Outline

- Association rules
- Create a descriptive data mining solution using Microsoft association rules.

Association Rules

- Rule-based machine learning method for discovering interesting relations between variables in large databases.
- Intended to identify strong rules discovered in databases using some measures of interestingness.
- Based on the concept of strong rules, Rakesh Agrawal, Tomasz Imieliński and Arun Swami [2] introduced association rules for discovering regularities between products in large-scale transaction data recorded by point-of-sale (POS) systems in supermarkets.
- For example, the rule {onions,potatoes} => {burger} found in the sales data of a supermarket would indicate that if a customer buys onions and potatoes together, they are likely to also buy hamburger meat.

Association Rules & Frequent Itemsets

Transactions

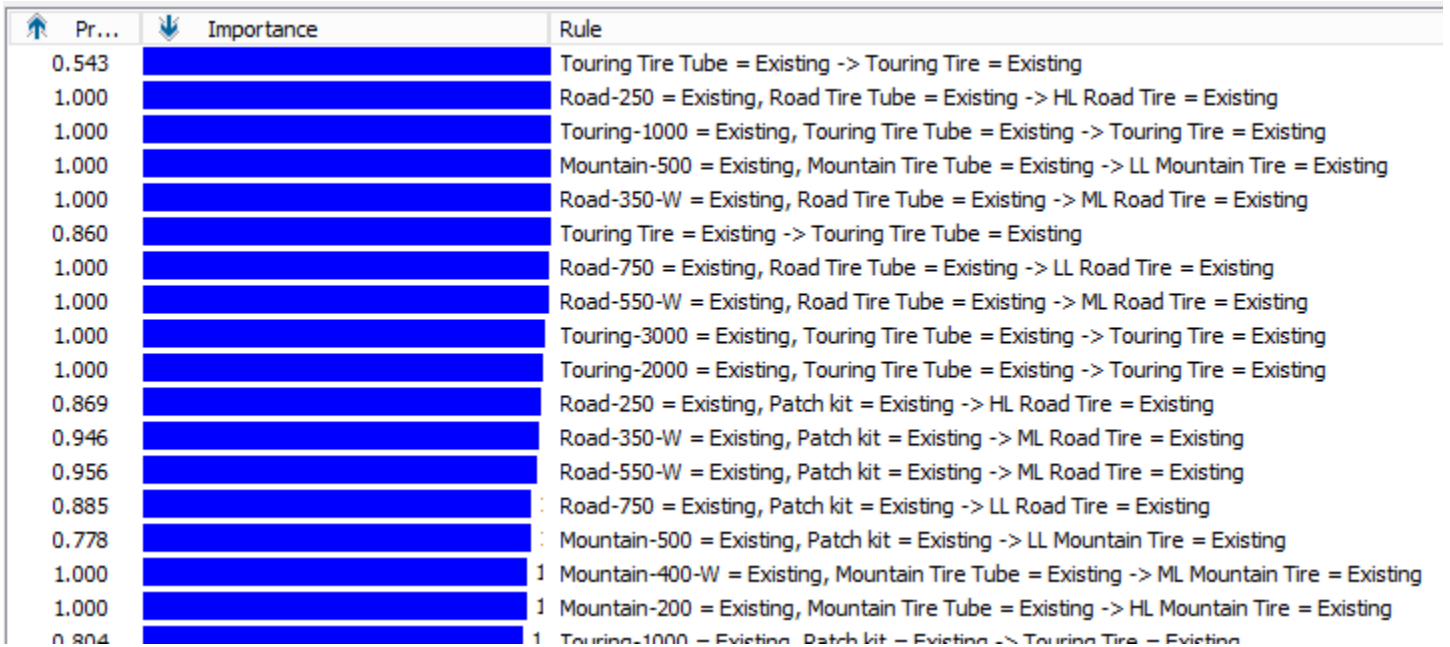
TID	Produce
1	MILK, BREAD, EGGS
2	BREAD, SUGAR
3	BREAD, CEREAL
4	MILK, BREAD, SUGAR
5	MILK, CEREAL
6	BREAD, CEREAL
7	MILK, CEREAL
8	MILK, BREAD, CEREAL, EGGS
9	MILK, BREAD, CEREAL

Frequent Itemsets:

~~Milk, Bread (4)~~
Bread, Cereal (3)
Milk, Bread, Cereal (2)
…

Rules:
Milk => Bread (66%)

