International Journal of Grid and Distributed Computing
Vol. 9, No. 4 (2016), pp.87-94
http://dx.doi.org/10.14257/ijgdc.2016.9.4.08

Research of Data-Aiming Mining Algorithm in Cloud Environment

Jiangang Jin

(Software Technology Vocational College, North China University of Water Resources and Electric Power, Zhengzhou 450045, China)
henanjjg@sina.com

Abstract

Cloud computing contains a huge amount of data, which are featured as being widely distributed, heterogeneous, and dynamic. Thus, aiming at how to mine useful parts in these information, this paper proposes an Apriori algorithm based on cloud computing and introduces cost-sensitive learning and non-filter matrix to find k frequency set and uses the method of generating association rules to improve effectiveness of data mining. Simulation experiments show that mining algorithm in this paper is highly effective and suitable for data mining in the context of cloud computing.

Keywords*: Cloud computing; Data mining; Apriori algorithm*

1. Introduction

With the emergence of the concept cloud computing, more and more information is shared and spread through the Internet. How can carry out data mining in cloud computing became an important means of access to information [1]. Literature [2] data mining solutions in the cloud computing environment is proposed, through the cloud computing and cloud computing services, describes the problem solving mechanism for data mining services. Literature [3] proposed a data processing method based on cloud computing, based on mining user browsing preference path. Experiments show that the mining algorithm for large amount of data logs, accuracy and efficiency than ordinary users preferred path mining algorithm based on frequency increased. Literature [4] proposed a new framework for data mining in the field of cloud platforms and cloud computing environments are described mechanism of data mining in the field of service-oriented. Literature [5] is a method of data mining based on cloud computing technology: large data sets and mining tasks on multiple computers in parallel, experiments show that use cloud computing technology to handle large data sets in a cluster, you can significantly improve the efficiency of data mining. Literature [6] for cloud computing resources prediction model based on data mining technology, results showed that this model not only improves the prediction accuracy of cloud computing resources, and reduce the complexity of modeling, improves the efficiency of models, to cloud computing resources provides a new way of building models. Literature [7] proposed data mining services view, cloud services task force has analyzed data mining advantages and challenges, data mining and prediction of cloud point data mining development trend of cloud services, providing data mining services framework for cloud services.

This paper proposes an Apriori algorithm based on cloud computing and introduces cost-sensitive learning and non-filter matrix to find k frequency set and uses the method of generating association rules to improve effectiveness of data mining. Simulation experiments show that mining algorithm in this paper is highly effective and suitable for data mining in the context of cloud computing with certain advantages.

ISSN: 2005-4262 IJGDC

2. Description of Basic Algorithm

Apriori algorithm is an algorithm for Mining Association rules, the algorithm is divided into two steps, first of all looking for solutions to meet frequently in the data set, find the minimum support of collection, and the second was found on the steps in front of the smallest confidence strong association rules. Steps are as follows:

Step1: scan all the record in cloud data base to find frequent item set-1 that meets the minimum support, and denoted it as L_1, and then denote all the C_1 as L_1 through summary;

Step 2: Find all the frequent item set-2 on the basis of C_1, and denote them as L_2. Combine all the L_2 to form C_2;

Step 3: Scan every record in cloud database and get frequent item set k to form the assembly C_k of item k (the algorithm ends based on $C_{k+1} = \varnothing$).

3. Apriori Algorithm based on Cost-Sensitive Non-frequent Filter Matrix

Apriori algorithm has some defects in time complexity as well as space complexity, so if data mining is carried out in the condition of cloud computing, the mining effects of each cloud node are not good. Based on Apriori algorithm, this paper introduces cost-sensitive learning and non-frequency filter matrix to improve this algorithm, and algorithm in this paper is generally described as followed: First, on cloud database collection for structure decision tree and cost sensitive learning, get related property of price, select these related property than in the of maximum of price; second through non-frequency set filter matrix looking for k -frequency set, then generated Apriori algorithm matrix of initial of results; to further generated non-frequency set filter matrix, judge matrix whether exists meet conditions of k -frequency set, if meet is algorithm end, the algorithm in the filter off which of minimum support degrees, to avoid all structure non-frequency set matrix.

