UNIT - II

* FREQUENT PATTERNS

the patterns that appear frequently in dataset
(include frequent data items, sequences, Substructures)

Example: Milk and bread.

market Basket Analysis:
process of Analysing customer buying habits by finding the associations b/w the diff. items that a customer will place in their baskets.
- mainly useful for sellers.

Strategies Used:

1. placing them together.
2. placing them at ② different Ends.

- This Analysis will help Sellers to plan their Shelf Space for increased Sales.
- frequent patterns are represented by assiociation Rule

Ex: Computer and anti-virus.

Support:
identifies how frequently a rule is applied to given dataset.

$$S(P \rightarrow Q) = \frac{\sigma(P \cup Q)}{N} \quad (\because N = \text{Total Transactions})$$

$P(A \cup B)$

Confidence:
define frequent occurrence of items of Q in transactions of

$$C(P \to Q) = P(B/A)$$

* MINING METHODS

- Apriori Algorithm
- Fp Growth Algorithm

Apriori Algorithm

- by R. Agarwal and R. Srikant.

- Shows how objects are associated with each other

objective: To generate an association.

Example:
minimum Support = 50%
Threshold confidence = 70%

T/D	items
100	1 3 4
200	2 3 5
300	1 2 3
400	2 5

Itemset	Support	miniSupport
1	2	2/4 = 50%
2	3	3/4 = 75%
3	3	3/4 = 75% (X)
4	1	1/4 = 25%
5	3	3/4 = 75%

Itemset: (1, 2, 3, 5)

— Form pairs

(1,2) (1,3) (1,5) (2,3) (2,5) (3,5)

itemset	Support	minimum Support
(1,2)	1	1/4 = 25% (x)
(1,3)	2	2/4 = 50%
(1,5)	1	1/4 = 25% (x)
(2,3)	2	2/4 = 50%

itemset = (1,3) (2,3) (2,5) (3,5)

- Form triplets

(1,2,3) (1,2,5) (1,3,5) (2,3,5)

itemset	Support	minimum Support.
(1,2,3)	1	1/4 = 25%
(1,2,5)	1	1/4 = 20%
(1,3,5)	1	1/4 = 25%
(2,3,5)	2	2/4 : 50%

Itemset : (2,3,5)

- now that lets calculate Support and confidence

Confidence = Support (A∪B) / Support of A

using (2,3,5) we can generate association Rules

Rules	Support	confidence.
$(2 \wedge 3) \to 5$	2	2/2 : 100%
$(3 \wedge 5) \to 2$	2	2/2 = 100%
$(2 \wedge 5) \to 3$	2	2/3 = 66% (x)
$2 \to (3 \wedge 5)$	2	2/3 = 66% (x)
$5 \to (2 \wedge 3)$	2	2/3 = 66% (x)
$3 \to (2 \wedge 5)$	2	2/3 = 66% (x)

$(2 \wedge 3) \rightarrow 5$ - confidence $= \frac{\text{Support}(A \cup B)}{\text{Support}(A)}$ 2,3,5

$S \frac{((2 \wedge 3) \cup 5)}{S(2 \wedge 3)} = \frac{2}{2} = 100\%$

$2 \rightarrow (3 \wedge 5) : \frac{S(2 \cup (3 \wedge 5))}{S(2)} = 2/3 = 66\%$

$\therefore (2 \wedge 3) \rightarrow 5$, $(3 \wedge 5) \rightarrow 2$ are association rules

* FP GROWTH ALGORITHM

Fp → frequent pattern

- is an efficient and scalable method for mining the complete set of Fp using a tree structure for storing information about fp called fp tree.

Example:

minimum support = 30%.

Trans id	items
1	E, A; D B
2	D, A, E, C, B
3	C, A, B, E
4	B, A, D
5	D
6	D, B
7	A, D, E
8	B, C

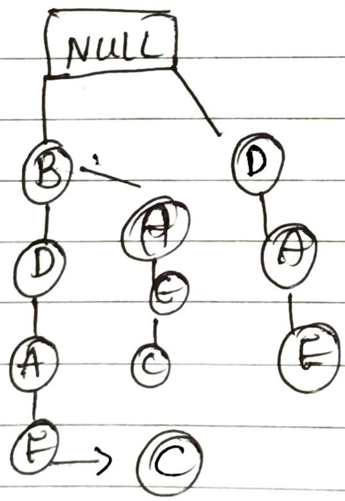
write priorities
More frequency → more priority
Same frequency → FCFS

- list out the priorities

itemset	frequency	priority
A	5	3
B	6	1
C	3	5
D	6	2
E	4	4

- Order items according to priority.

