Global Journals LATEX JournalKaleidoscope™

Artificial Intelligence formulated this projection for compatibility purposes from the original article published at Global Journals. However, this technology is currently in beta. *Therefore, kindly ignore odd layouts, missed formulae, text, tables, or figures.*

Research of Data Mining Algorithm based on Cloud Database

Xia Zhang[1]

[1] henyang University of Technology

Received: 12 December 2013 Accepted: 5 January 2014 Published: 15 January 2014

Abstract

There is an immense amount of data in the cloud database and among these data, much potential and valuable knowledge are implicit. The key point is to discover and pick out the useful knowledge, and to do so automatically. In this paper, the data model of the cloud database is analyzed. Through analyzing and classifying, the common features of the data are extracted to form a feature data set. The relationships among different areas in the data are then analyzed, from which the new knowledge can be found. In the paper, the basic data mining model based on the cloud database is defined, and the discovery algorithm is presented.

Index terms— cloud database, data mining, association rules, classification characteristic

1 Introduction

loud computing is derived from technologies such as distributed processing, parallel processing, grid computing, etc. It is an emerging approach to sharing the infrastructure architecture [1]. It distributes all the computing tasks on the resource pool that is made of many computers, making sure all the application systems can acquire desired computing power, memory space and software service according to their demand **??**2]. All the computing is provided to the terminal user by the form of service, and all the application software in the cloud as shared resources. A cloud database is a database deployed and virtualized in the cloud computing environment. It is predicated that as it develops overtime, more and more people and companies will store all their data in the cloud, which will make data mining based on the cloud computing one of the trends in the future data mining systems [3].

There is a massive amount of data in the cloud database, and among them, lives potentially valuable knowledge. How to discover such useful knowledge is the key point in database research. Data mining is the process of picking out the hidden knowledge and regulations, which possess potential value that could influence decision making [4]. Data mining namely refers to the knowledge discovery from a database and is comprised of the following procedures: data preprocessing, data alternating, data mining operation, rule expression and evaluation [5]. A data mining system includes: control unit -used to control all parts in a harmonious way; database interface -used to generate and process data according to the given query; database -used to store and manage relevant knowledge; focus -refers to the data extent that needs to be inquired; model extracting -refers to the various data mining algorithms; and finally, knowledge evaluation-used to evaluate the extracted conclusion [6].

2 II.

3 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 [7]. 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.

All data of the key-value model, including the rows and columns, 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.

4 a) Row Key

Data is 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 are distributed amongst data servers.

5 b) Column Key

Column keys are 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, which means data is compressed at a higher rate. Data can be stored in a column key of the column family.

6 c) Timestamp

Each cell contains multiple versions of the same data and these versions are indexed by the timestamp. Data model for key-value cloud database is as shown in Fig. **??**:

Figure **??** : Data model for key-value cloud database ”com.cnn.www” ”contents” ”anchor:cnnsi.com” ”anchor:mylook.ca” ”<html>” ”<html>” ”<html>” ”cnn” ”cnn.com”t5 t9 t3 t6 t8

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 are 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 Fig. 2: The normal target of the association rules is to discover the data relations among the data item set in the relationship type cloud database. Through mining based on the association rules, we can discover the relevance of the data.

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, milk, 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 bought milk also bought bread; 50% of the patients who have hyperlipidemia also have diabetes.

The common targets of the association rules are transaction databases with the characters of subjects oriented item set. In practice, most databases are relational, and many applications and the required knowledge are from many different item sets(or multiitem set for simplicity). For relational databases, it is difficult to describe the complicated association rules between the multi-item set with models of general association rules. We present the association rules model of multi-item set for the relational databases: Definition 1: I is the subject item set, J is the taxonomy item set, each transaction corresponds to a subset T of the subject item sets and a taxonomy item U of the taxonomy item sets, called T belonging to class U.

Model 1: it is supposed that R=(r1,r2,?, rn) is the rows group in the relational cloud database, rk 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?R. Each sample is marked with SID (sample identifier). As for the classifying item set X, only when X?d, the sample X belongs to d. association rules is a formula like X?d?Y?d, it can be X?Y, therein, X?R, Y?R and X?Y=?.

The rule X?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 C% in the transaction X in D also contains Y. Degree of support S is defined as transaction X?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 shown in the rule.

