IS422P - DATA MINING
6- MINING FREQUENT PATTERNS, ASSOCIATIONS, & CORRELATIONS

AMIRA REZK

INFORMATION SYSTEM DEPARTMENT

The Basics

Market Basket Analysis

Frequent Itemsets

Association Rules

Frequent Itemset Mining Methods

Apriori Algorithm

Generating Association Rules from Frequent Itemsets

FP-Growth

Pattern Evaluation Methods

The Basics

Market Basket Analysis

Frequent Itemsets

Association Rules

Frequent Itemset Mining Methods

Apriori Algorithm

Generating Association Rules from Frequent Itemsets

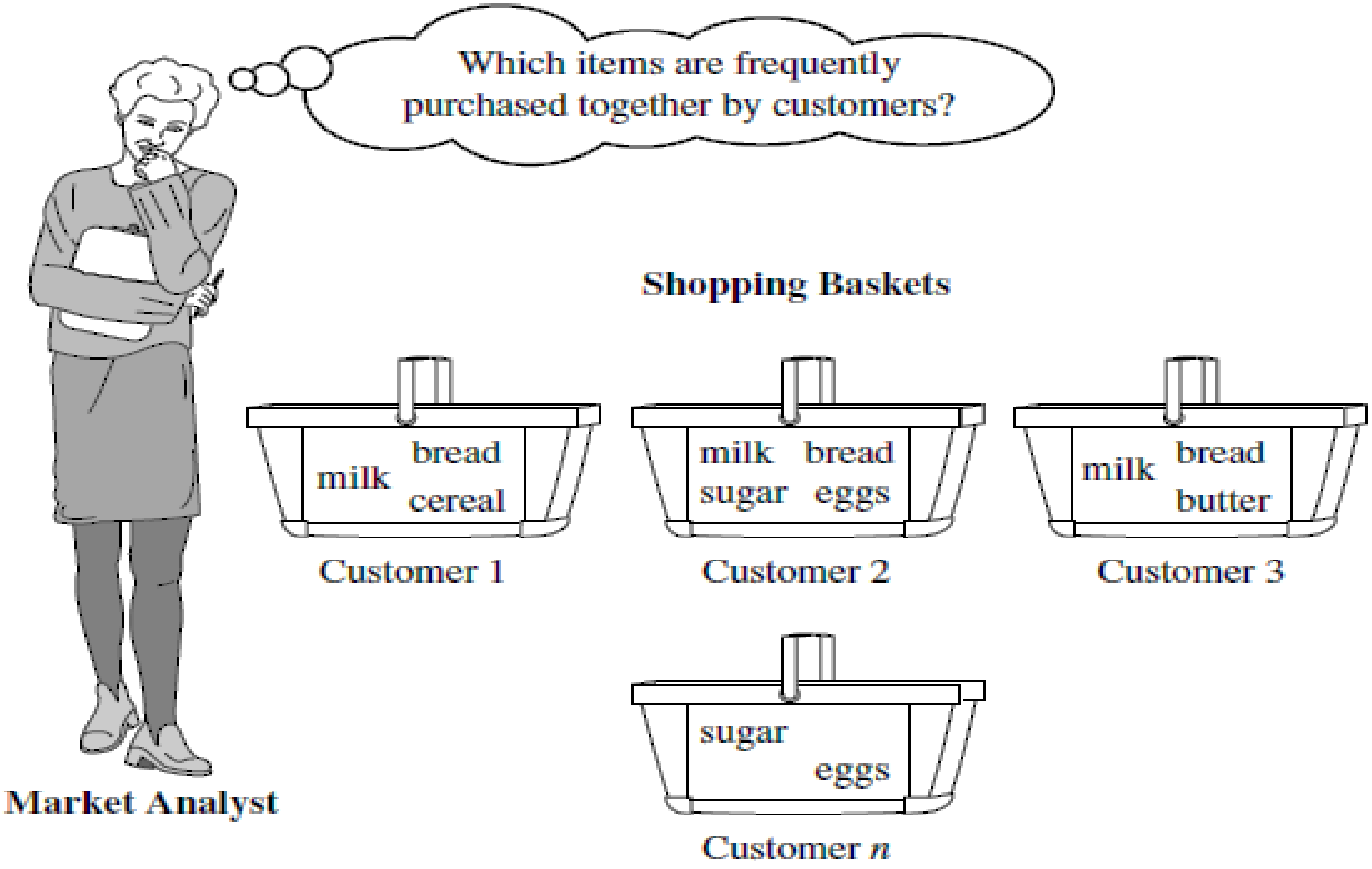
FP-Growth

Pattern Evaluation Methods

THE BASICS
WHAT IS FREQUENT PATTERN ANALYSIS?

