MapReduce Algorithms: Consecutive Retrieval of Clusters and Blackboard Database System

Venkata Subba Reddy Poli

May 4, 2020

MapReduce Algorithms: Consecutive Retrieval of Clusters and Blackboard Database System

Poli Venkata Subba Reddy
Department of Computer Science and Engineering,
Sri Venkateswara University, Tirupati
Email:vsrpoli@hotmail.com

Abstract. The objects of data mining are knowledge discovery process and reduce time complexity. Time taken for Information retrieval in big data is very high. Time complexity will be reduced through information retrieval techniques. Cluster is set of query-data item instances. Consecutive Retrieval(C-R) cluster Property is retrieval of data items in data set or cluster from the consecutive locations. This may be achieved through the consecutively retrieval (C-R) cluster property. C-R cluster property is retrieval information using query-data set incidence or clusters. MapReduce algorithms are Map and Reduce for cluster retrieval consecutively. The time will be reduced through the consecutive retrieval cluster property. Parallelism of clusters is designed through parallel clusters, distributed and concurrency of clusters. The parallel clusters are designed using vector approach and genetic algorithms approach. The distributed and parallel algorithms are designed through blackboard architecture. Time and space complexity shall be reduced using directly storage data items with the Blackboard Architecture. The blackboard architecture shall be used store and retrieve the data items of clusters.

Keywords: Data mining, MapReduce algorithms, Consecutive Retrieval, cluster analysis, Blackboard architecture, Blackboard database systems

Introduction

Data mining is knowledge discovery process. Some of the data mining methods are frequent, Association rules and Clustering to discover the knowledge. Data warehousing is the representation in relational dataset grouping data set for particular object. The blackboard architecture will provide retrieval of different objects as clusters. Data mining is to reduce the space complexity with consecutive storage of data warehousing.

The information is to be retrieved within a time for big data. This can be achieved through the consecutively retrieval of information. The consecutive retrieval (C-R) cluster property is retrieval of information consecutively. . The existence of C-R property will retrieve the data items for consecutively the data items. The C-R property will reduce the retrieval time for big data. The designing of Map Reduce algorithms will reduce time for big data retrieval.

The C-R property was first introduced by Gosh [2]. The C-R property is extended to statistical databases by Chin [1]. The C-R property extends to exising of CR-Property

[7]. The MapReduce algorithms are studied for consecutive retrieval cluster analysis. C-R cluster property may be represented through the Vector, graph, genetic and clustering approach. The data items may be stored consecutively with the quarries. The consecutive data items are used for parallel cluster analysis to reduce time complexity The C-R cluster property is studied for parallel cluster analysis using these representations. It is necessary to study relational databases and data mining.
C-R cluster property is consecutive retrieval of data items of clusters for queries.
Suppose C= $\{C_1,C_2,..,C_n)$ is set of clusters for queries $Q=\{Q_1,Q_2,\ldots,Q_n\}$.
Cluster set C is query-data items instances. The clusters are to be consecutive retrieval data items.
The clusters $C_1,C_2,..,C_n$ are set of clusters for pre- queries $Q=\{Q_1,Q_2,\ldots,Q_n\}$.
These clusters are consecutive retrieval data items. For instance , pre-sorted for searching.

2. MapReduce Algorithms

The Relational dataset is representation with domains and tuples [9]. The "Map" is reading datasets and "Reduce" is writing into databases.

Definition: A relational database or dataset is defined as collection of attributes A_1. A_2 ... A_m and is represented as

$R=A_1 \times A_2 \times \ldots \times A_m$
$t_i=a_{i1} \times a_{i2} \times. \ldots \times a_{im}$. i=1. …. n are tuples
or
$R(A_1.\ A_2.\ \ldots\ A_n)$. R is relation.

$R(t_i)= (a_{i1}.\ a_{i2} \ldots\ a_{im)}$. i=1. …. n are tuples

For instance, consider cluster dataset for Account are given by

Table 1. Account

Ac.No	Ac.Name	Ac.Bal
8347102	Rama	10000
8347103	Sita	15000
8347104	Jhon	20000
8347105	Khan	15000
8347106	Marry	18000
8347107	Krishna	25000

For instance, consider cluster dataset for Bank are given by

Table 2. Bank

Ac.No	Ac.Name	Bank
8347102	Rama	SBI
8347103	Sita	ANZ
8347104	Jhon	ICCI
8347105	Khan	AB
8347106	Marry	SBI
8347107	Krishna	AB

MapReduce lossless Join of Account and Bank is given by

Table 3. Account-Address

Ac.No	Ac.Name	Ac.Bal	Bank
8347102	Rama	10000	SBI
8347103	Sita	15000	ANZ
8347104	Jhon	20000	ICCI
8347105	Khan	15000	AB
8347106	Marry	18000	SBI
8347107	Krishna	25000	AB

MapReduce lossless decomposition of Account-Address is given by Table 1 and Table 2.
In the following some of the data mining methods are discussed for MapReduce algorithmsConsider the dataset Account-Address of Table 3.

