database frequent item list, according to the data distribution structure rearrangement association rules list, improve the efficiency of data index. Simulation results show that using the method of text information database index, it has better accuracy in the mining of association rules. It can improve the retrieval precision and recall, and the construction of the database access. It has very good application value.

KEYWORDS
Association rule; database; index; data mining; frequent item.

INTRODUCTION

With the expansion of the development of computer network technology and the demand for information storage, database access and security index problem attracted much attention, improve information retrieval ability by optimizing the index of text database, in order to make the database resources can effectively be utilized, rational management and development of database in information flow, to optimize the retrieval the efficiency of database optimization. The index is the key technology to realize intelligent text data storage and transfer data, network information has characteristics of large text association rules, resulting in data structure class interference and influence accuracy. The current ASP WEB database index commonly used database with ACCESS and SQL, the general virtual host the user or site users are using ACCESS as their site database. By adding script commands in the HTML page, can create a HTML user interface, experimental optimization of the database, research database optimization index algorithm has been widespread attention[1].

In the text retrieval process, text classification based on set of known categories, the process of judging category according to the content of the text, it is one of the key technologies of information retrieval and data mining. In order to make the computer...
understand natural language, commonly used vector space model (VSM) to express the text. The original text of high-dimensional vector space not only increased the time complexity of text classification, also has certain influence on the result of classification, so it needs to introduce the characteristics of dimensionality reduction of feature dimensionality reduction technique, reduce the complexity of database index and computational overhead in the traditional method, index technology of database network text mainly uses the virtual data distribution database reorganization index algorithm and support vector machine algorithm, scheduling algorithm and data fusion algorithm of phase space reconstruction. The recombinant distribution index algorithm through query expansion and information expanding to realize the optimization of information retrieval, the main idea is based on the original query as the foundation, through a variety of methods, to find the words or phrases related to the original query, this implementation is based on the text features of the database and optimize the index, however, the traditional method of text information network database index, makes text database query expansion will appear in the clustering feature, database indexing precision and efficiency is good.

In this regard, this paper proposes an improved association rule mining algorithm using multi index database, frequent item sets of association rules produce the desired set of features, the feature sub space narrowing of candidate item sets scale compression method based on repeated scan of the index database, an index node according to the minimum support threshold and filtering rules the constraint conditions, according to the association rules list data structure to achieve the distribution of rearrangement, improve the efficiency of the index. Finally, simulation tests show the superior performance of the proposed method.

CORRELATION CONCEPTS AND DATABASE ASSOCIATION RULES

FREQUENT ITEM SETS

Text information database association rules

Relationship between the pattern of a text database for R (A1, A2, Am), a relationship which is set to R, wherein, A1, A2, Am, M is the attribute name, the number of attributes, U={A1, A2, Am}, association rules is a form of implicit data are as follows the relationship between X, Y is X→Y, X⊆U, Y⊆U, and X∩Y=Φ, m is called the number. The number of records to support the project in X is denoted by Count (X). The total number of records in R is |R|, the degree of support for the project in X: Count (X) /|R| remember, Sup (X).

Definition 1: Support (number of degrees) of association rules X→Y. The number of records to support the project in X∪Y, X→Y in R called the number of support association rules from X to Y, denoted by Count(X→Y). The degree of support for X→Y rules: Count(X→Y)/|R|, remember Sup (X, Y).

Definition 2: Defines the confidence of the 2 association rule X→Y. The confidence of rule X→Y is defined as: Count(X→Y)/ Count(X), denoted as Conf(X→Y).

In order to dig out the valuable association rules, must be given two thresholds. The minimum support and minimum confidence (minsup) (minconf.Minsup) said that a group of data sets needed to meet the minimum requirements in the statistical sense, minconf reflects the user the low confidence of association rules.
The task of association rule mining is given in the table R, frequently found all association rules in R. The association rule refers to the rules of support, confidence of not less than the user specified minimum support and confidence.

In association rule mining algorithm, the Apriori algorithm proposed by Agrawal (including AprioriTid and AprioriHybrid algorithm) is the most famous, it is one of the most influential and the most commonly used association rule mining algorithm, the basic idea is to mining association rules is divided into two steps as follows:

The first step: find all frequent item support is not less than the user specified minimum support threshold from the relationship between R sets in. Support is not less than the user specified minimum support threshold set of items referred to as frequent item sets.

