Soft Computing

Association Analysis II

Dr. Chun-Hao Chen

Outline

1、Association Mining and it Problems

2、Multi-level AR Mining & Multiple Minimum Supports

3、Fuzzy Data Mining Approaches

4、Conclusions

Association Mining and it Problems

✓Proposed by Agrawal *et al.* (*SIGMOD'93*)

Apriori algorithm

Database	
TID	Items
100	A C D
200	B C E
300	A B C E
400	B E

Min. Sup.

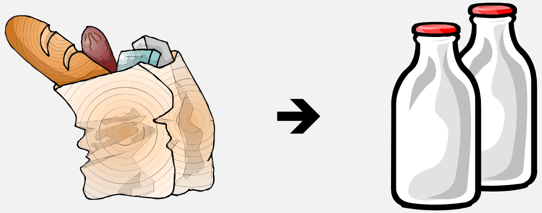
Min. Conf.

Generate Large Itemsets

Candidate Itemsets

Large Itemsets

Association Rule



Association Mining and it Problems (Cont.)

✓Basic problems are listed as follows

- A minimum support for all item is not appropriate
- Items may have taxonomy
- Quantitative value may exist in transaction
- Items may have lifespan

✓Business-oriented problems

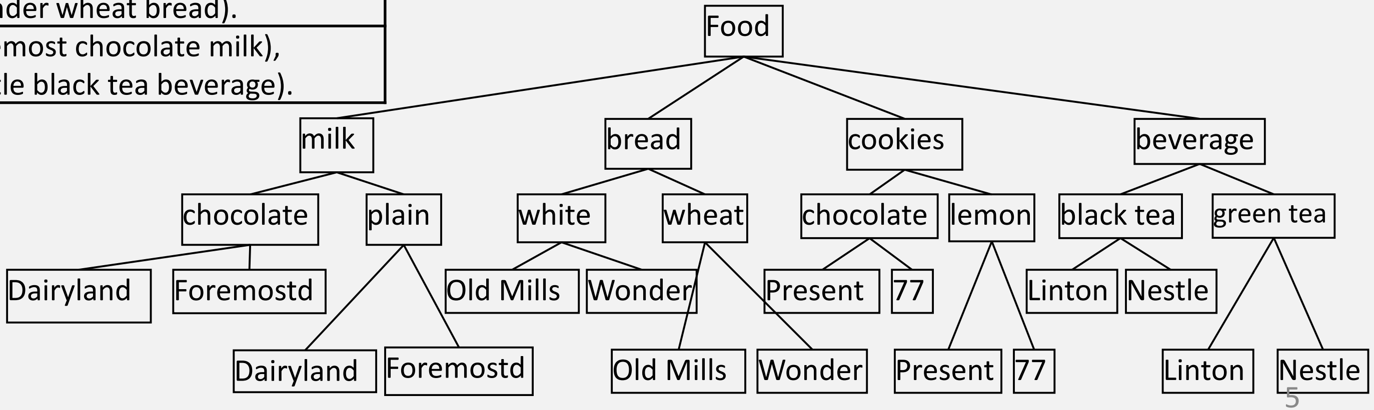
- How to find high utility itemsets
- How to reduce cost in manufacturing

Association Mining with Taxonomy

✓ What kinds of rules you can mine from transactions with taxonomy

TID	Items
D_1	(Dairyland chocolate milk), (Foremost chocolate milk), (Old Mills white bread),(Wonder white bread), (Present chocolate cookies),(Linton green tea beverage).
D_2	(Dairyland chocolate milk), (Foremost chocolate milk), (Dairyland plain milk), (Old Mills wheat bread), (Wonder wheat bread), (Present lemon cookies), (77 lemon cookies).
D_3	(Old Mills white bread), (Old Mills wheat bread), (77 chocolate cookies), (77 lemon cookies).
D_4	(Dairyland chocolate milk), (Old Mills white bread), (77 chocolate cookies).
D_5	(Old Mills white bread), (Wonder wheat bread).
D_6	(Dairyland chocolate milk), (Foremost chocolate milk), (Linton black tea beverage), (Nestle black tea beverage).

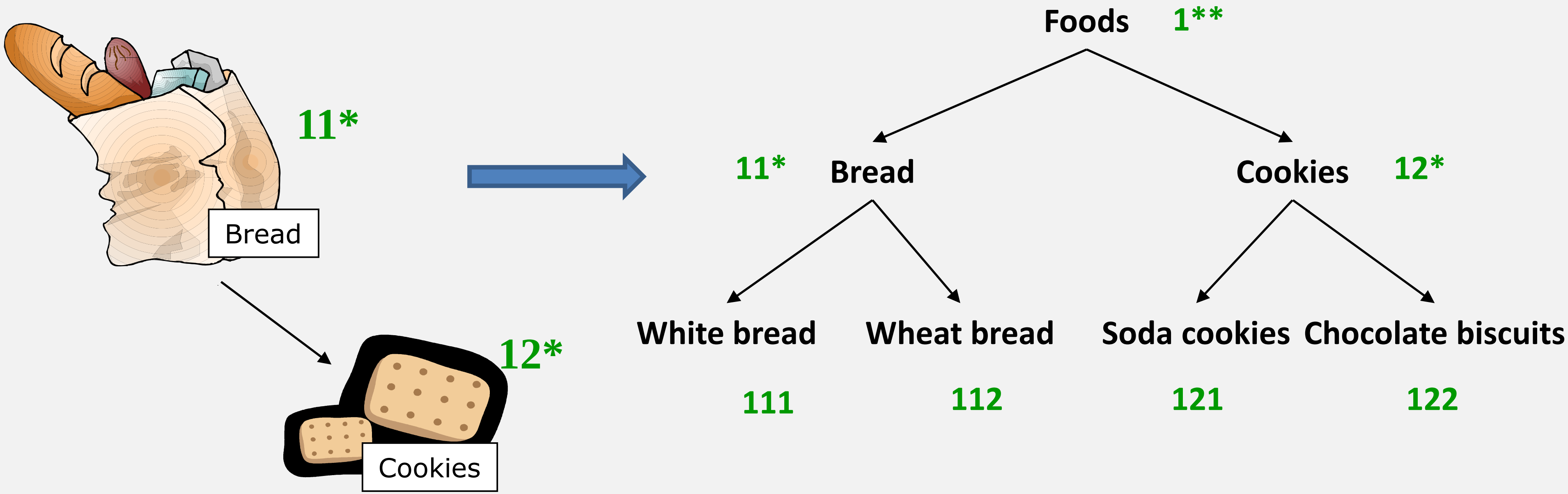
Taxonomy



Multiple-level Association Rules

✓Proposed by Han *et al.* (*VLDB'95*)

- Discovering large itemsets level by level
  - ✓A top-down searching approach

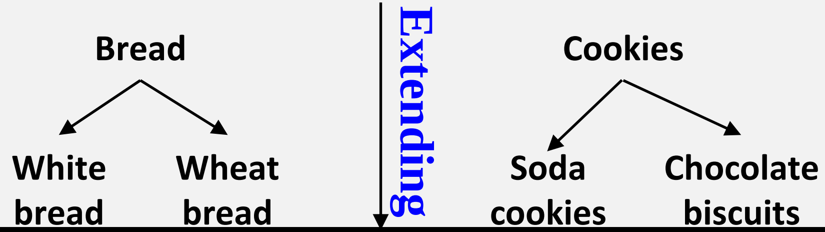


Generalized Association Rules

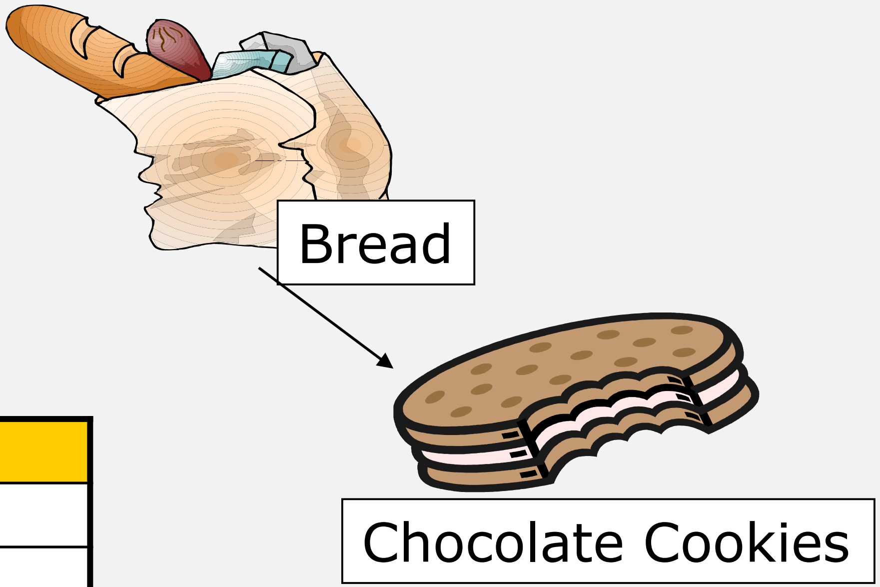
✓Proposed by Srikant et al. (VLDB'95)

- Extending Original Transactions
- Apriori-like Approach

TID	Items Bought
100	White bread, Chocolate biscuits
200	Wheat bread, Chocolate biscuits



TID	Items Bought
100	White bread, Bread, Chocolate biscuits, Cookies
200	Wheat bread, Bread, Chocolate biscuits, Cookies



Items with Different Minimum Supports

- ✓Proposed by Liu *et al.* (*KDD'99*)
  - An itemset ➔ **Minimum itemset support (MIS)**
    - ✓MIS: the lowest minimum supports among the items in an itemset
    - ✓ i.e. minsup{A} ≥ 20%
      minsup{B} ≥ 35%
      ➔ MIS {*A*, *B*} ≥ 20%
- ✓Proposed by Wang *et al.* (*VLDB'00*)
  - Specifying support constrain on itemsets
  - **Assigning the minimum supports with any function**
  - Too complicated, hard to follow

Generalized AR with Multiple Min. Sup.

