Data Mining of the Association Rules Based on the Cloud Database

Zhu Tianxiang
School of Software
Shenyang University of Technology
Shenyang, China
zhutianxiang@gmail.com

Sun Shuhui, Zhang Dan, Liu Xin
School of Software
Shenyang University of Technology
Shenyang, China
sunshuhui@126.com,danzhang6@163.com,lx@northlab.cn

***Abstract*—Cloud computing is the latest trend in IT technical development, the importance of cloud databases has been widely acknowledged. There are numerous data in the cloud database and among these data, much potential and valuable knowledge are implicit. The key point is to discover and pick up the useful knowledge automatically. An association rule is one of the main models in mining out these data, and it mainly focuses on the relationship among different areas in the data. This paper puts forward the basic model of data mining based on association rules in cloud database and introduces corresponding mining algorithms.**

Keywords-Cloud Database; Cloud computing; data mining; association rules

I. INTRODUCTION

Cloud computing is developed on the basis of distributed processing, parallel processing and grid computing, and it is a new method based on the shared architecture[1]. It distributed all the computing tasks on the resource pool that is made from many computers, making sure all the application systems can acquire desired computing power, memory space and software service according to its demand. All computing will be provided to the terminal user by the form of service; all the application software will be shared on the shared cloud as shared resources. In this sense, all the terminal users can get their desirable service as they want and needn't to pay for the software. It is predicated that as it develops overtime, more and more companies and people will save their own data in all the storage cloud, which will make the data mining based on the cloud computing become one of the trends in the future data mining systems.

There are mass data in the cloud database, and among these many of them are potential valuable knowledge. How to pick up and discover such useful knowledge is the key point in database research. The data mining is the process of picking up the hiding and unknown knowledge and regulations, which possess potential values for the decision making from the big amount of the databases [2].

It mainly has several steps: data pre-processing, data alternating, data mining operation, rule expression and evaluation. A data mining system includes: control unit, used to control all the parts in a harmonious way; database interface, used to generate and process database according to the given inquiry requirement; database, used to store and manage relevant knowledge; focus, which refers to the data extent that needs to be inquired; model extracting, which refers to the various data mining algorithms; knowledge evaluation, used to evaluate the extracted conclusion.

The association rule is one of the main model in data mining research; it is used to focus on the relationship among different areas while defining the data, and find the connection that can meet the given threshold value of both degree of confidence and degree of support[3]. The target of the association rules is transaction database. It can be used in many areas, such as sales, analysis of the transaction data, giving valuable information to the purchasing behavior, improving the retail industry, etc. It also can be used in the medical diagnosis, by analyzing various medical diagnosis cases, the symptom and reaction of the disease can be known. Besides, it can be used in the information processing in criminal cases, by analyzing the relationship between criminal means and cases to get more valuable clues and information[4].

II. CLOUD DATABASE

A distributed database is a logical set of the databases at various sites or nodes in a computer network and logically such databases belong to the same system. Different from the traditional distributed database, a cloud database contains isolated as well as shared data; a cloud database can be designed by using different data models, which mainly include the key-value model and relationship model[5].

All data of the key-value model, including the rows and columns and timestamps, are stored in the cells of a table, contents are partitioned by row the rows make up a tablet and the tablet is stored on a server node.

Row key: data are maintained in the lexicographic order on the row key. For a table, a row interval is dynamically partitioned according to the value of the row key and is the basic unit in which load balancing and data distribution are performed. Row keys will be distributed to different data servers.

Column key: Column keys will be grouped into sets of many “column families” and are the basic units in which access control is performed. All data stored in a column family usually belong to the same data type and this means data is compressed at a higher rate. Data can be stored in a column key of the column family.

Timestamp: Each cell contains multiple versions of the same data and these versions are indexed by the

Published by Atlantis Press, Paris, France.

timestamp. Data model for key-value cloud database is as shown in Figure 1:

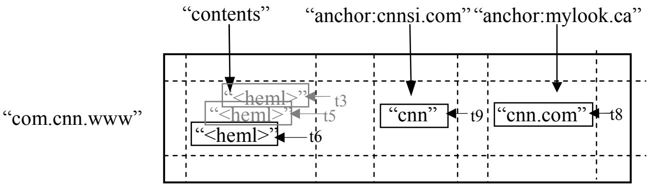

Figure 1. Data model for key-value cloud database

The data model for the relational cloud database involves such relevant terms as row group and table group. A table is a logical relationship and includes a partitioning key which is used for partitioning the table. The set of many tables with the same partitioning key is called a table group. In that table group, the set of rows with the same partitioning key value is called a row group. The rows in that row group are always allocated to the same data node. Each table group contains many row groups, which will be allocated to different data nodes. A data partition contains many row groups, so each data node stores all rows with a certain partitioning key value. The data model for the relational cloud database is as shown in Figure 2:

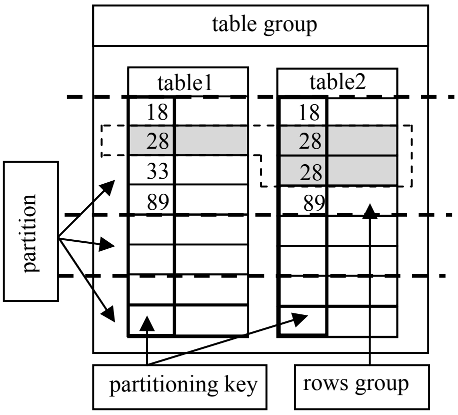

Figure 2. Data model for relational cloud database

III. Basic Model of the Association Rules Based on the Cloud Database

The normal target of the association rules is to discover the data relations in the relationship type cloud database, which has the features of facing the data item set. Through the mining of the association rules, we can discover the relevance of the data.

A. *Concept promoting*

In cloud databases, many property values can be classified, and form concept converging points, all property values and the concepts form a layer structure according to different abstract degrees. Usually, this layer structure is called concept tree. The lower layer' concepts are included in the higher layer's concepts, replacing the lower layer's concepts with the higher layer's concepts is called concept promote.

The knowledge through the data mining mostly appears in the higher layer's concepts, so it is very important to alternate the data with concept promoting. Through the concept promoting, the range of property values can be highly decreased, the data mining algorithm can be simplified, so the data mining effect is largely improved.

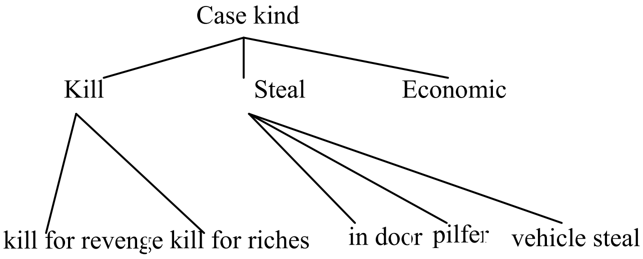

Figure 3. Concept tree of the case kind

B. *The model of the association rules*

In the subject item set, there are some target features in the relationship type cloud database. For instance, the commodity data item set in the commercial behavior analysis {T-shirt, coat, shoes, mild, bread...... }; data item set in the medical diagnosis analysis {hypertension, diabetes...... }.

Classifying item set has the similar features with the subject item set, for instance, customer data item set in the commercial behavior analysis {vocation, gender, age...... }; diagnosis behavior in medical diagnosis and signs and symptoms item set {smoking, polysaccharide, hyperlipidemia...... }.

Sample item set, which has both the features in the subject item set and the transaction data item set in the classifying item set, for instance, transaction data in the commercial activity analysis { {Zhangsan, milk}, {Zhangsan, bread}, {Lisi, T-shirt} ..... }, health check information in the medical diagnosis {{ Zhangsan, smoking, hypertension}, {Lisi, hyperlipidemia, diabetes} ...... }.

Through the mining based on the association rules, we can find that 90% of the customers who buy milk will buy bread; 50% of the patients who have hyperlipidemia have diabetes.

The common objects of the association rules are transaction databases with the characters of subjects oriented item set. In practice, most databases are the relational database, and many applications and the required knowledge are from many different item sets （or multi-item set for simplicity）. For the relational databases, it is difficult to describe the complicated association rules between the multi-item set with models of general association rules, and then we present the association rules model of multi-item set for the relational databases:

Model 1: it is supposed that $R=(r_1,r_2,\ldots, r_n)$ is the rows group in the relationship type cloud database, r_k is one of the rows item set, D is a sample item set relevant to R, and each sample d corresponds to one rows item set, i.e. $d\subseteq R$. Each sample is marked with SID (sample identifier). As for the classifying item set X, only when $X\subseteq d$, the sample X belongs to d. association rules is a formula like $X\subseteq d\Rightarrow Y\subseteq d$, it can be $X\Rightarrow Y$, therein, $X\subseteq R$, $Y\subseteq R$ and $X\cap Y=\Phi$.