Rules



Pr...	Importance	Rule
0.543		Touring Tire Tube = Existing -> Touring Tire = Existing
1.000		Road-250 = Existing, Road Tire Tube = Existing -> HL Road Tire = Existing
1.000		Touring-1000 = Existing, Touring Tire Tube = Existing -> Touring Tire = Existing
1.000		Mountain-500 = Existing, Mountain Tire Tube = Existing -> LL Mountain Tire = Existing
1.000		Road-350-W = Existing, Road Tire Tube = Existing -> ML Road Tire = Existing
0.860		Touring Tire = Existing -> Touring Tire Tube = Existing
1.000		Road-750 = Existing, Road Tire Tube = Existing -> LL Road Tire = Existing
1.000		Road-550-W = Existing, Road Tire Tube = Existing -> ML Road Tire = Existing
1.000		Touring-3000 = Existing, Touring Tire Tube = Existing -> Touring Tire = Existing
1.000		Touring-2000 = Existing, Touring Tire Tube = Existing -> Touring Tire = Existing
0.869		Road-250 = Existing, Patch kit = Existing -> HL Road Tire = Existing
0.946		Road-350-W = Existing, Patch kit = Existing -> ML Road Tire = Existing
0.956		Road-550-W = Existing, Patch kit = Existing -> ML Road Tire = Existing
0.885		Road-750 = Existing, Patch kit = Existing -> LL Road Tire = Existing
0.778		Mountain-500 = Existing, Patch kit = Existing -> LL Mountain Tire = Existing
1.000	1	Mountain-400-W = Existing, Mountain Tire Tube = Existing -> ML Mountain Tire = Existing
1.000	1	Mountain-200 = Existing, Mountain Tire Tube = Existing -> HL Mountain Tire = Existing
0.804	1	Touring-1000 = Existing, Patch kit = Existing -> Touring Tire = Existing

Itemset and Support count

- Support is an indication of how frequently the itemset appears in the dataset.
- For two items X, Y the support parameter is the number of cases in the dataset that contain the combination of items, X and Y.
- Size – number of items in set

$$\text{Support, } s(X \longrightarrow Y) = \frac{\sigma(X \cup Y)}{N};$$

$$\text{Confidence, } c(X \longrightarrow Y) = \frac{\sigma(X \cup Y)}{\sigma(X)}.$$

Calculating support

Table 6.1. An example of market basket transactions.

TID	Items
1	{Bread, Milk}
2	{Bread, Diapers, Beer, Eggs}
3	{Milk, Diapers, Beer, Cola}
4	{Bread, Milk, Diapers, Beer}
5	{Bread, Milk, Diapers, Cola}

- Consider the rule {Milk, Diapers} -> {Beer}.
- Since the support count for {Milk, Diapers, Beer} is 2 and the total number of transactions is 5, the rule's support is 2/5 = 0.4

Confidence

- Probability parameter, also named *confidence*, represents the fraction of cases in the dataset that contain X and that also contain Y.
- Importance measures the usefulness of the rule.
- How can we measure the usefulness of a rule?

  it is unexpected (surprising to the user); and/or

  actionable (the user can do something with it)

Calculating Confidence

- Confidence is obtained by dividing the support count for {Milk, Diapers, Beer} by the support count for {Milk, Diapers},
- Since there are 3 transactions that contain milk and diapers, the confidence for this rule is 2/3 : 0.67

Itemsets

Support	S.	Itemset
589	3	Mountain Bottle Cage = Existing, Mountain-200 = Existing, Water Bottle = Existing
485	3	Road-750 = Existing, Road Bottle Cage = Existing, Water Bottle = Existing
446	3	Mountain Bottle Cage = Existing, Water Bottle = Existing, Sport-100 = Existing
383	3	HL Mountain Tire = Existing, Mountain Tire Tube = Existing, Sport-100 = Existing
376	3	Road Bottle Cage = Existing, Water Bottle = Existing, Sport-100 = Existing
333	3	ML Mountain Tire = Existing, Mountain Tire Tube = Existing, Sport-100 = Existing
331	3	HL Mountain Tire = Existing, Mountain-200 = Existing, Mountain Tire Tube = Existing
309	3	Touring-1000 = Existing, Road Bottle Cage = Existing, Water Bottle = Existing
307	3	Mountain Bottle Cage = Existing, Fender Set - Mountain = Existing, Water Bottle = Existing
236	3	Touring Tire = Existing, Touring Tire Tube = Existing, Sport-100 = Existing
229	3	Mountain Bottle Cage = Existing, Cycling Cap = Existing, Water Bottle = Existing
228	3	HL Mountain Tire = Existing, Mountain Tire Tube = Existing, Patch kit = Existing
222	3	Road Bottle Cage = Existing, Cycling Cap = Existing, Water Bottle = Existing
221	3	ML Road Tire = Existing, Road Tire Tube = Existing, Sport-100 = Existing
220	3	Mountain Bottle Cage = Existing, Mountain-200 = Existing, Sport-100 = Existing
203	3	Touring Tire = Existing, Touring Tire Tube = Existing, Patch kit = Existing

