WHAT IS ASSOCIATION RULE MINING?

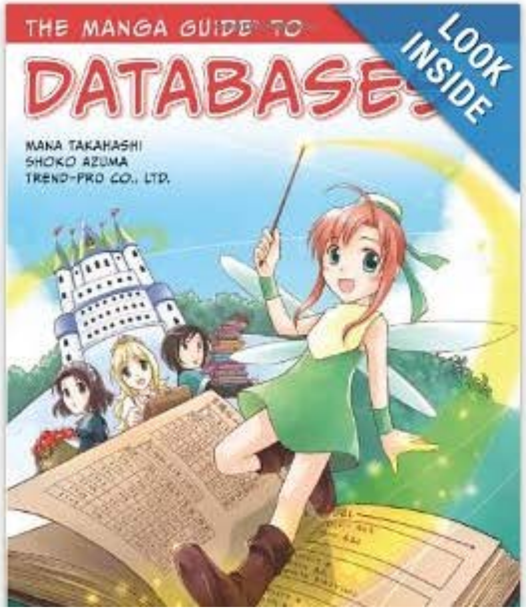
SYRACUSE UNIVERSITY
School of Information Studies

WHAT IS FREQUENT PATTERN ANALYSIS?

Frequent pattern

What products do people frequently buy together?

What other products would people buy if they bought a laptop?



The Manga Guide to Databases Paperback

by Mana Takahashi (Author) , Shoko Azuma (Author) , Trend-Pro Co. Ltd. (Author)

★★★★½ 31 customer reviews

See all 3 formats and editions

Kindle	Library Binding	Paperback
$9.99	$26.06	$13.87
	1 New from $26.06	47 Used from $6.39 44 New from $10.64

Want to learn about databases without the tedium? With its unique combination of Japanese-style comics and serious educational content, *The Manga Guide to Databases* is just the book for you.

Princess Ruruna is stressed out. With the king and queen away, she has to manage the Kingdom of

Frequently Bought Together

Price for all three: $44.14

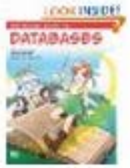
Add all three to Cart | Add all three to Wish List

Show availability and shipping details

- ☑ **This item:** The Manga Guide to Databases by Mana Takahashi Paperback $13.87
- ☑ The Manga Guide to Statistics by Shin Takahashi Paperback $14.76
- ☑ The Manga Guide to Linear Algebra by Shin Takahashi Paperback $15.51

SYRACUSE UNIVERSITY
School of Information Studies

Frequently Bought Together


+

+


Price for all three: $46.83

Add all three to Cart | Add all three to Wish List

Some of these items ship sooner than the others. Show details

- **This item:** The Manga Guide to Databases by Mana Takahashi Paperback $14.49
- The Manga Guide to Statistics by Shin Takahashi Paperback $15.80
- The Manga Guide to Linear Algebra by Shin Takahashi Paperback $16.54

Customers Who Bought This Item Also Bought

The Manga Guide to Electricity
› Kazuhiro Fujitaki
(24)
Paperback
$14.29 Prime

The Manga Guide to Calculus
› Hiroyuki Kojima
(28)
Paperback
$14.82 Prime

The Manga Guide to Physics
› Hideo Nitta
(31)
Paperback
$14.59 Prime

The Manga Guide to Statistics
› Shin Takahashi
(38)
Paperback
$15.80 Prime

ASSOCIATION RULE MINING

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} $\rightarrow$ {Beer}
{Milk, Bread} $\rightarrow$ {Eggs, Coke}
{Beer, Bread} $\rightarrow$ {Milk}

Implication means co-occurrence, not causality!

ASSOCIATION RULE (AR) MINING

Textbook chapter 6 requires some background knowledge in undergraduate computer science courses such as data structure.

Requirement for this class: Learn the basic concepts about AR and the main idea of the Apriori Algorithm.

MORE APPLICATIONS

Product recommendation

- E.g., Amazon.com

Catalog design

Web log (clickstream) analysis

DNA sequence analysis

BASIC CONCEPTS IN AR MINING

SYRACUSE UNIVERSITY
School of Information Studies

FREQUENT ITEMSET

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

Can you answer the following questions?
Which two items are frequently bought together?
Which three items are often bought together?
...

