Master of Technology in Enterprise Business Analytics
(Web Analytics Elective)

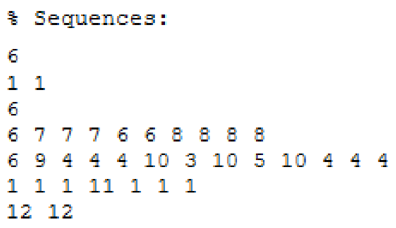
Web Usage Mining

Association Mining Familarisation Workshop

Dr. Barry Shepherd
Institute of Systems Science
National University of Singapore
Email: barryshepherd@nus.edu.sg

Workshop Goal

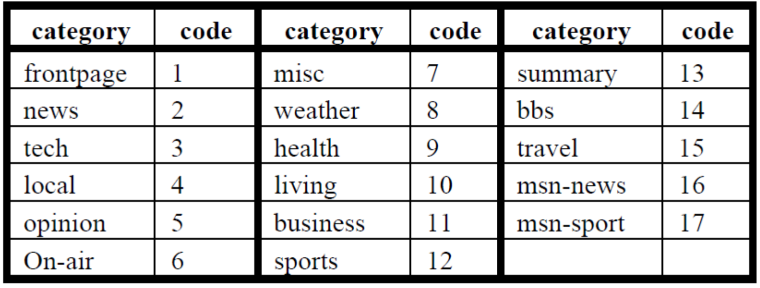
- Use SPSS Modeler and or R/Rattle to detect pairs or triplets of MSNBC webpages that are commonly visited by the same user in one day
- Contrast results with those obtained using the SPSS Modeler sequence mining node and/or the R "Spade" algorithm implementation



```
% Sequences:

6
1 1
6
6 7 7 7 6 6 8 8 8 8
6 9 4 4 4 10 3 10 5 10 4 4 4
1 1 1 11 1 1 1
12 12
```

The data shows the categories of all MSNBC webpages viewed by users on one specific day

Codes for the msnbc.com page categories



category	code	category	code	category	code
frontpage	1	misc	7	summary	13
news	2	weather	8	bbs	14
tech	3	health	9	travel	15
local	4	living	10	msn-news	16
opinion	5	business	11	msn-sport	17
On-air	6	sports	12		

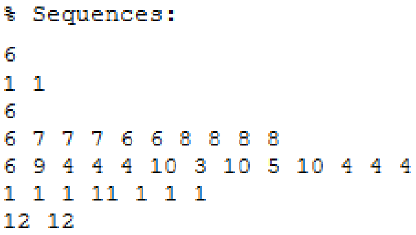
Tools we can use

- Association Mining
  - SPSS Modeler ~ classroom 2-1 & 3-12, breakout rooms
  - R (or Rattle)
  - *Weka*
  - *RapidMiner*
- Sequence Mining
  - SPSS Modeler
  - Spade (or other) library in R

Data Formats for Association Finding

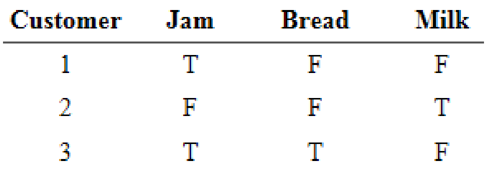
In the raw data (as downloaded from the web) each record contains the MSNBC web page categories visited by one user on one day. *(Approx. 1 million records).*

This must be converted to one of the two common formats that association rule tools accept. (For this workshop I have already converted to transaction format)



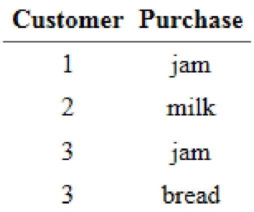
```
% Sequences:
6
1 1
6
6 7 7 7 6 6 8 8 8 8
6 9 4 4 4 10 3 10 5 10 4 4 4
1 1 1 11 1 1 1
12 12
```

<u>Tabular format</u>: rows represents item-sets. Each item is a separate column



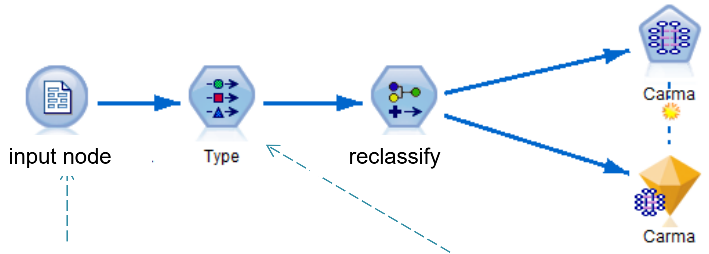
Customer	Jam	Bread	Milk
1	T	F	F
2	F	F	T
3	T	T	F

<u>Transaction format</u>: each row is a single item. An item-set id is required for each row.