In the cloud database containing cases information, 66% of the crime site in the theft cases happened in factories, so the C is 66%. Theft cases and factory cases 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 considered strong association rules, otherwise it is considered weak association rules. 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 is as follows:

1. Divide transaction D into several transaction subset D’={D1’,D2’,?Dn’} according to taxonomy item sets.

7 For all D1'<D' Do

Find the strong sets of the main subject item Derive the association rules using the strong set 3. Next These association rules of the multi-item set possess a feature where only one value is available in each sample (SID) set. With this method, mining the data's association rules is applicable for one-to-many relational databases. This is more practical and expands the mining range for the association rules.

In practice, most of the applications and knowledge is 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, several suspects, as well as several methods by which the crime is committed. 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.

There are association rules with several multiitem sets, the association rules model can be termed as:

Model 2: It is supposed that I=(i 1 ,i 2 ,?, i n) is a classifying item set, J=(j 1 ,j 2 ,?,j m) is another one, D is a sample item set, each sample has two classifying item sets T(T?I) and T'(T'?J), and each sample is marked with SID. The formula is X?I?Y?J, degree of confidence C can be termed as that in sample where D contains X?I, C% has Y?J, degree of support S can be defined as transaction with X?I and Y?J accounts for S% in D.

8 b) 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, 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 Ck, using the apriori-gen function. Next, the database is scanned and the support of candidates in Ck is counted. For fast counting, we need to efficiently determine the candidates in Ck that are contained in a given samples.

As for the association rules of multiple data item sets, we need to have strong item sets L1 with item 1, and then we can have C2 from L1 with the item 2, after this we can have L2, based on this method we can finally have Ck, and get Lk from the database.

Classifying item set D into m classifying item sets D1, D2, ... Dm 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. Since Model 2 corresponds to two classifying item sets and each sample S?D includes classifying item set I and J, 1-itemsets represent the strong item sets we select from I and J, which is Li,j. From Li,j we can have C1,2 from L1,2, done with the similar manner, and then get L1,k. From L1,1, we can have C2,1 from L2,1, the algorithm is: In management information systems, the relational database is widely used; the connection among different data is one-to-many and many-tomany, so it is universal to discover knowledge in the database. As the cloud age is coming, 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 L 1,

9 Data Mining for Classification Characteristic Rule

Knowledge discovered from a database with massive data is diversified in variety. Knowledge classification refers to clustering or classifying tuples in the database to divide these tuples into different categories by characteristic rules extracted from a certain target class, and thus achieve the purpose of describing the characteristics of the tuples of that class.

Clustering refers to categorizing a group of individuals into several categories, which means those with the same characteristic are classified as one category. Clustering is a process in which a data object with multiple attributes is continuously classified. In such process, classification is automatically executed by the classification algorithm to divide the data into several classes by identifying data features. A relational database mainly containing character information may be equivalently partitioned into equivalence classes according to the concept of equivalence class. The resulting equivalence classes are a group of classes. The characteristics of each class are further analyzed and this can lead to the determination of the classification characteristic rules. Such analysis process is of practical significance. For example, symptoms and reaction characteristics of various diseases can be determined by analyzing a great amount of medical diagnosis cases.

10 a) Classification Model For Key-Value Model Based

Cloud Database

Let D be a key-value model based cloud database, K represents the set of all row keys in D with the formula K={k1, k2 ,?,kn}, At represents the set of all column keys in D with the formula A ={a1,a2,?,am}, V represents the dataset of certain attribute characters of the column keys with the formula V={v11,v12,?,vmn} and f represents a function of a and k with the formula Vi,j=f(ai,kj). Definition 2: For ?a?A t (A t is the dataset of column keys, A t ?A), if k i ?K, k j ?K, i?j and f(a, ki)= f(a, kj), then ki is said to be equivalent to kj based on the dataset of column key attributes At Example 1: Let D be a key-value model based cloud database, K is the set of all row keys in D, A is the set of all column keys in D and At is a subset of A. V is the dataset of certain attribute characters of the column keys and each data has the latest timestamp. K={ k1, k2, k3, k4, k5, k6, k7, k8, k9, k10, k11, k12 } A ={a1,a2,a3} At ={a1} V={ v10,v11, v12,v20, v21,v30, v31,v32} The values of f(k,a) are as shown in Table 1. In the above mentioned database, the field {a1} in the column key set At can be classified into three classes: K1 = {k1, k4, k5, k8, k9, k11}?K2 = {k2,k7,k10,k12}? K3={k3,k6}. {K1,K2,K3} is a class based on the column key set At, the name of that class is {v11?v10?v12} and its classification support degree is {50%?33.33%?16.67%}.

