**Abstract:**

*Association rules are rules which are derived from existing dataset. These rules can be used to improve many things in real life world. This report is about Paper Author dataset in which we use association rules to tell which authors are contributing papers together by using R statistical tool.*

**Introduction:**

In real world there are many events occurring continuously with different objects. Many time we are interested to discover the relations between different objects of events. Associations rules is an technique to discover some kind of relation between them, that relation may also called rule. That relation is also measured by some scales, which helps to define strength.

**Association Rule Learning:**

  It s a method for discovering interesting relations between variables in large databases. It is intended to identify strong rules discovered in databases using some measures of interestingness. n order to select interesting rules from the set of all possible rules, constraints on various measures of significance and interest are used. The best-known constraints are minimum thresholds on support and confidence.

**Support:**

It is contribution of a single item or set of items in total dataset.

Example: Lets take scenario of market purchasing example in which we assume that every customer from a super mart purchases some items. If flour is been purchased by 50 customer from 100 customer;

Support = = 0.5

**Confidence:**

It is measure of surety that a item is purchased with another item.

Example: With same market purchasing scenario, flour is been purchased 50 customer and sugar is purchased 25 times with flour then,

Confidence = = 0.5

We can say it is 50 percent sure that sugar is been purchased with floor.

**Training Data:**

We have taken a dataset in which there are two columns; paper ID and author ID. This defines that which paper is written by which author(s). So it not prepared for our association rule, because association rule requires data in shape of transaction. So for preprocessing, we have used a Java program to clean data by deleting duplicate records and place author IDs against every paper ID. Dataset contains 97994 papers.

public class DMEffiencient {

/\*\*

\* @author Jay Kumarr

\*/

public static void main(String arg[]) throws IOException {

File file = new File("D:\\paper\_author\_dataset.csv");

File outputFile = new File("D:\\dataset\_processed\_duplicatefree.csv");

PrintStream out = new PrintStream(new FileOutputStream(outputFile));

DataInputStream dos = new DataInputStream(new FileInputStream(file));

String str = null;

int records = 1;

int total= 453745;

String paperId = "";

StringBuilder builder = new StringBuilder("AAAA");

Set<String> authorSet = new HashSet<>();

while( (str = dos.readLine()) != null){

String[] strArray = str.split(",");

String tempPaperId = ( strArray[0].trim());

String authorId = (strArray[1].trim());

if(paperId.equals(tempPaperId)){

if(!authorSet.contains(authorId))

{

builder.append(authorId).append(",");

authorSet.add(authorId);

}else{

String p = applyCustomDecrypt(paperId);

String a = applyCustomDecrypt(authorId);

System.out.println(p+" has Duplicate: "+a);

}

}else{

if(authorSet.size()>1)

out.println(builder.substring(0, builder.length()-2));

paperId = tempPaperId;

authorSet.clear();

builder = new StringBuilder().append(authorId).append(",");

authorSet.add(authorId);

}

double d = ((double)((double)records/total)\*100);

if( d%1 == 0.0 ){

System.out.println("Percentage : "+d);

}

records++;

}

}

**Solving in R:**

For association rule we have to install package ‘arules’. And load libraries ‘arules’ and ‘Matrix’. Then we load output CSV file in form of transaction. Apriori alogorithm is used for association.

We have to set some threshold in which more interesting rules can been found.

install.packages("arules")

library(Matrix)

library(arules)

PA <- read.transactions("D:\\dataset\_processed\_duplicatefree.csv", sep=",")

m1 <- apriori(Groc, parameter = list(support = 0.0001,confidence=1, minlen=2))

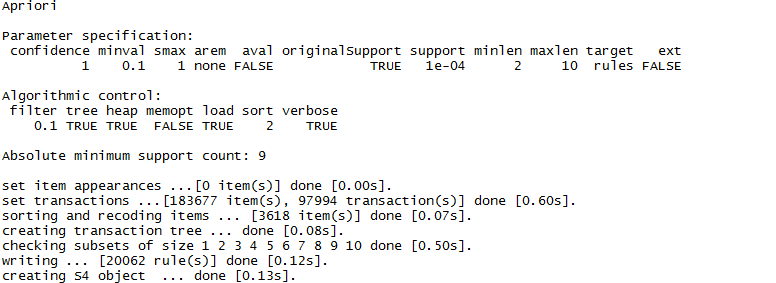


Figure: status after applying apriori algorithm

So it found 20062 rules, below some rules are shown in figure.

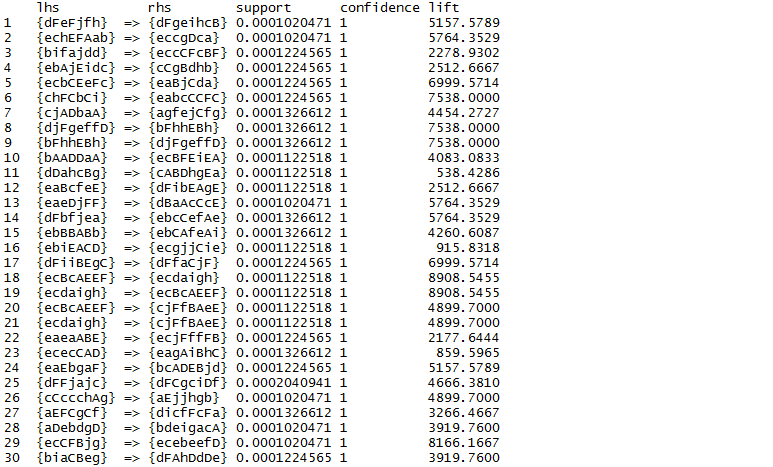


Figure: Association Rules