3.1 Cost-Sensitive Learning

Cost-sensitive learning (CSL for short) is a diagnostic method considering both classification accuracy in the classification building and the cost of property. Its main use in the process of construction of decision trees, it needed to in order to be in the wrong category costs and seek a balance between cost, trying to find property as a Division with the highest performance properties, reference standard is the wrong category cost value cost ratios, the property value $Cost\_radio(A_i)$ of property is A_i defined as:

$$\cos t\_ratio(A_i) = \frac{Mc - \left(n \times FP - \left(FP * \sum_{i=0}^{r} n_i - FN * \sum_{i=r+1}^{n} p_i \right) \right)}{TestCost(A_i) + 1} \quad (1)$$

Represents the test costs of property A_i, and the denominator refers to the decrease of misclassification costs brought by property A_i. In order to avoid that the denominator is 0, 1 is added here in this paper.

3.2 Seek k –Frequency Set by Using Non-Frequency Filter Matrix

In the process of seeking k –frequency set, in order to ensure that the filter matrix can

have good effects, some records with actual effects should be deleted in cloud database. Methods to construct filter matrix are as follows:

(1) Sequential scanning every cloud database of all records. Each completed in the scanning process a record in generating an intermediate matrix row vector, when the first record in a database appears in the scan, the position vector where it is "1", otherwise referred to as "0 ";
(2) Scan complete unit database values, in the middle of all the construction of matrix rows and column vectors. In order to calculate and fill in each column and the last row vector, followed by calculation of each line and fill in the last column vector.
(3) "Cut" the intermediate matrix, and when you need to look for the k –frequency set, delete the final value of row vector less than k, and the modify the final line of data in the matrix; if the required data are $\min\_\sup port$, the final column vector whose value is less than $\min\_\sup port$ should be deleted, modify the last column of data in the matrix, and constantly repeat the above two processes until the matrix is "cut out".

3.3 Generate Strong Correlation Rules

Strong correlation rules are mainly related to better generate frequent k - predicate set at based on support and confidence of the framework in order to be able to meet frequently to find the predicate term k with the minimum support. Thereby to generate strong association rules, the flowchart shown in Figure 1.