Potential Rules governing the association

Rule
Road Bottle Cage = Existing, Cycling Cap = Existing -> Water Bottle = Existing
Mountain-200 = Existing, Mountain Tire Tube = Existing -> HL Mountain Tire = Existing
Mountain-200 = Existing, Water Bottle = Existing -> Mountain Bottle Cage = Existing
Touring-1000 = Existing, Water Bottle = Existing -> Road Bottle Cage = Existing
Road-750 = Existing, Water Bottle = Existing -> Road Bottle Cage = Existing
Touring Tire = Existing, Sport-100 = Existing -> Touring Tire Tube = Existing

Challenges of analyzing Associations

- Even for a small number of items there is a huge number of possibilities.

Table 6.2. A binary $0/1$ representation of market basket data.

TID	Bread	Milk	Diapers	Beer	Eggs	Cola
1	1	1	0	0	0	0
2	1	0	1	1	1	0
3	0	1	1	1	0	1
4	1	1	1	1	0	0
5	1	1	1	0	0	1

In general, the number of combinations is given by the formula below

$$R = 3^d - 2^{d+1} + 1.$$

For 6 items we have 602 rules

Apriori Algorithm

- Is monotonic (decreasing) leads to smaller list of transaction thru every pass of the data.
- Prunes the original dataset
- Good for analyzing large datasets

Apriori Algorithm

alpha	beta	epsilon	
alpha	beta	theta	a
alpha	beta	epsilon	a
alpha	beta	theta	b

Consider the following database, where each row is a transaction and each cell is an individual item of the transaction:
The association rules that can be determined from this database are the following:
1. 100% of sets with alpha also contain beta
2. 50% of sets with alpha, beta also have epsilon
3. 50% of sets with alpha, beta also have theta

following itemsets:

Transactions in the database

- The numbers are sku numbers of the products

Itemsets
{1,2,3,4}
{1,2,4}
{1,2}
{2,3,4}
{2,3}
{3,4}
{2,4}

Bottom up approach

- Counting sets of size 1 , with a threshold (support) of 3

Item	Support
{1}	3
{2}	6
{3}	4
{4}	5

Apriori cont’d

- Form pairs and find frequency, select the ones with support >= 3

Item	Support
{1,2}	3
{1,3}	1
{1,4}	2
{2,3}	3
{2,4}	4
{3,4}	3

Apriori

- Look for triples – nothing above the threshold.

Item	Support
{2,3,4}	2

Number of possible rules for size - d

$$R = 3^d - 2^{d+1} + 1.$$

TID	Items
1	{Bread, Milk}
2	{Bread, Diapers, Beer, Eggs}
3	{Milk, Diapers, Beer, Cola}
4	{Bread, Milk, Diapers, Beer}
5	{Bread, Milk, Diapers, Cola}

- For the above table with 6 items
- – 602 rules

Apriori principle

If an itemset is frequent, than all of its subsets must also be frequent.