The rule $X\Rightarrow Y$ in the sample item set D is constrained by degree of confidence C and degree of support S. Degree of confidence C is defined as that C% in the transaction X in D also contains Y. Degree of support S is defined as transaction $X\cup Y$ accounts for S% in D. Degree of confidence represents the strength of the rule, while Degree of support means the frequency of the model, which is showed in the rule.

In the cloud database of the cases information, 66% of the crime site in the theft cases in factory, so the C is 66%, the theft case and factory case account for 17% of the total cases, so the S is 17%.

The data frequency item set can be defined as the data item set where the degree of support S is over the pre-defined minimum degree of support S. The association rules with high degree of support S and degree of confidence C is strong association rules, otherwise it is called weak association rule Association rules mining means to find the line group that accord to the strong association rules in the database.

The procedure for mining these kinds of association rules of multi-item set as fallow:

(1) Dividing transaction D into several transaction subset $D'=\{D_1',D_2',\ldots D_n'\}$ according to taxonomy item sets.

(2) For all $D_l'<D'$ Do

  Finding the strong sets of the main subject item

  Deriving the association rules using the strong set

(3) Next

These association rules of the multi-item set possess such a feature of only one value available in each sample (SID) set. With this method, mining the data's association rules is applicable for one-to-many relational databases, more practical and expanding the mining range for the association rules.

In practice most of the application and knowledge are from the multiple data item set. For example, we regard a criminal case as the sample item set, for each case there is one mark SID, and in each case there are several suspects, as well as several methods in committing crime, so we can first take the education level of the suspects as one data item set, and the methods of committing crime as another data item set, such as:

TABLE I. THESE TWO CRIMINAL CASES.

	Case kind	***The education level***	***The methods of committing***
SID 1	indoor steal	high school	door lock prize
SID 2	Indoor steal	middle school	door smash

Definition: I is the subject itemset,

There are association rules with several multiple data item sets, the association rules model of multiple data item sets can be termed as:

Model 2: it is supposed that $I=(i_1,i_2,\ldots, i_n)$ is a classifying item set, $J=(j_1,j_2,\ldots,j_m)$ is another one, D is a sample item set, each sample has two classifying item sets $T(T\subseteq I)$ and $T'(T'\subseteq J)$, and each sample is marked with SID, the formula is $X\subseteq I\Rightarrow Y\subseteq J$, degree of confidence C can be termed as that in sample where D contains $X\subseteq I$, C% has $Y\subseteq J$, degree of support S can be defined as transaction with $X\subseteq I$ and $Y\subseteq J$ accounts for S% in D.

IV. MINING ALGORITHM

There are many algorithms in the association rules, and the representative Apriori Algorithm follows the rule that the sub-item sets of all the strong item sets are classified to the strong item sets, while the super item sets of the weak item sets are weak item sets.

The first pass of the algorithm simply counts item occurrences to determine the strong 1-itemsets. A subsequent pass say pass k, consists of two phases. First, the strong item sets L found in the (k-1)th pass are used to generate the candidate item sets C_k, using the apriori-gen function. Next, the database is scanned and the support of candidates in C_k is counted. For fast counting, we need to efficiently determine the candidates in C_k that are contained in a given sample s.

As for the association rules of multiple data item sets, we need to have strong item sets L_l with item 1, and then we can have C_2 from L_l with the item 2, after this we can have L_2, based on this method we can finally have C_k, and get L_k from the database.

Classifying item set D into m classifying item sets D_1, D_2, ... D_m according to the separating item set J, then we can find out the association rules after using Apriori Algorithm to each sub-sample item set D.

```
for(j=1;j<=m;j++) do
 begin
    $L_{j1}$={large 1-items};
    for (k=2;$L_{j,k-1}\neq\Phi$;k++) do
      begin
        $C_k$=apriori-gen($L_{j,k-1}$);
        forall samples $s\in D_j$ do
          begin
            Cs=subset($C_k$,s);
            forall candidates c∈Cs do
               c.count++;
            end
        $L_{j,k}$={c∈$C_k$|c.count>=minsup}
      end
    Answer=∪$_{j,k}$ $L_{j,k}$;
end;
```

$L_{j,1}$ represents the strong item set in D_j sub sample item set, which will generate K item in D_j, scan the database to have $L_{j,k}$, we finally can have D_1,D_2,…,D_m strong item set from the sub sample item set.

