DATA MINING

Contents:

About Data set:

I used “Anonymous Microsoft Web Data” from UCI Machine Learning Repository. Below is the link of my data set:
http://archive.ics.uci.edu/ml/machine-learning-databases/anonymous/

Summary of Data:

The data is being created by sampling and processing the www.microsoft.com logs. The data records the use of www.microsoft.com by 38000 anonymous, randomly-selected users. For each user, the data lists all the areas of the web site that user visited in a one-week timeframe.

Data format:

```
C,"10164",10164
V,1123,1
V,1009,1
V,1052,1
```

Where:
’C’ marks this as a case line,
’10164’ is the case ID number of a user,
’V’ marks the vote lines for this case,
’1123’,’ 1009’, 1052’ represents relative URL that a user visited. For instance, 1009 represents "Windows Family of OSs","/windows" URL.

Mean number of visits per user: 3.0

Data Mining Task：

In this dataset, the data mining task is to find rules based on user’ s website visit. Such rule can help Microsoft to improve its website suggestion algorithms. It can also give users website visit behavior. For example, users who visit search webpage also visits news webpage.

This assignment shows how we can find association rules using Apriori algorithm in R.

R Code：

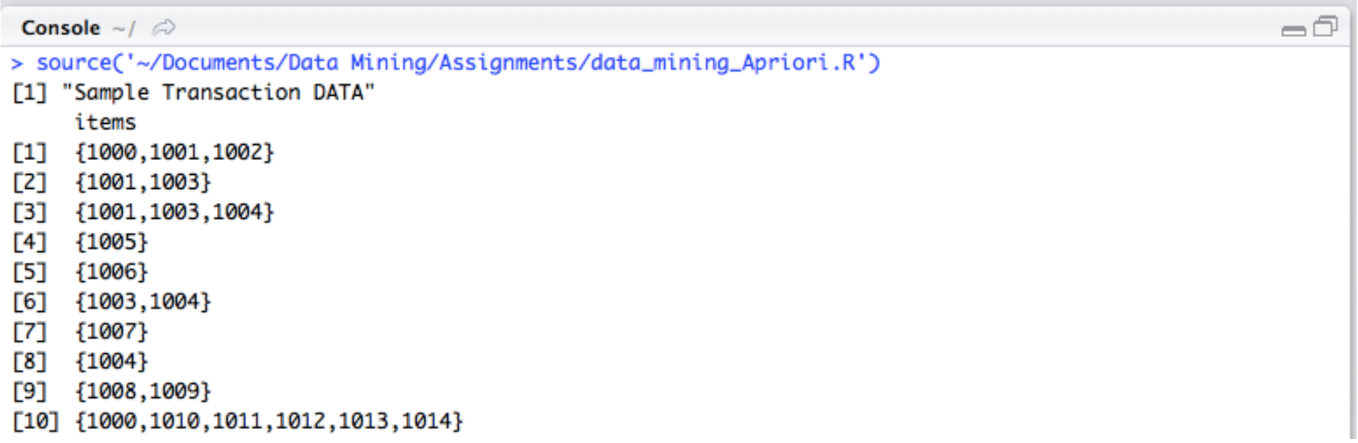
Below is the R code for finding patterns by Apriori algorithm.

```
#loading library aruples
#install.packages(″arules″);
#install.packages(″arulesViz″);
library(arules);
library(arulesViz);
transactionData<-read.transactions(″/Users/Ravi/Documents/Data
Mining/Assignments/transactions.txt″,sep=″,″)

# display Sample DATA
print(″Sample Transaction DATA″)
inspect(transactionData[1:10])

# Displaying Frequency Count
itemFrequencyPlot(transactionData,type=″absolute″,topN=15)

# Get all rules applying apriori with support = 0.15 % confidence  = 75 %
all_rules <- apriori(transactionData, parameter = list(supp = 0.01, conf = 0.75))

# all rules will have list of the rules which is found by apriori algorithms
# first 10 rules
inspect(all_rules[1:10])

#sorting rules based on confidece which will letter help us to remove redundant rules
```

```
all_rules<-sort(all_rules, by=″confidence″, decreasing=TRUE)

#top rules 10 with high confidence
print(″top rules 10 with high confidence ″)
inspect(all_rules[1:10])

#To find redundant rules, first find all the subsets by using is.subset function
subSetOfRules<- is.subset(all_rules, all_rules)

#set lower triangle to NA as upper triangle is sufficiant for each pair
subSetOfRules[lower.tri(subSetOfRules, diag=T)] <- NA
#taking sum of the column for redundant rules
redundantRules <- colSums(subSetOfRules, na.rm=T) >= 1

#finding redundant rules
AllredundantRules<-all_rules[redundantRules]

#display all redundant rules
print(″--------------Redundant rules-------------″)
inspect(AllredundantRules)

#removing redundant rules
finalRules <- all_rules[!redundantRules]

#display all final rules
print(″-------------final rules-------------″)
inspect(finalRules)

#display graph of rules
#plot(finalRules)
plot(finalRules,method=″graph″)
plot(finalRules, method=″paracoord″, control=list(reorder=TRUE))
plot(finalRules, method=″graph″, control=list(type=″items″))
```