11 b) Classification Model for Relational Model Based Cloud Database

Let D be a relational model based cloud database and T be a table group of D, P represents the set of partitioning keys in T with the formula P={ p1, p2,?,pn } and R represents the set of row groups of the partitioning key Pi with the formula R={r1,r2,?,rn}. database and T be a table group of D, S t represents the record count of all row groups, the row group set R is the class based on the partitioning key P and S y represents the record count of Y row groups in R, then S=S y /S t *100% is said to be the classification support degree of the class Y.

Example 2: Let D be a relational model based cloud database and T be a table group of D, P is the partitioning key and the value of P is {p1,p2,p3}, then the corresponding row group is {r1,r2,r3}, namely: P ={p1,p2,p3} R={r1,r2,r3} r1={v11,v12,v13,v14,v15,v16} r2={v21,v22,v23} r3={v31}

The above mentioned database can be partitioned into three classes based on the partitioning key P: r1={v11, v12, v13, v14, v15, v16}?r2={v21,v22,v23}? r4={v31}?the name of the class is {r1,r2,r3} and the classification support degree is {60%?30%?10%}.

For the cloud database D, all classification support degrees {S1,S2,S3,?} can be obtained according to a certain classification R={R1,R2, R3,?}. Definition 7: Let Sp be a given threshold, 0?S p ?1. Those classes with a classification support degree S?S p are called strong class and those with a classification support degree S<S p are called weak classes.

In mining knowledge from massive data, we usually are concerned about and interested in data classes with higher classification support degree, namely the strong classes. Strong classes contain more representative knowledge.

12 c) Classification Characteristic Rule Model

According to the definitions as mentioned above, data in a database can be classified and the characteristics of the strong classes need to be further analyzed.

Definition 8: Let E be a At-based class, A t is the complementary set of A t against the attribute A, A t ?A, B is the subset of A t , the equivalence class T based on B is called the characteristic domain in the class E, and the value {b 1 ,b 2 ,?} of the attribute in B is called the characteristics in the class E. Definition 9: Let ec be the record count of the class E, and tc be the record count of the characteristic domain T, then C= t c /e c *100% is said to be the confidence degree of characteristic.

Definition 10: Let C p be a given threshold, 0?C p ?1, a characteristic domain with the confidence degree of characteristic C?C p is called a strong characteristic domain and a characteristic domain with the confidence degree of characteristic C<C p is called a weak characteristic domain. The value of the field with strong characteristic domain is called a strong characteristic, while the value of the field with weak characteristic domain is called a weak characteristic.

Strong characteristics in a strong class are usually representative knowledge and can be expressed as: (

13 Application of the Classification Characteristic Rules in Case Information Systems

Suppose the related property is the case type, selected site and the way of commit, the related degree of the door smashed versus picked is 0.8, the given threshold of the related degree is 2.5, two cases as shown in Table 2: Based on the above definitions, as long as the related degrees of the related properties are known, the related cases can be discovered. The values of the related degrees are provided by the field experts according to the field knowledge. In order to express the related degrees, a related degree matrix Ma is defined as follows:Ma= ? ? ? ? ? ? ? ? ? ? ? ? ? mn m m n n C C C C C C C C C ? ? ? ? ? ? ? ? 2 1 2 22 21 1 12 11

(1) Cij: related degree of element j to element i of property a. M a is a symmetrical matrix, so only consider the lower triangle.

The related degree matrix of the way of commit is as shown in Table **??**: During case analysis, the related degree matrix of all related properties must be known. The sum of case related degrees can be found based on the related properties and the related degree matrix and all the related cases can be found based on the given threshold Cp.

Input U, an information system, has n records, Ma is a symmetrical matrix, Mak[i,j] seeks the correlation degree in the correlation matrix. Output L, a set of case related to it.

For i=1 To n-1 (n is the record number) For j=i+1 TO n A[i, j] =0 For k=1 TO m A[i,j]=A[i,j]+Mak[i,j] Next If A [i,j]>=Cp uj?{u-u relates to ui } End If Nexts L=?{u-u relates to ui } Next Output L

Following the idea of converging classes, the case information is divided into many kinds with the equivalent dividing method. Define the base value of kinds as classifying the support degrees. The kinds can be divided into strong class ones and weak class ones. The weak class has too small classifying support degrees, no practical meanings and can be neglected. For the strong class, Rough set theory can be used to analyze their common features and form the classifying characteristic regulations.