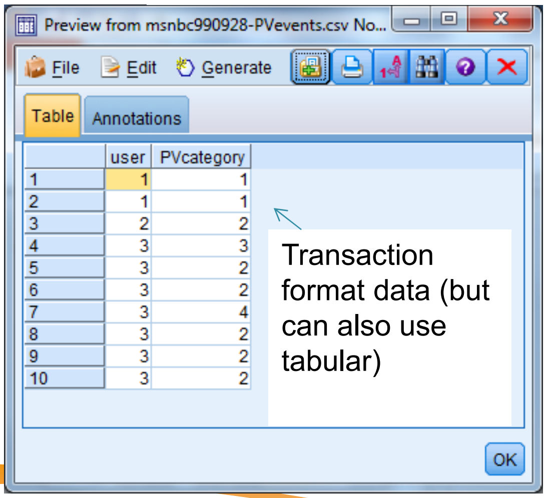
Customer	Purchase
1	jam
2	milk
3	jam
3	bread

Association finding using SPSS Modeler



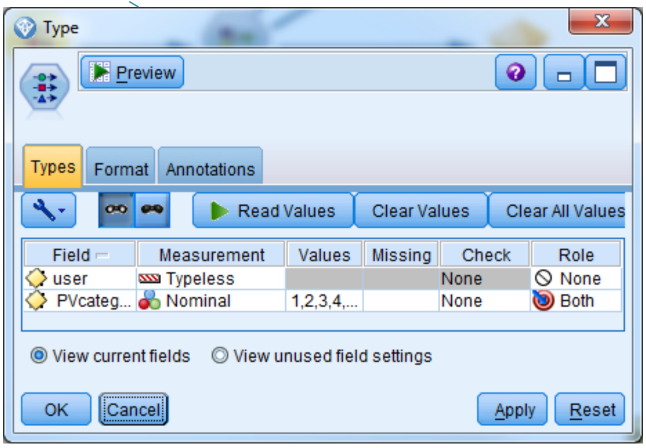
The association finding algorithm

The found association rules



	user	PVcategory
1	1	1
2	1	1
3	2	2
4	3	3
5	3	2
6	3	2
7	3	4
8	3	2
9	3	2
10	3	2

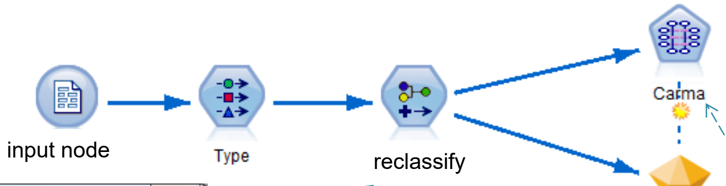
Transaction format data (but can also use tabular)

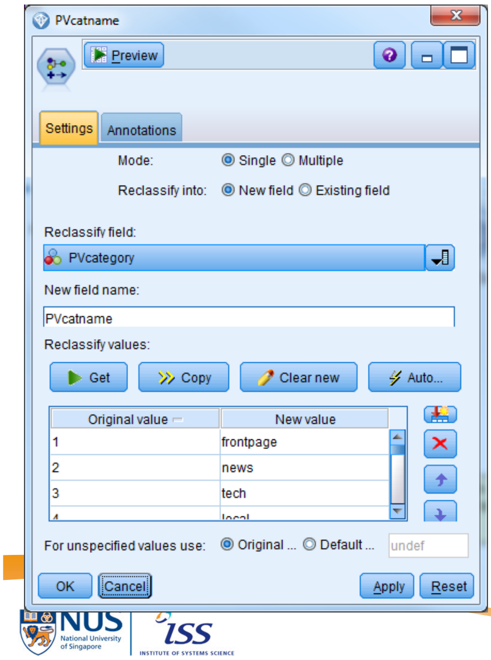


Field	Measurement	Values	Missing	Check	Role
user	Typeless			None	None
PVcateg...	Nominal	1,2,3,4,...		None	Both

● View current fields ○ View unused field settings

Association finding using SPSS Modeler





Mode: ● Single ○ Multiple

Reclassify into: ● New field ○ Existing field

Reclassify field: PVcategory

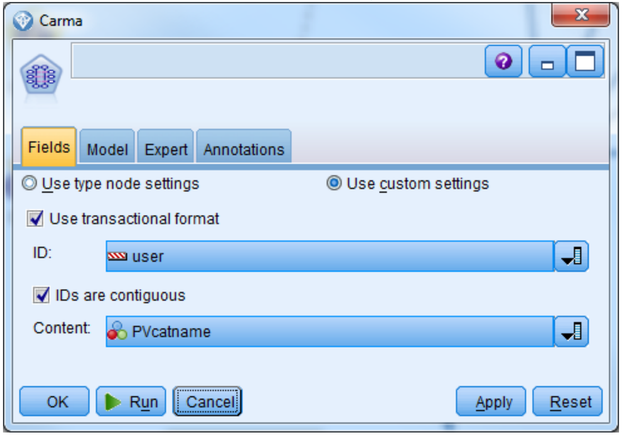
New field name: PVcatname

Reclassify values:

Original value	New value
1	frontpage
2	news
3	tech
4	local

For unspecified values use: ● Original ... ○ Default ... undef

Use a reclassify node to turn the category codes (integers) into more meaningful strings



○ Use type node settings ● Use custom settings

☑ Use transactional format

ID: user

☑ IDs are contiguous

Content: PVcatname

Edit the association node (Carma or Apriori) to accept transaction format data

Association algorithms in SPSS

Apriori

- Generally faster than Carma
- Input and target fields must be symbolic.
- Set fields as “both” if not sure which should be target or input