- Frequent pattern: a pattern (a set of items, subsequences, substructures, etc.) that occurs frequently in a data set
- First proposed by Agrawal, Imielinski, and Swami [AIS93] in the context of frequent itemsets and association rule mining

THE BASICS


Which items are frequently purchased together by customers?
Shopping Baskets
milk
bread
cereal
Customer 1
milk
bread
sugar
eggs
Customer 2
milk
bread
butter
Customer 3
sugar
eggs
Customer n
Market Analyst

THE BASICS

- Motivation: Finding inherent regularities in data
  - What products were often purchased together?— Beer and diapers?!
  - 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?
- Applications
  - Basket data analysis, cross-marketing, catalog design, sale campaign analysis, Web log (click stream) analysis, and DNA sequence analysis

THE BASICS

- Frequent Pattern are itemsets that appear frequently in a data set (e.g. Transaction record)
- Items that are frequently associated (e.g purchased) together can be represented as association rules

Computer → antivirus_SW [Support = 2% , Confidence =60%]

- Support and Confidence are measures of rule interestingness
- 2% Support means 2% of Transactions Show that computers and antivirus_SW are bought Together
- 60% Confidence means 60 % of customers who bought a computer also bought antivirus_SW

THE BASICS
FREQUENT ITEM-SETS

- Itemset X = {x1, …, xk}   ex: X={A, B, C, D, E, F}
- Find all the rules X → Y with minimum support and confidence
  - support, s, probability that a transaction contains X ∪ Y
  - confidence, c, conditional probability that a transaction having X also contains Y

$$support\ X \to Y = P(X \cup Y) = \frac{n(X \cup Y)}{N}$$

$$confidence\ (X \to Y) = P(Y|X) = \frac{n(X \cup Y)}{n(X)}$$

THE BASICS
ASSOCIATION RULES

Ex: Let min_Sup. = 50%, min_conf. = 50%

Frequent Patterns:

{*A:3, B:3, D:4, E:3, AD:3*}

Association rules:

A → *D* (60%, 100%)

D → *A* (60%, 75%)

$$conf\ (A \rightarrow D) = \frac{3}{3} = 100\ \%$$

$$conf\ (D \rightarrow A) = \frac{3}{4} = 75\ \%$$

Transaction-id	Items bought
10	A, B, D
20	A, C, D
30	A, D, E
40	B, E, F
50	B, C, D, E, F

THE BASICS
ASSOCIATION RULES

- If frequency of itemset I satisfies min_support count then I is a frequent itemset
- If a rule satisfies min_support and min_confidence thresholds, it is said to be strong
  - problem of mining association rules reduced to mining frequent itemsets
- Association rules mining becomes a two-step process:
  - Find all frequent itemsets that occur at least as frequently as a predetermined min_support count
  - Generate strong association rules from the frequent itemsets that satisfy min_support and min_confidence

The Basics

Market Basket Analysis

Frequent Itemsets

Association Rules

Frequent Itemset Mining Methods

Apriori Algorithm

Generating Association Rules from Frequent Itemsets

FP-Growth

Pattern Evaluation Methods

MINING FREQUENT ITEMSETS
APRIORI ALGORITHM

- Goes as follows:
  - Find frequent 1-itemsets → L1
  - Use L1 to find frequent 2-itemsets → L2
  - … until no more frequent k-itemsets can be found
- Each Lk itemset requires a full dataset scan
- To improve efficiency, use the Apriori property:
  - "All nonempty subsets of a frequent itemset must also be frequent" – if a set cannot pass a test, all of its supersets will fail the same test as well – if $P(I) <$ min_support then $P(I \cup A) <$ min_support

MINING FREQUENT ITEMSETS
APRIORI ALGORITHM

Transactional data example
N=10, min_supp count=2

TID	List of items
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
T1000	I1, I2

Scan dataset for count of each candidate

C_1

Itemset	Support count
{I1}	7
{I2}	8
{I3}	6
{I4}	2
{I5}	2

Compare candidate support with min_support

L_1

Itemset	Support Count
{I1}	7
{I2}	8
{I3}	6
{I4}	2
{I5}	2

MINING FREQUENT ITEMSETS
APRIORI ALGORITHM

Itemset	Support Count
{I1}	7
{I2}	8
{I3}	6
{I4}	2
{I5}	2

C_2

Itemset
{I1, I2}
{I1, I3}
{I1, I4}
{I1, I5}
{I2, I3}
{I2, I4}
{I2, I5}
{I3, I4}
{I3, I5}
{I4, I5}

Generate C_2 candidates from L_1 by joining $L_1 \bowtie L_1$

C_2

Itemset	Support count
{I1, I2}	5
{I1, I3}	4
{I1, I4}	1
{I1, I5}	3
{I2, I3}	4
{I2, I4}	2
{I2, I5}	2
{I3, I4}	0
{I3, I5}	1
{I4, I5}	0

Scan dataset for count of each candidate

L_2

Itemset	Support count
{I1, I2}	5
{I1, I3}	4
{I1, I5}	3
{I2, I3}	4
{I2, I4}	2
{I2, I5}	2