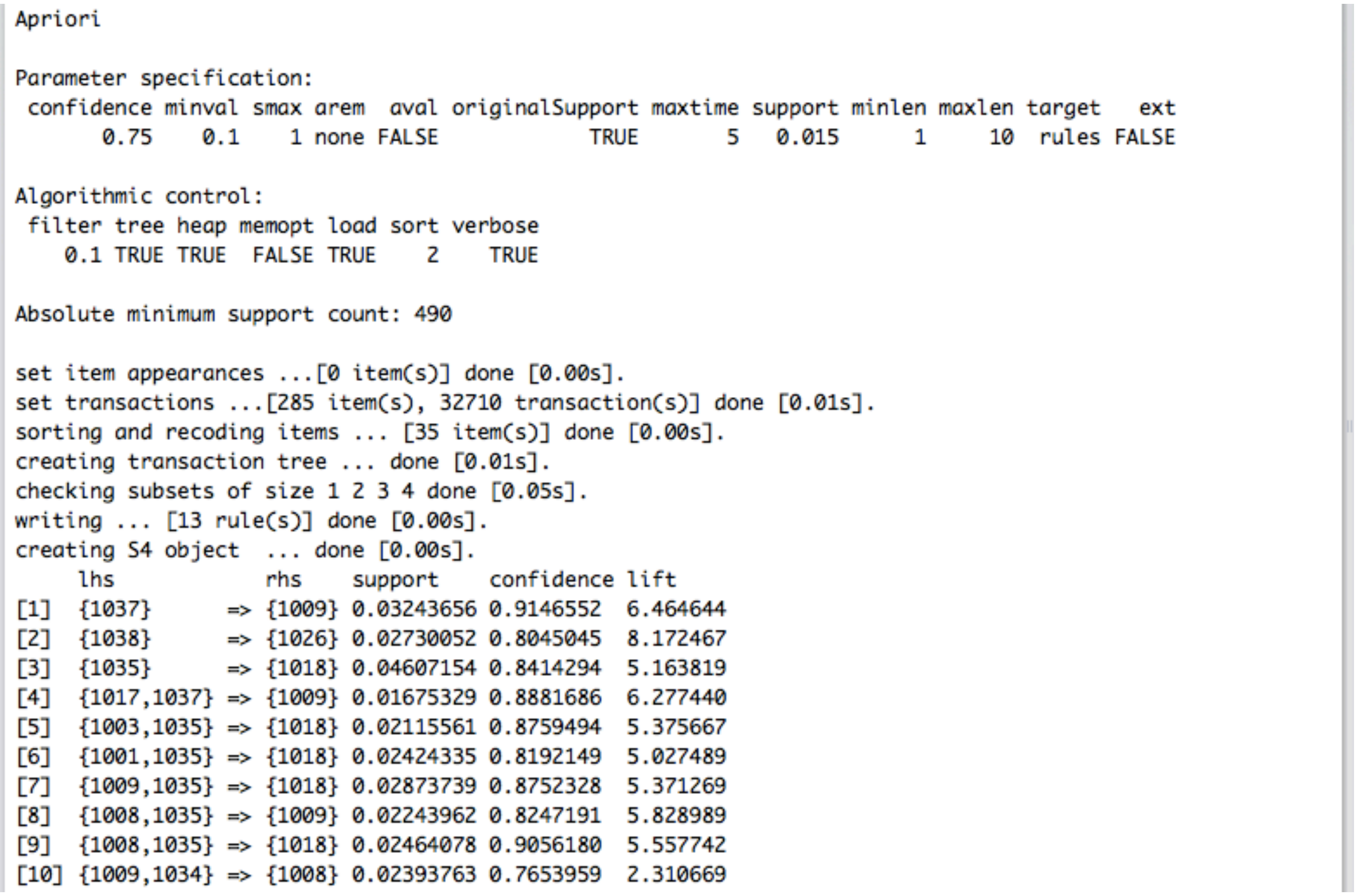
Output:

Sample transactional Data output:



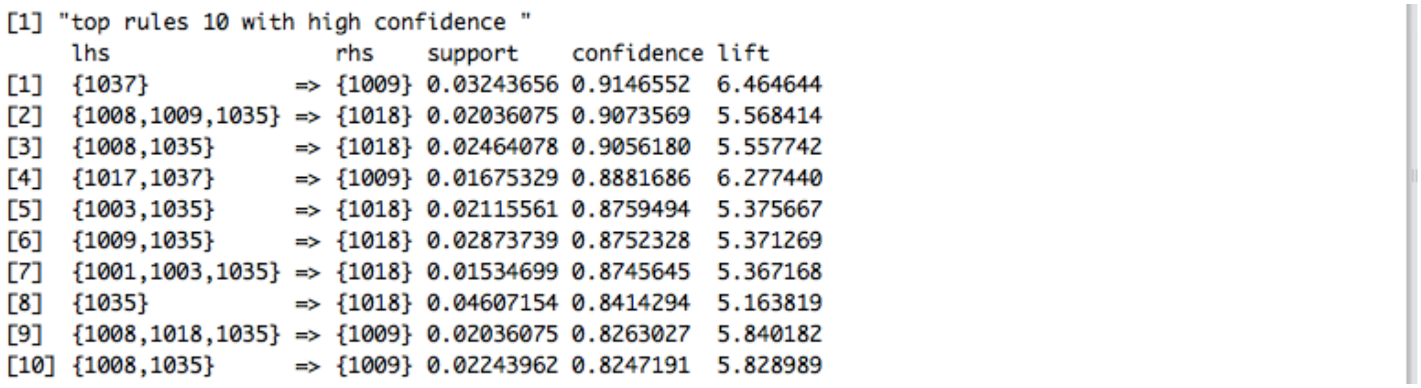
```
> source('~/Documents/Data Mining/Assignments/data_mining_Apriori.R')
[1] "Sample Transaction DATA"
     items
[1]  {1000,1001,1002}
[2]  {1001,1003}
[3]  {1001,1003,1004}
[4]  {1005}
[5]  {1006}
[6]  {1003,1004}
[7]  {1007}
[8]  {1004}
[9]  {1008,1009}
[10] {1000,1010,1011,1012,1013,1014}
```

After running Apriori Algorithm:



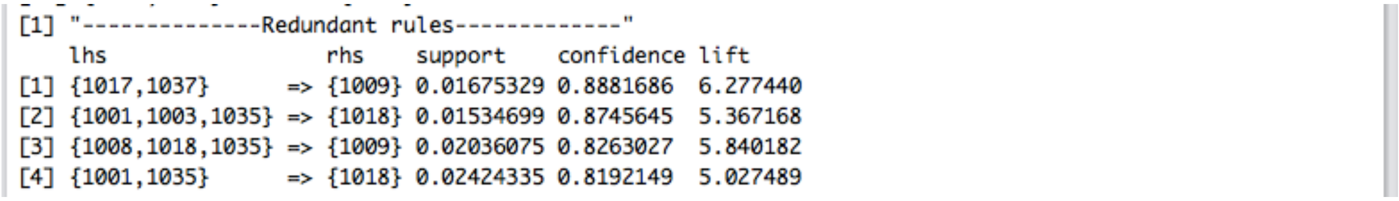
```
Apriori

Parameter specification:
 confidence minval smax arem  aval originalSupport maxtime support minlen maxlen target   ext
       0.75    0.1    1 none FALSE            TRUE       5   0.015      1     10  rules FALSE

Algorithmic control:
 filter tree heap memopt load sort verbose
    0.1 TRUE TRUE  FALSE TRUE    2    TRUE

Absolute minimum support count: 490

set item appearances ...[0 item(s)] done [0.00s].
set transactions ...[285 item(s), 32710 transaction(s)] done [0.01s].
sorting and recoding items ... [35 item(s)] done [0.00s].
creating transaction tree ... done [0.01s].
checking subsets of size 1 2 3 4 done [0.05s].
writing ... [13 rule(s)] done [0.00s].
creating S4 object  ... done [0.00s].
     lhs            rhs    support    confidence lift
[1]  {1037}      => {1009} 0.03243656 0.9146552  6.464644
[2]  {1038}      => {1026} 0.02730052 0.8045045  8.172467
[3]  {1035}      => {1018} 0.04607154 0.8414294  5.163819
[4]  {1017,1037} => {1009} 0.01675329 0.8881686  6.277440
[5]  {1003,1035} => {1018} 0.02115561 0.8759494  5.375667
[6]  {1001,1035} => {1018} 0.02424335 0.8192149  5.027489
[7]  {1009,1035} => {1018} 0.02873739 0.8752328  5.371269
[8]  {1008,1035} => {1009} 0.02243962 0.8247191  5.828989
[9]  {1008,1035} => {1018} 0.02464078 0.9056180  5.557742
[10] {1009,1034} => {1008} 0.02393763 0.7653959  2.310669
```

Top 10 rules with high confidence:



```
[1] "top rules 10 with high confidence "
     lhs                 rhs    support    confidence lift
[1]  {1037}           => {1009} 0.03243656 0.9146552  6.464644
[2]  {1008,1009,1035} => {1018} 0.02036075 0.9073569  5.568414
[3]  {1008,1035}      => {1018} 0.02464078 0.9056180  5.557742
[4]  {1017,1037}      => {1009} 0.01675329 0.8881686  6.277440
[5]  {1003,1035}      => {1018} 0.02115561 0.8759494  5.375667
[6]  {1009,1035}      => {1018} 0.02873739 0.8752328  5.371269
[7]  {1001,1003,1035} => {1018} 0.01534699 0.8745645  5.367168
[8]  {1035}           => {1018} 0.04607154 0.8414294  5.163819
[9]  {1008,1018,1035} => {1009} 0.02036075 0.8263027  5.840182
[10] {1008,1035}      => {1009} 0.02243962 0.8247191  5.828989
```

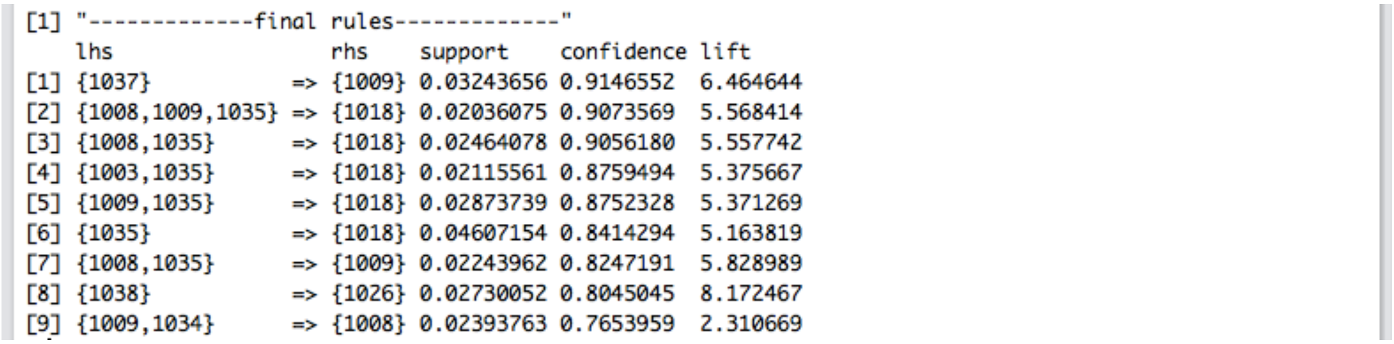
Redundant Rules:



```
[1] "--------------Redundant rules-------------"
    lhs                  rhs    support    confidence lift
[1] {1017,1037}       => {1009} 0.01675329 0.8881686  6.277440
[2] {1001,1003,1035} => {1018} 0.01534699 0.8745645  5.367168
[3] {1008,1018,1035} => {1009} 0.02036075 0.8263027  5.840182
[4] {1001,1035}       => {1018} 0.02424335 0.8192149  5.027489
```