DEFINITION: FREQUENT ITEMSET

Itemset:

A collection of one or more items

k-itemset contains k items

1-itemset:

{A}:3, {B}:3, {C}:2, {D}:4, {E}:3, {F}:2

2-itemset:

{A,B}:1, {A,D}:3

3-itemset:

{A,B,C}:0, {B,E,F}:2

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

METRICS TO EVALUATE FREQUENT LEVEL OF ITEMSETS

How frequent is an itemset?

Support count:

Number of transactions that contain an itemset

support_count({ D, E }) = 2

Support percentage:

Fraction of transactions that contain an itemset

support({D, E}) = 2/5

Frequent itemset:

An itemset with support ≥ threshold

DEFINITION: ASSOCIATION RULE

Association rule:

An implication of the form X → Y,

where X and Y are itemsets,

e.g., {E, F} → {B}

Example Rules:

LHS: Left-Hand Side

{B, E} → {F}

{E, F} → {B}

{B, F} → {E}

{B} → {E, F}

{E} → {B, F}

{F} → {B, E}

RHS: Right-Hand Side

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

METRICS TO EVALUATE THE RULE'S STRENGTH

Rule evaluation metrics

Support P(X, Y)

Fraction of transactions that contain both X and Y

Support({E, F} → {B}) = support_count({B,E,F})/N = 2/5

Confidence P(Y|X) = P(X, Y)/P(X)

How frequently items in Y appear in transactions that contain X

confidence({E,F} → {B})
= support({B,E,F})/support({E,F})
= support_count({B,E,F})/support_count({E,F})
= 2/2 = 1

CONFIDENCE

confidence({E,F} → {B})
= support({B,E,F})/support({E,F})
= support_count({B,E,F})/support_count({E,F})
= 2/2 = 1

Switching LHS and RHS results in different rules with different confidences.

APRIORI ALGORITHM

SYRACUSE UNIVERSITY
School of Information Studies

HOW TO MINE ASSOCIATION RULES?

Given a set of transactions T, the goal of association rule mining is to find all rules having:

- support ≥ *minsup* threshold
- confidence ≥ *minconf* threshold

Brute-force approach:

- List all possible association rules.
- Compute the support and confidence for each rule.
- Prune rules that fail the *minsup* and *minconf* thresholds.
- ⇒ Computationally prohibitive!

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} $\rightarrow$ {Beer} (s = 0.4, c = 0.67)
{Milk, Beer} $\rightarrow$ {Diaper} (s = 0.4, c = 1.0)
{Diaper, Beer} $\rightarrow$ {Milk} (s = 0.4, c = 0.67)
{Beer} $\rightarrow$ {Milk, Diaper} (s = 0.4, c = 0.67)
{Diaper} $\rightarrow$ {Milk, Beer} (s = 0.4, c = 0.5)
{Milk} $\rightarrow$ {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 confidences.

Thus, we may decouple the support and confidence requirements.

MINING ASSOCIATION RULES

Two-step approach:

Frequent itemset generation

Generate all itemsets whose support ≥ *minsup*.

Rule generation

Generate high-confidence rules from each frequent itemset, where each rule is a binary partitioning of a frequent itemset.

Frequent itemset generation is still computationally expensive.