Apriori – Item Set Lattice

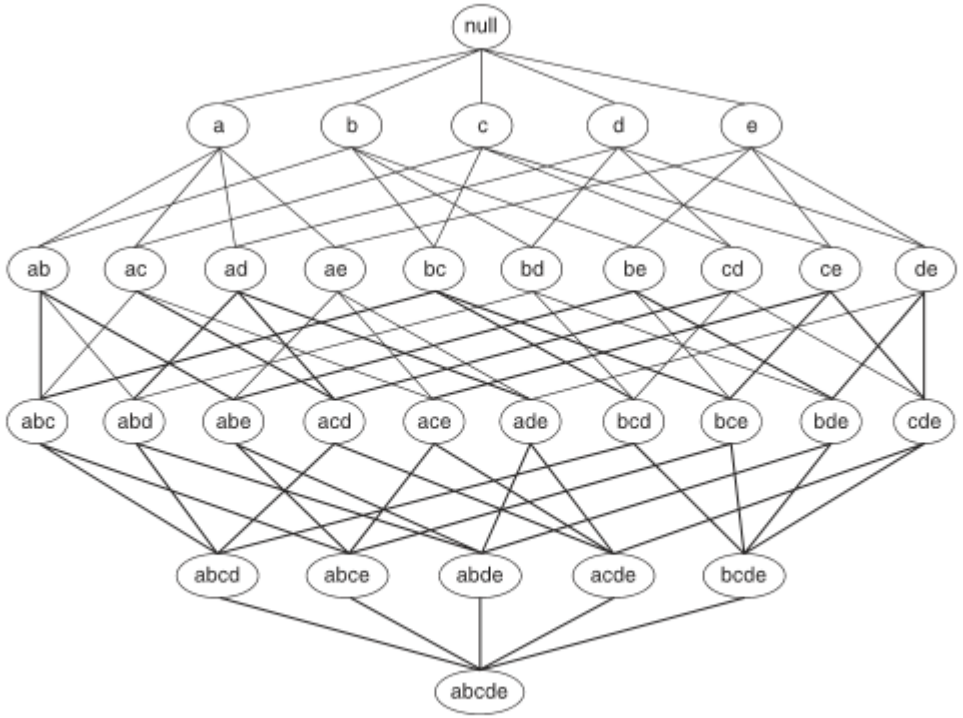


Figure 6.1. An itemset lattice.

Support based pruning

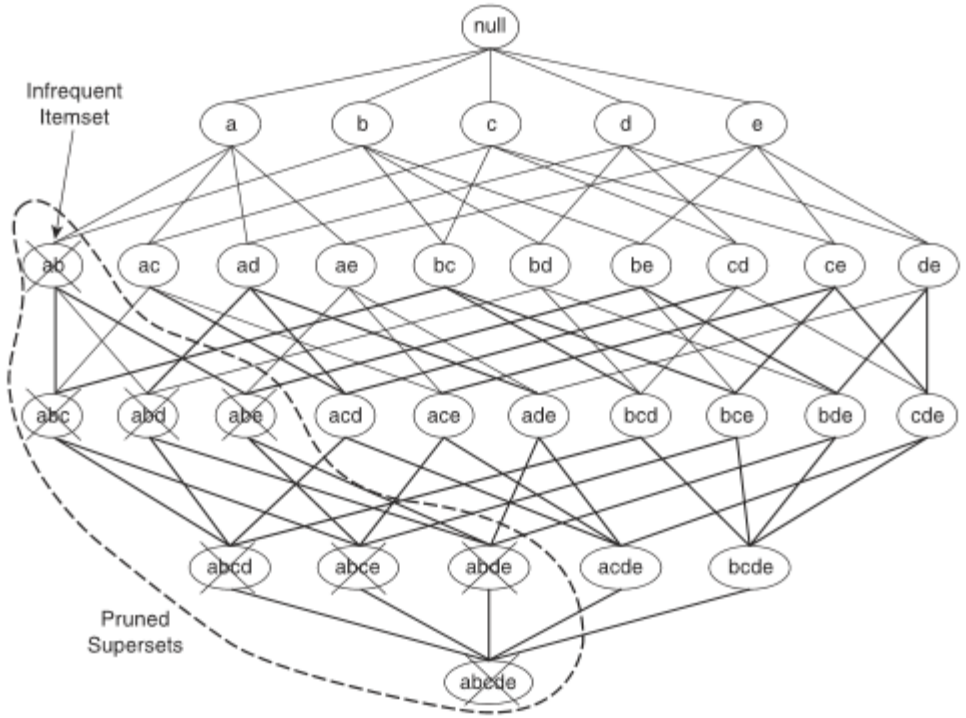
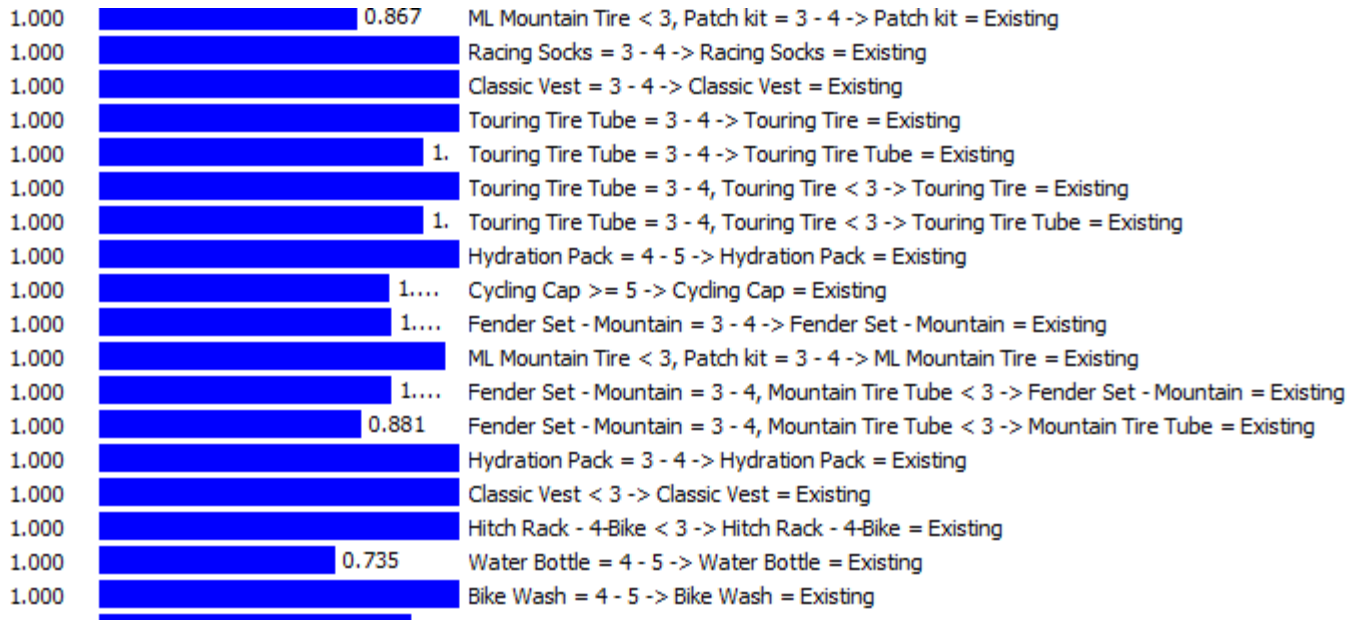


