Association Rule Problem

Given a database of transactions:

Transaction	Items
t_1	Bread,Jelly,PeanutButter
t_2	Bread,PeanutButter
t_3	Bread,Milk,PeanutButter
t_4	Beer,Bread
t_5	Beer,Milk

- Find all the association rules:

$X \Rightarrow Y$	s	α
Bread $\Rightarrow$ PeanutButter	60%	75%
PeanutButter $\Rightarrow$ Bread	60%	100%
Beer $\Rightarrow$ Bread	20%	50%
PeanutButter $\Rightarrow$ Jelly	20%	33.3%
Jelly $\Rightarrow$ PeanutButter	20%	100%
Jelly $\Rightarrow$ Milk	0%	0%

Association Rule Definitions

- $I=\{i_1, i_2, ..., i_n\}$: a set of all the items
- Transaction T: a set of items such that $T \subseteq I$
- Transaction Database D: a set of transactions
- A transaction $T \subseteq I$ contains a set $X \subseteq I$ of some items, if $X \subseteq T$
- *An Association Rule*: is an implication of the form $X \Rightarrow Y$, where $X, Y \subseteq I$

Association Rule Definitions

- A set of items is referred as an itemset. A itemset that contains k items is a k-itemset.
- The support s of an itemset X is the percentage of transactions in the transaction database D that contain X.
- The support of the rule $X \Rightarrow Y$ in the transaction database D is the support of the items set $X \cup Y$ in D.
- The confidence of the rule $X \Rightarrow Y$ in the transaction database D is the ratio of the number of transactions in D that contain $X \cup Y$ to the number of transactions that contain X in D.

Example

- **Given a database of transactions:**

Transaction	Items
t_1	Bread,Jelly,PeanutButter
t_2	Bread,PeanutButter
t_3	Bread,Milk,PeanutButter
t_4	Beer,Bread
t_5	Beer,Milk

- **Find all the association rules:**

$X \Rightarrow Y$	s	α
Bread $\Rightarrow$ PeanutButter	60%	75%
PeanutButter $\Rightarrow$ Bread	60%	100%
Beer $\Rightarrow$ Bread	20%	50%
PeanutButter $\Rightarrow$ Jelly	20%	33.3%
Jelly $\Rightarrow$ PeanutButter	20%	100%
Jelly $\Rightarrow$ Milk	0%	0%

Association Rule Problem

- Given:
  - — a set I of all the items;
  - — a database D of transactions;
  - — minimum support s;
  - — minimum confidence c;
- Find:
  - — all association rules $X \Rightarrow Y$ with a minimum support s and confidence c, i.e. Find all Frequent Itsemsets.

Important Properties to Exploit

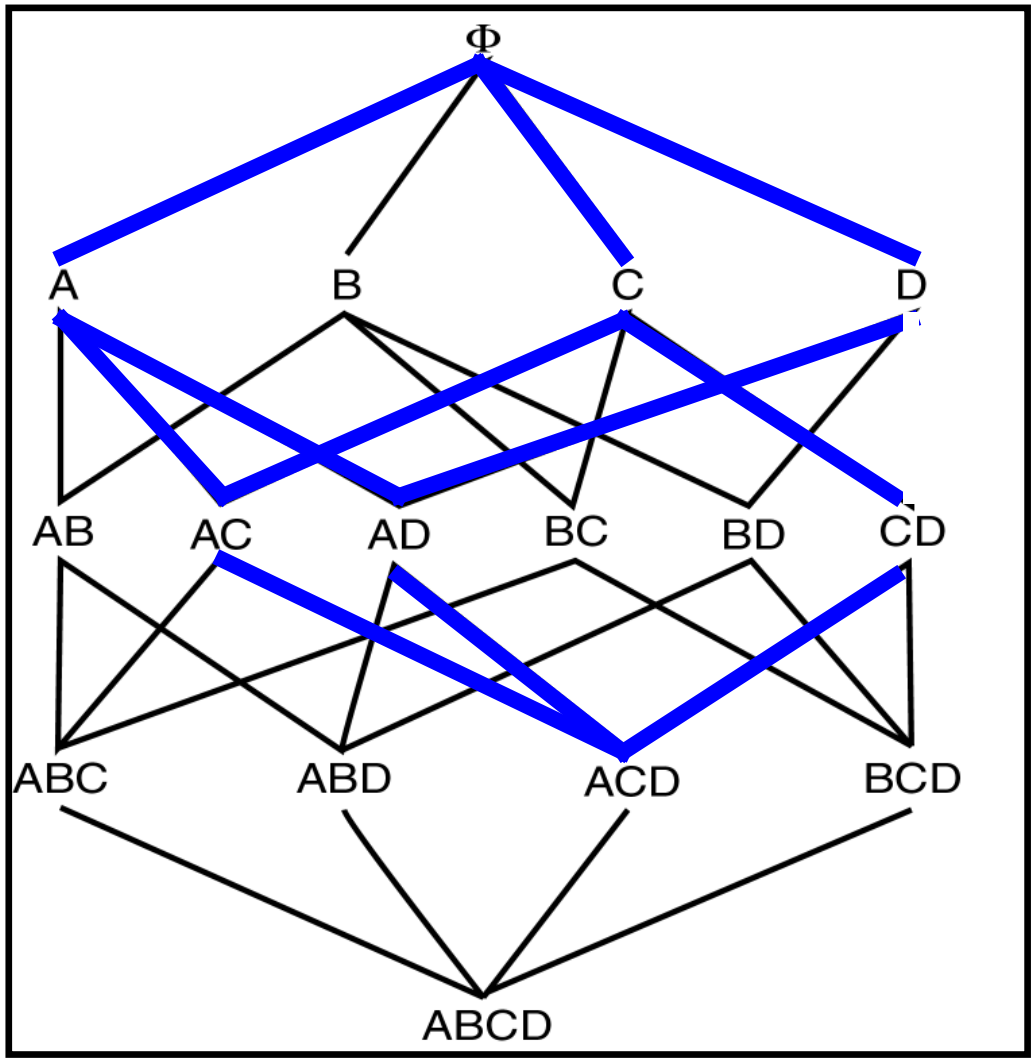
- *Frequent Itemset Property:*

Any subset of a frequent itemset is frequent.

- *Contrapositive:*

If an itemset is not frequent, none of its supersets are frequent.

Frequent Itemset Property



Sequential Algorithm

- L_k: Set of frequent itemsets of size k
- C_k: Set of candidate itemset of size k

L_1 = {frequent items};
for (k = 1; L_k !=∅; k++) do
    C_{k+1} = candidates generated from L_k;
    for each transaction t
        increment the count of all candidates in C_{k+1}
            that are contained in t
    L_{k+1} = frequent candidates in C_{k+1} with min_support

How do you generate C_{k+1}

If we have large itemsets of length k, we can do joins on these (after lexicographically sorting them) to get possible itemsets of length k.

Equivalence Classes

L2 = { AB, AC, AD, BC, BD, CD, DE }

C3 = {ABC, ABD, ACD, BCD}

Parallel Implementation

Say we have Large Itemsets of Size 2

Reorganize the Database as follows:

{ISx, T1, T2, … }, {ISy, T1, T3, ….}, ….

Balanced assignment of Equivalence Classes Amongst the Processors

Each Processor can independently check for larger itemsets.

Can lead to some imbalance because of mismatch in equivalence classes for larger itemsets.