Matters of Discussion

Associative Prediction:

Frequent pattern Mining[rules],

Utility item set mining,

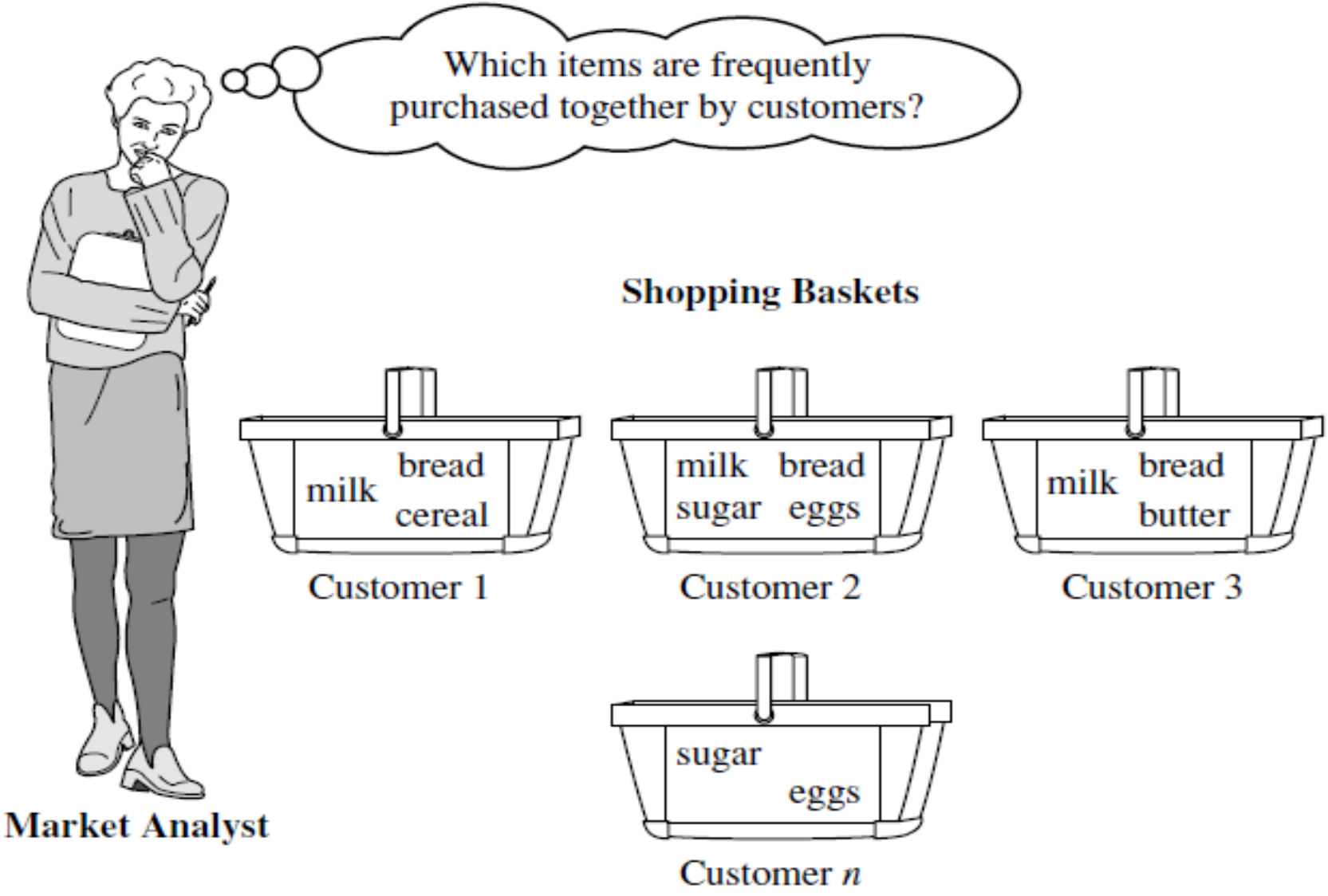
Association Rules – Association Algorithms

Compiled By: Dr. Nilamadhab Mishra [(PhD- CSIE) Taiwan]

Association Rule Mining

- Association Mining searches for frequent items in the data-set.
- In frequent mining usually the interesting associations and correlations between item sets in transactional and relational databases are found.
- In short, Frequent Mining shows which items appear together in a transaction or relation.