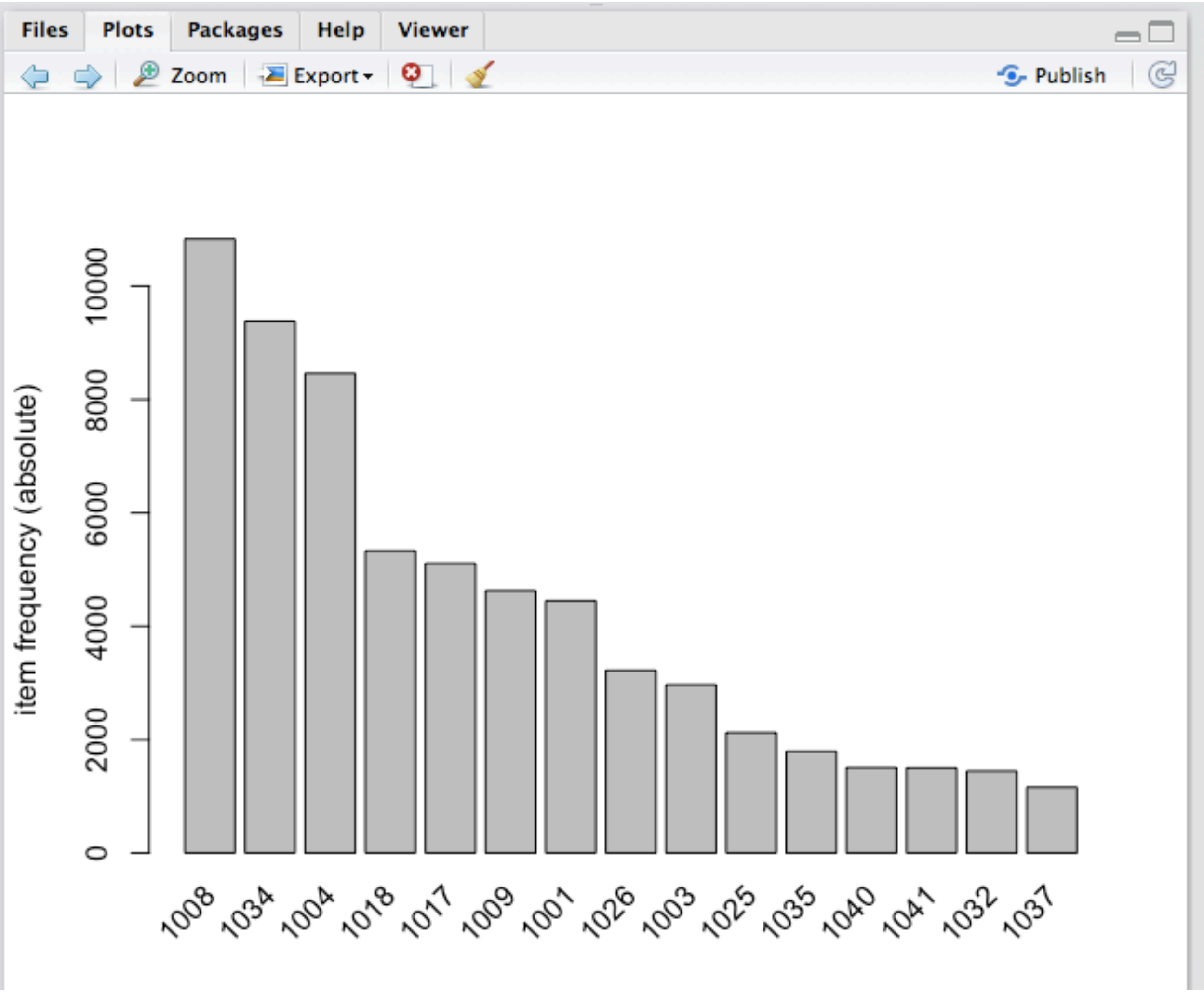
Here rule [1] is redundant because {1017,1037}=> {1009} has same information as rule {1037} = > {1009} with lower confidence and support. So the former rule is redundant.

After Removing Redundant Rules:



```
[1] "-------------final rules-------------"
    lhs                  rhs    support    confidence lift
[1] {1037}            => {1009} 0.03243656 0.9146552  6.464644
[2] {1008,1009,1035} => {1018} 0.02036075 0.9073569  5.568414
[3] {1008,1035}       => {1018} 0.02464078 0.9056180  5.557742
[4] {1003,1035}       => {1018} 0.02115561 0.8759494  5.375667
[5] {1009,1035}       => {1018} 0.02873739 0.8752328  5.371269
[6] {1035}            => {1018} 0.04607154 0.8414294  5.163819
[7] {1008,1035}       => {1009} 0.02243962 0.8247191  5.828989
[8] {1038}            => {1026} 0.02730052 0.8045045  8.172467
[9] {1009,1034}       => {1008} 0.02393763 0.7653959  2.310669
```

Below is frequency graph of data:



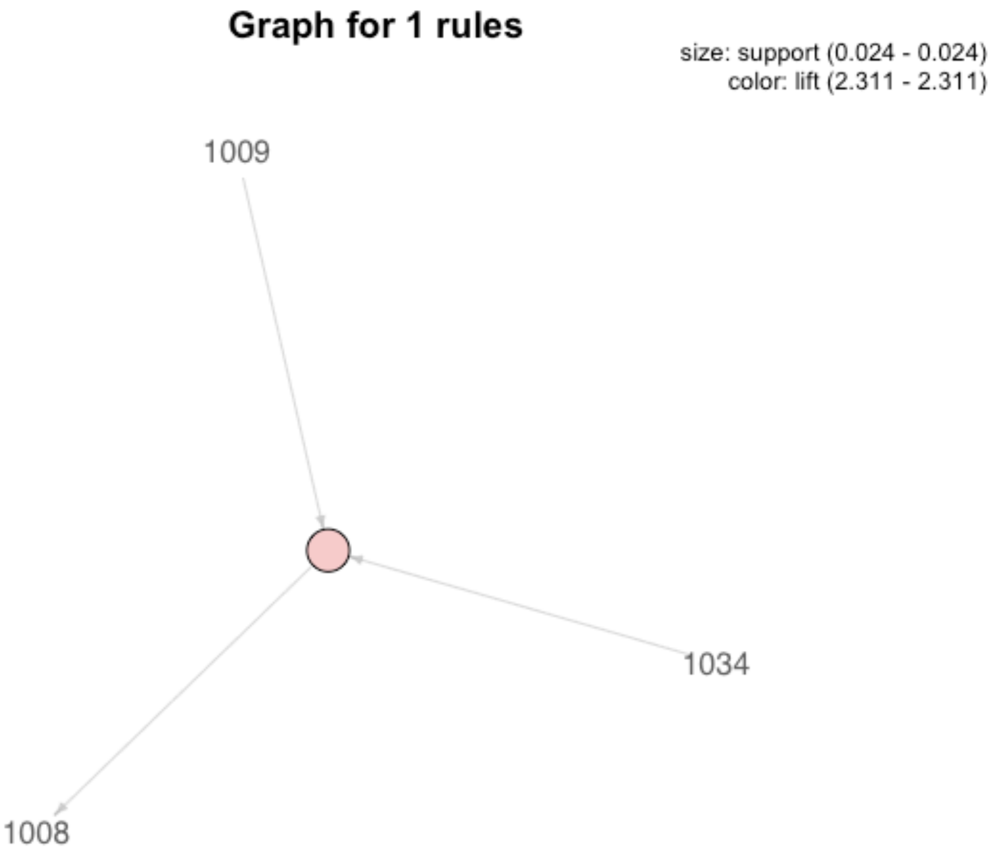
Here, 1008 show the highest visited webpage which is visited more that 10,000 times.