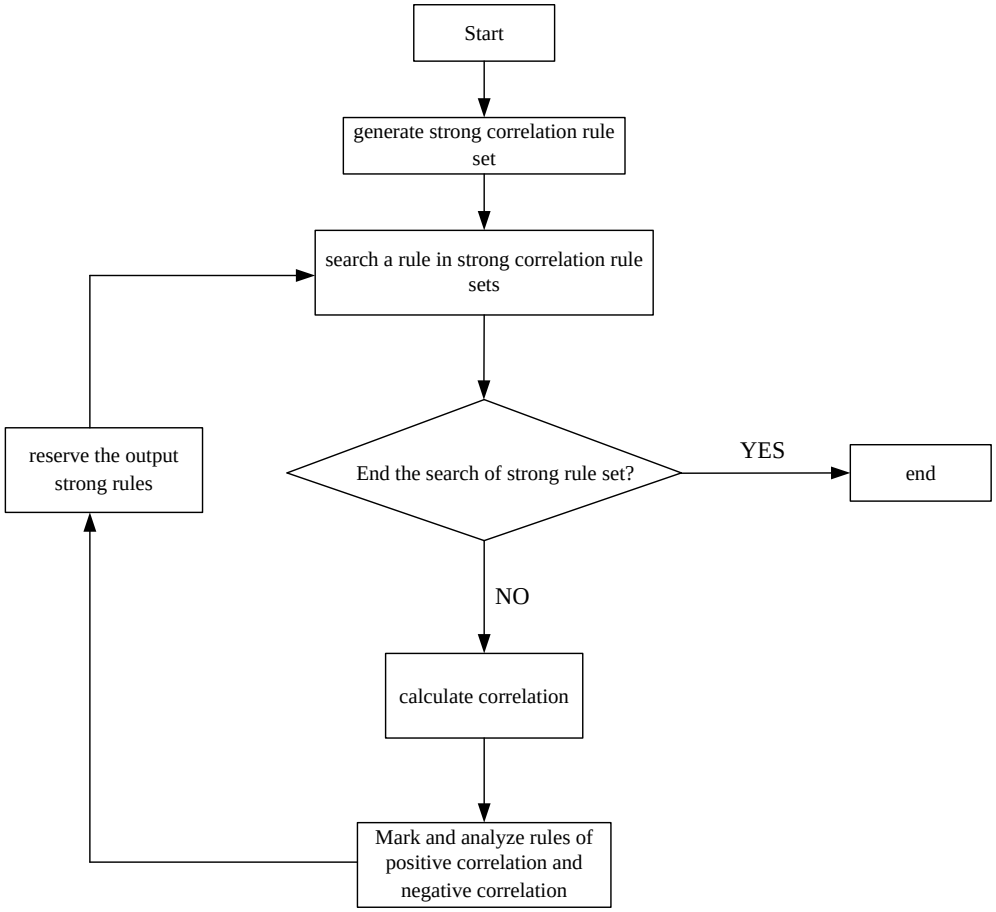

Figure 1. Diagram of Generating Strong Correlation Rules

Step 1: Generate k-frequent set according to non-frequent filter matrix of Apriori algorithm;

Step 2: Select content in the k-frequent set as the result of correlation rules, and set corresponding property of this item as the range to inspect results of this correlation rules.

Step 3: Divide the previous frequent set into two frequent item sets $cause\_set$ and $result\_set$. Calculate the ratio of the occurrence of this k-frequent set's $cause\_set$ and $result\_set$ according to the formula $confidence(x \Rightarrow y) = \frac{\sup port(X \bigcup Y)}{\sup port(X)} \times 100\%$ ( ); if this ratio is greater than the minimum confidence, it indicates that a mining rule is generated.

Step 4: Calculate the occurrence of k-frequent set.

3.4 Generate Non-Frequent Filter Matrix

After the cost-sensitive learning, choose a property of interest, according to association rules, statutes matrix is constructed, the system will check out data tables to generate a frequency filter set matrix. Algorithm is shown in Figure 2.