- ✓Proposed by Tzeng et al. (LNCS, 2001)
  - ●Based on Liu's algorithm
  - ●Minsups of items and higher level items
    - ✓Different minimum supports are assigned
  - ● Maintaining the discovered multi-support, and generalized association rules

Discussions

Do you find any problem about using Minimum itemset support (MIS) for association rule mining?

e.g., minsup{A} ≥ 20%

minsup{B} ≥ 35%

➔ MIS {*A*, *B*} ≥ 20%

Problems of Liu's Approach (MIS)

✓ 1. Unsatisfying the Downward Closure Property

- e.g.

$minsup\{A\} \geq 20\%$
$minsup\{B\} \geq 35\%$
$\text{MIS}\{A, B\} \geq 20\%$
assume $\sup\{A, B\} = 15\%$
➔ {A, B} is pruned

⇨

$\text{minsup}\{C\} \geq 10\%$
➔ $\text{MIS}\ \{A, B, C\} \geq 10\%$
assume $\sup\{A, B, C\} = 12\%$
{A, B, C} is large
➔ Prune {A, B} (X)

Cannot efficiently prune itemsets!

✓ 2. Too Many Rules May be Derived

- e.g.

$minsup\{A\} \geq 20\%$
$minsup\{B\} \geq 35\%$
$minsup\{C\} \geq 1\%$
➔ $\text{MIS}\{A, B, C\} \geq 1\%$

⇨

{A, B, C}
➔ become large itemsets easily

Maximum Support Constraint

✓minsup{A, B}

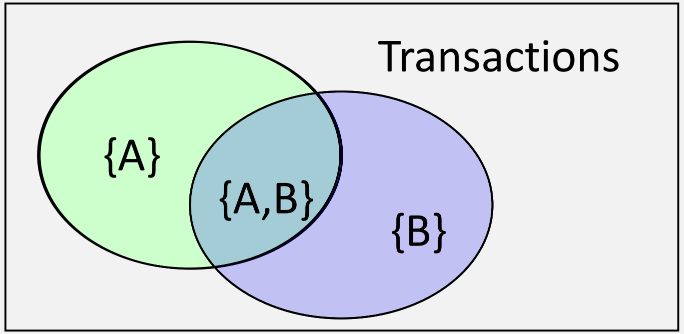
→ minsup{A} ∩ minsup{B} (Liu's)

→ max(minsup{A}, minsup{B})

✓i.e.

✓minsup{A}: sup{A} ≥ 0.15

✓minsup{B}: sup{B} ≥ 0.2



Advantages of the Maximum Support Constraint

✓ 1. Satisfying the downward closure property

$minsup\{A\} \geq 20\%$
$minsup\{B\} \geq 35\%$
$\mathrm{MIS}\{A, B\} \geq 35\%$
assume $\sup\{A, B\} = 15\%$
➔ {A, B} is pruned

➔

$\mathrm{minsup}\{C\} \geq 10\%$,
assume $\sup\{C\} = 12\%$
➔ $\mathrm{minsup}\{A, B, C\} \geq 35\%$,
$\sup\{A, B, C\} = 12\%$
➔ {A, B, C} is also small

✓ 2. An additional pruning

- If an itemset contains a small item ➔ It will be pruned

minsup{A} ≧ 20%;
minsup{D} ≧ 15%
sup{A} = 22%, sup{D} = 18%
➔ both are large

➔

minsup{A, D} ≧ 20%
sup{A, D} ≦ 18%
{A, D} is impossible to be large
{A, D} ➔ pruned

Discussions

Which strategy is better?

1. Minimum Support Constraint, or

2. Maximum Support Constraint?

Brief Summery

Single Minimum Support
Agrawal *et al.* (*SIGMOD'93*)

- Han *et al.* (*SIGMOD'00*)
  FP-growth algorithm

Taxonomy

- Han *et al.* (*VLDB'95*)
  Level-by-Level Approach
- Srikant *et al.* (*VLDB'95*)
  Generalized Approach

Apriori

Multiple Minsup

Multiple Minimum Support
Liu *et al.* (*KDD'99*) -MIS

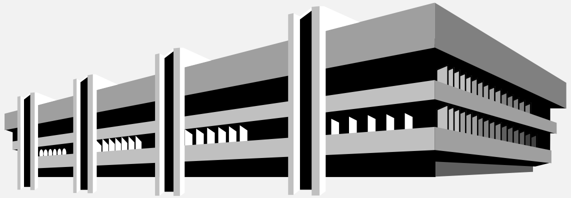
- Lee *et al.* (*IJAR, 2005*)
  Maximum constraints

Taxonomy

- Lui *et al.* (*LNAI*, 2000)
  Generalized Approach
- Tzeng *et al.* (*LNCS*, 2001)
  Generalized Approach

Quantitative Transactions

- ✓In Real World Applications
  - ●Quantitative value may exist in transaction



TID	Items
T1	(milk, 6); **(bread, 4); (cookies, 7), (beverage, 7).**
T2	(milk, 7); **(bread, 7); (cookies, 12).**
T3	**(bread, 8); (cookies, 12); (beverage, 6).**
T4	(milk, 2); **(bread, 3).**
T5	(milk, 3); **(bread, 8).**

Quantitative value

How to mine rules from Quantitative Transactions?

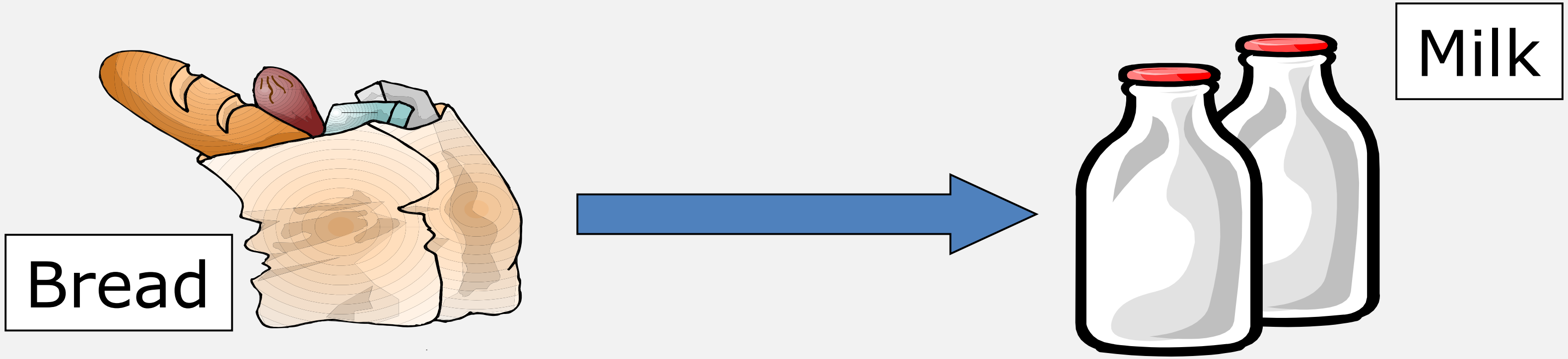
Fuzzy Data Mining

- ✓Solving quantitative values
  - ●e.g. John buys 10 bread, 2 butter and 3 Milk
- ✓Fuzzy Data Mining
  - ●Middle Bread and Little Butter
    - ✓Middle Bread ➔ Linguistic Term
- ✓Three Advantages (Kuok et al.)
  - ●Understandable to Human
  - ●Handling Quantitative Value Well (Shape boundary problem)
  - ●Deriving Extra Information

Fuzzy Association Rules

✓Proposed by Kuok *et al.* (*SIGMOD Record*, 1998)

- Another approach to handle quantitative value



IF middle amount of bread is bought
Then high amount of milk is bought

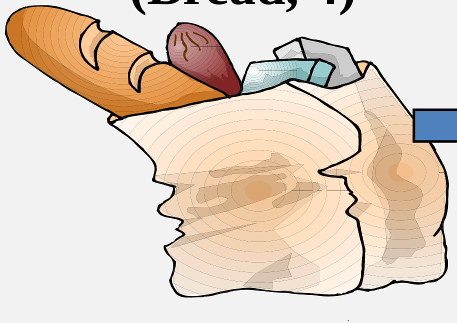
Linguistic Term

Fuzzy Association Rules (Cont.)