2.1 Frequency

Frequency is the repeatedly accrued data.
Find the frequently customers purchase more than one Item.

Table 4.Frequency

Bank	Frequency
SBI	2
ANZ	1
ICCI	1
AB	2

2.2 Association rule

Association is of the <Ac.No ⇔Bank> is given by

Table 5. Association

Ac.No ⇔Bank	
831	SBI
832	ANZ
833	ICCI
834	AB

2.3 Clustering

Clustering is grouping the particular data.
Group the customers who are account in Bank

Table 6. Clustering

Ac.No	Ac.Name	Ac.Bal	Bank
8347102	Rama	10000	SBI
8347106	Marry	18000	
8347103	Sita	15000	ANZ
8347104	Jhon	20000	ICCI

8347105	Khan	15000	AB
8347107	Krishna	25000	

3. MapReduce for Join C-R clusters

Suppose R= {r_1, r_2, r_n} is data set of records and C= {C_1, C_2, C_m} is set of clusters . The best type of file organization on a linear storage is one in which records pertaining to Clusters are stored in consecutive locations without redundancy storing data of R. If there exists on such organization of R for C said to have the Consecutive Retrieval property or C-R cluster property with respect to data set R. Thus C-R cluster property applicable to linear storage.

The C-R cluster property is a binary relation between a cluster set and data set. Suppose if a cluster in a cluster set C is relevant to the data in a data set R, than the relevancy is denoted by 1 and the irrelevancy is denoted by 0.Thus the relevancy between cluster set C and data set R can be represented as (n x m) matrix. The matrix is called data item- Cluster Incidence Matrix(DCIM).

Table 7. Data-cluster incidence matrix

R	C_1	C_2	….	C_m
r_1	1	0	…	1
r_2	0	1	---	0
-	-	-	…	-
-	-	-	…	-
-	-	-	…	-
r_n	1	1	…	1

Consider the data set for Custer Account

Table 8. Account

R	*Ac.No*	*Ac.Name*	*Ac.Bal*
r_1	8347102	Rama	10000
r_2	8347103	Sita	16000
r_3	8347104	Jhon	20000
r_4	8347105	Khan	15000
r_5	8347106	Marry	18000
r_6	8347107	Krishna	25000

Reorganization for C-R cluster property is given by

Table 9. Consecutive cluster

R	Ac.No	Ac.Name	Ac.Bal
r_6	8347107	Krishna	25000
r_3	8347104	Jhon	20000
r_5	8347106	Marry	18000
r_2	8347103	Sita	16000
r_4	8347105	Khan	15000
r_1	8347102	Rama	10000

Consider the following clusters of queries

$C_1$is Q_1=Find the customers whose average balance greater than equal to 18000.
C_2 is Q_2= Find the customers whose average balance less than   18000.
C_3 is Q_3=Find the customers whose Balance is >16000.
C_4 is Q_4=Find the customers whose Balance is <15000.

The DCIM is given by

Table 10. DCIM

R	C_1	C_2	C_3	C_4
r_6	1	0	1	0
r_3	1	0	1	0
r_5	1	0	1	0
r_2	0	1	0	1
r_4	0	1	0	1
r_1	0	1	0	1

```
 SQL> create table account(acno integer, acname varchar(10), acbal real);
SQL> insert into account values(8347107 , 'Krishna', 25000);
SQL> insert into account values(8347104 , 'John', 20000);
SQL> insert into account values(8347106 , 'Marry', 18000);
SQL> insert into account values(8347103 , 'Sita', 16000);
SQL> insert into account values(8347105 , 'Khan', 15000);
SQL> insert into account values(8347102 , 'Rama', 10000);
```

The clusters are given by SQL queries.

```
SQL> select acno from account group by acno having avg(acbal)>=18000;
SQL> select acno from account group by acno having avg(acbal)<18000;
SQL> select acno from account where acbal>16000;
SQL> select acno from account where acbal<=16000;
```

The dataset is given for $C_1 \bowtie C_2$ has C-R cluster property.