  If coke then ice

  If temp=cold then ice

  If temp=cold then buy=coat

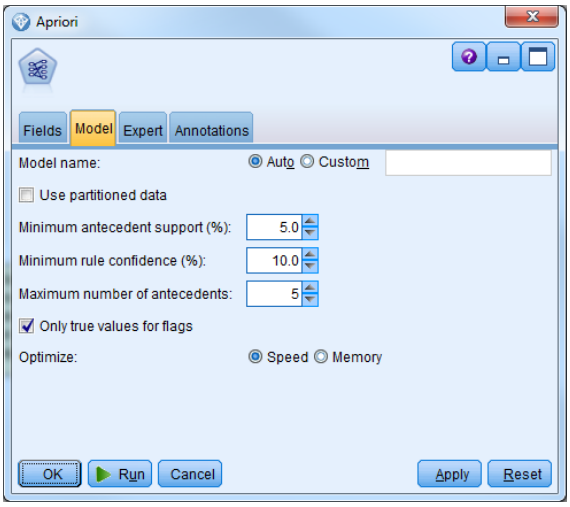
Carma

- Does not require you to define input and output fields.
- All fields should be boolean
- Can generate rules with multiple consequents

  If whiskey then ice and coke

Both accept tabular or transaction data. The rules generated using each format are identical

Building Assoc. Rules in SPSS



Model name: ● Auto ○ Custom

□ Use partitioned data

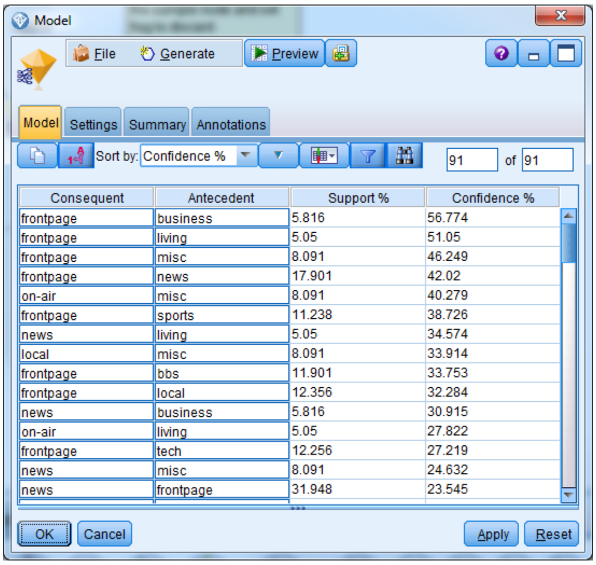
Minimum antecedent support (%): 5.0

Minimum rule confidence (%): 10.0

Maximum number of antecedents: 5

☑ Only true values for flags

Optimize: ● Speed ○ Memory



Sort by: Confidence %

91 of 91

Consequent	Antecedent	Support %	Confidence %
frontpage	business	5.816	56.774
frontpage	living	5.05	51.05
frontpage	misc	8.091	46.249
frontpage	news	17.901	42.02
on-air	misc	8.091	40.279
frontpage	sports	11.238	38.726
news	living	5.05	34.574
local	misc	8.091	33.914
frontpage	bbs	11.901	33.753
frontpage	local	12.356	32.284
news	business	5.816	30.915
on-air	living	5.05	27.822
frontpage	tech	12.256	27.219
news	misc	8.091	24.632
news	frontpage	31.948	23.545

Apriori

Edit the model node before executing, you may have to reduce the expected rule support and confidence

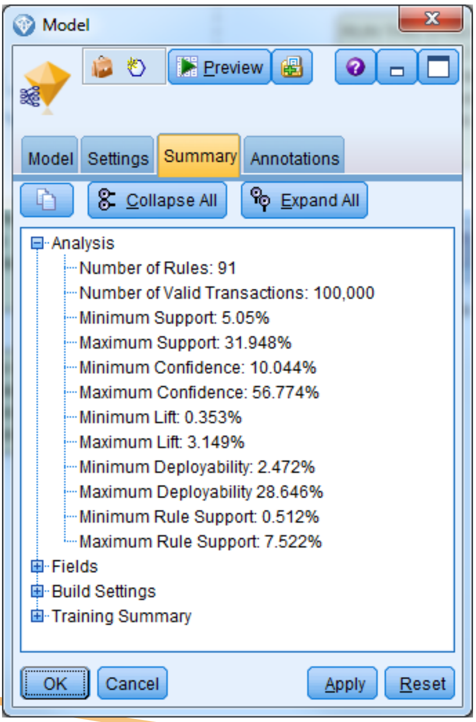
Model

Edit the nugget node to view the built rule set and to set the execution settings (see next slides)

Building Assoc. Rules in SPSS

- The summary tab in the nugget node shows information about the built ruleset

Model



Analysis
- Number of Rules: 91
- Number of Valid Transactions: 100,000
- Minimum Support: 5.05%
- Maximum Support: 31.948%
- Minimum Confidence: 10.044%
- Maximum Confidence: 56.774%
- Minimum Lift: 0.353%
- Maximum Lift: 3.149%
- Minimum Deployability: 2.472%
- Maximum Deployability 28.646%
- Minimum Rule Support: 0.512%
- Maximum Rule Support: 7.522%