Graph to understand rules:

For example:

Rule [9] {1009, 1034} => {1008}

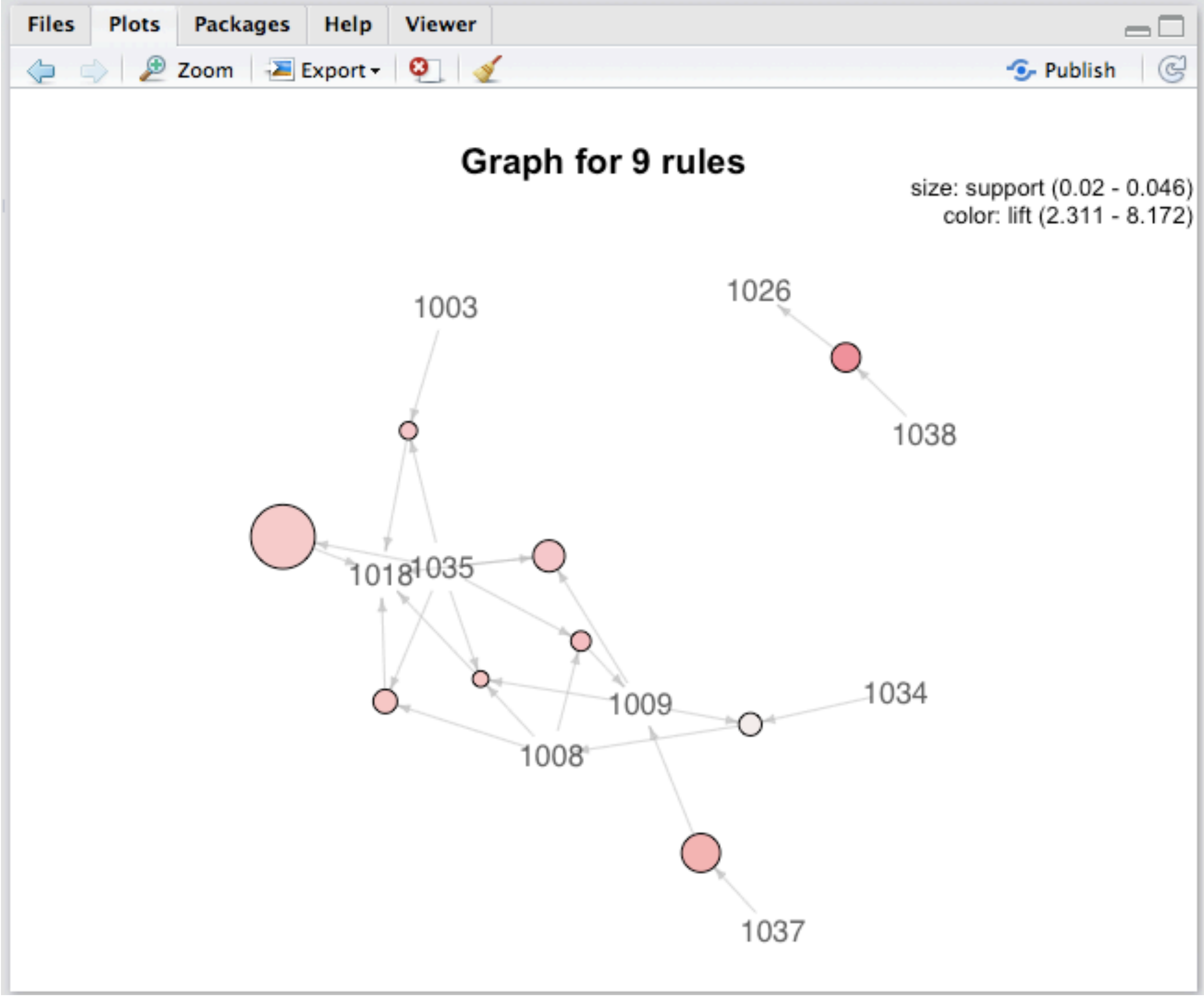
```
plot(finalRules[9],method="graph")
```



Above graph indicates that user who visits 1009 web page and 1034 web page, also visits 1008 web page. This rule has 76% confidence and 2.4 % support.

Here, 1009, 1034 and 1008 indicates webpages. Webpages strings like www.microsoft.com/news are converted into some numbers, because it becomes easy and fast to work with numbers than strings.

Below is the graph for all 9 rules:

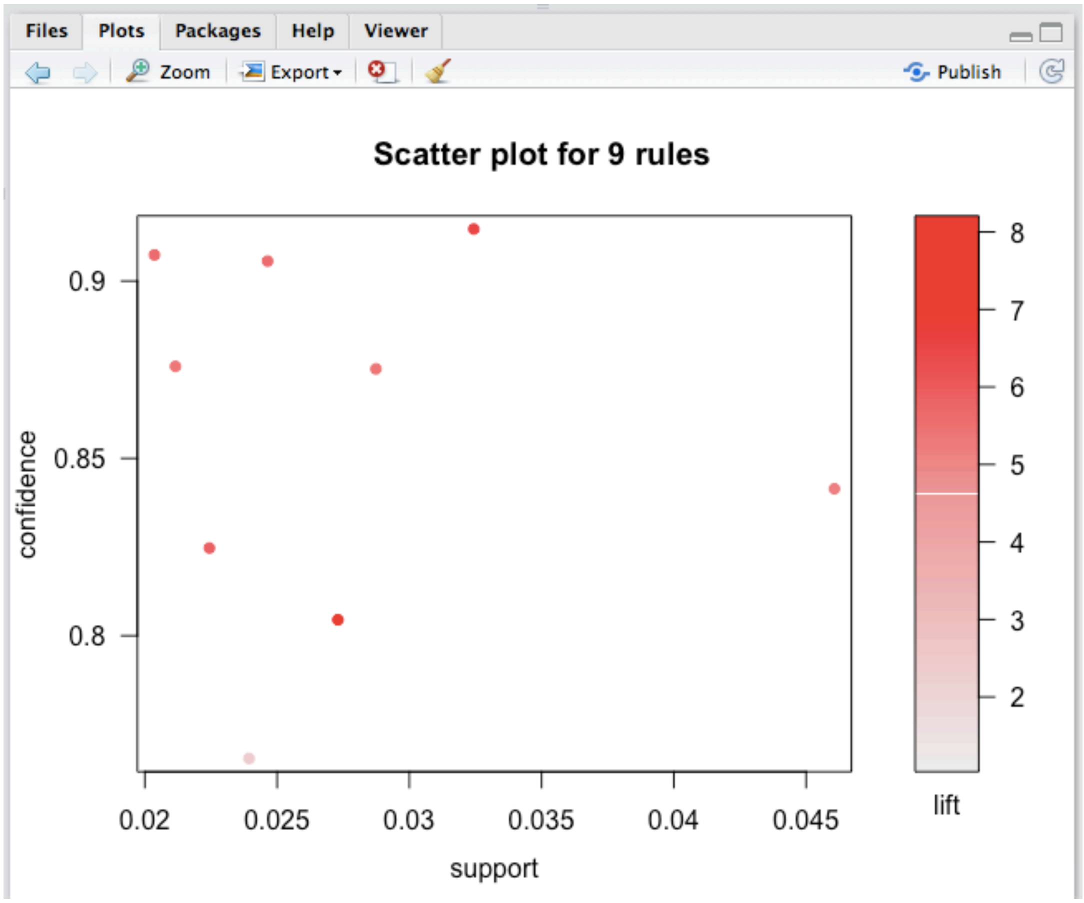


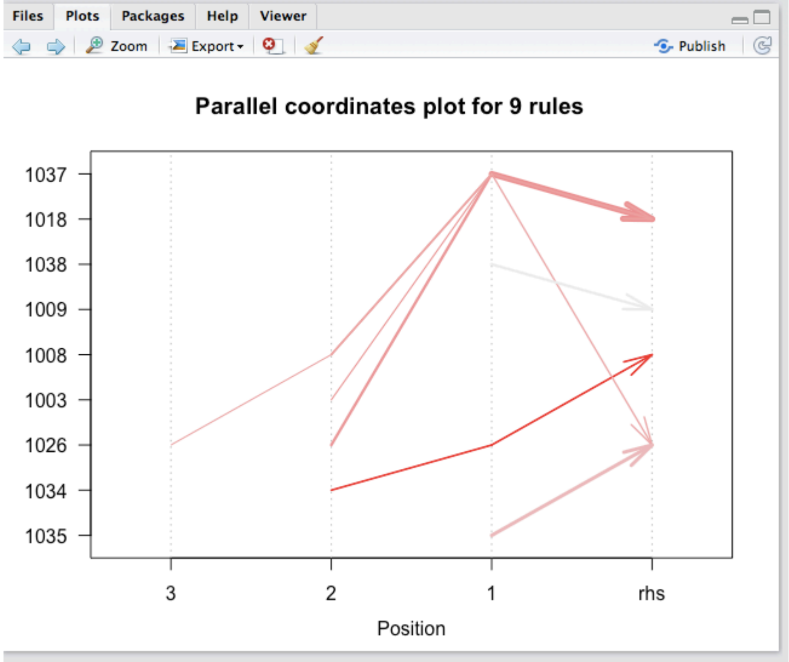
In above graph, bubble size shows the support for the rule, and color of the bubble shows lift measure for that rule.

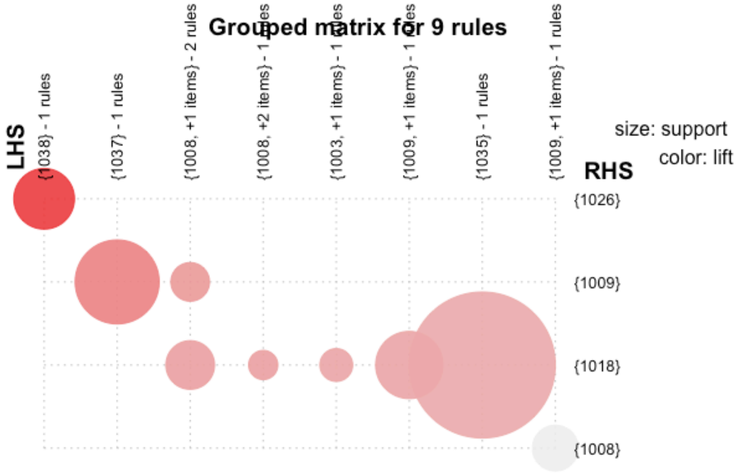
So, it is clear from the above graph that rule **1035 => 1018** has the **highest support** and rule **1038 => 1026** shows it has **highest lift value.**

Scatter Plot:

Below Scatter plot gives information about all 9 rules. Here, color of the points indicates lift value, X-axis indicates support and Y-axis indicates confidence.







```
> plot(finalRules, method=”group”)
```

In Above graph, it displays left side item(LHS) and right side items(RHS). Point size in the graph shows support and color shows lift value. So, we can observe that {1038} => {1026} has the **highest lift value** and rule {1035} => {1018} **has highest support**.

Conclusion:

By using this Apriori algorithm on user’s web history with support = 0.15 % confidence = 75 % , 9 rules are found after removing redundancy which is shown below.

