Census Data analysis with Multi Relational Data Mining in Clusters

Gowthami N
Dept of CSE, Cit, Gubbi,Tumkur

Harish T A
Assistant professor .
Dept of CSE, Cit, Gubbi,Tumkur.

ABSTRACT: In this method we propose a model for census data analysis to establish relation between the tables by using association rules. And this method is based on a Multi relational Data Mining (MRDM) and multi-node clustering, to reduce the space complexity and time constraints. Census is the procedure of systematically collecting and analyzing the statistical data about the members of a given population and it gives the complete details of the peoples within the country. And by using MRDM technique will helpful for the users to get the complete details about the census from the large datasets in easy way within matter of a seconds.

II. INTRODUCTION

Data Analysis is the process of systematically applying statistical and/or logical techniques to describe and illustrate, condense and recap, and evaluate data. Data is useless without the skills to analyze it so we need data analysis to analyzing the large, messy and unstructured data**.**

Census is the procedure of systematically collecting and analyzing the statistical data about the members of a given population. Census is a big task and it a expensive so it is usually done only once in every decade. The census gives the complete details of the people within the country[4].We all are using public services such as roads, schools, libraries and health services, These all services need to be planned first to develop and improve the residents quality of life. So the census counts the number of people living in each city, town and country.

The multi relational data mining approach has developed as an alternative way for handling the structured data such that RDBMS. This will provides the mining in multiple tables directly. In MRDM the patterns are available in multiple tables (relations) from a relational database. As the data are available over the many tables which will affect the many problems in the practice of the data mining. To deal with this problem, one either constructs a single table by Propositionalisation, or uses a Multi-Relational Data Mining algorithm[5].

clustering is the process of grouping a set of similar objects. It is a main task of exploratory data mining and common technique for statistical data analysis. The multi node clustering consists of two nodes, those are master and slave node.

III. LITERATURE REVIEW

Census gives the complete picture of the nation, this will helps to the government to improve the residential quality of people. And here the dataset

table is divided into multiple tables to establish the relation between the tables also use the multi node clusters.

A. Multi Relational Data Mining (MRDM)

Multi relational data mining (MRDM) is a form of data mining operating on data stored in multiple database tables[7]. It is an alternative way for handling the structured data such RDBMS [1]. There are many approaches are supported by MRDM those are:

- Inductive Logic Programming(ILP)
- Multi-relational Clustering
- Probabilistic Relational Models

The analyzed data is stored in relational database and it consists of a set of tables and a set of associations between pairs of table, it also describes how records in one table relate to the records in another table[2]. Both tables and associations are called relations, so we will use former terminology to be able to distinguish between two concepts. The association between two tables describes the relationships between records in both tables. And the nature of this relationship is characterized by the multiplicity of the association.

B Multi node clusters

In this multi-node cluster master node will run the "master" daemons in each layer: the job tracker for the mapReduce layer and namenode for the HDFS storage layer. And both machines will run the "slave" daemons: tasktracker for mapreduce processing layer and datanode for the HDFS layer. Basically, the "master"daemons are responsible for coordination and management of the "slave" daemons while the latter will do the actual data storage and data processing work. The master node will also act as a slave because we have only two machines available in our cluster but still want to spread data storage and processing to multiple machines. Starting the cluster is done in two steps. First, the HDFS daemons are started: the namenode daemon is started on master, and datanode daemons are started on all slaves. Second, the MapReduce daemons are started: the jobtracker is started on master, and tasktracker daemons are started on all slaves[6].

IV.EXISTING SYSTEM

Central Statistics Office (CSO) is a central agency for the collection, consolidation, processing, analyzing and publication of census data. In CSO a team of computer operators do the analysis on the data which generally takes about 6 months to publish the preliminary reports .The classification and tabulation is Complicated process, Every data cannot be put into tables due to Lack of flexibility . Analysis of population from a huge amount of raw data requires processing of raw data at first, which takes a lot of time i.e. it is computationally expensive in terms of time and storing such a huge amount of data also brings space complexity.