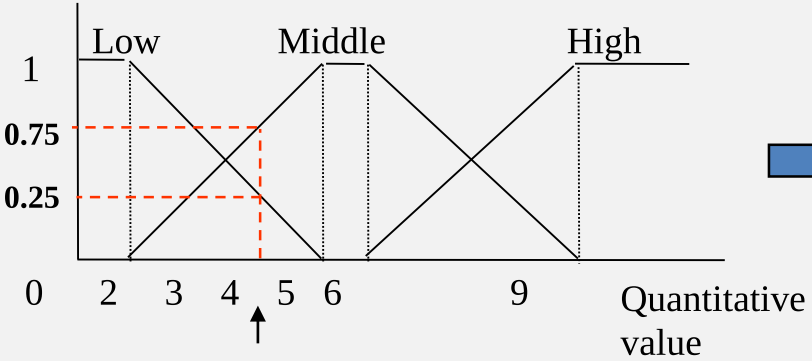
- ✓How to handle quantitative data?
- ✓e.g.

Quantitative data

(Bread, 4)



Membership Functions



Linguistic term

$$\frac{0.25}{Bread.Low} + \frac{0.75}{Bread.Middle}$$

Fuzzy Data Mining Algorithm

- ✓Input
  - *n* quantitative transaction data
  - *m* attributes (items)
  - A set of membership functions
  - Two thresholds
    - ✓Minimum support = α
    - ✓Minimum confidence = λ
- ✓Output
  - A set of fuzzy association rules

An Example

- ✓ Generating association rules for course grades
  - According to historical data concerning students' course scores
- ✓ The data set
  - 10 transactions

An Example

- ✓ Each case consists of five course scores (Five items)
  - Object-Oriented Programming (denoted OOP),
  - Database (denoted DB),
  - Statistics (denoted ST),
  - Data Structure (denoted DS),
  - Management Information System (denoted MIS).

Transactions

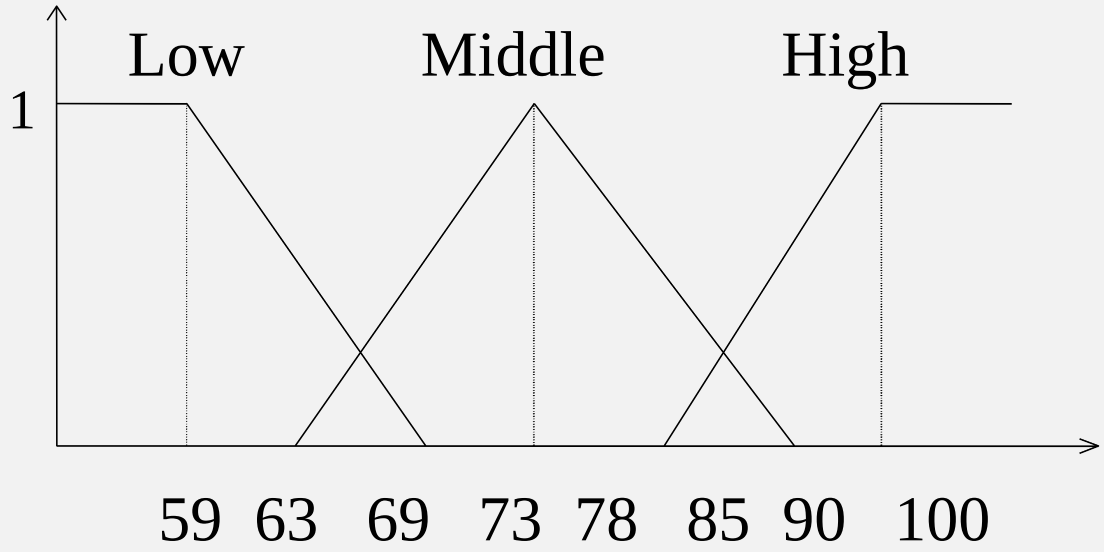
✓The set of students' course scores in the example

- *Minsup=1.5, Minconf=0.7*

Case	OOP	Database	Statistics	Data structure	MIS
1	86	77	86	71	68
2	61	87	89	77	80
3	84	89	86	79	89
4	73	86	79	84	62
5	70	85	87	72	79
6	65	67	86	61	87
7	71	87	75	71	80
8	86	69	64	84	88
9	75	65	86	86	79
10	83	68	65	85	89

Membership Functions

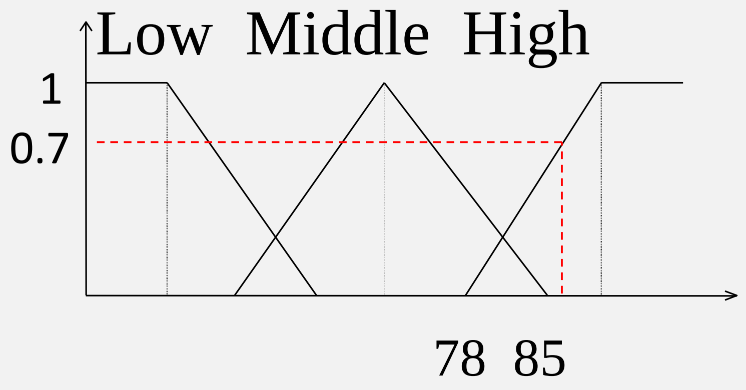
✓Three linguistic terms: Low, Middle, High



STEP 1

- ✓ Transform the quantitative values of each transaction datum into fuzzy sets.
- ✓ e.g., Score=86

$$86 = \frac{0.0}{Low} + \frac{0.0}{Middle} + \frac{0.7}{High}$$



STEP 2

✓ Calculate the count of each attribute region

Case	OOP			Database			Statistics			Data Structure			MIS		
	L	M	H	L	M	H	L	M	H	L	M	H	L	M	H
1	0.0	0.0	0.7	0.0	0.7	0.0	0.0	0.0	0.7	0.0	0.8	0.0	0.1	0.5	0.0
2	0.8	0.0	0.0	0.0	0.0	0.8	0.0	0.0	0.9	0.0	0.7	0.0	0.0	0.4	0.2
3	0.0	0.1	0.5	0.0	0.0	0.9	0.0	0.0	0.7	0.0	0.5	0.1	0.0	0.0	0.9
4	0.0	1.0	0.0	0.0	0.0	0.7	0.0	0.5	0.1	0.0	0.1	0.5	0.7	0.0	0.0
5	0.0	0.7	0.0	0.0	0.0	0.6	0.0	0.0	0.8	0.0	0.9	0.0	0.0	0.8	0.1
6	0.4	0.2	0.0	0.2	0.4	0.0	0.0	0.0	0.7	0.8	0.0	0.0	0.0	0.0	0.8
7	0.0	0.8	0.0	0.0	0.0	0.8	0.0	0.8	0.0	0.0	0.8	0.0	0.0	0.4	0.2
8	0.0	0.0	0.7	0.0	0.6	0.0	0.5	0.1	0.0	0.0	0.1	0.5	0.0	0.0	0.8
9	0.0	0.8	0.0	0.4	0.2	0.0	0.0	0.0	0.7	0.0	0.0	0.7	0.0	0.5	0.1
10	0.0	0.2	0.4	0.1	0.5	0.0	0.4	0.2	0.0	0.0	0.0	0.6	0.0	0.0	0.9
count	1.2	**3.8**	**2.3**	0.7	**2.4**	**3.8**	0.9	**1.6**	**4.6**	0.8	**3.9**	**2.4**	0.8	**2.3**	**4.0**

STEP 3

- ✓ Find L_1
- ✓ Example: *minimum support count* = 1.5

Itemset	Support
OOP.Middle	3.8
OOP.High	2.3
DB.Middle	2.4
DB.High	3.8
ST.Middle	1.6
ST.High	4.6
DS.Middle	3.9
DS.High	2.5
MIS.Middle	2.3
MIS.High	4.0

STEP 4

- ✓ STEP 4
  - Generate the candidate set C_{r+1} from L_r
  - e.g.
    - ✓ (OOP.Middle, DB.Middle), (OOP.Middle, DB.High)
    - ✓ (OOP.Middle, ST.Middle), (OOP.Middle, ST.High)
    - ✓ (OOP.Middle, DS.Middle), (OOP.Middle, DS.High)
    - ✓ (OOP.Middle, MIS.Middle), (OOP.Middle,MIS.High)
    - ✓ …,
    - ✓ (DS.High, MIS.Middle), and (DS.High, MIS.High)

STEP 5

- ✓ Calculate the fuzzy membership value of each candidate
- ✓ e.g.