The second step: using frequent item sets to generate desired association rules, the basic principle of generating association rules is that their confidence is not less than the minimum confidence threshold specified by the user.

**Frequent item sets extraction from database index**

The frequent item sets of association rules is used to produce the desired feature set, reflecting the text stored in the database of mutual information of two random variables correlated variables in text classification reflects the mutual dependency relation between features and categories \( t \) and \( C_i \), mutual information feature \( t \) and the category \( C_i \) is calculated as follows:

\[
MI(t,C_i) = \sum_{i=1}^{\phi} \log \frac{p(t,C_i)}{p(t)p(C_i)} = \sum_{i=1}^{\phi} \log \frac{p(t|C_i)}{p(t)} \tag{1}
\]

Wherein, \( p(t) \) is the document frequency of the word \( t \), \( p(t,C_i) \) is the probability of feature words \( t \) and categories \( C_i \) appear at the same time, \( p(C_i) \) is appear probability of categories \( C_i \), \( p(t|C_i) \) is categories \( C_i \) characteristic, if feature \( t \) in the category \( C_i \) has the greater probability in other categories, other categories has small mutual information, mutual information \( C_i \) and category \( MI(t,C_i) \) showed that the correlation characteristics of \( t \) and \( C_i \) categories is stronger, more conducive to the identification of category \( C_i \). If the characteristics of \( t \) and \( C_i \) categories are independent of each other, namely \( p(t,C_i) = p(t)p(C_i) \), \( t \) and \( C_i \) categories without relevance.

Let \( X \) and \( Y \) is a subset of \( U \), if the \( R(A1, A2, ......, Am) \) any possible relationship between \( R \), there are two records of the \( X \) property on the same value may not be \( R \), in the multi category labels, mutual information feature of \( t \) and each class is weighted as global the weights of features of \( t \), it can be expressed as:

\[
MI(t) = \sum_{i=1}^{\phi} p(C_i) \log \frac{p(t,C_i)}{p(t)p(C_i)} = \sum_{i=1}^{\phi} p(C_i) \log \frac{p(t|C_i)}{p(t)} \tag{2}
\]
Where \( r \) represents the total number of categories, and \( MI(t) \) represents the feature \( t \) in the whole text set, the classification discrimination capability of \( MI(t) \) is higher, and the \( n \) feature words consist of a set of feature vectors.

A group of \( U \) attribute dependence among attributes of the data set is \( F, R = \{ r_1, r_2, \ldots, r_n \} \), where \( n \) is the number of records, write \( \text{mimX} = \text{minsup} \times n \). database index thus obtained frequent item sets. Resource scheduling and text indexing in the database text management information database, the computing resources the physical resources, knowledge resources, data resources and logic resources semantic fusion, \( \{(s_1, a_1), (s_2, a_2), \ldots, (s_n, a_n)\} \) is two dimensional set of functions describing the semantic features of text database management information access\(^5\), frequent project database index set:

\[
(\vec{x}, \vec{a}) = \phi((s_1, a_1), (s_2, a_2), \ldots, (s_n, a_n))
\]
\[
= \Delta \left( \sum_{j=1}^{n} \omega_j \Delta^{-1}(s_j, a_j) \right)
\]

(3)

Wherein, \( \sum_{j=1}^{n} \omega_j = 1, \vec{x} \in S, \vec{a} \in [-0.5, 0.5] \).

OPTIMIZATION OF DATABASE INDEX ALGORITHM BASED ON ASSOCIATION RULES

Mining association rules

This paper proposes an improved association rule mining multilayer database indexing algorithm, analyze the mapping of semantic relevance in text database management information base, construct the semantic ontology model to reflect the text information index database, \( \{(s_1, a_1), (s_2, a_2), \ldots, (s_n, a_n)\} \) data distribution node semantic features of index nodes based on the minimum support threshold filtering rules and constraints,\(^6\) in a class of probabilistic feature words are equal, mutual information feature selection algorithms tend to choose the rare words, but the classification of text classification of high frequency characteristics are more conducive to the text. In addition, when \( p(t \mid C_i) > p(t) , MI(t,C_i) > 0 \), text \( t \) may belong to class \( C_i \), two were positive; when \( p(t \mid C_i) < p(t) , MI(t,C_i) < 0 \), text \( t \) may not belong to class \( C_i \), two for the negative correlation.