In single data mining applications the assumption of single table turns out to be a great limitation. The data to be mined are represented in a single table (or relation) of a relational database, such that each row (or tuple) represents an independent unit of the sample rural population and columns correspond to properties of units. Two relations are required to represent object interactions[3].

V.PROPOSED SYSTEM

In this proposed system multi relational data mining(MRDM) typically consists of several tables and not just a single one. MRDM can perform the linkage Knowledge discovery in multi

relational environments, Patterns found by these approaches are called *relational*. It is frameworks which deals with gathering the data about the data (metadata) from a database and choose the best approach to get the optimal results.

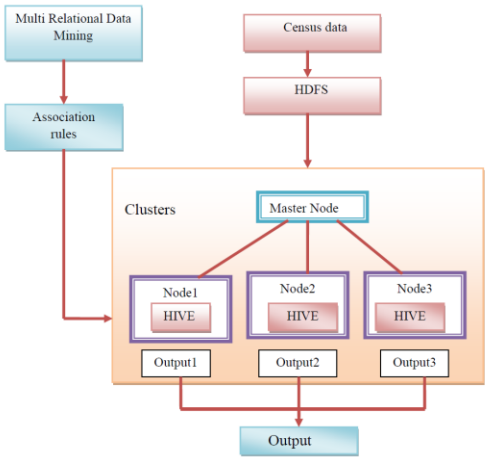

Fig1. System architecture

Input raw census data in "tsv" format that is loaded into HDFS (hadoop distributed file system). The cluster is composed with a master node and multiple data nodes. The data is divided among the slave/data nodes by the supervision of the Namenode(master node) which keeps track of all the jobs associated with the slave nodes. Once the data division is complete we provide the association rules formed on the basis of Multi Relational Data Mining (MRDM) technique. The associated rules are given as input to all the data nodes, then the processing begins at each data nodes. After the completion of the processing at each datanode, the obtained partial results from each datanode is combined by the namenode and it outputs the final result.

VI .CONCLUSION

Census gives the complete details of all the people within the country. It is useful for the government to improve the residential quality of life. The Multi relational Data Mining (MRDM) is an alternative way for handling structured data. It will directly provides the mining in multiple tables, also provides relation establishment between the tables. And this paper present multi-node clustering to reduce the time and space complexity.

VII. REFERENCES

[1]. Annalisa Appice, Michelangeol Ceci and Donato malerba**, "Mining Model Trees: A Multi-relational Approach"**, Dipartmento di Informatica, University Dedli Studi via Orabona, 4-70126 bari,Italy.

[2]. Neelamadhab padhy, Asst,Professor, GIET, Gunupur, Researc school, CMJ University, shilong (Meghalaya) and Rasmith panigrahi, Lecturer, G.I.E.T, Orissa,India, **"Multi Relational Data Mining approachs: A Data Miningh technique"**.

[3]. Donato Malerba, Francesca A. Lisi, Annalisa Appice, Francesco Sblendorio **"Multi Spatial association rule in Census data: A relational approach"**, Dipartimento di Informatica, Universita degli Studi di Bari, via Orabona 4, 70126 Bari, Italy

[4]. A.Vijayaraj and P.DineshKumar, **"Design and implementation of census data collection system using PDA"**, *International Journal of Computer Applications (0975–8887)Volume9– No.9, November 2010.*

[5]. Saso Dzeroski , **"Multi Relational Data Mining : An introduction"**, Jozef Stefan Institute Jamova 39, SI1000 Ljubljana, Slovenia.

[6]. Arno Jan Knobbe, Geboren op 20 oktober 1970, te Willemstad, Curaçao, **"Multi Relational Data Mining",** SIKS Dissertation Series No. 2004-15.

[7]. Neelamadhab Padhy and Rasmita Panigrahi, **" Multi Relational Data Mining Approaches: A Data Mining Technique",** *International Journal of Computer Applications (0975 – 8887) Volume 57– No.17, November 2012.*