Figure 6.4. An illustration of support-based pruning. If $\{a, b\}$ is infrequent, then all supersets of $\{a, b\}$ are infrequent.

Association rule from exercise

- .543 – percentage of datasets containing Touring Tire Tube also contain Touring Tire.



Association rule from exercise

- .410 percentage of datasets containing Road-550-W < 3 also contain Sport-100

0.410	0.160	Road-550-W < 3 -> Sport-100 = Existing
0.414	0.158	Touring-2000 < 3 -> Sport-100 = Existing
0.415	0.783	Water Bottle = 3 - 4 -> Mountain Bottle Cage = Existing
0.419	0.164	LL Mountain Tire < 3, Mountain Tire Tube < 3 -> Sport-100 = Existing
0.424	0.171	HL Mountain Tire < 3, Mountain Tire Tube < 3 -> Sport-100 = Existing
0.428	0.198	Mountain Tire Tube < 3 -> Sport-100 = Existing
0.437	0.192	Touring-1000 < 3 -> Sport-100 = Existing
0.440	0.712	Sport-100 = 4 - 5, Water Bottle = 3 - 4 -> Mountain Bottle Cage = Existing
0.456	0.202	Half-Finger Gloves = 3 - 4 -> Sport-100 = Existing
0.471	0.219	ML Mountain Tire < 3, Mountain Tire Tube < 3 -> Sport-100 = Existing
0.472		Touring Tire Tube < 3 -> Touring Tire = Existing
0.478	0.223	Long-Sleeve Logo Jersey = 3 - 4 -> Sport-100 = Existing
0.488	0.747	LL Road Tire < 3 -> Road Tire Tube = Existing