Case	OOP.Middle	DB.High	OOP.Middle ∩DB.High
1	0.0	0.0	0.0
2	0.0	0.8	0.0
3	0.1	0.9	0.1
4	1.0	0.7	0.7
5	0.7	0.6	0.6
6	0.2	0.0	0.0
7	0.8	0.8	0.8
8	0.0	0.0	0.0
9	0.8	0.0	0.0
10	0.2	0.0	0.0

STEP 6

✓ Find large itemsets

- (OOP.Middle, DB.High), (OOP.Middle, ST.Middle)
- (OOP.Middle, ST.High), (OOP.Middle, DS.Middle)
- …

STEP 7

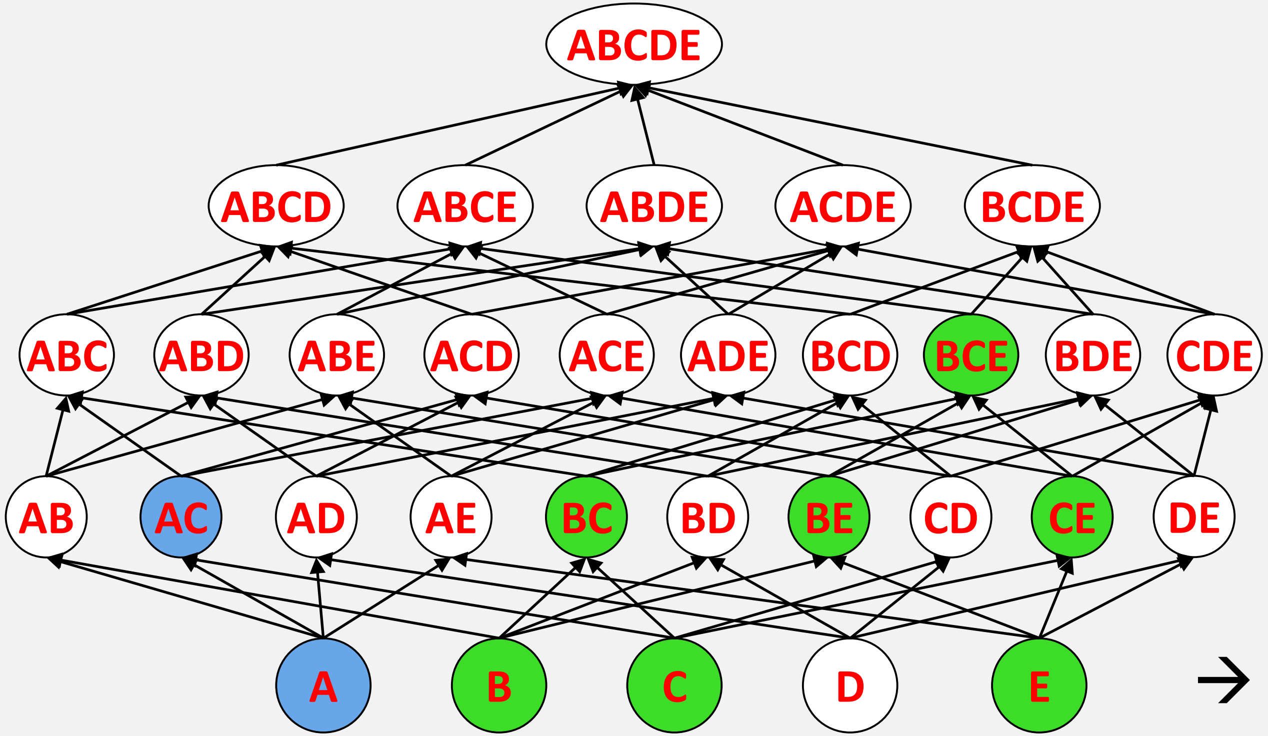
✓Construct the association rules

- If OOP.Middle and DS.Middle then DB.High, conf = 0.94
- If ST.Middle then OOP.Middle, conf = 0.94
- …

Execution Time of FAR Mining

✓Five items will become 15 fuzzy regions in Fuzzy rule mining

- Number of linguistic terms → 3



Time-consuming process

↑

→ E.Low, E.Middle, E.High

Discussion

Do you have any idea to make the fuzzy data mining process more efficient (快速) or effective (有用)?

An Efficient Approach

✓ Use only the region with the maximum fuzzy cardinality for each item

Case	OOP			Database			Statistics			Data Structure			MIS		
	L	**M**	H	L	M	**H**	L	M	**H**	L	**M**	H	L	M	**H**
1	0.0	**0.0**	0.7	0.0	0.7	**0.0**	0.0	0.0	**0.7**	0.0	**0.8**	0.0	0.1	0.5	**0.0**
2	0.8	**0.0**	0.0	0.0	0.0	**0.8**	0.0	0.0	**0.9**	0.0	**0.7**	0.0	0.0	0.4	**0.2**
3	0.0	**0.1**	0.5	0.0	0.0	**0.9**	0.0	0.0	**0.7**	0.0	**0.5**	0.1	0.0	0.0	**0.9**
4	0.0	**1.0**	0.0	0.0	0.0	**0.7**	0.0	0.5	**0.1**	0.0	**0.1**	0.5	0.7	0.0	**0.0**
5	0.0	**0.7**	0.0	0.0	0.0	**0.6**	0.0	0.0	**0.8**	0.0	**0.9**	0.0	0.0	0.8	**0.1**
6	0.4	**0.2**	0.0	0.2	0.4	**0.0**	0.0	0.0	**0.7**	0.8	**0.0**	0.0	0.0	0.0	**0.8**
7	0.0	**0.8**	0.0	0.0	0.0	**0.8**	0.0	0.8	**0.0**	0.0	**0.8**	0.0	0.0	0.4	**0.2**
8	0.0	**0.0**	0.7	0.0	0.6	**0.0**	0.5	0.1	**0.0**	0.0	**0.1**	0.5	0.0	0.0	**0.8**
9	0.0	**0.8**	0.0	0.4	0.2	**0.0**	0.0	0.0	**0.7**	0.0	**0.0**	0.7	0.0	0.5	**0.1**
10	0.0	**0.2**	0.4	0.1	0.5	**0.0**	0.4	0.2	**0.0**	0.0	**0.0**	0.6	0.0	0.0	**0.9**
count	1.2	**3.8**	2.3	0.7	2.4	**3.8**	0.9	1.6	**4.6**	0.8	**3.9**	2.4	0.8	2.3	**4.0**

An Efficient Approach (Cont.)

- ✓ Advantages and problems
  - Less number of large itemsets
  - Not complete
  - Less computation time
- ✓ Other approaches
  - FP-tree-based approaches

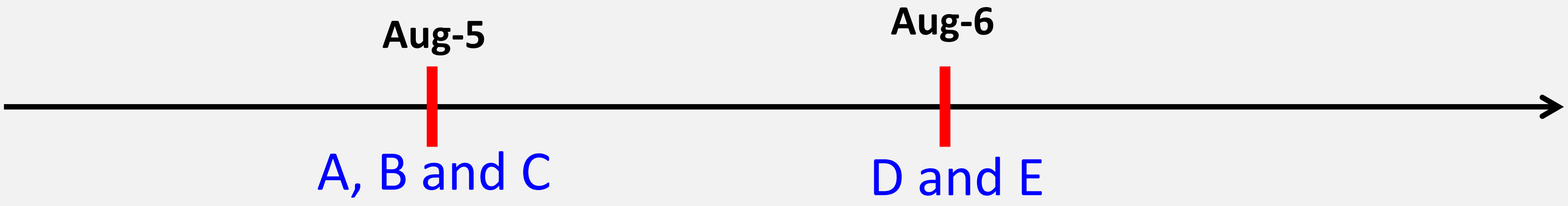
Items Have Lifespan

✓ e.g.

Period	TID	Items
P_1 (Aug-5)	$Trans_1$	$(A, 5), (C,4);$
	$Trans_2$	$(A, 3), (B,2);$
	$Trans_3$	$(C,4);$
P_2 (Aug-6)	$Trans_4$	$(A, 3), (B, 2), (D,4);$
	$Trans_5$	$(A, 3), (B, 5), (D,4), (E,2);$

Lifespan of A, B and C starts from Aug-5

Lifespan of D and E starts from Aug 6



What should we do to mine fuzzy temporal association rules from temporal transactions?

Fuzzy Temporal ARM Algorithm

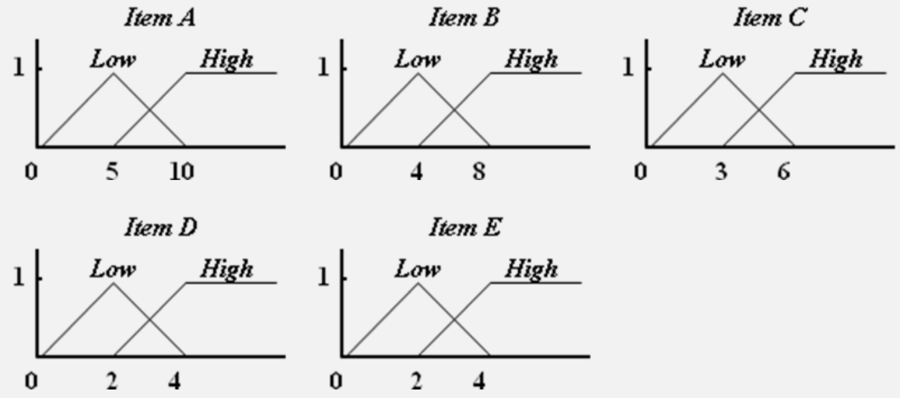
- ✓Input
  - ●Temporal quantitative transactions *TempTrans*
  - ●A set of membership functions *MF*
  - ●Two thresholds
    - ✓Minimum support threshold α
    - ✓Minimum confidence threshold λ
- ✓Output
  - ●Fuzzy temporal association rule set *temportalFuzzyRuleSet*