SCALABLE METHODS FOR MINING FREQUENT PATTERNS

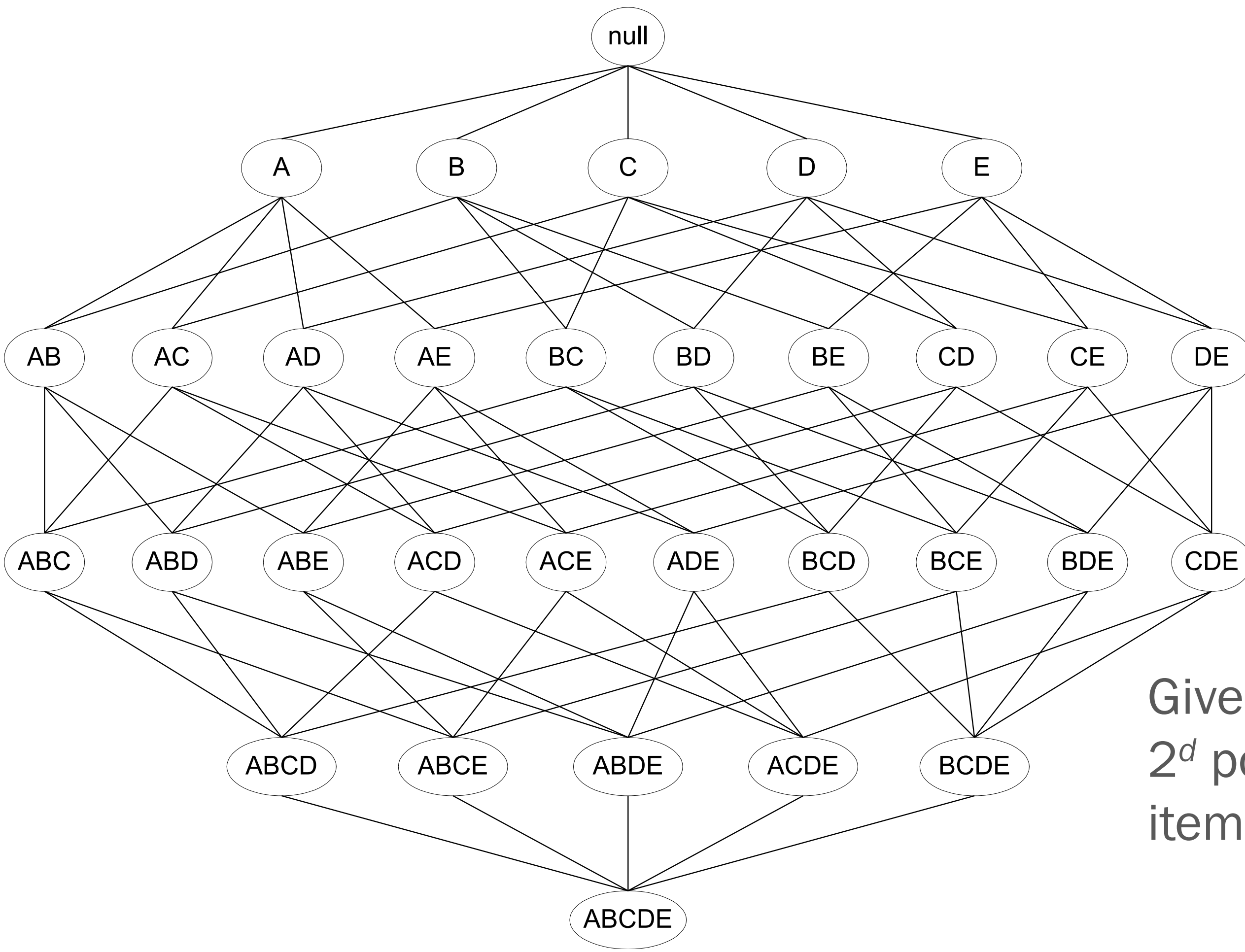
Scalable mining methods: Three major approaches

Apriori (Agrawal & Srikant@VLDB'94)

Frequent pattern growth (FPgrowth—Han, Pei, & Yin @SIGMOD'00)