Fields

Build Settings

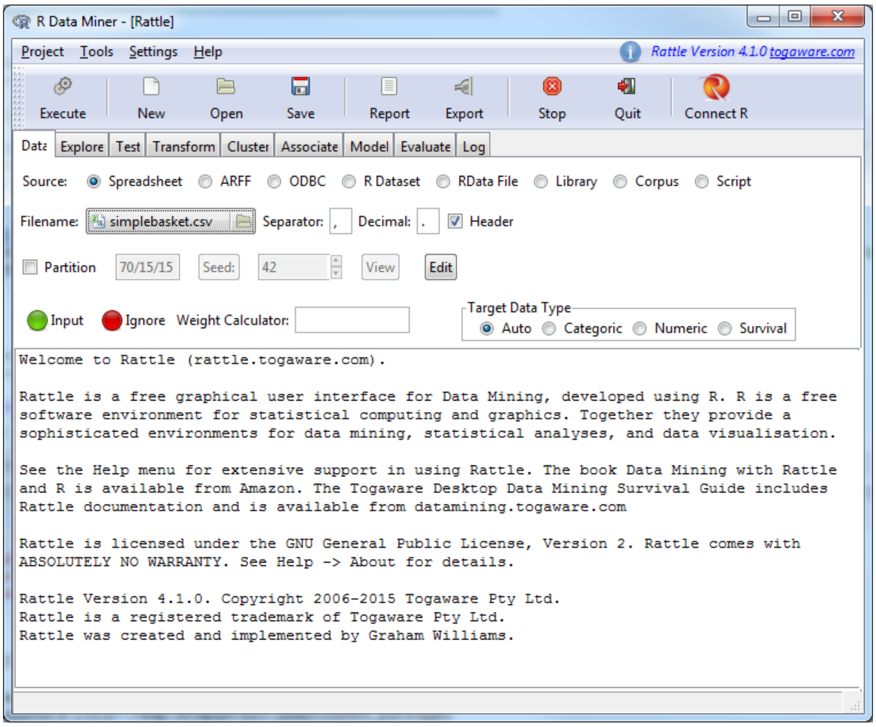
Training Summary

Association Mining Using R/Rattle

In R…

library(rattle)

rattle()



Source: ● Spreadsheet ○ ARFF ○ ODBC ○ R Dataset ○ RData File ○ Library ○ Corpus ○ Script

Filename: simplebasket.csv Separator: , Decimal: . ☑ Header

□ Partition 70/15/15 Seed: 42

Target Data Type: ● Auto ○ Categoric ○ Numeric ○ Survival

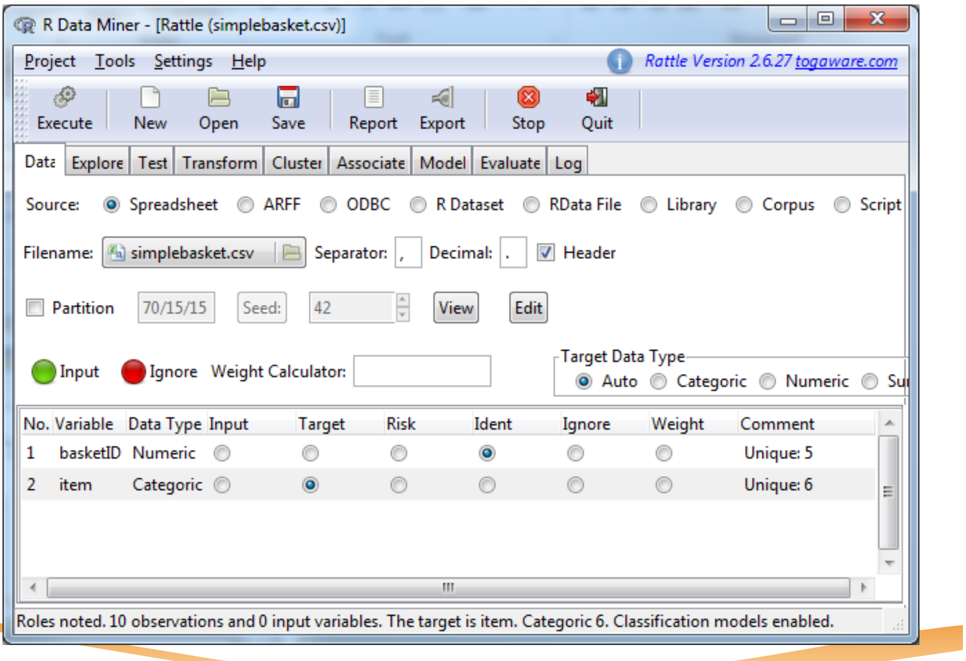
```
Welcome to Rattle (rattle.togaware.com).

Rattle is a free graphical user interface for Data Mining, developed using R. R is a free
software environment for statistical computing and graphics. Together they provide a
sophisticated environments for data mining, statistical analyses, and data visualisation.

See the Help menu for extensive support in using Rattle. The book Data Mining with Rattle
and R is available from Amazon. The Togaware Desktop Data Mining Survival Guide includes
Rattle documentation and is available from datamining.togaware.com

Rattle is licensed under the GNU General Public License, Version 2. Rattle comes with
ABSOLUTELY NO WARRANTY. See Help -> About for details.

Rattle Version 4.1.0. Copyright 2006-2015 Togaware Pty Ltd.
Rattle is a registered trademark of Togaware Pty Ltd.
Rattle was created and implemented by Graham Williams.
```

Using R/Rattle

- Click **Execute** to import the training data, then select the basket identifier (Ident) and set the items as target. Ensure **Partition** is deselected.