Since the Model 2 corresponds to two classifying item sets, and each sample S⊆D includes classifying item set I and J, so 1-itemsets represents that we select one strong item sets from I and J, which is $L_{i,j}$, from $L_{i,j}$ we can have $C_{1,2}$ from $L_{1,2}$, it can be done with the similar manner, and then get $L_{1,k}$. From $L_{1,1}$, we can have $C_{2,1}$ from $L_{2,1}$, the algorithm is:

```
$L_{1,1}$={1-itemsets x∈I, y∈J};
for (i=2;i=n;i++) do
 begin
  for(j=2;j<=m;j++) do
     begin
       $C_{i,j}$=apriori-gen($L_{i,j-1}$);
       forall samples s∈D do
          begin
             Cs=subset($C_{i,j}$,s);
             forall candidates c∈Cs do
                c.count++;
          end
       $L_{i,j}$={c∈$C_{i,j}$|c.count>=minsup}
     end
    Answer=∪$_{i,j}$$L_{i,j}$;
End
```

V. CONCLUSION

The data mining technique is new facing the information society. Many subjects need to be studied in this field. In many professions, a certain amount of databases have been accumulated, in which some hiding knowledge needs to be discovered. The basic study on the data mining technique in this paper has practical meanings.

In management information systems, the relational database is widely used; the connection among the different data is one-to-many and many-to-many, so it is universal to discover knowledge in the database. As the cloud age is coming, the data mining from the cloud data is more practical. The mining method that is used in the association rules is applied to the cloud database, making the association rules more practical and universal. This paper also extends the Apriori Algorithm into association rules mining model, which realize the mining multi-item set association rules.

REFERENCES

[1] Lin Ziyu, Lai Yongxuan, Lin Chen, Xie Yi, Zou Quan. Research on cloud databases. Journal of Software, 2012,23(5):124-137.

[2] Zhu Tianxing, W Peng, Zhang Dan. Application of the data mining technique in case information systems. ICCSEE 2012, 3(1):43-46.

[3] Zhu Tianxiang, Li Li, Xu Zhan Wen, Technology for Mining Classification-Characteristic Rules, Journal of Shenyang Polytechnic University, 1999.(x)x:22-22 .

[4] Feng DG, Zhang M, Zhang Y, Xu Z. Study on cloud computing security. Journal of Software, 2011,22(1):71−83

[5] Xu M, Gao D, Deng C, Luo ZG, Sun SL. Cloud computing boosts business intelligence of telecommunication industry. In: Jaatun MG, Zhao GS, Rong CM, eds. Proc. of the 1st Int'l Conf. on Cloud Computing (CloudCom 2009). Berlin: Springer-Verlag, 2009. 224−231.

[6] Dash D, Kantere V, Ailamaki A. An economic model for self-tuned cloud caching. In: Ioannidis YE, Lee DL, Ng RT, eds. Proc. Of the 25th Int'l Conf. on Data Engineering (ICDE 2009). New York: IEEE Computer Society Press, 2009. 1687−1693.

[7] He Jianjia, Ye Chunming, Wang Xiangbin, Huang Zaixin, Liu Qiuling. Cloud Computing-Oriented Data Mining System Architecture, Application Researche of Computers, 2011,28(4):1372-1374.

[8] Wang Jun, Study on knowledge discovery of databases, Software Research Institute of Chinese Academy of Science, Doctor thesis, 1997.

[9] Graham Cormode, S Muthukrishnan, Summarizing and mining inverse distributions on data streams via dynamic inverse sampling, [C] Proc of the 31st International Conference on VLDB, Trondheim, Norway: VLDB Endowment, 2005: 25-36.

[10] S.K. Gupta,V. Bhatnagar,S.K. Wasan, Architecture for knowledge discovery and knowledge management[J]. Knowledge and Information Systems, 2005,7(3) .

[11] IBM Almaden Research Center, Quest synthetic data generation code [CP/OL].United States: IBM, [2007-01-20], http://www.admaden. ibm.com/cs/projects/iis/hdb/Projects/datat-mining/mining.shtml.

[12] Han J,Conference Tutorial Notes:Data Mining Techniques,In Proceedings of ACM SIGMOD International Conference'96 on Management of Data (SIGMOD'96),Montreal,Canads, 1996.

[13] Fayyed U, Hausller D, Stolorz P, Mining Scientific Data, comm. ACM 39(11), 1996.

[14] Li Li, Xu Zhanwen, Liu Guifang, The Algorithm for Mining Classification Characteristic Rules and Trend Rules, Journal of Chinese Computer Systems, 2000.3.