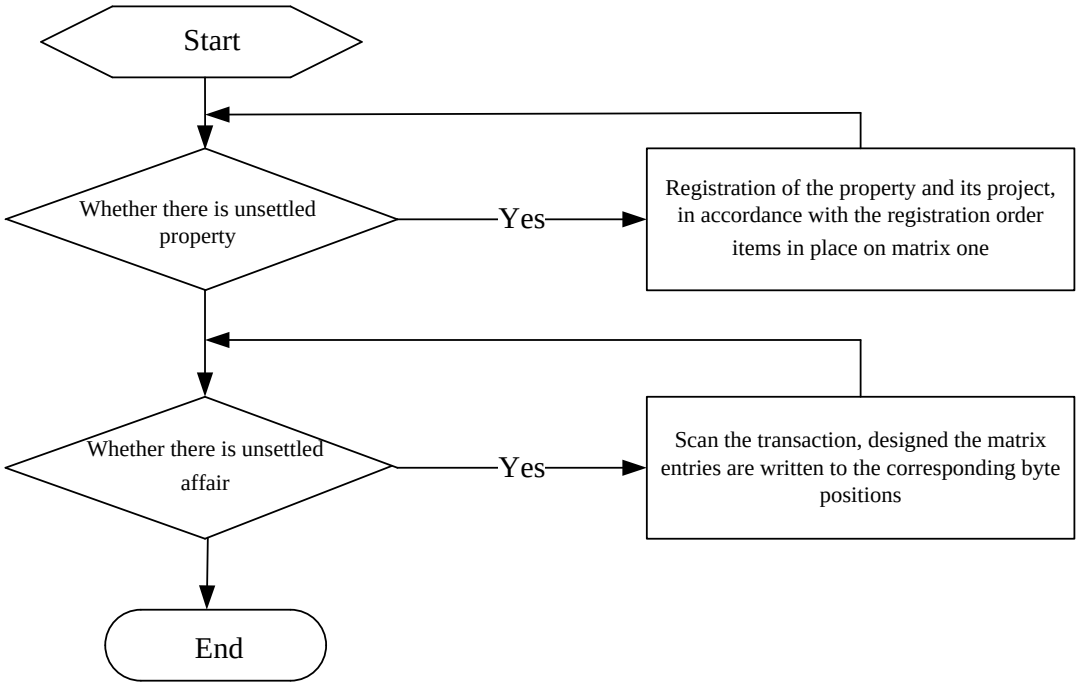

Figure 2. Process to Generate Initial Matrix

4. Data Mining in Cloud Computing

Step1: Generate the non-filter frequency set matrix FF_M


```
FF _ M
min_ sup   k _
Procedure  FF _ M (Mid _ A, min_ sup, k)
{
 FF _ M = null
 Mid _ B = Mid _ A
 While FF _ M ≠ Mid _ A
{  for (i = 1; i < m; i++)
       if (A[i, n] < k) then
         {   delete(Mid _ A, line[i])
             m − −
}
  for (i = 1; i < n; i++)
            if (A[m, i] < min_ sup) then
{   delete(Mid _ A, row[i])
    n − −
    FF _ M = Mid _ A
    Mid _ A = Mid _ B
    Mid _ B = FF _ M
}
    return   FF _ M
}
}
```

Step2: find k_frequency set's algorithm according to non-filter frequency set matrix FF_M(m×n orders)

```
FF _ M (m * n)   k _
Ck = ∅
Desorting (rows, n − 1)
Desorting (lines, m − 1)
flag = ture
procedure   FS _ search(FF _ M , min_ sup, k)
{  if  k = 0 then
        flag =  false
  else {
  Mid _ M == FF _ M
  Lk = ∅
  for (i = 1; i < n; i++)
  for ( j = i + 1; j < n;)
 {Lk = Lk  ⋃ { item(row[i])}
   Delete(Mid _ M , row[i])
   while (|Lk| ≠ k)   and   flag
      FS _ search(Mid _ M , min_ sup, k − 1)
      if   flag = 1  then
     { Ck = Ck  ⋃  Lk
        j++
      }else
      break
       }
       return   Ck
   }
```

5. Analysis of Improved Apriori Algorithm in Cloud Computing

5 PC machines are used for this article (the PC1 to PC5) to build a Hadoop distributed computing platform, where PC1 as Master, run the Jobtracker; the other four machines running Tasktrcker. Machine configuration is as follows: Cpu for Inter Core2.2Ghz, 4GDDR3, 500G hard disk, software environment for Ubuntu12, Hadoop 0.20.3, and OpenSSH.

This paper compared Literature [6] and Literature [7], data source Heritrix [11] open source crawler attack] data, cloud computing data query in 100,000, 300,000, 500,000, 1 million allocated, set 5 cluster nodes, respectively, 100,000, 200,000, 300,000, 400,000. From the comparison of algorithms under different data, different amount of nodes, accuracy aspects are compared, comparison is shown in Figure 1-3.