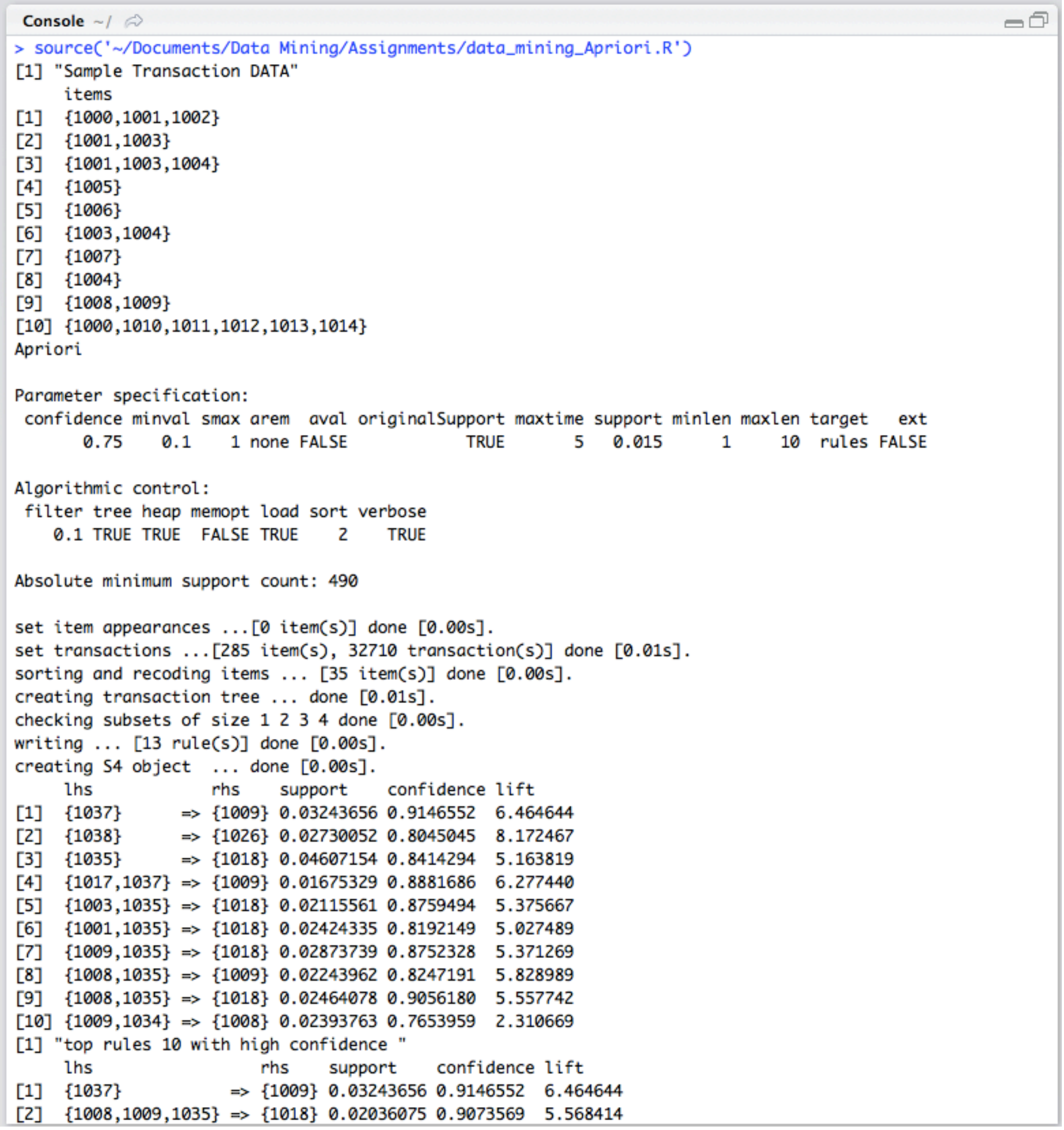
```
    Lhs                  Rhs    Support    Confidence Lift
[1] {1037}            => {1009} 0.03243656 0.9146552  6.464644
[2] {1008,1009,1035}  => {1018} 0.02036075 0.9073569  5.568414
[3] {1008,1035}       => {1018} 0.02464078 0.9056180  5.557742
[4] {1003,1035}       => {1018} 0.02115561 0.8759494  5.375667
[5] {1009,1035}       => {1018} 0.02873739 0.8752328  5.371269
[6] {1035}            => {1018} 0.04607154 0.8414294  5.163819
[7] {1008,1035}       => {1009} 0.02243962 0.8247191  5.828989
[8] {1038}            => {1026} 0.02730052 0.8045045  8.172467
[9] {1009,1034}       => {1008} 0.02393763 0.7653959  2.310669
```

For rule **[8] {1038} => {1026}**
Above number indicates
1038 ->"SiteBuilder Network Membership","/sbnmember"
1026 ->"Internet Site Construction for Developers","/sitebuilder"

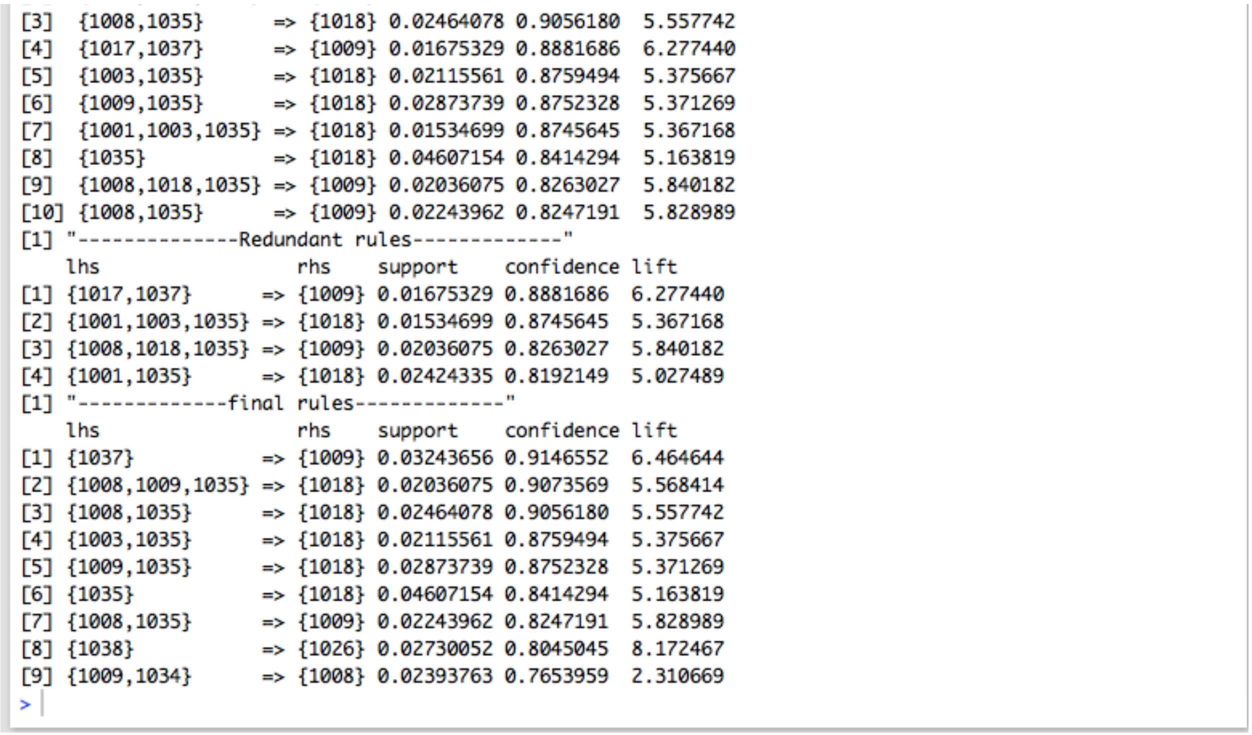
So the users who visits “/sbnmember” also visits “/sitebuilder” web pages of the Microsoft.com website. Such association/patterns help Microsoft to understand user visit behaviors, and also helps to improve webpage suggestions.