Association Rule Mining Task



Cont..

- **Association Rule (AR) Mining**: Finding frequent patterns, associations, correlations, or causal structures among sets of items or objects.
- **Frequent pattern**: a pattern (a set of items, subsequences, substructures, etc.) that occurs frequently in a data set.
- Association Rule (AR) discovery is often referred to as Market Basket Analysis (MBA), and is also referred to as Affinity Grouping

Motivation

Finding inherent regularities in data

- What products were often purchased together?
- What are the subsequent purchases after buying a PC?
- What kinds of DNA are sensitive to this new drug?
- Can we automatically classify web documents?

Association Rule- Basic Concepts

- Given a set of transactions, find rules that will predict the occurrence of an item based on the occurrences of other items in the transaction.
- Ex. transaction

Tid	Items bought
10	Beer, Nuts, Diaper
20	Beer, Coffee, Diaper
30	Beer, Diaper, Eggs
40	Nuts, Eggs, Milk
50	Nuts, Coffee, Diaper, Eggs, Milk

Association Rule- Basic . . . (cont'd)

- Given a set of transactions, find rules that will predict the occurrence of an item based on the occurrences of other items in the transaction

Market-Basket transactions

TID	*Items*
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

Example of Association Rules

{Diaper} → {Beer},
{Milk, Bread} → {Eggs,Coke},
{Beer, Bread} → {Milk},

Implication means co-occurrence, not causality!

Association Rule- Basic . . . (cont'd)

- **Support** or **utility** for an association rule $X \Rightarrow Y$ is the percentage of transactions in the database that contains X AND Y.
- **Confidence** or **Certainty** for an association rule $X \Rightarrow Y$ is the ratio of the number of transactions that contain X (AND) Y to the number of transactions that contain X.
- **Association Rule form :**

  Antecedent → Consequent [support, confidence]

Association Rule- Basic . . . (cont'd)

Support count("X → Y") = Number of transactions that contain both X and Y

$$\text{Support } (X \rightarrow Y) = \frac{\#\_of\_tuples\ containing\ both\ X\ and\ Y}{Total\_Number\_of\_tuples}$$

$$\text{Confidence } (X \rightarrow Y) = \frac{\#\_of\_tuples \text{ containing both X and Y}}{\text{Number of tuples containing X}}$$

Definition: Association Rule

- **Association Rule**
  - An implication expression of the form $X \rightarrow Y$, where X and Y are itemsets
  - Example: {Milk, Diaper} → {Beer}

TID	Items
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

- **Rule Evaluation Metrics**
  - Support (s)
    - Fraction of transactions that contain both X and Y
  - Confidence (c)
    - Measures how often items in Y appear in transactions that contain X

Example:

$$\{\text{Milk}, \text{Diaper}\} \Rightarrow \text{Beer}$$

$$s = \frac{\sigma(\text{Milk}, \text{Diaper}, \text{Beer})}{|\text{T}|} = \frac{2}{5} = 0.4$$

$$c = \frac{\sigma(\text{Milk}, \text{Diaper}, \text{Beer})}{\sigma(\text{Milk}, \text{Diaper})} = \frac{2}{3} = 0.67$$

Mining Association Rules

TID	Items
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

Example of Rules:

{Milk,Diaper} → {Beer} (s=0.4, c=0.67)
{Milk,Beer} → {Diaper} (s=0.4, c=1.0)
{Diaper,Beer} → {Milk} (s=0.4, c=0.67)
{Beer} → {Milk,Diaper} (s=0.4, c=0.67)
{Diaper} → {Milk,Beer} (s=0.4, c=0.5)
{Milk} → {Diaper,Beer} (s=0.4, c=0.5)

Observations:

- All the above rules are binary partitions of the same itemset:
  {Milk, Diaper, Beer}
- Rules originating from the same itemset have identical support but can have different confidence
- Thus, we may decouple the support and confidence requirements

Apriori Algorithm

- **Apriori** is a seminal algorithm proposed by R. Agrawal and R. Srikant in 1994 for mining frequent item sets for Boolean association rules.
- It uses prior knowledge of frequent item set properties,
- Apriori employs an iterative approach known as a level-wise search, where k-item-sets are used to explore k+1 item-sets.
- **Apriori pruning principle**: If there is any item set which is infrequent, its superset should not be generated/tested!