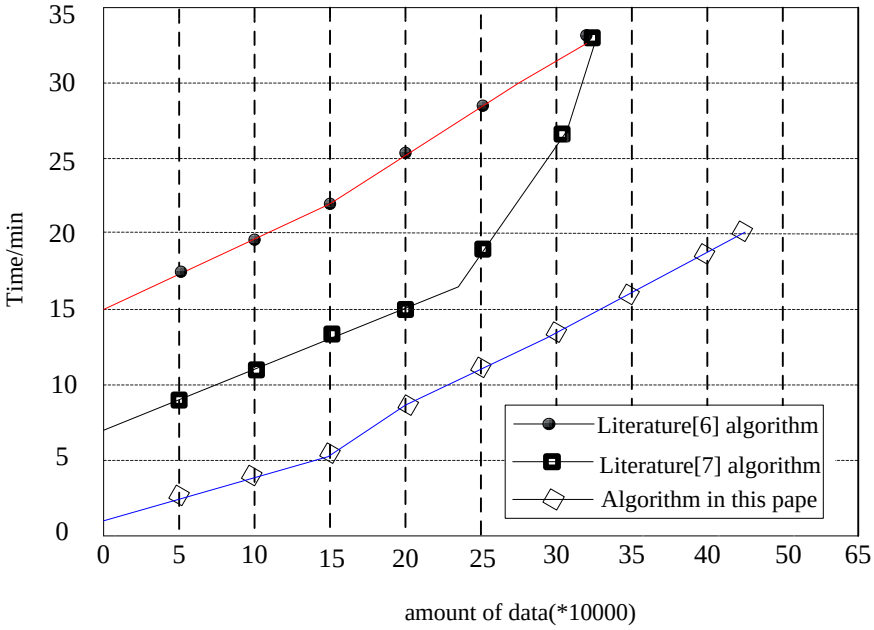

Figure 1. Comparison of Three Algorithms in Different Amount

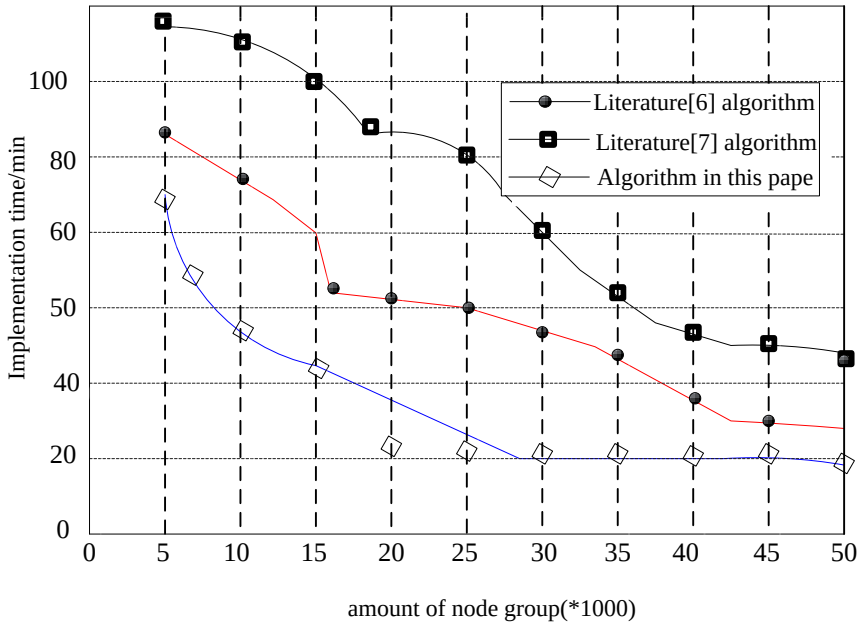

Figure 2. Comparison of Algorithm with Different Amount of Node Groups

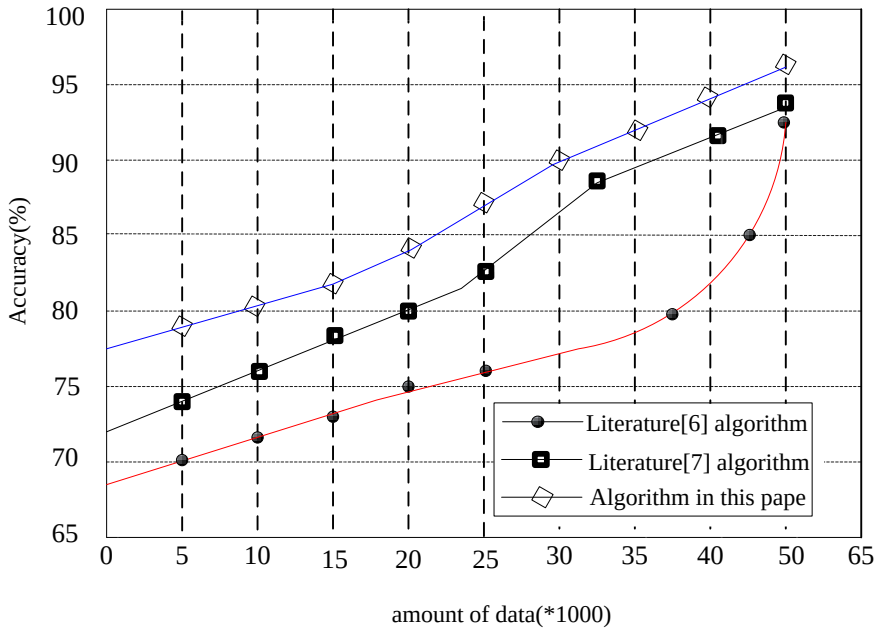

Figure 3. Inquiry Accuracy of Different Data

It can be found from Figure 1-3 that with the constant increase of data, algorithm in this paper is superior to the above two algorithms for reference in terms of operation time, indicating that algorithm in this paper has improved its implementation efficiency, and it can be found through comparing the implementation time of different node groups that algorithm in this paper has stable operation time, indicating that algorithm in this paper is relatively stavle, thus it can ensure the stability of mining results.

6. Conclusion

Faced with the vast amount of information in cloud computing, this paper proposes an Apriori algorithm based on cloud computing and introduces cost-sensitive learning and non-filter matrix to find k frequency set and uses the method of generating association rules to improve effectiveness of data mining. Simulation experiments show that mining algorithm in this paper is highly effective and suitable for data mining in the context of cloud computing with good guidance and practical significance to the development of data mining in cloud computing.