Used Transactions and MF

- ✓Two periods and five transactions

Period	TID	Items
P_1 (Aug-5)	$Trans_1$	(A, 5), (C,4);
	$Trans_2$	(A, 3), (B,2);
	$Trans_3$	(C,4);
P_2 (Aug-6)	$Trans_4$	(A, 3), (B, 2), (D,4);
	$Trans_5$	(A, 3), (B, 5), (D,4), (E,2);

- ✓Membership functions for Items

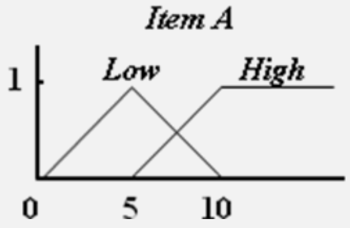


Step 1.1

✓Transform quantitative value into fuzzy sets

Period	TID	Items
P_1 (Aug-5)	$Trans_1$	(1/*A.Low*), (0.66/*C.Low* + 0.33/ *C.High*);
	$Trans_2$	(0.5/*A.Low*), (0.5/*B.Low*);
	$Trans_3$	(0.66/*C.Low* + 0.33/*C.High*);
P_2 (Aug-6)	$Trans_4$	(0.5/*A.Low*), (0.5/*B.Low*), (1/*D.High*);
	$Trans_5$	(0.5/*A.Low*), (0.75/*B.Low* + 0.25/*B.High*), (1/*D.High*), (1/*E.High*);

(*A*, 5) ➔



➔ (1/*A.Low* + 0/*A.High*)

Step 1.2

✓Build Temporal Information (TI) table

Time Period	Number of Transactions	Temporal Items
P_1	3	A, B, C
P_2	2	D, E

Step 2

- ✓ Get scalar cardinality of each fuzzy region (linguistic term)
  - e.g., Fuzzy region *A.Low*
    - ✓ Scalar cardinality is 2.5 (= 1.0 + 0.5 + 0.0 + 0.5 + 0.5)

Period	TID	Items
P_1 (Aug-5)	$Trans_1$	**(1/*A.Low*)**
	$Trans_2$	**(0.5/*A.Low*)**
	$Trans_3$	**0**
P_2 (Aug-6)	$Trans_4$	**(0.5/*A.Low*)**
	$Trans_5$	**(0.5/*A.Low*)**

Items	Count	Items	Count
A.Low	**2.5**	*C.High*	0.66
A.High	0	*D.Low*	0
B.Low	1.75	*D.High*	2
B.High	0.25	*E.Low*	0
C.Low	1.33	*E.High*	1

Step 3

✓ Calculate fuzzy temporal support of fuzzy regions according to temporal item starting periods

Period	# Transactions	Temporal Items
P_1	3	*A, B, C*
P_2	2	*D, E*

Items	Count
A.Low	**2.5**

tFuzzySup(***A.Low***) = 2.5/5

Items	Count
D.High	**2**

tFuzzySup(***D.High***) = 2 / 2

Items	tFuzzy Support	Items	tFuzzy Support
A.Low	**0.5**	*C.High*	0.132
A.High	0	*D.Low*	0
B.Low	0.35	***D.High***	**1**
B.High	0.05	*E.Low*	0
C.Low	0.266	*E.High*	0.5

Step 4

- ✓ Minimum support was set at 0.3,
- ✓ Four large 1-itemsets are derived as follows

Items	tFuzzy Support	Items	tFuzzy Support
A.Low	**0.5**	~~*C.High*~~	~~0.132~~
~~*A.High*~~	~~0~~	~~*D.Low*~~	~~0~~
B.Low	**0.35**	***D.High***	**1**
~~*B.High*~~	~~0.05~~	~~*E.Low*~~	~~0~~
~~*C.Low*~~	~~0.266~~	***E.High***	**0.5**

Steps 5 to 9: Find All Large Itemsets

- ✓ Candidate 2-itemset and its tFuzzySupport
  - Case1: Items have **the same starting period**
    - ✓ Take {*A.Low*, *B.Low*} as an example

Period	# Transactions	Temporal Items
P_1	3	***A***, ***B***, *C*
P_2	2	*D*, *E*

Period	TID	*A.Low*	*B.Low*	*A.Low* ∩ *B.Low*
P_1 (Aug-5)	$Trans_1$	1.0	0.0	0.0
	$Trans_2$	0.5	0.5	0.5
	$Trans_3$	0.0	0.0	0.0
P_2 (Aug-6)	$Trans_4$	0.5	0.5	0.5
	$Trans_5$	0.5	0.75	0.5

tFuzzySup(*A.Low*, *B.Low*)
= (0.0 + 0.5 + 0.0 + 0.5 + 0.5) / 5
= 1.5 / 5
= 0.3

Steps 5 to 9: Find All Large Itemsets (Cont.)

Period	# Transactions	Temporal Items
P_1	3	**A**, B, C
P_2	2	**D**, E

- ✓ Candidate 2-itemset and its tFuzzySupport
  - Case 2: Items have **different starting period**
    - ✓ Take {*A.Low*, *D.High*} as an example

Period	TID	*A.Low*	*D.High*	*A.Low* ∩ *D.High*
P_1 (Aug-5)	$Trans_1$	1.0	-	0.0
	$Trans_2$	0.5	-	0.0
	$Trans_3$	0.0	-	0.0
P_2 (Aug-6)	$Trans_4$	0.5	1.0	0.5
	$Trans_5$	0.5	1.0	0.5

⇨

tFuzzySup(*A.Low*, *D.High*)
= (0.0 + 0.0 + 0.0 + 0.5 + 0.5) / 2
= 1 / 2
= 0.5

Steps 5 to 9: Find All Large Itemsets (Cont.)

- ✓ Minimum support is 0.3
- ✓ 5 large 2-itemset are generated

Itemset	Count	tFuzzySupport
(*A.Low*, *B.Low*)	1.5	0.3 (=1.5/5)
(*A.Low*, *D.High*)	1	0.5 (=1/2)
(*B.Low*, *D.High*)	1.25	0.625 (=1.25/2)
(*B.Low*, *E.High*)	0.75	0.375 (=0.75/2)
(*D.High*, *E.High*)	1	0.5 (=1/2)

Steps 10 to 11: Generate Temporal FAR

✓ Take temporal large *2*-itemset (*A.Low*, *D.High*) as an example

- 2 candidate rules are listed as follows:

CRule1: If *D.High*, Then *A.Low*, and
CRule2: If *A.Low*, Then *D.High*.

- Confidence of CRule2 = 0.5 (0.5 / 1)

Itemset	tFuzzySupport
(*A.Low*, *D.High*)	0.5 (=1/2)

P_2 (Aug-6)	$Trans_4$	**(0.5/*A.Low*)**
	$Trans_5$	**(0.5/*A.Low*)**

⇨ Conf.(CRule2) = 0.5 / 1

Brief Summery (Cont.)

Single Minimum Support

Agrawal *et al.* (*SIGMOD'93*)

- Han *et al.* (*SIGMOD'00*)
  FP-growth algorithm

Taxonomy

- Han *et al.* (*VLDB'95*)
  Level-by-Level Approach
- Srikant *et al.* (*VLDB'95*)
  Generalized Approach

Quantitative Value

- Hong et al. (IDA)
  Fuzzy AR mining
- Chen et al. (Applied soft computing)
  Fuzzy temporal AR mining

Apriori

Multiple MinSup

Multiple Minimum Support

Liu *et al.* (*KDD'99*) -MIS

- Lee *et al.* (*IJAR'05*)
  Maximum constraints

Taxonomy

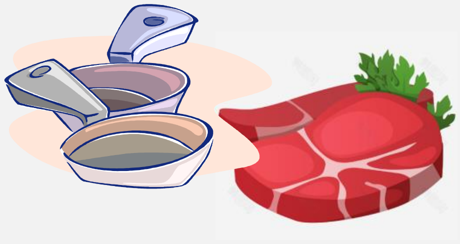
- Lui *et al.* (*LNAI*, 2000)
  Generalized Approach
- Tzeng *et al.* (*LNCS*, 2001)
  Generalized Approach

Advanced Fuzzy Data Mining Approaches

✓Business-oriented problems

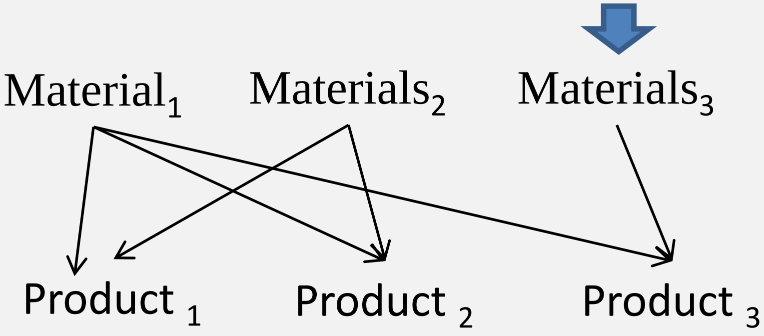
- How to find high utility itemsets
- How to reduce cost in manufacturing

High Utility Itemsets



Profit: NT$2500

Erasable Pattern



1. High Coherent Utility Fuzzy Itemsets Mining

2. Erasable Pattern Mining

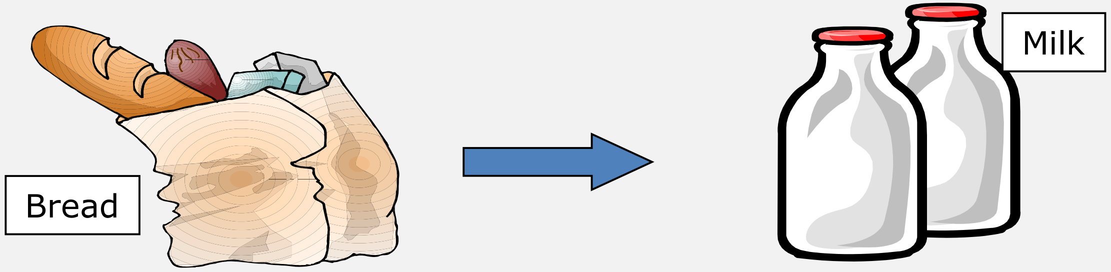
Two Fuzzy Data Mining Approaches

✓High coherent utility fuzzy itemset mining

✓Erasable-itemset mining

Actionable Association Rule?