Computational Steps of Apriori Algorithm

- **Method:**
  1. Initially, scan DB once to get frequent 1-itemset
  2. Generate length (k+1) candidate item sets from length k frequent item sets
  3. Test the candidates against DB
  4. Terminate when no frequent or candidate set can be generated

Apriori Algorithm

C_k: Candidate itemset of size k
L_k : frequent itemset of size k

L_1 = {frequent items};

for (k = 1; L_k !=∅; k++) do begin

C_{k+1} = candidates generated from L_k;

for each transaction t in database do

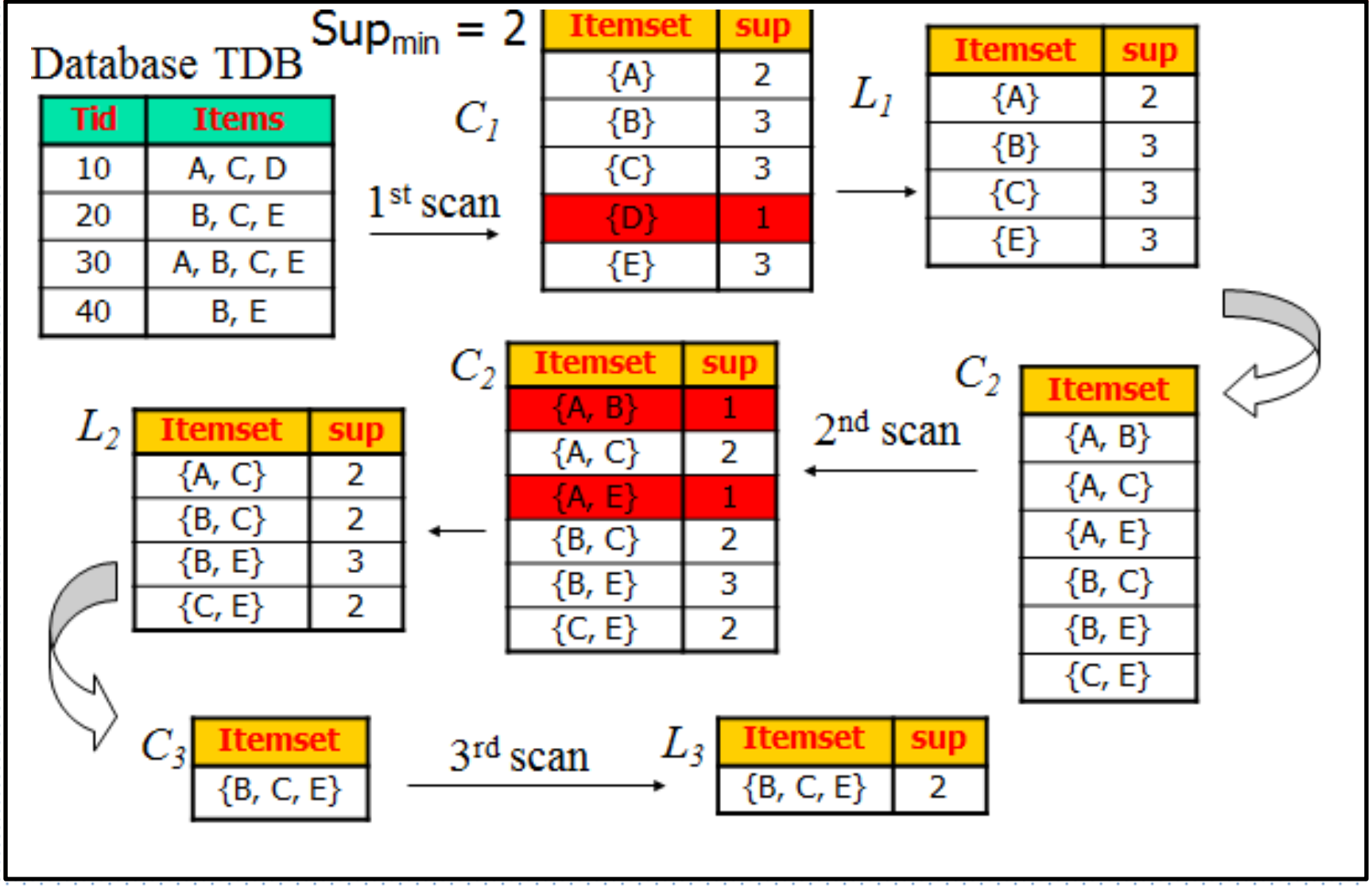
increment the count of all candidates in C_{k+1} that are contained in t