Compare candidate support with min_supp

MINING FREQUENT ITEMSETS
APRIORI ALGORITHM

$C_3 = L2 \rhd\lhd L2$ = {{I1, I2, I3}, {I1, I2, I5}, ~~{I1, I3, I5}, {I2, I3, I4}, {I2, I3, I5}, {I2, I4, I5}~~}

Not all subsets are frequent → **Prune** (Apriori property)

Itemset	Support count
{I1, I2}	5
{I1, I3}	4
{I1, I5}	3
{I2, I3}	4
{I2, I4}	2
{I2, I5}	2

Generate C_3 candidates from L_2 by **joining** $L_2 \rhd\lhd L_2$

Itemset
{I1, I2, I3}
{I1, I2, I5}

Scan dataset for count of each candidate

C_3

Itemset	Support count
{I1, I2, I3}	2
{I1, I2, I5}	2

Compare candidate support with min_supp

L_3

Itemset	Support count
{I1, I2, I3}	2
{I1, I2, I5}	2

Two joining (lexicographically ordered) k-itemsets must share first k-1 items →{I1, I2} is not joined with {I2, I4}

MINING FREQUENT ITEMSETS
APRIORI ALGORITHM

Itemset	Support count
{I1, I2, I3}	2
{I1, I2, I5}	2

→

Itemset
{I1, I2, I3, I5}

→

Not all subsets are frequent → **Pruning**

$C_4 = \phi$ → Terminate

APRIORI ALGORITHM

Algorithm: Apriori. Find frequent itemsets using an iterative level-wise approach based on candidate generation.

Input:

- D, a database of transactions;
- *min_sup*, the minimum support count threshold.

Output: L, frequent itemsets in D.

Method:

Generate C_k using L_{k-1} to find L_k

```
L_1 = find_frequent_1-itemsets(D);
for (k = 2; L_{k-1} ≠ φ; k++) {
    C_k = apriori_gen(L_{k-1});
    for each transaction t ∈ D { // scan D for counts
        C_t = subset(C_k, t); // get the subsets of t that are candidates
        for each candidate c ∈ C_t
            c.count++;
    }
    L_k = {c ∈ C_k | c.count ≥ min_sup}
}
return L = ∪_k L_k;
```

Join

Prune

```
procedure apriori_gen(L_{k-1}:frequent (k − 1)-itemsets)
    for each itemset l_1 ∈ L_{k-1}
        for each itemset l_2 ∈ L_{k-1}
            if (l_1[1] = l_2[1]) ∧ (l_1[2] = l_2[2])
                ∧ ... ∧ (l_1[k − 2] = l_2[k − 2]) ∧ (l_1[k − 1] < l_2[k − 1]) then {
                c = l_1 ⋈ l_2; // join step: generate candidates
                if has_infrequent_subset(c, L_{k-1}) then
                    delete c; // prune step: remove unfruitful candidate
                else add c to C_k;
            }
    return C_k;
```

```
procedure has_infrequent_subset(c: candidate k-itemset;
            L_{k-1}: frequent (k − 1)-itemsets); // use prior knowledge
    for each (k − 1)-subset s of c
        if s ∉ L_{k-1} then
            return TRUE;
    return FALSE;
```

MINING FREQUENT ITEMSETS
GENERATING ASSOCIATION RULES FROM FREQUENT ITEMSETS

- Association rules van be generated using the confidence equation, as follows"

$$confidence\ (A \Rightarrow B) = P(B|A) = \frac{support\_count(A \cup B)}{support\_count(A)}$$

- For each frequent itemset L, generate all nonempty subset of L
- For every nonempty subset S of L, out put rule: S → L-S
- $\frac{Support_{count}(L)}{Support_{count}(S)}$ ≥min_conf,

where min_conf is the minimum confidence threshold.

MINING FREQUENT ITEMSETS
GENERATING ASSOCIATION RULES FROM FREQUENT ITEMSETS

Itemset	Support count
{I1, I2, I3}	2
$\boxed{\text{\{I1, I2, I5\}}}$	$\boxed{2}$

Nonempty subsets	Association Rules	Confidence
{I1, I2}	{I1, I2}→ I5	2/5 = 40%
{I1, I5}	{I1, I5}→ I2	2/2 = 100%
{I2, I5}	{I2, I5}→ I1	2/2 = 100%
{I1}	I1→ {I2, I5}	2/7 = 28%
{I2}	I2→ {I1, I5}	2/8 = 25%
{I5}	I5→ {I1, I2}	2/2 = 100%

For a *min_confidence* = 70%

MINING FREQUENT ITEMSETS
FP-GROWTH

- To avoid costly candidate generation
- Divide-and-conquer strategy:
- Compress database representing frequent items into a frequent pattern tree (FP-tree) – 2 passes over dataset
- Divide compressed database (FP-tree) into conditional databases, then mine each for frequent itemsets – traverse through the FP-tree