✓ An example



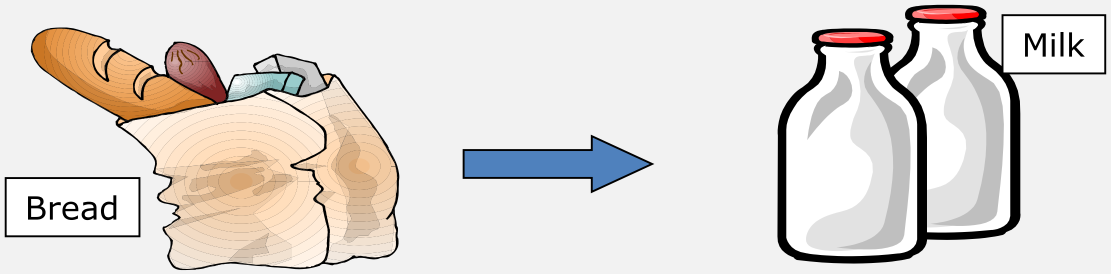
Support is 30% and Confidence is 80%

$$\underbrace{\hspace{10em}}$$

How to make sure the rule is more actionable?

How about the Following Solution

✓A rule: Bread → Milk



✓The following three rules should also be held

- Not Bread → Not Milk
- Not Bread → Milk
- Bread → Not Milk

Actionable Rule - Coherent Concept

- ✓A rule: X imply Y (X → Y)
  - Calculate contingency table of X → Y

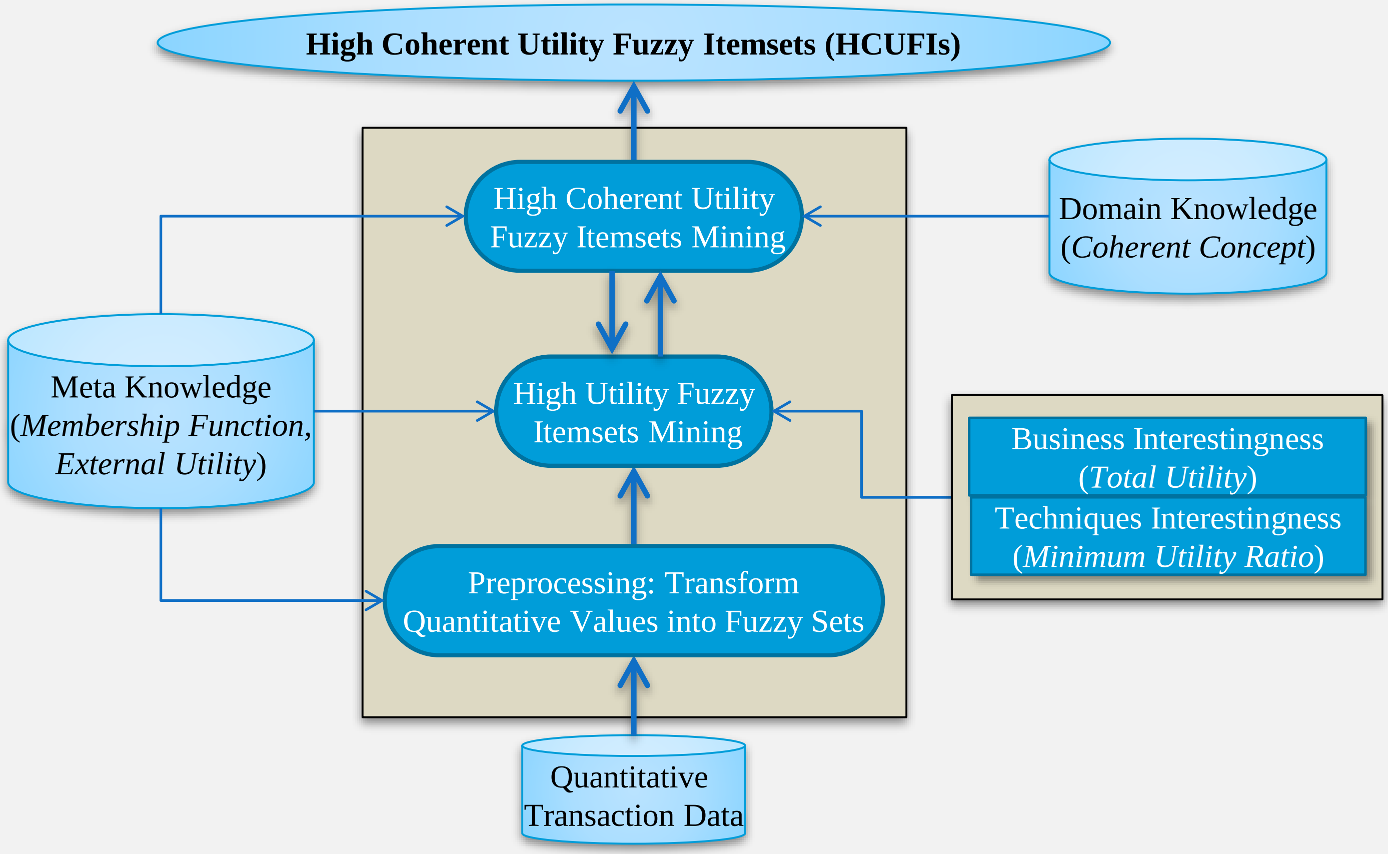
Contingency Table of X → Y

Frequency of co-occurrences		Consequence Y	
		Y	¬Y
Antecedent X	X	(X → Y) Q_1	(X → ¬Y) Q_2
	¬X	(¬X → Y) Q_3	(¬X → ¬Y) Q_4

- ✓Reach the Four conditions:
  - $Q_1 > Q_2$, $Q_1 > Q_3$, $Q_4 > Q_2$ and $Q_4 > Q_3$
  - X → Y is a coherent rule

High Coherent Utility Fuzzy Itemsets Mining

✓Flowchart



An Example

✓Six Transactions

TID	Items
T1	(milk, 10); (bread, 10); (cookies, 7), (beverage, 7).
T2	(milk, 12); (bread, 14); (cookies, 12).
T3	(bread, 2); (cookies, 12).
T4	(milk, 2); (bread, 4); (cookies, 5).
T5	(milk, 9) ;(bread, 9).
T6	(milk, 5); (beverage, 12).

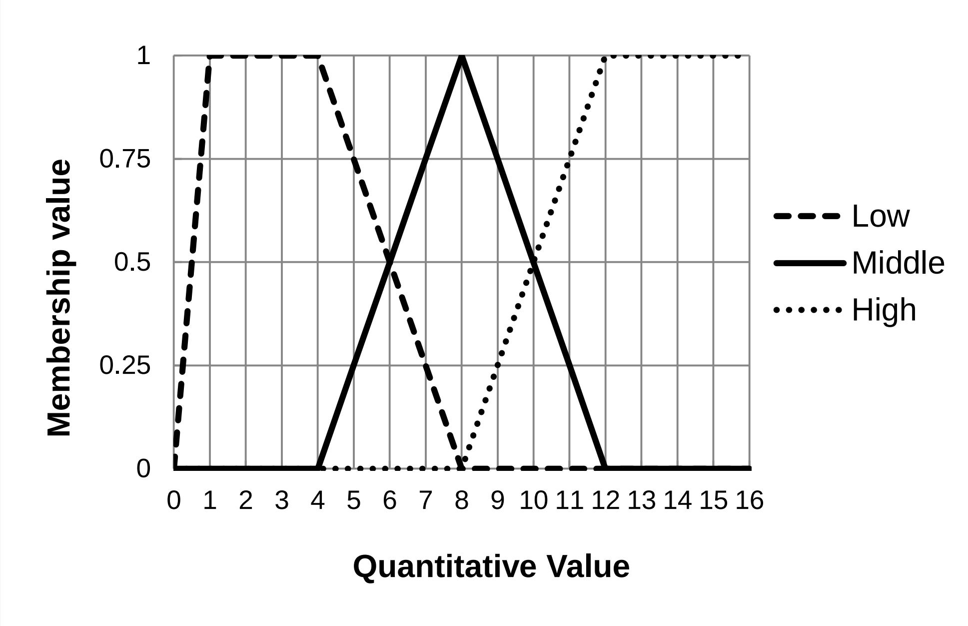
✓ Minimum utility ratio α = 10%

✓External Utility

Items	milk	bread	cookies	beverage
EU	5	3	1	7

Membership Functions

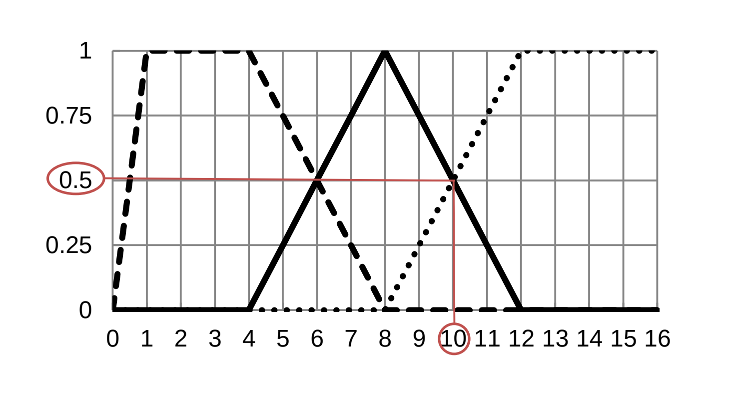
✓Three Linguistic Terms: Low, Middle and High