For rule [5] {1009,1035} => {1018}
1009 -> "Windows Family of OSs","/windows"
1035 -> "Windows95 Support","/windowssupport"
1018 -> "isapi","/isapi"

Output of the R program:



```
> source('~/Documents/Data Mining/Assignments/data_mining_Apriori.R')
[1] "Sample Transaction DATA"
     items
[1]  {1000,1001,1002}
[2]  {1001,1003}
[3]  {1001,1003,1004}
[4]  {1005}
[5]  {1006}
[6]  {1003,1004}
[7]  {1007}
[8]  {1004}
[9]  {1008,1009}
[10] {1000,1010,1011,1012,1013,1014}
Apriori

Parameter specification:
 confidence minval smax arem  aval originalSupport maxtime support minlen maxlen target   ext
       0.75    0.1    1 none FALSE            TRUE       5   0.015      1     10  rules FALSE

Algorithmic control:
 filter tree heap memopt load sort verbose
    0.1 TRUE TRUE  FALSE TRUE    2    TRUE

Absolute minimum support count: 490

set item appearances ...[0 item(s)] done [0.00s].
set transactions ...[285 item(s), 32710 transaction(s)] done [0.01s].
sorting and recoding items ... [35 item(s)] done [0.00s].
creating transaction tree ... done [0.01s].
checking subsets of size 1 2 3 4 done [0.00s].
writing ... [13 rule(s)] done [0.00s].
creating S4 object  ... done [0.00s].
     lhs            rhs    support    confidence lift
[1]  {1037}      => {1009} 0.03243656 0.9146552  6.464644
[2]  {1038}      => {1026} 0.02730052 0.8045045  8.172467
[3]  {1035}      => {1018} 0.04607154 0.8414294  5.163819
[4]  {1017,1037} => {1009} 0.01675329 0.8881686  6.277440
[5]  {1003,1035} => {1018} 0.02115561 0.8759494  5.375667
[6]  {1001,1035} => {1018} 0.02424335 0.8192149  5.027489
[7]  {1009,1035} => {1018} 0.02873739 0.8752328  5.371269
[8]  {1008,1035} => {1009} 0.02243962 0.8247191  5.828989
[9]  {1008,1035} => {1018} 0.02464078 0.9056180  5.557742
[10] {1009,1034} => {1008} 0.02393763 0.7653959  2.310669
[1] "top rules 10 with high confidence "
     lhs                 rhs    support    confidence lift
[1]  {1037}           => {1009} 0.03243656 0.9146552  6.464644
[2]  {1008,1009,1035} => {1018} 0.02036075 0.9073569  5.568414
```



```
[3]  {1008,1035}      => {1018} 0.02464078 0.9056180  5.557742
[4]  {1017,1037}      => {1009} 0.01675329 0.8881686  6.277440
[5]  {1003,1035}      => {1018} 0.02115561 0.8759494  5.375667
[6]  {1009,1035}      => {1018} 0.02873739 0.8752328  5.371269
[7]  {1001,1003,1035} => {1018} 0.01534699 0.8745645  5.367168
[8]  {1035}           => {1018} 0.04607154 0.8414294  5.163819
[9]  {1008,1018,1035} => {1009} 0.02036075 0.8263027  5.840182
[10] {1008,1035}      => {1009} 0.02243962 0.8247191  5.828989
[1] "--------------Redundant rules-------------"
    lhs                  rhs    support    confidence lift
[1] {1017,1037}      => {1009} 0.01675329 0.8881686  6.277440
[2] {1001,1003,1035} => {1018} 0.01534699 0.8745645  5.367168
[3] {1008,1018,1035} => {1009} 0.02036075 0.8263027  5.840182
[4] {1001,1035}      => {1018} 0.02424335 0.8192149  5.027489
[1] "-------------final rules-------------"
    lhs                  rhs    support    confidence lift
[1] {1037}           => {1009} 0.03243656 0.9146552  6.464644
[2] {1008,1009,1035} => {1018} 0.02036075 0.9073569  5.568414
[3] {1008,1035}      => {1018} 0.02464078 0.9056180  5.557742
[4] {1003,1035}      => {1018} 0.02115561 0.8759494  5.375667
[5] {1009,1035}      => {1018} 0.02873739 0.8752328  5.371269
[6] {1035}           => {1018} 0.04607154 0.8414294  5.163819
[7] {1008,1035}      => {1009} 0.02243962 0.8247191  5.828989
[8] {1038}           => {1026} 0.02730052 0.8045045  8.172467
[9] {1009,1034}      => {1008} 0.02393763 0.7653959  2.310669
>
```