Trans ID	Items	Ordered Items
1	E A D B	B D A E
2	D A E C B	B D A E C
3	C A B E	B A E C
4	B A D	B D A
5	D	D
6	D B	B D
7	A D E	D A E
8	B C	BC



B – 1, 2, 3
D – 1, 2
A – 1, 2
E – 1, 2
C – 1
A – 1
E – 1
C – 1
D – 1, 2
A – 1

* MINING VARIOUS KINDS OF Association RULES

- 4 types

1. mining Multilevel association Rules
   - ~~mining~~ Uniform Support for all levels
   - Using reduced minimum Support at lower levels.
   - using item or group based minimum Support
2. mining Multidimensional association rules from relational database or data warehouse.
3. mining Multi dimensional association Rules using static discretisation of Quantitave attributes
4. minining quantitative association Rules.

* CORRELATION ANALYSIS

used to measure the relationship b/w 2 variables.

$$r_{A,B} = \frac{\Sigma (A - A')(B - B')}{(n-1)\sigma_A \sigma_B}$$

$r_{A,B}$ = karle pearson correlation cofficient.

A', B' = mean of A and B.

σ_A, σ_B = Standard deviation of A and B.

n = no. of tuples in db.

$r \rightarrow$ 3 values $(0, -1, +1)$

$r \rightarrow +1 \Rightarrow$ perfect positive correlation

$r \rightarrow 0 \Rightarrow$ No correlation (no dependence)

$r \rightarrow -1 \Rightarrow$ perfect Negative correlation.

Example:

A	B
20	8
12	34
	4

$$r_{A,B} = \frac{\Sigma (A - A')(B - B')}{(n-1)\sigma_A \sigma_B}$$

$$A' = \frac{20+12+9}{3} = 13.66 \quad B' = \frac{8+34+4}{3} = 15.33$$

$$\sigma A = \sqrt{\frac{\Sigma(A-A')^2}{n-1}}$$

$$= \sqrt{\frac{(20-13.66)^2+(12-13.66)^2+(9-13.66)^2}{2}} = 5.68$$

$$\sigma B = \sqrt{\frac{\Sigma(B-B')^2}{n-1}}$$

$$= \sqrt{\frac{(8-15.33)^2+(34-15.33)^2+(4-15.33)^2}{2}} = 16.28$$

$$r_{A,B} = \frac{(20-13.66)(8-15.33)+(12-13.66)(34-15.33)+(9-13.66)(4-15.33)}{2 \times 5.68 \times 16.28}$$

$= -1 \ldots$

≈ -1

i.e negative correlation

* CONSTRAINT BASED ASSOCIATION MINING:

Constraint – Condition
- association Rules are generated based on conditions

* Types of constraints.

1. Knowledge Type:

- Specifies the types of knowledge you want to do mining – association, correlation, Regression etc,

2. Data Constraints

- Specifies the type of data on which you want to generate the Rules.
  - only task relevant data.

3. dimension level Constraints.
Specifies the dimension or level concept hierarchy.

4. Interestingness Constraints
Support, confidence are used to Identify.

5. Rule Constraints.
Specifies the form of rules to be mined.
2 ways

1. metarules guided mining
2. constraint pushing.

* GRAPH PATTERN MINING

set of tools techniques used to mine frequent Subgroups Subgraphs.

- Used to Analyse the properties of real world graphs

- used to Analyse how structure of graph will effect the rules

2 ways

1. Apriori based approaches
2. pattern growth approaches

Algorithms used:

1. Gspan → all types
2. closed Graph → closed Subgraphs.

Applications
1. in xmL Structures
2. anomaly detection
3. Network Analysis
4. control flow Analysis
5. Biological Structures etc,

moodbanao.net

* SEQUENTIAL PATTERN MINING: (Spm)

sequence = set of ordered events

Ex: S = { e_1, e_2, e_3, e_4, e_5 }

Spm → process of finding frequent Subsequences from a set of sequences.

Sequences are represented by "< >"

Normal transaction data

CID	TID	Transactions
1	100	a, b, c, d
3	111	a, f, d, e
1	122	d, e, f
3	133	b, f, s, a
1	144	b, c, d, e
3	155	a, f, d, c
1	166	a, e, p

Sequential data.

CID	Sequences
1.	< (abcd), (dep), (bcde), (aep) >
3.	< (afde), (bfsa), (afdc) .

Challenges in Spm:

- finding all Subsequences

min-Sup = 2

Sid	Sequence
10	< a (abc) (ac) d (cf) >
20	< (ad) c (bc) (ae) >
30	< (ef) (ab) (dfcb) >
40	< eg (af) (bc >

min-Sup = 2

< (ab) c > (✓)

< eg > (✗)

Algorithms used:

1. GSP (Generalised Sequential patterns)
2. SPADE (vertical format based mining)
3. prefixspan
4. clospan - for closed patterns.