References

[1] H qing, “The Internet of things and the data mining cloud service[J]”,Transactions on Intelligent Systems, vol. 7, no. 3, **(2012)**, pp. 189-193.
[2] D Jing,Yang, S-l,Luo He. “Data Mining Serviece Model in Cloud Computing Environment[J]”,Computer Science, vol. 39, no. 6A, **(2012)**, pp. 217-219
[3] C Miao, “Algorithm of discovering preferred browsing paths based on cloud-computing[J]”,Computer Engineering and Applications, vol. 47, no. 29, **(2011)**, pp. 85-89.
[4] C xiao-chun,Z an,Pan dan, “Research on the Domain-oriented Data Mining Service Platform under Cloud Computation Environment[J]”, Process Automation Instrumentation, vol. 39, no. 5, **(2014)**, pp. 9-12.
[5] Y yi, R kai, ,L zheng-tao, “Data Mining Based on Cloud-Computing Technology[J]”, Microelectronics & Computer, vol. 30, no. 2, **(2013)**, pp. 161-164.
[6] Z Ling-wei,W zhen-xing,D Wen-cai, “Applicaiton of data mining in cloud computing resources prediction”,Laser Journal[J], vol. 36, no. 4, **(2015)**, pp. 185-188.
[7] Z you-ting, “Data mining analysis of cloud services[J]”,Information Studies:Theory & Application, vol. 34, no. 9, **(2012)**, pp. 33-36.

Author

Jiangang Jin (1972.11-) Lecturer, Master, Research Orientation: Computing Network.