MINING FREQUENT ITEMSETS
FP-GROWTH

Transactional data example
N=10, min_supp count=2

TID	List of items
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
T1000	I1, I2

Scan dataset for count of each candidate

C_1

Itemset	Support count
{I1}	7
{I2}	8
{I3}	6
{I4}	2
{I5}	2

Compare candidate support with min_supp

L_1 - Reordered

Itemset	Support count
{I2}	8
{I1}	7
{I3}	6
{I4}	2
{I5}	2

MINING FREQUENT ITEMSETS
FP-GROWTH – FP-TREE CONSTRUCTION

L_1 - Reordered

Itemset	Support count
{I2}	8
{I1}	7
{I3}	6
{I4}	2
{I5}	2

FP-tree

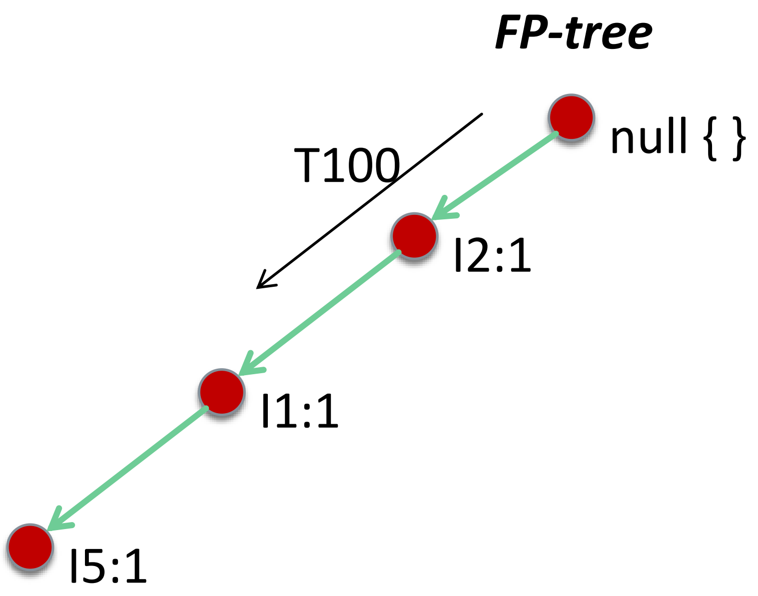
null { }

TID	List of items
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
T1000	I1, I2

MINING FREQUENT ITEMSETS
FP-GROWTH – FP-TREE CONSTRUCTION

L_1 - Reordered

Itemset	Support count
{I2}	8
{I1}	7
{I3}	6
{I4}	2
{I5}	2



Order of items is kept throughout path construction, with common prefixes shared whenever applicable

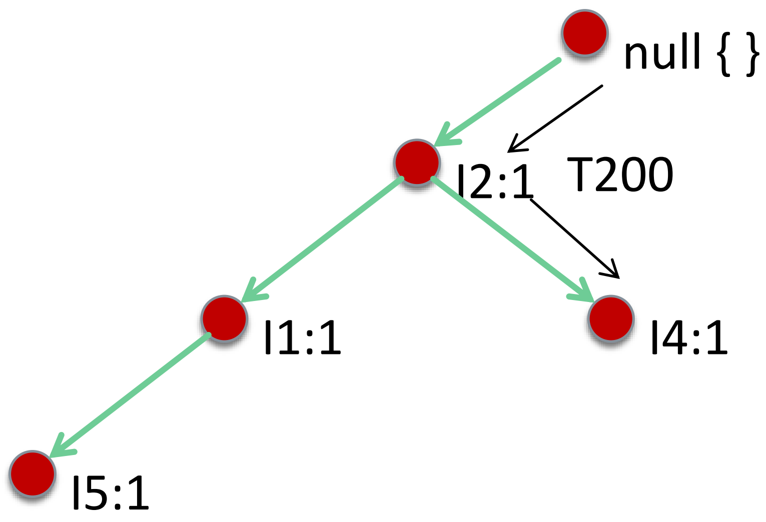
TID	List of items
T100	I1, I2, I5
T200	I2, I4
T300	I2, I3
T400	I1, I2, I4
T500	I1, I3
T600	I2, I3
T700	I1, I3
T800	I1, I2, I3, I5
T900	I1, I2, I3
T1000	I1, I2