Table 11. $C_1 \bowtie C_2$

R	$C_1 \bowtie C_2$
r_6	1
r_3	1
r_5	1
r_2	1
r_4	1
r_1	1

The dataset is given for $C_3 \bowtie C_4$ has C-R cluster property.

Table 12. $C_3 \bowtie C_4$

R	$C_3 \bowtie C_4$
r_6	1
r_3	1
r_5	1
r_2	1

r_4	1
r_1	1

m

The dataset is given for $C_1 \bowtie C_3$ has C-R cluster property.

Table 13. C_1 U C_3

R	$C_1 \bowtie C_3$
r_6	1
r_3	1
r_5	1
r_2	0
r_4	0
r_1	0

The dataset is given for $C_2 \bowtie C_4$ has C-R cluster property.

Table 14 $C_2 \bowtie C_4$

R	$C_2 \bowtie C_4$
r_6	0
r_3	0
r_5	0
r_2	1
r_4	1
r_1	1

The dataset is given for $C_2 \bowtie C_3$ has C-R cluster property.

Table 15. $C_2 \bowtie C_3$

R	$C_2 \bowtie C_3$
r_1	1
r_3	1
r_6	1
r_2	1
r_4	1
r_5	1
r_7	1

The cluster sets { $C_1 \bowtie C_2$, $C_3 \bowtie C_4$, $C_1 \bowtie C_3$, C_2 U$\bowtie$ C_4 , $C_2 \bowtie C_3$ } has C-R cluster property .
Thus the cluster sets has C-R cluster property with respect to dataset R

4. MapReduce for Parallel C-R Clusters Property

The design of Parallel cluster shall be studied through the C-R cluster property , It can be studied in two ways. The Parallel cluster design through Graph theoretical approach and The Parallel cluster design through Response vector approach

4.1 Parallel C-R Cluster Property using Response Vector approach

The C-R cluster property between cluster set C and dataset R can be stated in terms of the properties of vectors. The data cluster Incidences of cluster set C with C-R cluster property may be represented as Response Vector set V. For instance the cluster set { C_1, C_2, C_3, C_4} has response vector set {V1=(1,1,1,0,0,0), V2=(0,0,0,1,1,1), V3=(1,1,1,0,0,0), V4=(0,0,0,,1,1,1)

For instance, the Response Vector of the cluster C1 is given by column vector (1,1,1,0,0,0).

Suppose C_i and C_j are two clusters . If the two vectors Vi, Vj of C_i and Cj and the intersection $V_i \cap Vj=\Phi$ then the cluster set {Ci, Cj} has Parallel cluster property

Consider the vectors V_1 and V2 of C_1 and C_2. The intersection of $V_1 \cap V_2=\Phi$, so that the cluster set {C_1, C_2} has Parallel cluster property .

4.3 Parallel C-R Cluster Property using Genetic approach

Genetic Algorithms(GA) introduce by Darwin[18].. GA's are used to learn, and optimize the problem[8]. There are four evaluation processes.

Selection

Reproduction

Mutation

Competition

Consider crossover with two cuts

Parent #1 0000000

Parent #2 1111111

The parent #1 and #2 match by mutation.

Parent #1 111111

Parent #2 111111

The parallel cluster property exists if G(Ci) and G(Cj) matches with mutation.

Consider cluster C1 and C2

Parent #1 111000

Parent #2 000111

The parent #1 and #2 match by crass over

The parallel cluster property exists if G(Ci) and G(Cj) matches with crossover.

5. Consecutive Retrieval using Blackboard database System

Usually in database systems, the entire data has to taken into main memory for operation. There is no need to take entire data in main memory in Blackboard Architecture, Blackboard Architecture used to store and retrieve knowledge sources. Data mining is a knowledge discovery process. Blackboard Arctitecture may used to store and retrieve data sources. Parallel, distributed and concurrent retrieval of data items shall be achieved through the Blackboard architecture.

Blackboard database system approach is storage and retrieval of databases. The blackboard database technique is to store database, retrieve the database and performing transaction for very large databases or big data. The data items of database are data sources. These data sources are shared and processes independently.