L_{k+1} = candidates in C_{k+1} with min_support

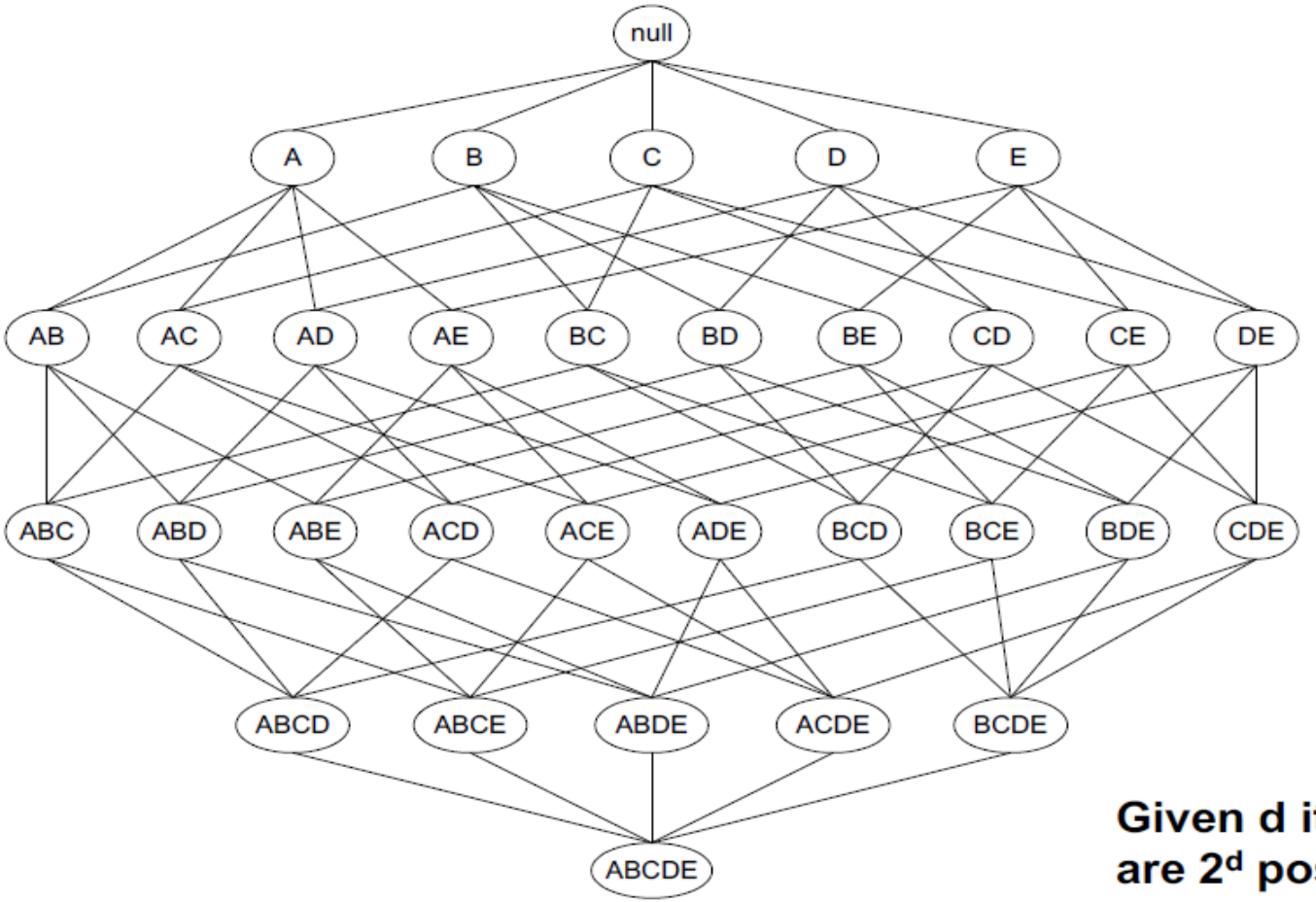
end

return $\cup_k L_k$;