basketID	item
1	A
1	B
1	C
2	A
2	B
3	A
3	D
4	D
4	E
5	F

Simplebasket.csv



R Data Miner - [Rattle (simplebasket.csv)]
Project Tools Settings Help
Rattle Version 2.6.27 togaware.com
Execute New Open Save Report Export Stop Quit
Data Explore Test Transform Cluster Associate Model Evaluate Log
View Edit

Source: ● Spreadsheet ○ ARFF ○ ODBC ○ R Dataset ○ RData File ○ Library ○ Corpus ○ Script

Filename: simplebasket.csv Separator: , Decimal: . ☑ Header

□ Partition 70/15/15 Seed: 42

● Input ● Ignore Weight Calculator:

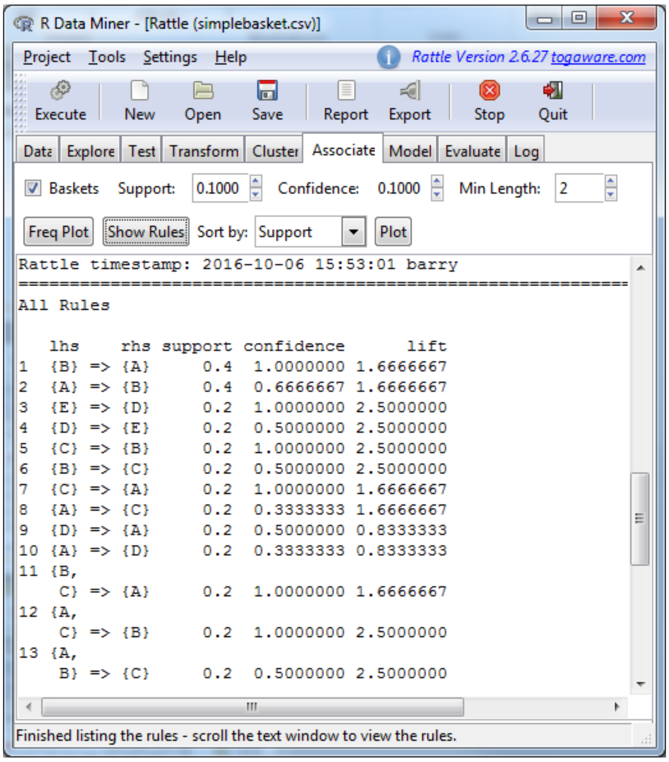
Target Data Type: ● Auto ○ Categoric ○ Numeric ○ Su

No.	Variable	Data Type	Input	Target	Risk	Ident	Ignore	Weight	Comment
1	basketID	Numeric	○	○	○	●	○	○	Unique: 5
2	item	Categoric	○	●	○	○	○	○	Unique: 6

Roles noted. 10 observations and 0 input variables. The target is item. Categoric 6. Classification models enabled.

Using R/Rattle

- Go to the Associate tab
- Select ***Baskets***, then click ***Execute*** and then ***Show rules***
- Scroll down to see the rules



R Data Miner - [Rattle (simplebasket.csv)]
Project Tools Settings Help
Rattle Version 2.6.27 togaware.com
Execute New Open Save Report Export Stop Quit
Data Explore Test Transform Cluster Associate Model Evaluate Log
Freq Plot Show Rules Plot

☑ Baskets Support: 0.1000 Confidence: 0.1000 Min Length: 2

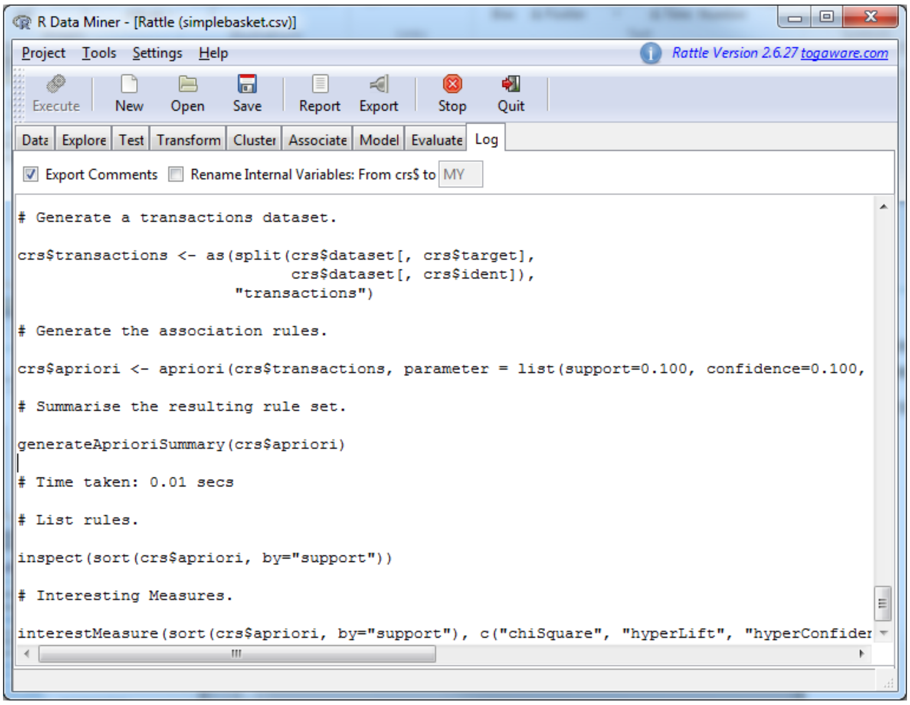
Sort by: Support

```
Rattle timestamp: 2016-10-06 15:53:01 barry
======================================================
All Rules

   lhs    rhs support confidence      lift
1  {B} => {A}     0.4  1.0000000 1.6666667
2  {A} => {B}     0.4  0.6666667 1.6666667
3  {E} => {D}     0.2  1.0000000 2.5000000
4  {D} => {E}     0.2  0.5000000 2.5000000
5  {C} => {B}     0.2  1.0000000 2.5000000
6  {B} => {C}     0.2  0.5000000 2.5000000
7  {C} => {A}     0.2  1.0000000 1.6666667
8  {A} => {C}     0.2  0.3333333 1.6666667
9  {D} => {A}     0.2  0.5000000 0.8333333
10 {A} => {D}     0.2  0.3333333 0.8333333
11 {B,
    C} => {A}     0.2  1.0000000 1.6666667
12 {A,
    C} => {B}     0.2  1.0000000 2.5000000
13 {A,
    B} => {C}     0.2  0.5000000 2.5000000
```