In order to eliminate the positive and negative correlation between reduction effect on feature words and reflects the positive, important degree of negative impact, \( I \), was introduced to quantify the weight difference factor:

\[
\alpha = \frac{p(t \mid C_i) - p(t)}{p(t)}
\]

(4)

The formula of adjusted mutual information is expressed as follows:

\[
MI(t) = \sum_{i=1}^{r} \alpha \times p(C_i) \log \frac{p(t \mid C_i)}{p(t)}
\]

(5)
While \( p(t \mid C_i) > p(t) \), \( \log \frac{p(t \mid C_i)}{p(t)} > 0 \), \( \alpha > 0 \), \( MI(t, C_i) > 0 \), it reflects the degree of the positive correlation between the feature words and the category; While \( p(t \mid C_i) < p(t) \), \( \log \frac{p(t \mid C_i)}{p(t)} < 0 \), \( \alpha < 0 \), \( MI(t, C_i) > 0 \). The reaction characteristics and categories negatively correlated with the degree of \( \alpha \) not only differences in the degree of probability and probability of global response characteristics in the \( t \) class in \( C_i \), and eliminate the influence of positive and negative factors related to each other in the subtractive feature weight weighted\(^7\).

Based on repeated scanning of index database, index node according to the minimum support threshold filtering rules and constraints, the distribution characteristics in the within class and between class were used within \( D_{ac} \) dispen and \( D_{ac} \) divergence between class said:

\[
D_{ac} = \sqrt{\frac{1}{m} \sum_{j=1}^{m} (df_j(t) - \overline{df_j(t)})^2} \quad \quad (6)
\]

\[
D_{ac} = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n-1} (df_i(t) - \overline{df_i(t)})^2} \quad \quad (7)
\]

Where \( df_j(t) \) denotes the probability in the \( C_i \) class feature \( t \), \( \overline{df_j(t)} \) said \( t \) feature in the \( C_i \) class average frequency, \( df(t) \) said the feature word in the average frequency of each class, \( m \) is the total number of documents in the \( C_i \). According to the above analysis, the depth first traversal of the index database frequent item list data is used, according to the distribution structure rearrangement, association rules list mining association rule is obtained.

**Data multi-layer index implementation**

Considering the distribution of the class, the class association factor \( \beta \) is defined as follows:

\[
\beta = 1 - D_{ac} + D_{ac} \quad \quad (8)
\]

Based on the above factors, the association rule mining model is improved as:

\[
MI(t) = \sum_{i=1}^{C} \alpha \times p(C_i) \log \frac{p(t \mid C_i)}{p(t)} + \beta \quad \quad (9)
\]
Class association factor $\beta$ distribution characteristics of both within class and between class and within class scatter is small, between class scatter is larger, the higher the proportion of global features, the classifying ability is stronger. The depth first traversal of the index database frequent item list, For the limited bandwidth center information through pre derived entity related knowledge base, using the characterization of related knowledge effectively guide the network information text database resources integration and reorganization, constructing a solid model representation as:

$$Q_{rr}(r) = \frac{1}{N-r} \sum_{i=1}^{N-r} (x_i - x_{r,i})^2$$

(10)

In the formula, $N$ is a heterogeneous data sequence formed by semantic information. It is assumed that the association rule item $q_i$ has $q_i$ nearest words $d_i = (d_{i1}, d_{i2}, ..., d_{in})$:

$$W(q_i, d_i) = \frac{g(q_i, d_i) \times \log_2[f(d_i) + 1]}{\sum_{j=1}^{m} [g(q_i, d_j) \times \log_2[f(d_j) + 1]]}$$

(11)

The index distributed database system predicts the sequence of time $t_0$ has the optimal set of syndrome solution, then there is a normal number:

$$\xi = 2\rho_{max} \lambda_{max}(Q)N\delta^2KT$$

(12)

The $A = \{a_1, a_2, ..., a_n\}$ stream preprocessing attribute set for training sequence database, $B = \{b_1, b_2, ..., b_m\}$ query attribute set, the $\{c_1, c_2, ..., c_k\}$ is property value of $a_i$. Association rules data structure rearrangement process, $C(r)$ is larger, it means that $x(t)$ and $x(t+r)$ are more similar, the data matching effect is good, good effect of database index, further the design of system function of Fuzzy subtractive clustering transfer operator $h_i(t)$, $n_m(t)$ is query interference for the flow of information, the first generation of discrete subtractive clustering:

$$X_n(t) = X(t) * h_i(t) + n_m(t)$$

(13)