The C-R of cluster may be retrieval from distributed datasets. The blackboard architecture contains data items sources. The data item sources shall be directly retrievable. Retrieval of clusters from blackboard system is directly retrieval of data sources. When query being processing, the entire database has to bring to main memory bit in blackboard architecture, the data item source is directly from blackboard structure . For retrieval of information for query. Data item directly retrieved from the Blackboard which contains data item sources.

The blackboard systems may construct with the creation of data item sources in Oracle. Here is algorithm is given to create blackboard architecture, store and retrieve for data item sources.

For instance, each account is a table for banking information systems.

```
Algorithm:
Begin
Create table with account number
Insert data item into account number table
Retrieve data item  from account number table
End
```

Each data item is data source which is created by h(x) account number table.

The blackboard structure is created with each account.

```
 SQL> create table ab8347102(acno int, acname varchar(10), acbal real);
 SQL> create table ab8347103(acno int, acname varchar(10), acbal real);
 SQL> create table ab8347104(acno int, acname varchar(10), acbal real);
 SQL> create table ab8347105(acno int, acname varchar(10), acbal real);
 SQL> create table ab8347106(acno int, acname varchar(10), acbal real);
 SQL> create table ab8347107(acno int, acname varchar(10), acbal real);
```

Inserted accounts into blackboard structure.

```
SQL> insert into ab8347102 values(8347102,'Rama',10000);
SQL> insert into ab8347103 values(8347103,'Sita',16000);
SQL> insert into ab8347104 values(8347104,'John',20000);
SQL> insert into 8347105 values(8347105,'Khan',15000);
SQL> insert into ab8347106 values(8347106,'Marry',18000);
SQL> insert into ab8347107 values(8347107,'Krishna',25000);
```

Select each account number from blackboard structure.

```
SQL> select * from  ab8347102 where acno=8347102;

      ACNO ACNAME  ACBAL
---------- ------------------------------ ----------
   8347102  Rama      10000

SQL> select * from  ab8347103 where acno=8347103;
```

```
    ACNO  ACNAME     ACBAL
---------- ------------------------------ ----------
  8347103    Sita          16000
```

These data items are stored in blackboard data structure.

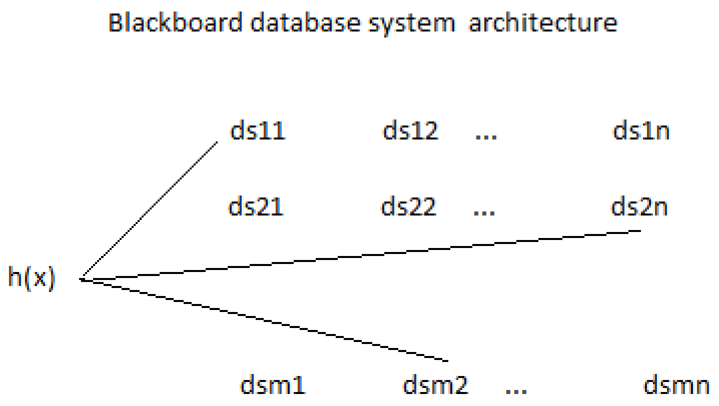

Figure 1. Blackboard database system

h(x) is create, store and retrieval of data sources (ds). When transaction being possessing, there is no need to take entire database into main memory. Just it is sufficient to retrieval of particular data item of particular transaction from the blackboard system.

The advantage of blackboard database architecture is directly operated on data sources.

The blockchain technology is also operates on data sources or data items.

Acknowledgements

The author expressing thanks to Conference Chair for accepting this paper.

References

[1] F.Y Chin, Effective inference control for range SUM queries, Theoretical Computer Science, 32(1974)77-86.

[2] S.P. Ghosh, File Organization: the Consecutive Retrieval Property, Communications of ACM, 15, 9 ,(1972)802-808.

[3] Mircea Eremia ; Chen-Ching Liu ; Abdel-Aty (Edris), Genetic Algorithms ,IEEE, 2018

[4] Robert Englemore, Tony Morgan, Blackboard Systems, Addison-Wesley,1988.

[5] Ramakrishnan,R. Gehrike,J, datasets Management Systems, McGraw-Hill, 2003.

[6] Tan,P.N., Steinbach, V. Kumar,V., Introduction to Data Mining, Addison-Wesley, 2006.

[7] Poli Venkata Subba Reddy, On Existence of C-R Property, Proceedings of Mathematical Society, B.H.U, 5(1989)167-71.

[8] J.D, Ullman, Principles of Datasets Systems, Galgotia Publications, 1999.