MINING FREQUENT ITEMSETS
FP-GROWTH – FP-TREE CONSTRUCTION

L_1 - Reordered

Itemset	Support count
{I2}	8
{I1}	7
{I3}	6
{I4}	2
{I5}	2

FP-tree



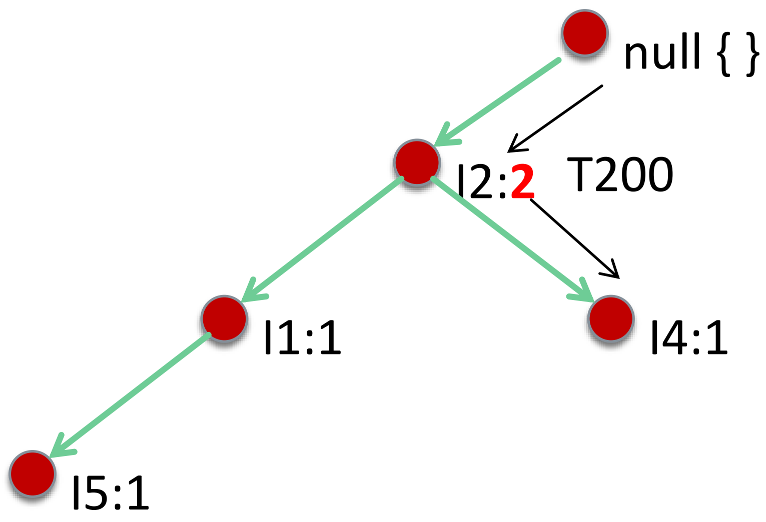
TID	List of items
T100	I1, I2, I5
T200	I2, I4
T300	I2, I3
T400	I1, I2, I4
T500	I1, I3
T600	I2, I3
T700	I1, I3
T800	I1, I2, I3, I5
T900	I1, I2, I3
T1000	I1, I2

MINING FREQUENT ITEMSETS
FP-GROWTH – FP-TREE CONSTRUCTION

L_1 - Reordered

Itemset	Support count
{I2}	8
{I1}	7
{I3}	6
{I4}	2
{I5}	2

FP-tree



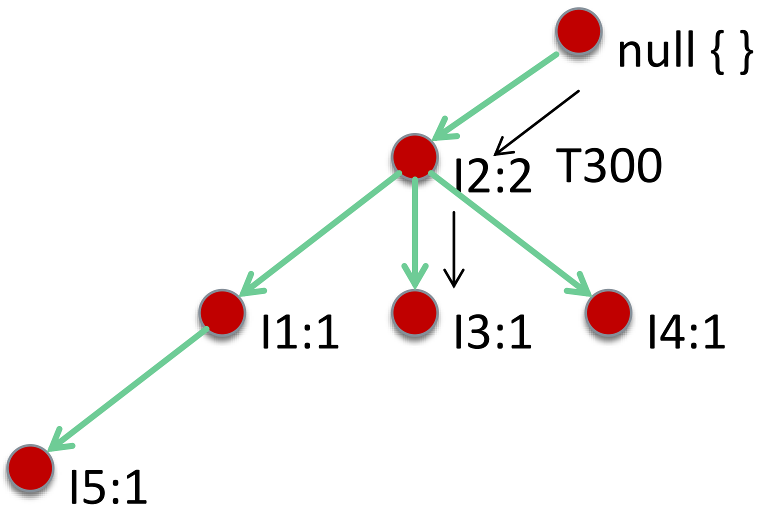
TID	List of items
T100	I1, I2, I5
T200	I2, I4
T300	I2, I3
T400	I1, I2, I4
T500	I1, I3
T600	I2, I3
T700	I1, I3
T800	I1, I2, I3, I5
T900	I1, I2, I3
T1000	I1, I2

MINING FREQUENT ITEMSETS
FP-GROWTH – FP-TREE CONSTRUCTION

L_1 - Reordered

Itemset	Support count
{I2}	8
{I1}	7
{I3}	6
{I4}	2
{I5}	2

FP-tree



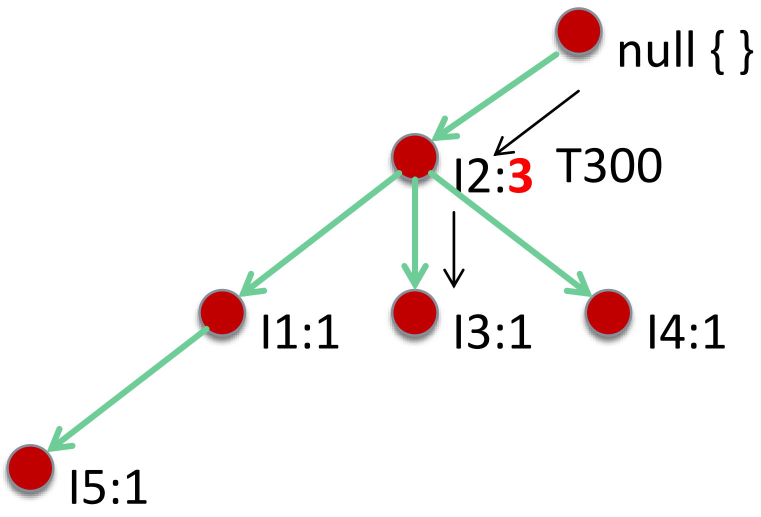
TID	List of items
T100	I1, I2, I5
T200	I2, I4
T300	I2, I3
T400	I1, I2, I4
T500	I1, I3
T600	I2, I3
T700	I1, I3
T800	I1, I2, I3, I5
T900	I1, I2, I3
T1000	I1, I2