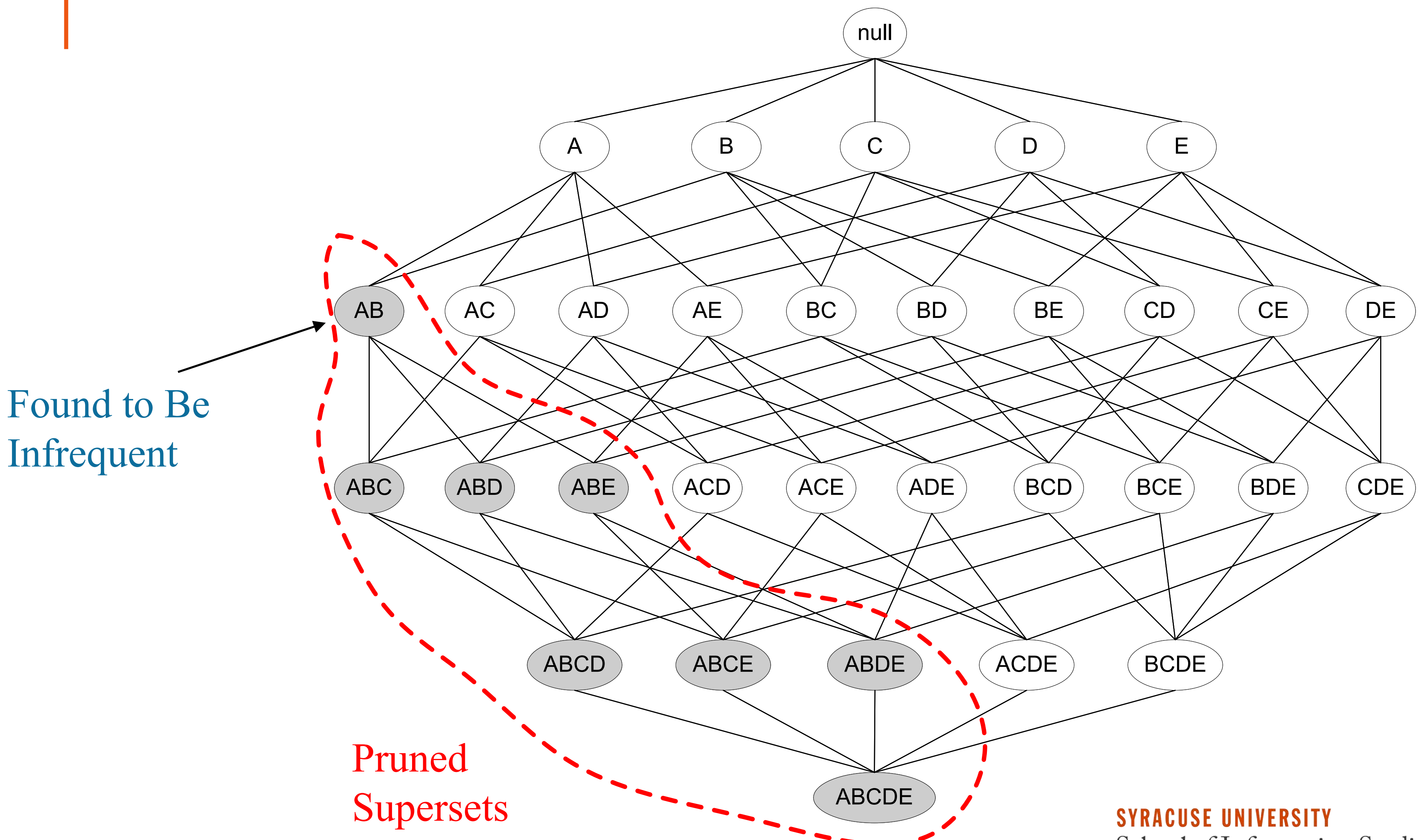
Vertical data format approach (Charm—Zaki & Hsiao @SDM'02)

FREQUENT ITEMSET GENERATION



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

ILLUSTRATING APRIORI PRINCIPLE



APRIORI: A CANDIDATE GENERATION-AND-TEST APPROACH

Apriori pruning principle: If there is any itemset that is infrequent, its superset should not be generated or tested!

Method:

- Initially, scan database once to get frequent 1-itemset.
- Generate length (k + 1) candidate itemsets from length k frequent itemsets.
- Test the candidates against the database.
- Terminate when no frequent or candidate set can be generated.

THE APRIORI ALGORITHM: GENERATE FREQUENT ITEMSET

$Sup_{min} = 2$

Database TDB

TID	Items
10	A, C, D
20	B, C, E
30	A, B, C, E
40	B, E

1st scan

C_1

Itemset	sup
{A}	2
{B}	3
{C}	3
{D}	1
{E}	3

L_1

Itemset	sup
{A}	2
{B}	3
{C}	3
{E}	3

C_2

Itemset
{A, B}
{A, C}
{A, E}
{B, C}
{B, E}
{C, E}

2nd scan

C_2

Itemset	sup
{A, B}	1
{A, C}	2
{A, E}	1
{B, C}	2
{B, E}	3
{C, E}	2

L_2

Itemset	sup
{A, C}	2
{B, C}	2
{B, E}	3
{C, E}	2

C_3

Itemset
{B, C, E}

3rd scan

L_3

Itemset	sup
{B, C, E}	2

RULE GENERATION

Given a frequent itemset L, find all nonempty subsets f, such that f → (L – f) satisfies the minimum confidence requirement.

If {A, B, C, D} is a frequent itemset, candidate rules:

ABC → D, ABD → C, ACD → B, BCD → A

AB → CD, AC → BD, …

A → BCD, B → ACD, C→ ABD, D → ABC

Compute the confidence for each rule, and keep the ones that are greater than min_conf.

RULE GENERATION

How to efficiently generate rules from frequent itemsets?