When the information flow subtraction clustering, multi-level clustering, central convolution calculation, the corresponding $r(t)$ is:

$$r(t) = S(t) * \delta(t) * \sum_{i=1}^{M} \delta_i(t) + \sum_{i=1}^{N} n_i(t)$$

$$= MS(t) + \sum_{i=1}^{N} n_i(t)$$

(14)
For the multi index iteration process, the association rules and the data structure distribution rearrangement are obtained:

\[
\hat{h}(t) = f(d_i) \sum_{i=1}^{M} h_i(t) \ast h_i(-t)
\]  

Wherein, \( f(d_i) \) is the degree of the word \( d_i \), the denominator is the normalization factor, and the logarithm of the degree \( f(d_i) \) is the product of balancing the product of \( g(q,d_i) \):

\[
X(t) \ast X(-t) \equiv \delta(t)
\]

In order to traverse the parameters in the estimation of the mixing matrix, the index node split each time, the two split node versions in the version based on the parent node 1, visible is to play a leading role in classification, negatively related to play a supporting role in the whole text, feature words \( t \) information is known under the condition of data index based on association rules data index list, improve efficiency.

**EXPERIMENTAL ANALYSIS**

In order to test the performance of the algorithm in optimizing the network text information in the index database, simulation test is taken, data is sampled from the Internet, the hardware design is: 4 ordinary PC machine consists of a cluster of 1 sets of Namenode machines, 3 sets of Datanode machine, the experimental data for 4G data memory set the length of the data to forecast time series data 100200, association rule features a minimum support threshold for the \( \minsup_{1} = 30.0\% \) SFF database, the minimum support threshold is \( \minsup_{2} = 2.0\% \), the maximum number of iterations is \( NP = 30 \), the spatial dimension set thesaurus distribution is 20, while the precision ratio (Precision, P), and recall ratio (Recall, R) is the most direct performance evaluation index data indexes, so the use of recall and precision as the evaluation index, the actual combination often used to test the comprehensive evaluation index F1 value on class effect the evaluation is defined as follows:

\[
F_{i,test} \text{ data(F}_{i}\text{-Measure}) = \frac{2 \times P \times R}{P + R}
\]

All kinds of sample set is \( S_i (i = 1,2,\cdots,9) \), the number of samples is a subset of SDFF 210, 420, 248, 201, 480, 216, 189, 312, 192. All kinds of samples were randomly divided into 3 equal samples \( S_{ij} (j = 1,2,3) \) respectively, a subset of them as test samples, the remaining 2 subset as the training sample set is calculated using 3 fold cross validation method, average results of 3 experiments as the evaluation index. By using this method (IMI-2) and the traditional association rules method (IMI-1) and cross-correlation method (TMI), get the performance comparison of database index results shown in Figure 1.
According to Figure 1, the individual class of the recall and precision of the existence of the phenomenon of the shift, the overall classification performance in order to study in different weight classification results, considering the two aspects of recall and precision influence factors, take the F1 test value of three under the weight of the classification results were examined, as can be seen, adding weight adjustment factor after the data index is much better than the traditional method, and then introduces the category correlation factor data index, recall ratio and performance remained stable, than traditional methods in more than 5 percentage points.

Figure 1. Performance comparison of database indexes.
CONCLUSION

This paper proposes an improved association rule mining algorithm using multi index database, frequent item sets of association rules produce the desired set of features, the feature sub space narrowing of candidate item sets scale compression method based on repeated scan of the index database, an index node according to the minimum support threshold filtering rules and constraints the depth first traversal of the index database, frequent item list data structure according to the distribution of rearrangement of association rules list, improve data indexing efficiency. Simulation results show that using this method in text information database index, it has the better the accuracy of mining association rules, and it can improve the retrieval precision and recall, through the research and implementation of the improved algorithm, it improves the association rules in database indexing. The next step after the index data we will make improvement of classification and information fusion algorithm achieves the combination of feature selection and indexing algorithm, and further improves the comprehensive performance of database construction.

ACKNOWLEDGEMENTS


REFERENCES