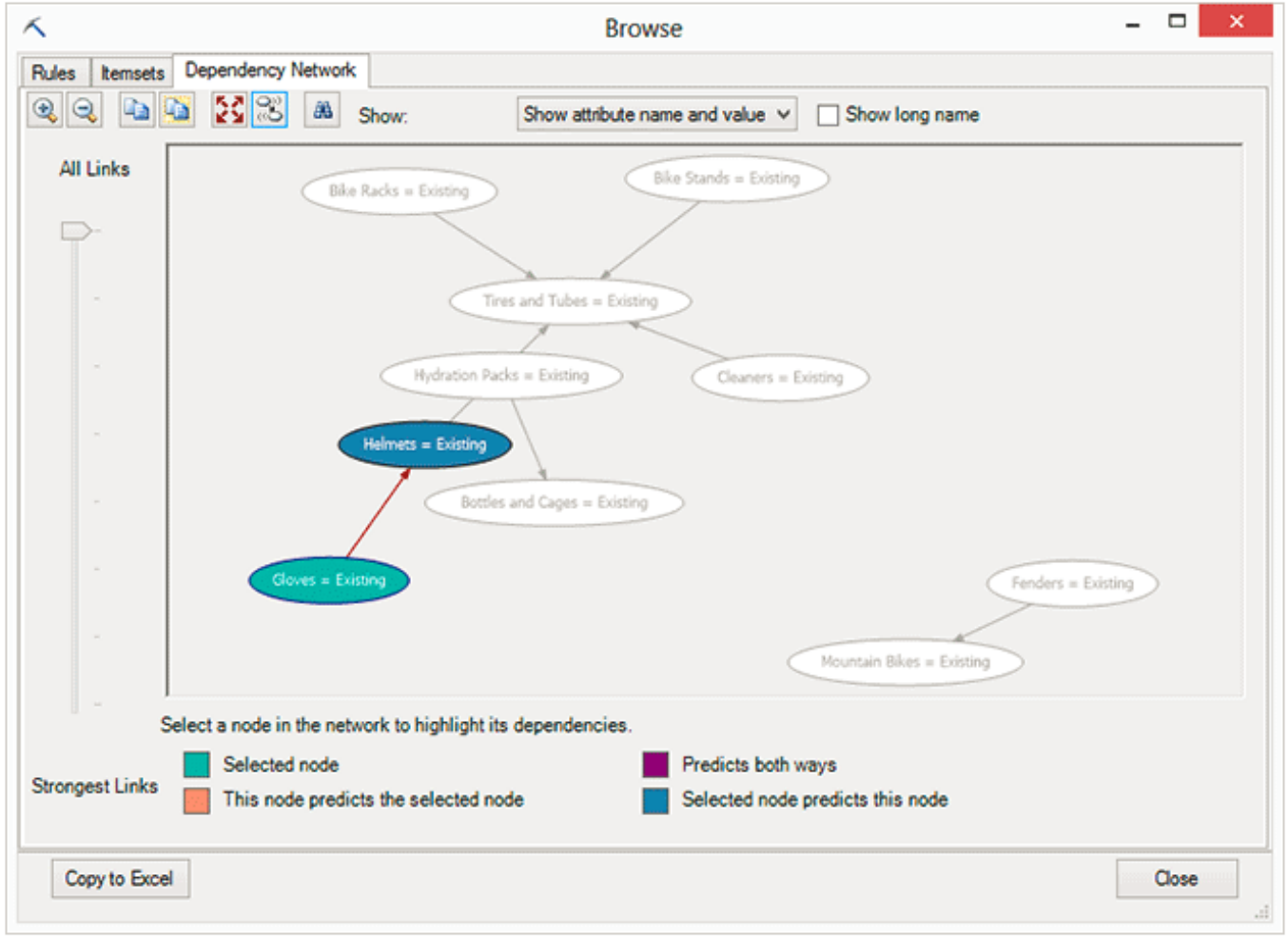
Dependency Network

The **Dependency Network** tab is a visual map of the correlations among items.

Each oval in the graph (referred to as a *node*) represents an attribute-value pair, such as "Vest = Existing" or "Age = 1-30".

Each line connecting the ovals (referred to as an *edge*) represents a type of correlation

Dependency graph



Dependency graph

- The node that contains the item is highlighted
- The arrows pointing to the node represent a rule that connects the items.
- The direction of the arrow tells you the direction of the rule.
- For example, if someone who buys gloves is also likely to buy a vest, the arrow will start from the “glove” node and terminate on the “vest” node.

Limitations of Support Confidence

Beverage preferences among a group of 1000 people.