MINING FREQUENT ITEMSETS
FP-GROWTH – FP-TREE CONSTRUCTION

L_1 - Reordered

Itemset	Support count
{I2}	8
{I1}	7
{I3}	6
{I4}	2
{I5}	2

FP-tree



TID	List of items
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
T1000	I1, I2

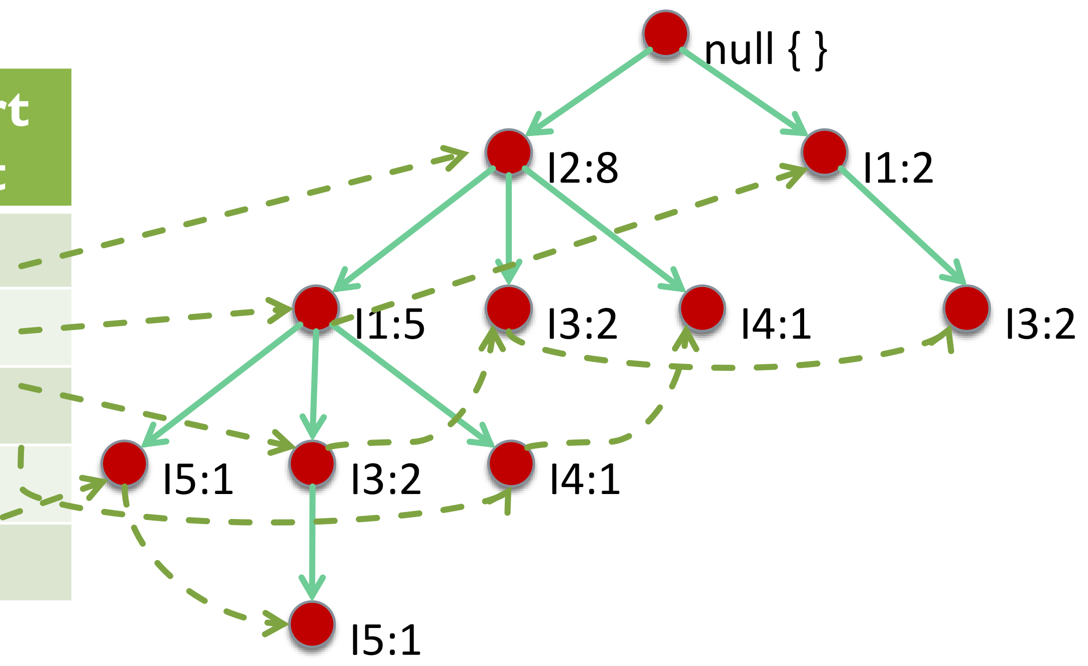
MINING FREQUENT ITEMSETS
FP-GROWTH – FP-TREE CONSTRUCTION

L_1 - Reordered

Itemset	Support count
{I2}	8
{I1}	7
{I3}	6
{I4}	2
{I5}	2

For Tree Traversal

FP-tree



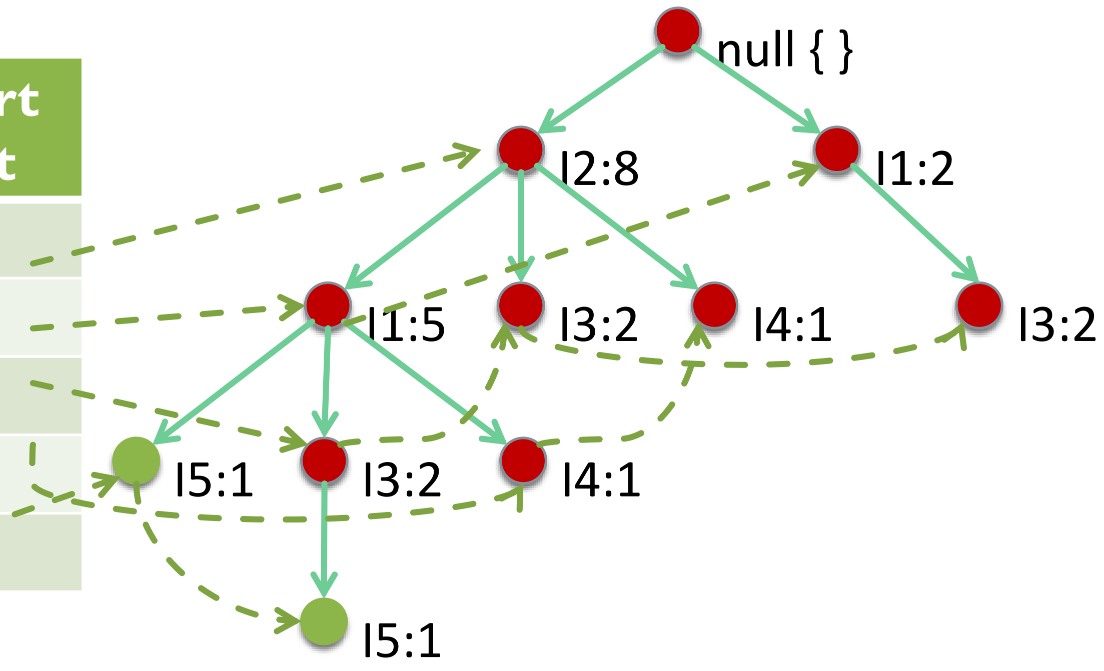
TID	List of items
T100	I1, I2, I5
T200	I2, I4
T300	I2, I3
T400	I1, I2, I4
T500	I1, I3
T600	I2, I3
T700	I1, I3
T800	I1, I2, I3, I5
T900	I1, I2, I3
T1000	I1, I2