Data was mined from the database for the experimental criminal case information system by using the aforementioned algorithm. Taking the crime approach table group as an example, the table contains 100762 records. Given a classification support degree threshold of 10%, and a characteristic confidence degree threshold of 20%, 359 classification characteristic rules were mined, for example:

(Residential house, night, 23.4%) (Rubbery, less than RMB10000, 93.3%) VI.

14 Conclusions

The data mining technique is new to 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 hidden knowledge needs to be discovered. Starting with the concept of set theory, the data model for the cloud database was analyzed; the model and algorithm for mining classification characteristic rules from cloud database were designed to make data mining of classification characteristic rule more practical.

The abstracted related knowledge models presented in this paper can be put into practice in the public security affairs, such as case chaining, which is one of the highly demanded, complex tasks in the public security affairs. The presented methods about the related case data mining in this paper promote the work effect of the chained cases. On the case material analysis, the mining of the classifying characteristic regulations help users with their classifying work and overcome the weaknesses that exist in the old statistics method, in which repeated experimentation are required. [1] [2]

2

Figure 1: Figure 2 :

[1] © 2014 Global Journals Inc. (US)

[2] © 2014 Global Journals Inc. (US) Global Journal of Computer Science and Technology

```
for(j=1;j<=m;j++) do
begin
L j,1 ={large 1-items};
for (k=2;L j,k-1 ??;k++) do
begin
C k =apriori-gen(L j,k-1 );
forall samples s?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;
Lj,1
```

Figure 2:

1

A	A1	A2	A3
k 1	v 11	v 20	v 32
k 2	v 10	v 21	v 30
k 3	v 12	v 20	v 30
k 4	v 11	v 21	v 30
k 5	v 11	v 20	v 32
k 6	v 12	v 20	v 30
k 7	v 10	v 21	v 31
k 8	v 11	v 21	v 31
k 9	v 11	v 20	v 32
k 10	v 10	v 21	v 31
k 11	v 11	v 20	v 32
k 12	v 10	v 21	v 31

Figure 3: Table 1 :

E?T?Cp)
E: class
T: characteristic
Cp: confidence degree of characteristic
Discovery Algorithm for Classification
Characteristic Rules.
D is a cloud database and A is the set of classification attributes of D.

```
For all A t For j=1 To k Do
if C i ?C p Then
(E i , T j , C j ) ?result base
Endif
Next
Endif
Next
Next
V.
```

Figure 4:

2

	Case kind	Selected site	Way of commit
Case1	Burglary	residence	door picked
Case2	Burglary	residence	door smashed

Figure 5: Table 2 :

[Zhu and Zhang] 'Application of the data mining technique in case information systems'. T X Zhu , W P Zhang , D . *ICCSEE* 2012 (1) p. .

[He et al. ()] 'Cloud Computing-Oriented Data Mining System Architecture'. J J He , C M Ye , X B Wang , Z X Huang , Q L Liu . *Application Researche of Computers* 2011. 28 (4) p. .

[Abouzeid et al. ()] 'HadoopDB: An architectural hybrid of MapReduce and DBMS technologies for analytical workloads'. A Abouzeid , K Bajda-Pawlikowski , D J Abadi , A Rasin , A Silberschatz . *PVLDB* 2009. 2 (1) p. .

[Jaatun Zhao Rong (ed.) ()] *Proc. of the 1st Int'l Conf. on Cloud Computing*, M G Jaatun, G S Zhao, C M Rong (ed.) (of the 1st Int'l Conf. on Cloud ComputingBerlin) 2009. 2009. Springer-Verlag. p. .

[Lin et al. ()] 'Research on cloud databases'. Z Y Lin , Y X Lai , C Lin , Y Xie , Q Zou . *Journal of Software* 2012. 23 (5) p. .

[Feng et al. ()] 'Study on cloud computing security'. D G Feng , M Zhang , Y Zhang , Z Xu . *Journal of Software* 2011. 22 (1) p. .

[Zhu et al. ()] 'Technology for Mining Classification-Characteristic Rules'. T X Zhu , L Li , Z W Xu . *Journal of Shenyang Polytechnic University* 1999. p. .