Step 1: Transform to Fuzzy Value

✓Take (milk, 10) as an example

TID	Fuzzy Set
T1	$(\frac{0.5}{milk.Middle}+\frac{0.5}{milk.High})(\frac{0.5}{bread.Middle}+\frac{0.5}{bread.High})$ $(\frac{0.25}{cookies.Low}+\frac{0.75}{cookies.Middle})(\frac{0.25}{beverage.Low}+\frac{0.75}{beverage.Middle})$
T2	$(\frac{1}{milk.High})(\frac{1}{bread.High})(\frac{1}{cookies.High})$
T3	$(\frac{1}{bread.Low})(\frac{1}{cookies.High})$
T4	$(\frac{1}{milk.Low})(\frac{1}{bread.Low})(\frac{0.75}{cookies.Low}+\frac{0.25}{cookies.Middle})$
T5	$(\frac{0.75}{milk.Middle}+\frac{0.25}{milk.High})(\frac{0.75}{bread.Middle}+\frac{0.25}{bread.High})$
T6	$(\frac{0.75}{milk.Low}+\frac{0.25}{milk.Middle})(\frac{1}{beverage.High})$



(milk, 10)

$$\frac{0.5}{milk.Middle}+\frac{0.5}{milk.High}$$

Step 2: Generate Complement Fuzzy Set

✓Fuzzy value of (milk, 10) is (0/milk.Low + 0.5/milk.Middle + 0.5/milk.High)

$$\frac{0}{milk.Low}+\frac{0.5}{milk.Middle}+\frac{0.5}{milk.High}$$

TID	Complement Fuzzy Set
T1	$\boxed{(\frac{1}{milk.Low}+\frac{0.5}{milk.Middle}+\frac{0.5}{milk.High})}(\frac{1}{bread.Low}+\frac{0.5}{bread.Middle}+\frac{0.5}{bread.High})$ $(\frac{0.75}{cookies.Low}+\frac{0.25}{cookies.Middle}+\frac{1}{cookies.High})(\frac{0.75}{beverage.Low}+\frac{0.25}{beverage.Middle}+\frac{1}{beverage.High})$
T2	$(\frac{1}{milk.Low}+\frac{1}{milk.Middle})(\frac{1}{bread.Low}+\frac{1}{bread.Middle})$ $(\frac{1}{cookies.Low}+\frac{1}{cookies.Middle})(\frac{1}{beverage.Low}+\frac{1}{beverage.Middle}+\frac{1}{beverage.High})$
T3	$(\frac{1}{milk.Low}+\frac{1}{milk.Middle}+\frac{1}{milk.High})(\frac{1}{bread.Middle}+\frac{1}{bread.High})$ $(\frac{1}{cookies.Low}+\frac{1}{cookies.Middle})(\frac{1}{beverage.Low}+\frac{1}{beverage.Middle}+\frac{1}{beverage.High})$
T4	$(\frac{1}{milk.Middle}+\frac{1}{milk.High})(\frac{1}{bread.Middle}+\frac{1}{bread.High})$ $(\frac{0.25}{cookies.Low}+\frac{0.75}{cookies.Middle}+\frac{1}{cookies.High})(\frac{1}{beverage.Low}+\frac{1}{beverage.Middle}+\frac{1}{beverage.High})$
T5	$(\frac{1}{milk.Low}+\frac{0.25}{milk.Middle}+\frac{0.75}{milk.High})(\frac{1}{bread.Low}+\frac{0.25}{bread.Middle}+\frac{0.75}{bread.High})$ $(\frac{1}{cookies.Low}+\frac{1}{cookies.Middle}+\frac{1}{cookies.High})(\frac{1}{beverage.Low}+\frac{1}{beverage.Middle}+\frac{1}{beverage.High})$
T6	$(\frac{0.25}{milk.Low}+\frac{0.75}{milk.Middle}+\frac{1}{milk.High})(\frac{1}{bread.Low}+\frac{1}{bread.Middle}+\frac{1}{bread.High})$ $(\frac{1}{cookies.Low}+\frac{1}{cookies.Middle}+\frac{1}{cookies.High})(\frac{1}{beverage.Low}+\frac{1}{beverage.Middle})$

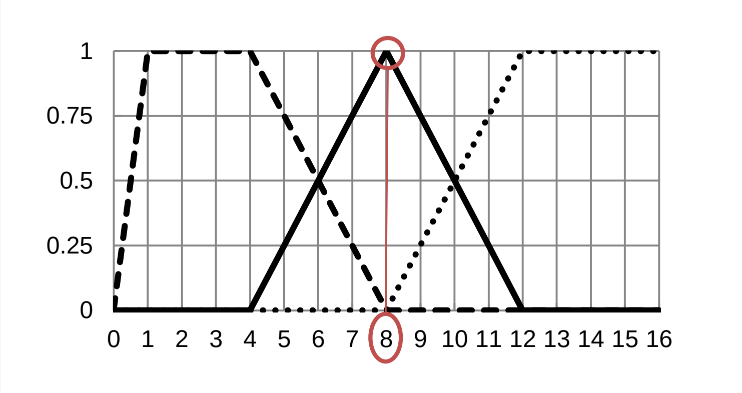
Step 3: Collect Fuzzy Regions

✓ 12 fuzzy regions can be collected

All of Fuzzy Regions	
milk.Low	cookies.Low
milk.Middle	cookies.Middle
milk.High	cookies.High
bread.Low	beverage.Low
bread.Middle	beverage.Middle
bread.High	beverage.High

Step 4: Calculate Utility of Fuzzy Itemsets Value (UFI)

✓Take milk.Middle as an example



Milk.M
0.5
0
0
0
0.75
0.25

Items	milk	bread	cookies	beverage
EU	5	3	1	7

$(0.5+0+0+0+0.75+0.25) \times (8 \times 5)$

Fuzzy regions	UFI	Fuzzy regions	UFI
milk.Low	8.75	cookies.Low	1
milk.Middle	60	cookies.Middle	8
milk.High	105	cookies.High	24
bread.Low	6	beverage.Low	1.75
breadMiddle	30	beverage.Middle	42
bread.High	63	beverage.High	84

Step 5: Calculate Total Utility (TU)

✓Total utility = 8.75+60+105+6+30+63+1+8+24+1.75+42+84 = 433.5

Fuzzy regions	UFI	Fuzzy regions	UFI
milk.Low	8.75	cookies.Low	1
milk.Middle	60	cookies.Middle	8
milk.High	105	cookies.High	24
bread.Low	6	beverage.Low	1.75
breadMiddle	30	beverage.Middle	42
bread.High	63	beverage.High	84

Steps 6 & 7: Put into the high utility fuzzy itemsets

✓Minimum utility ratio α = 10%

✓Minimum utility value = Total utility × Minimum utility ratio = 433.5 × 10% = 43.35

Fuzzy regions	UFI	Fuzzy regions	UFI
~~milk.Low~~	8.75	~~cookies.Low~~	1
milk.Middle	60	~~cookies.Middle~~	8
milk.High	105	~~cookies.High~~	24
~~bread.Low~~	6	~~beverage.Low~~	1.75
~~bread.Middle~~	30	~~beverage.Middle~~	42
bread.High	63	beverage.High	84

HUFI
milk.Middle
milk.High
bread.High
beverage.High

HUFI is not empty, continue

Step 8.1: Join Candidate High Utility Fuzzy Itemsets

✓Generate candidate High Utility Fuzzy Itemsets (CHUFI)

- Similar to generate candidate 2-itemsets

Fuzzy Regions
milk.Middle
milk.High
bread.High
beverage.High

CHUFI	
bread.High	milk.Middle
beverage.High	milk.Middle
bread.High	milk.High
beverage.High	milk.High
bread.High	beverage.High