	$Coffee$	$\overline{Coffee}$	
Tea	150	50	200
$\overline{Tea}$	650	150	800
	800	200	1000

- Rule {Tea} -> {Coffee}
- Support = 150/1000 = 15%
- Confidence = 150/200 = 75%
- Implies if you drink tea , you are likely to drink coffee
- On closer examination the probability of drinking coffee is 80% , while the probability of a tea drinker drinking coffee is 75%.

Research Issues in Mining association patterns

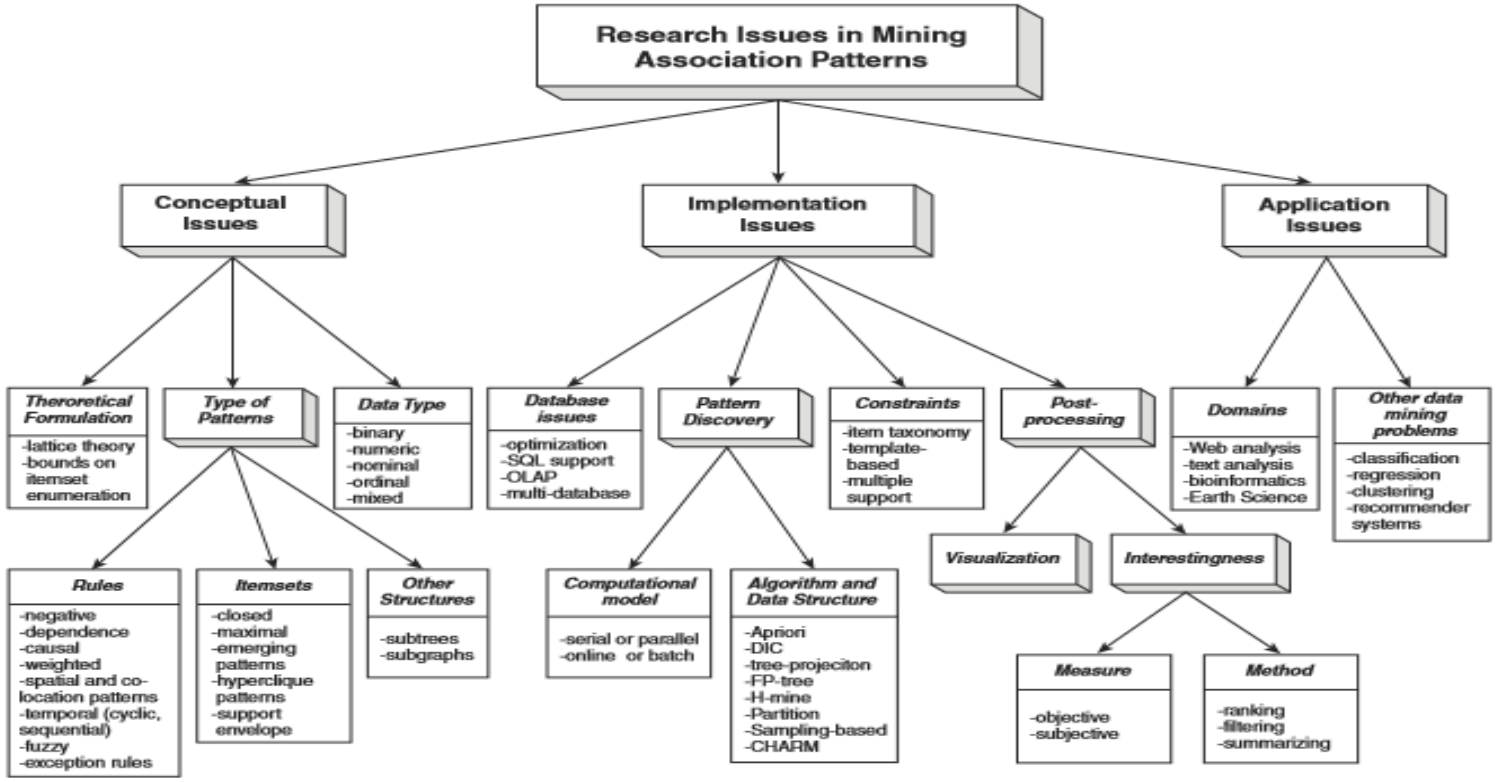


Figure 6.31. A summary of the various research activities in association analysis.