Finished listing the rules - scroll the text window to view the rules.

Using R/Rattle

- Select the log tab to see the underlying R code that was executed


R Data Miner - [Rattle (simplebasket.csv)]
Project Tools Settings Help
Rattle Version 2.6.27 togaware.com
Execute New Open Save Report Export Stop Quit
Data Explore Test Transform Cluster Associate Model Evaluate Log

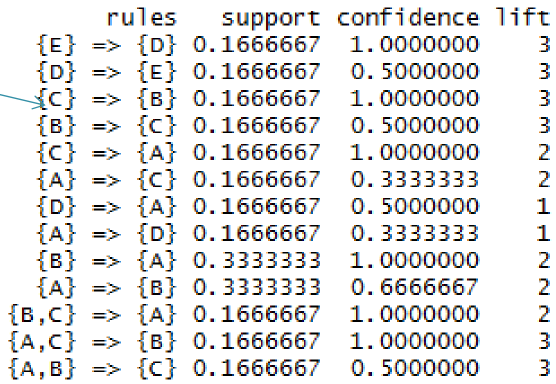
☑ Export Comments □ Rename Internal Variables: From crs$ to MY

```
# Generate a transactions dataset.

crs$transactions <- as(split(crs$dataset[, crs$target],
                             crs$dataset[, crs$ident]),
                       "transactions")

# Generate the association rules.

crs$apriori <- apriori(crs$transactions, parameter = list(support=0.100, confidence=0.100,

# Summarise the resulting rule set.

generateAprioriSummary(crs$apriori)

# Time taken: 0.01 secs

# List rules.

inspect(sort(crs$apriori, by="support"))

# Interesting Measures.

interestMeasure(sort(crs$apriori, by="support"), c("chiSquare", "hyperLift", "hyperConfider
```

Using R Directly

library("arules");

# for transaction format data
egs = read.transactions(file=*filename*,rm.duplicates=TRUE,format="single",sep=",",cols=c(1,2));
rules = apriori(egs, parameter = list(supp=0.1, conf=0.1, minlen=2))
summary(rules)
inspect(rules)
as(rules,”data.frame”)



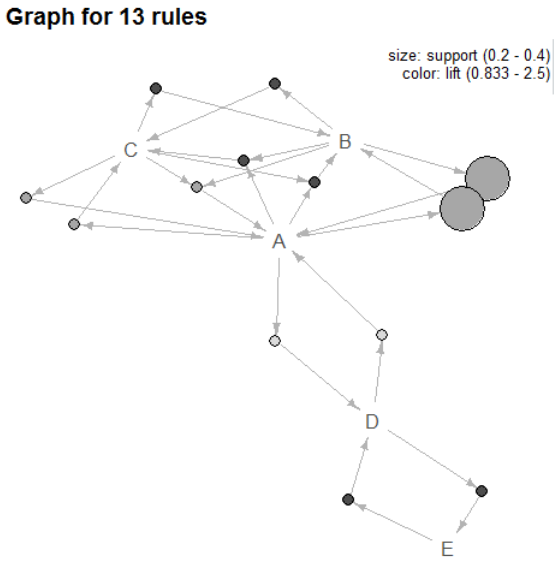
```
         rules    support confidence lift
  {E} => {D} 0.1666667  1.0000000    3
  {D} => {E} 0.1666667  0.5000000    3
  {C} => {B} 0.1666667  1.0000000    3
  {B} => {C} 0.1666667  0.5000000    3
  {C} => {A} 0.1666667  1.0000000    2
  {A} => {C} 0.1666667  0.3333333    2
  {D} => {A} 0.1666667  0.5000000    1
  {A} => {D} 0.1666667  0.3333333    1
  {B} => {A} 0.3333333  1.0000000    2
  {A} => {B} 0.3333333  0.6666667    2
{B,C} => {A} 0.1666667  1.0000000    2
{A,C} => {B} 0.1666667  1.0000000    3
{A,B} => {C} 0.1666667  0.5000000    3
```