MINING FREQUENT ITEMSETS
FP-GROWTH – FREQUENT PATTERNS MINING

L_1 - Reordered

Itemset	Support count
{I2}	8
{I1}	7
{I3}	6
{I4}	2
{I5}	2

FP-tree



TID	List of items
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
T1000	I1, I2

Bottom-up algorithm – start from leaves and go up to root

MINING FREQUENT ITEMSETS
FP-GROWTH – CONDITIONAL FP-TREE CONSTRUCTION

For I5

L_1 - Reordered

Itemset	Support count	Linked link
{I2}	8	
{I1}	7	
{I3}	6	
{I4}	2	
{I5}	2	

FP-tree

null { }

Eliminate I5

Eliminate transactions not including I5

TID	List of items
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~~
~~T1000~~	~~I1, I2~~

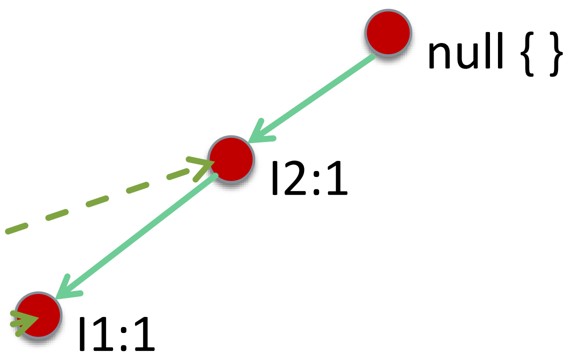
MINING FREQUENT ITEMSETS
FP-GROWTH – CONDITIONAL FP-TREE CONSTRUCTION

For I5

***L_1* - Reordered**

Itemset	Support count	Linked link
{I2}	8	
{I1}	7	
{I3}	6	
{I4}	2	
{I5}	2	

FP-tree



TID	List of items
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~~
~~T1000~~	~~I1, I2~~

Eliminate I5

Eliminate transactions not including I5

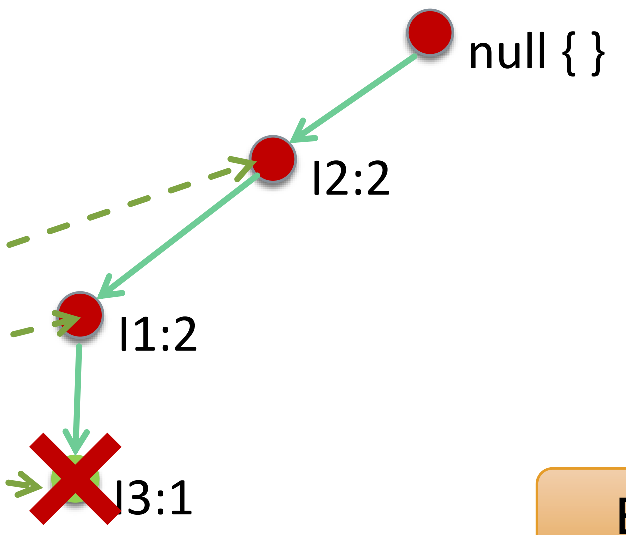
MINING FREQUENT ITEMSETS
FP-GROWTH – CONDITIONAL FP-TREE CONSTRUCTION

For I5

L_1 - Reordered

Itemset	Support count	Linked link
{I2}	8	
{I1}	7	
{I3}	6	
{I4}	2	
{I5}	2	

FP-tree



TID	List of items
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~~
~~T1000~~	~~I1, I2~~