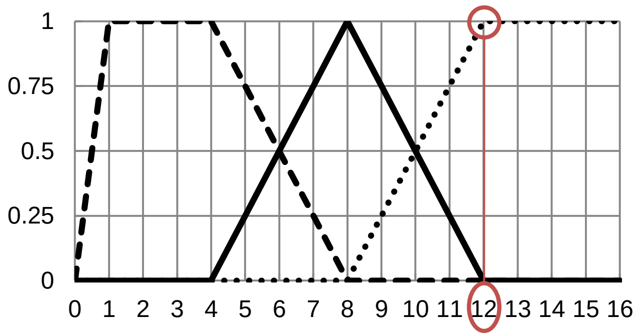
Step 8.2: Calculate Utility of Fuzzy Itemsets Value

✓Take (milk.High, bread.High) as an example

- $UFI_1(\text{milk.High}, \text{bread.High}) = (0.5+1+0.25) \times (5\times12) + (0.5+1+0.25) \times (3\times12)$

  $= 105 + 63 = 168$

Items	milk	bread	cookies	beverage
EU	5	3	1	7



milk.H	bread.H	TID
0.5	0.5	Y
1	1	Y
0	0	
0	0	
0.25	0.25	Y
0	0	

Step 8.3: Put into High Utility Fuzzy Itemsets (HUFI)

✓Continue previous example

Itemsets		UFI_1
bread.High	milk.Middle	77
beverage.High	milk.Middle	94
bread.High	milk.High	168
beverage.High	milk.High	0
bread.High	beverage.High	0

<43.35

<43.35

HUFI	
bread.High	milk.Middle
beverage.High	milk.Middle
bread.High	milk.High

Step 8.4: Calculate the Contingency Table

✓Take (milk.High, bread.High) as an example

milk.H	bread.H	TID
0.5	0.5	Y
1	1	Y
0	0	
0	0	
0.25	0.25	Y
0	0	

~milk.H	bread.H	TID
0.5	0.5	Y
0	1	
1	0	
1	0	
0.75	0.25	Y
1	0	

milk.H	~bread.H	TID
0.5	0.5	Y
1	0	
0	1	
0	1	
0.25	0.75	Y
0	1	

~milk.H	~bread.H	TID
0.5	0.5	Y
0	0	
1	1	Y
1	1	Y
0.75	0.75	Y
1	1	Y

Step 8.4: Calculate the Contingency Table (Cont.)

✓ Contingency tables of milk.High → bread.High and bread.High → milk.High should be calculated

(milk.High→bread.High)

	Consequence bread.H	Consequence ~bread.H
Antecedent milk.H	Q1: 168	Q2: 102
Antecedent ~milk.H	Q3: 90	Q4: 408

(bread.High →milk.High)

	Consequence milk.H	Consequence ~milk.H
Antecedent bread.H	Q1: 168	Q2: 90
Antecedent ~bread.H	Q3: 102	Q4: 408

Step 8.5: Check HUFI meets the four conditions

✓Check the fuzzy itemset meets the four conditions:

- Q1 > Q2 (168 > 90)
- Q1 > Q3 (168 > 102)
- Q4 > Q2 (408 > 90)
- Q4 > Q3 (408 > 102) …… put (milk.High, bread.High) into High **Coherent** UFI!

	Consequence bread.H	Consequence ~bread.H
Antecedent milk.H	Q1: 168 >	Q2: 102
	∨	∧
Antecedent ~milk.H	Q3: 90 <	Q4: 408

	Consequence milk.H	Consequence ~milk.H
Antecedent bread.H	Q1: 168 >	Q2: 90
	∨	∧
Antecedent ~bread.H	Q3: 102 <	Q4: 408

Steps 8.6 to 9

- ✓Step 8.6
  - ●Check the HCUFI set is empty or not
  - ●No, goto step 9.
- ✓Step 9
  - ●Parameter $k + 1$
  - ●Lengthen itemsets

Step 10: Output HCUFI

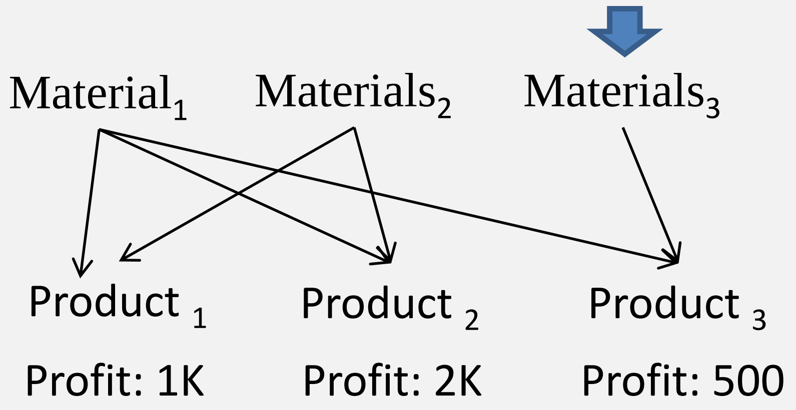
✓In this example, only a HCUFI is generated

HCUFI	
bread.High	milk.High

Erasable-Itemset Mining

✓ Possible Reasons for erasable-itemset mining

- Control the balance between cost and profit
- Meet financial crisis or other business reasons



META Algorithm

- ✓ Proposed by Deng et al. in 2009
  - A level-wise approach
  - Based on the *Apriori* algorithm
- ✓ Goal
  - Find materials can be eliminated and still have balance between cost and profit
  - Provide information for manager make a better production plan

An Example for META Algorithm

- ✓ Input dataset contain information of product, materials and profit

PID	Materials	Profit Value
P_1	*ABE*	200
P_2	*AB*	1000
P_3	*CD*	500
P_4	*BDEF*	50

- ✓ Parameter
  - A maximum erasable ratio $\alpha = 35\%$

Loss of Deleting An Itemset

✓Given the following information

PID	Items	Profit Value
P_1	*ABE*	200
P_2	*AB*	1000
P_3	*CD*	500
P_4	*BDEF*	50

✓Delete Itemset {*D*} affect P_3, P_4

- Loss(*D*) = 500 + 50 = 550

✓Delete Itemset {*DE*} affect P_1, P_3, P_4

- Loss(*DE*) = 200 + 500 + 50 = 750

An Erasable Itemset

✓Materials can be deleted

- Loss(X) ≤ total profit value (T) × maximum erasable ratio (α)

PID	Items	Profit Value
P_1	*ABE*	200
P_2	*AB*	1000
P_3	*CD*	500
P_4	*BDEF*	50

$T = 200 + 1000 + 500 + 50 = 1750$

Threshold $= 1750 \times 35\% = 612.5$

Loss(D) $= 550 < 612.5$, D is an erasable itemset

Loss(DE) $= 750 > 612.5$, DE is a non-erasable itemset

Mining Erasable Itemsets

Database		
PID	**Items**	**Profit Value**
P1	*ABE*	200
P2	*AB*	1000
P3	*CD*	500
P4	*BDEF*	50
Gain threshold		612.5

Scan database

Candidate 1-itemset	
A	1200
B	1250
C	500
D	550
E	250
F	50

By threshold

Erasable 1-itemset	
C	500
D	550
E	250
F	50

Mining Erasable Itemsets (Cont.)

Erasable 1-itemset	
C	500
D	550
E	250
F	50

Apriori-like approach →

Candidate 2-itemset
CD
CE
CF
DE
DF
EF

Scan database ↙

Candidate 2-itemset	
CD	550
CE	750
CF	550
DE	750
DF	550
EF	250

By threshold →

Erasable 2-itemset	
CD	550
CF	550
DF	550
EF	250

Mining Erasable Itemsets (Cont.)

Erasable 2-itemset	
CD	550
CF	550
DF	550
EF	250

Apriori-like approach

Candidate 3-itemset
CDF

Scan database

Candidate 3-itemset	
CDF	550

By threshold

Erasable 3-itemset	
CDF	550

Mining Erasable Itemsets (Cont.)

✓All the erasable itemsets

Final erasable itemsets	
C	500
D	550
E	250
F	50
CD	550
CF	550
DF	550
EF	250
CDF	550

Brief Summery (Cont.)

Single Minimum Support

Agrawal *et al.* (*SIGMOD'93*)

- Han *et al.* (*SIGMOD'00*)
  FP-growth algorithm

Taxonomy

- Han *et al.* (*VLDB'95*)
  Level-by-Level Approach
- Srikant *et al.* (*VLDB'95*)
  Generalized Approach

Quantitative Value

- Hong et al. (IDA)
  Fuzzy AR mining
- Chen et al. (Applied soft computing)
  Fuzzy temporal AR mining

Apriori

Multiple MinSup

Multiple Minimum Support

Liu *et al.* (*KDD'99*) -MIS

- Lee *et al.* (*IJAR, 2005*)
  Maximum constraints

Taxonomy

- Lui *et al.* (*LNAI*, 2000)
  Generalized Approach
- Tzeng *et al.* (*LNCS*, 2001)
  Generalized Approach

Utility Itemset Mining

- Chen *et al.* (*Soft Computing'14*)
  High Coherent Utility Fuzzy Itemsets Mining

Erasable Itemset Mining

- Deng et al. (ICMLC'09)
  Erasable Itemset Mining

Conclusions

- ✓Single minimum support
  - ●Utility itemset mining with taxonomy
  - ●Erasable itemset mining with taxonomy
  - ●Erasable itemset mining with quantitative value
  - ●Genetic-fuzzy data mining
  - ●etc
- ✓ Multiple minimum support
  - ●Topics listed above can take multiple minimum support into consideration