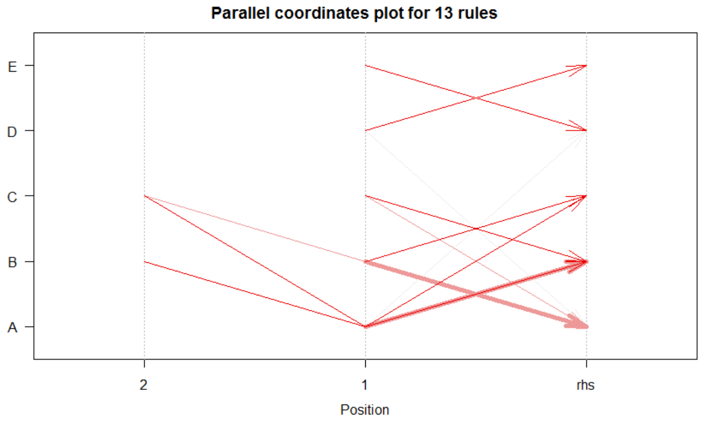
Also try…

itemsets <- eclat(egs, parameter = list(supp = 0.01, maxlen = 5))
rules2 <- ruleInduction(itemsets, egs, confidence = .1)

Visualising Rules

```
library(arulesViz)
plot(rules)
plot(rules, method="graph",nodeCol=grey.colors(10),edgeCol=grey(.7),alpha=1)
plot(rules, method="paracoord", control=list(reorder=TRUE))
```





https://cran.r-project.org/web/packages/arulesViz/vignettes/arulesViz.pdf
https://cran.r-project.org/web/packages/arulesViz/arulesViz.pdf

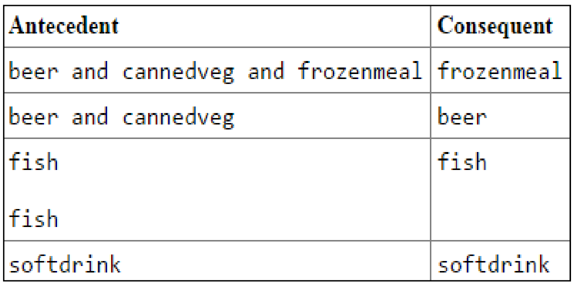
Sequence finding using SPSS Modeler

- Use the Sequence modelling node (based on Carma algorithm)
- Finds sequences of item-sets.
  - e.g. {coffee, milk, sugar} => {bread , butter}
- The item-set can be a row in tabular format data or a single transaction in transaction format data.
- Requires an item-set ID field PLUS a timestamp field for each item-set.
  - If no timestamp field is given then it uses the row number to indicate the sequence (assumes rows are in temporal order in the database)
  - e.g. (userID, date-time, news)
    (userID, date-time, sports)
    (userID, date-time, finance) …. *for transaction format*
  - e.g. (userID, date-time, news, sports, finance) …*for tabular format*
- Variable types must be specified in the Sequence node (not Type node)

Sequence

Examining SPSS Modeler Sequence Rules

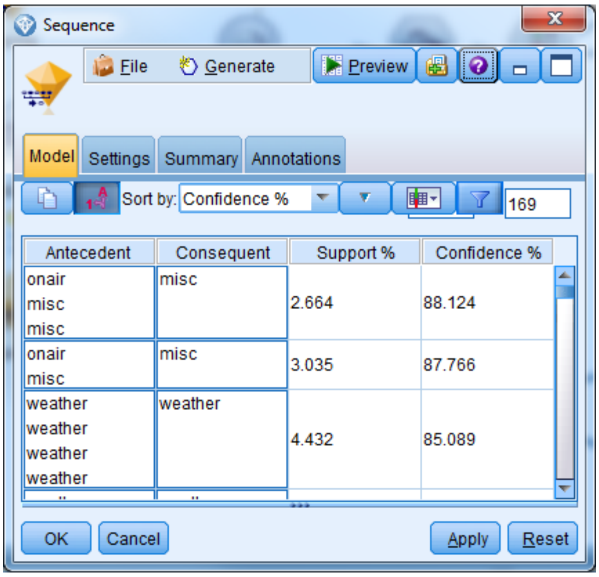
- Sequence rules in modeler look like:



Antecedent	Consequent
beer and cannedveg and frozenmeal	frozenmeal
beer and cannedveg	beer
fish fish	fish
softdrink	softdrink

fish->fish->fish (not very useful)

For example:

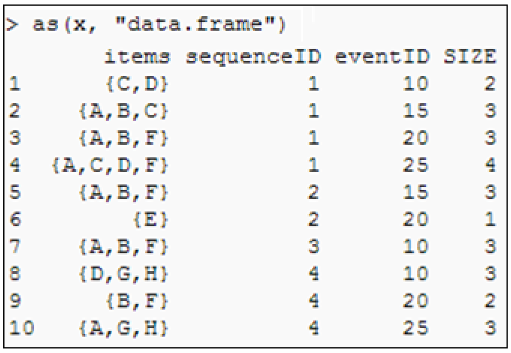


Antecedent	Consequent	Support %	Confidence %
onair misc misc	misc	2.664	88.124
onair misc	misc	3.035	87.766
weather weather weather weather	weather	4.432	85.089