Start from long LHS:

For itemset {ABCD}, c(x) means confidence of rule x
c(ABC → D) ≥ c(AB → CD) ≥ c(A → BCD)

Proof:

C(ABC->D) = support(ABCD)/support(ABC)

C(AB->CD) = support (ABCD)/support (AB)

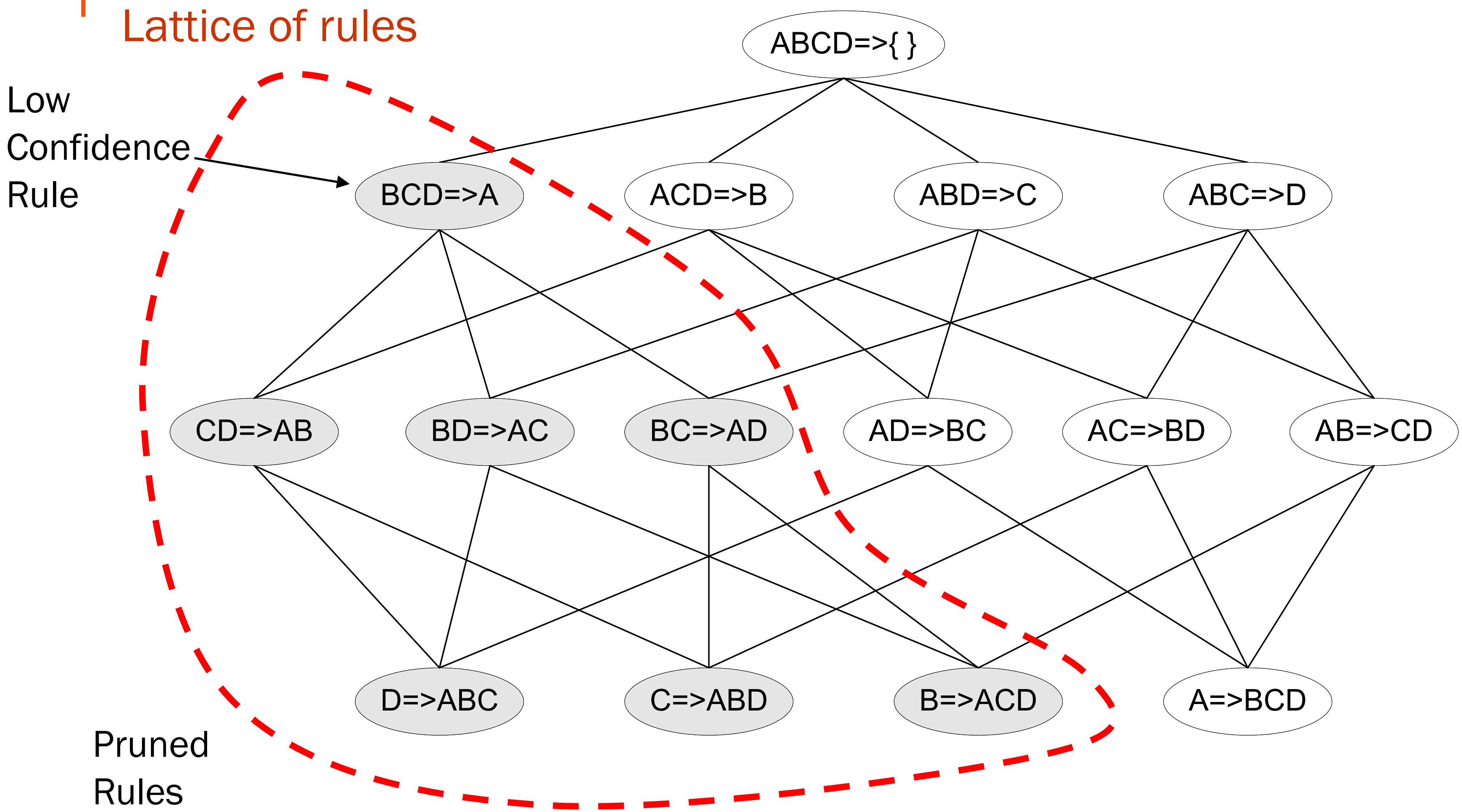
support(AB) ≥ support (ABC)

So C(ABC->D) ≥ C(AB->CD)

If min_conf is not satisfied, no need to generate rules with larger right-hand side (RHS).

THE APRIORI ALGORITHM: RULE PRUNING

Lattice of rules



RULE EVALUATION

SYRACUSE UNIVERSITY
School of Information Studies

LIMITATION OF CONFIDENCE MEASURE

100 transactions:

- 75 bought movies
- 60 bought games
- 40 bought both

Both seem to be strong rules.