Eliminate I5

Eliminate transactions not including I5

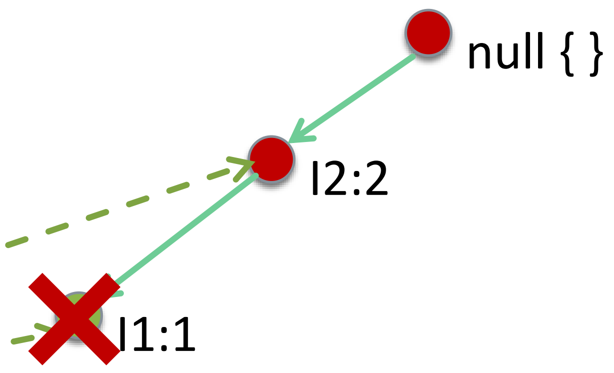
MINING FREQUENT ITEMSETS
FP-GROWTH – CONDITIONAL FP-TREE CONSTRUCTION

For I4

***L_1* - Reordered**

Itemset	Support count	Linked link
{I2}	8	
{I1}	7	
{I3}	6	
{I4}	2	
{I5}	2	

FP-tree



TID	List of items
~~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~~
~~T1000~~	~~I1, I2~~

Eliminate I4

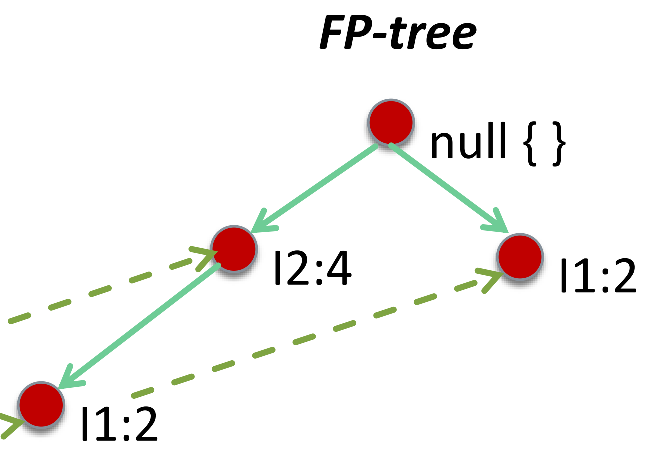
Eliminate transactions not including I4

MINING FREQUENT ITEMSETS
FP-GROWTH – CONDITIONAL FP-TREE CONSTRUCTION

For I3

***L_1* - Reordered**

Itemset	Support count	Linked link
{I2}	8	
{I1}	7	
{I3}	6	
{I4}	2	
{I5}	2	



TID	List of items
~~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~~
~~T1000~~	~~I1, I2~~

Eliminate I3

Eliminate transactions not including I3

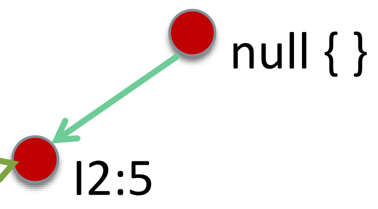
MINING FREQUENT ITEMSETS
FP-GROWTH – CONDITIONAL FP-TREE CONSTRUCTION

For I1

L_1 - Reordered

Itemset	Support count	Linked link
{I2}	8	
{I1}	7	
{I3}	6	
{I4}	2	
{I5}	2	

FP-tree



TID	List of items
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
T1000	~~I1~~, I2

Eliminate I1

Eliminate transactions not including I1

MINING FREQUENT ITEMSETS
FP-GROWTH

Item	Conditional Pattern Base	Conditional FP-tree	Frequent Patterns Generated
I5	{{I2, I1: 1}, {I2, I1, I3: 1}}	<I2:2, I1:2>	{I2, I5: 2}, {I1, I5: 2}, {I2, I1, I5: 2}
I4	{{I2, I1: 1}, {I2: 1}}	<I2:2>	{I2, I4: 2}
I3	{{I2, I1: 2}, {I2: 2}, {I1: 2}}	<I2:4, I1:2>, <I1:2>	{I2, I3: 4}, {I1, I3: 4}, {I2, I1, I3: 2}
I1	{{I2: 5}}	<I2:5>	{I2, I1: 5}

Paths ending with item

I5 → I2

I2 → I5

The Basics

Market Basket Analysis

Frequent Itemsets

Association Rules

Frequent Itemset Mining Methods

Apriori Algorithm

Generating Association Rules from Frequent Itemsets

FP-Growth

Pattern Evaluation Methods

PATTERN EVALUATION METHODS

- Not all association rules are <u>interesting</u>
  - Buys(X,"Computer games" → buys(X,"Videos") [40%, 66%]
  - P("videos") is 75% > 66%
  - The two items are negatively associated means buying one decreases the likelihood of buying the other
  - We need to measure "real strength" of rule
- **Correlation analysis**
  - A→ B [support , confidence , correlation]

PATTERN EVALUATION METHODS

1. Lift = $\frac{P(A \cup B)}{P(A)P(B)}$
   - A and B are independent if P(A U B) = P(A)P(B)
   - Otherwise, dependent and correlated occurrence
   - If lift <1, A is Negatively correlated with B
   - If lift >1, A is Positively correlated with B …… A's occurrence "lifts" the occurrence of B
2. $\chi 2$ → already discussed in previous lecture

QUESTIONS?

NEXT …