Apriori Algorithm Example



Frequent Item set Generation



Given d items, there are 2^d possible candidate itemsets

Apriori Algorithm Example

- Exercise 1:

  Find the frequent Itemset where the minimum support count=2

TID	Items purchased
T100	I1,I2,I5
T200	I2,I4
T300	I2,I3
T400	I1,I2,I4
T500	I1,I3
T600	I2,I3
T700	I1,I3
T800	I1,I2,I3,I5
T900	I1,I2,I3

Practice FOR YOU------

Extra Investigation Task For You

Investigate the Computational Steps of any Associative prediction algorithm through an appropriate problem scenario in context to an application.

LAB

Investigate any real world problem to Implement the Apriori Algorithm for Association Rule Learning and prepare your Lab Report……

LAB CONT..

- **1. Start the Tool**
- **2. Load the Datasets**
- **3. Discover Association Rules**

The “Apriori” algorithm will already be selected. This is the most well known association rule learning method because it may have been the first (Agrawal and Srikant in 1994) and it is very efficient.

LAB CONT..

❖ 4. Analyze Results

The real work for association rule learning is in the interpretation of results.

You have to be very careful about interpreting association rules. They are associations (think correlations), not necessary causally related.

Consider the results snap short in Report along with the steps analysis of step-1,2,3.

Best rules found:

1. biscuits=t frozen foods=t fruit=t total=high 788 ==> bread and cake=t 723 <**conf:(0.92**)> lift:(1.27) lev:(0.03) [155] conv:(3.35)
2. baking needs=t biscuits=t fruit=t total=high 760 ==> bread and cake=t 696 <**conf:(0.92**)> lift:(1.27) lev:(0.03) [149] conv:(3.28)
3. baking needs=t frozen foods=t fruit=t total=high 770 ==> bread and cake=t 705 <**conf:(0.92**)> lift:(1.27) lev:(0.03) [150] conv:(3.27)
4. biscuits=t fruit=t vegetables=t total=high 815 ==> bread and cake=t 746 <**conf:(0.92**)> lift:(1.27) lev:(0.03) [159] conv:(3.26)
5. party snack foods=t fruit=t total=high 854 ==> bread and cake=t 779 <**conf:(0.91**)> lift:(1.27) lev:(0.04) [164] conv:(3.15)
6. biscuits=t frozen foods=t vegetables=t total=high 797 ==> bread and cake=t 725 <**conf:(0.91**)> lift:(1.26) lev:(0.03) [151] conv:(3.06)
7. baking needs=t biscuits=t vegetables=t total=high 772 ==> bread and cake=t 701 <**conf:(0.91**)> lift:(1.26) lev:(0.03) [145] conv:(3.01)
8. biscuits=t fruit=t total=high 954 ==> bread and cake=t 866 <**conf:(0.91**)> lift:(1.26) lev:(0.04) [179] conv:(3)
9. frozen foods=t fruit=t vegetables=t total=high 834 ==> bread and cake=t 757 <**conf:(0.91**)> lift:(1.26) lev:(0.03) [156] conv:(3)
10. frozen foods=t fruit=t total=high 969 ==> bread and cake=t 877 <**conf:(0.91**)> lift:(1.26) lev:(0.04) [179] conv:(2.92)

Lift, conviction,leverage

Cheers For the Great Patience!

Query Please?

https://www.kirenz.com/post/2020-05-14-r-association-rule-mining/#:~:text=Association%20rule%20mining%20is%20one,the%20arules%20and%20arulesViz%20packages

R-Implements

Compiled By: Dr. Nilamadhab Mishra [(PhD- CSIE) Taiwan]