Sequence finding using R

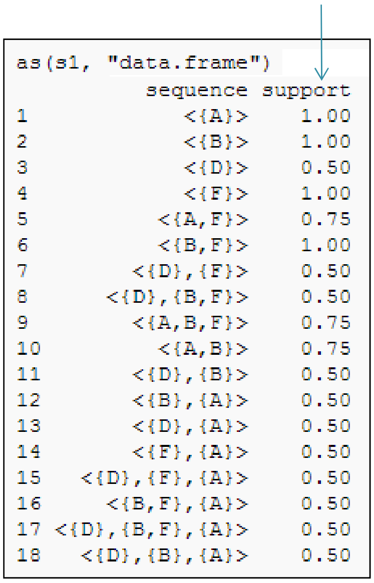
- The Spade algorithm is a popular sequence mining algorithm

```
library(arules)
library(arulesSequences)
x <- read_baskets(con=system.file("misc", "zaki.txt",
     package = "arulesSequences"),info =c("sequenceID","eventID","SIZE"))
s1 <- cspade(x, parameter = list(support = 0.4), control = list(verbose = TRUE))
as(s1, “data.frame”)
```



```
> as(x, "data.frame")
      items sequenceID eventID SIZE
1     {C,D}          1      10    2
2   {A,B,C}          1      15    3
3   {A,B,F}          1      20    3
4 {A,C,D,F}          1      25    4
5   {A,B,F}          2      15    3
6       {E}          2      20    1
7   {A,B,F}          3      10    3
8   {D,G,H}          4      10    3
9     {B,F}          4      20    2
10  {A,G,H}          4      25    3
```

# sequences found
#total sequences (i.e. 4)



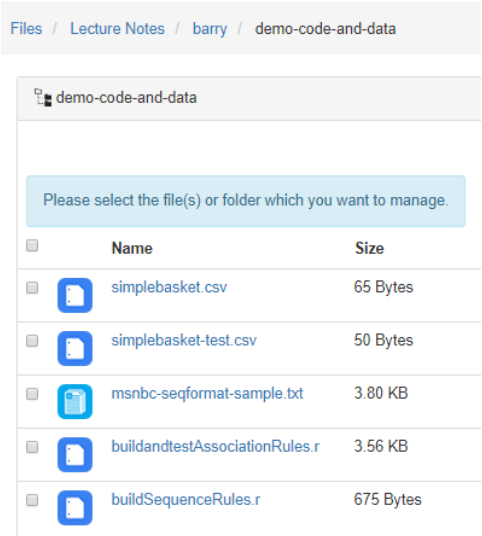
```
as(s1, "data.frame")
           sequence support
1             <{A}>    1.00
2             <{B}>    1.00
3             <{D}>    0.50
4             <{F}>    1.00
5           <{A,F}>    0.75
6           <{B,F}>    1.00
7         <{D},{F}>    0.50
8       <{D},{B,F}>    0.50
9         <{A,B,F}>    0.75
10          <{A,B}>    0.75
11        <{D},{B}>    0.50
12        <{B},{A}>    0.50
13        <{D},{A}>    0.50
14        <{F},{A}>    0.50
15    <{D},{F},{A}>    0.50
16      <{B,F},{A}>    0.50
17 <{D},{B,F},{A}>    0.50
18    <{D},{B},{A}>    0.50
```

Note: the read_baskets() above is not necessary since this dataset is preloaded with the arulesSequences library and called zaki. Use data(“zaki”) then inspect(zaki) to view it.

http://en.wikibooks.org/wiki/Data_Mining_Algorithms_In_R/Sequence_Mining/SPADE

Location of data files

Sample code to explore



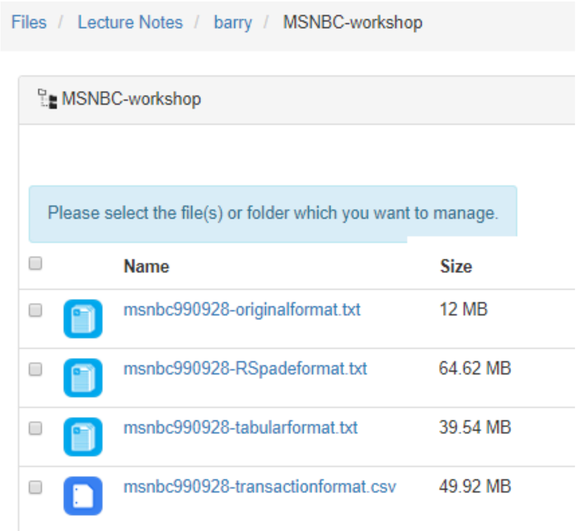
Files / Lecture Notes / barry / demo-code-and-data

demo-code-and-data

Please select the file(s) or folder which you want to manage.

□	Name	Size
□	simplebasket.csv	65 Bytes
□	simplebasket-test.csv	50 Bytes
□	msnbc-seqformat-sample.txt	3.80 KB
□	buildandtestAssociationRules.r	3.56 KB
□	buildSequenceRules.r	675 Bytes

The MSNBC data - I have converted into the various formats already



Files / Lecture Notes / barry / MSNBC-workshop

MSNBC-workshop

Please select the file(s) or folder which you want to manage.

□	Name	Size
□	msnbc990928-originalformat.txt	12 MB
□	msnbc990928-RSpadeformat.txt	64.62 MB
□	msnbc990928-tabularformat.txt	39.54 MB
□	msnbc990928-transactionformat.csv	49.92 MB