{movies}->{games}
support 40/100 = 0.4
confidence 40/75 =0.53

{games}->{movies}
support 40/100 = 0.4
confidence 40/60 = 0.67

HOWEVER …

100 transactions:
- 75 bought movies
- 60 bought games
- 40 bought both

P(movies) = 75/100 = 0.75

P(games) = 60/100 = 0.6

P(movies and games) = 40/100 = 0.4

So people tend not to buy movies and games together!

Correlation(movies, games) = P(movies and games) /[P(movies) × P(games)] = 0.4/(0.75 × 0.6) = 0.89

The confidence measure is sometimes misleading.

Correlation <1 means negative correlation.

METRIC: LIFT (CORRELATION)

Measure of dependent or correlated events: Lift

Lift (A => B) = support({A,B})/(support(A) × support(B))

$$lift(A \quad > B) = \frac{P(A \quad B)}{P(A)P(B)}$$

Association rules should have >1 lift to be meaningful.

THE LIFT (CORRELATION) MEASURE

	Game	Not Game	Total
Movie	40	35	75
Not movie	20	5	25
Total	60	40	100

P(buy game) = 0.6
P(not buy movie) = 0.25
P(buy game and not buy movie) = 0.20
Lift (buy game -> not buy movie)
=0.20/(0.6 × 0.25) = 1.33 > 1

Strong rule

ALTERNATIVE MEASURES

Association rule algorithms tend to produce too many rules.

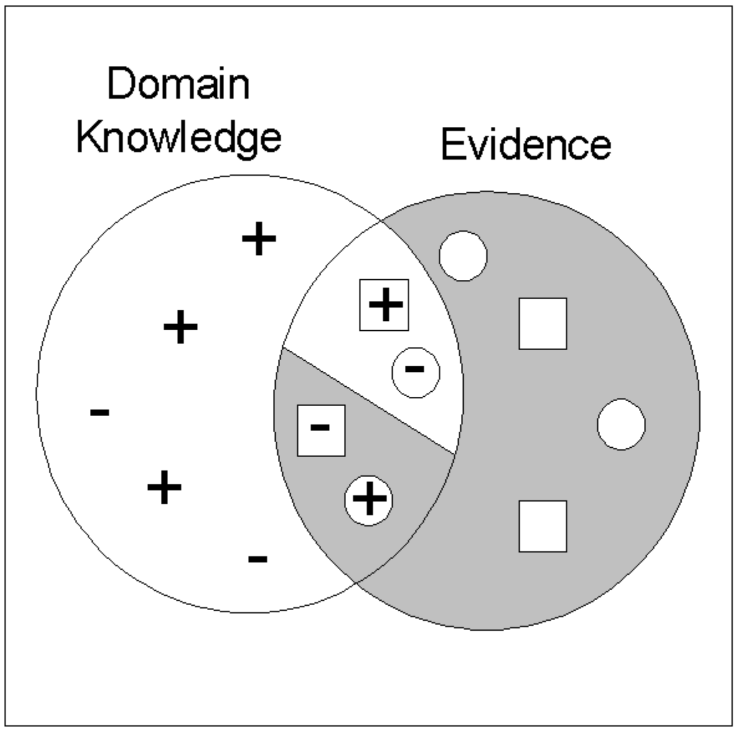
Many of them are uninteresting or redundant.

Uninteresting if it is known knowledge

Redundant if {A,B,C} $\rightarrow$ {D} and {A,B} $\rightarrow$ {D} have same support and confidence

INTERESTINGNESS VIA UNEXPECTEDNESS

Need to model expectation of users (domain knowledge)



+ Pattern expected to be frequent

- Pattern expected to be infrequent

Pattern found to be frequent

Pattern found to be infrequent

+ - Expected Patterns

- + Unexpected Patterns

Need to combine expectation of users with evidence from data (i.e., extracted patterns)

WEKA ASSOCIATION RULES

Implemented a variation of Apriori Algorithm that iteratively reduces the minimum support until it finds the required number of rules with the given minimum confidence

Allows mining of “class association rules”: If the data have a class label attribute, the right hand side of a rule can be restricted to that label.

ASSOCIATION RULE MEASURES

In practice, what levels of support, confidence, and lift should we aim for?

Support:

Depends on data set and business problem

Common setting is 20–40% of the transactions

Confidence:

Strong confidence rules ≥ .9, but .6 to .8 range might be OK

Lift:

Should be above 1.0, the higher the better

Levels of 2 and above can occasionally be seen but more likely to see around 1.3 to 1.5